frontiers | Research Topics

Insights in statistical genetics and methodology 2022

Edited by
Simon Charles Heath and Rongling Wu

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-3645-2
DOI 10.3389/978-2-8325-3645-2

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

Insights in statistical genetics and methodology: 2022

Topic editors

Simon Charles Heath — National Center for Genomic Analysis, Center for Genomic Regulation (CRG), Spain
Rongling Wu — The Pennsylvania State University (PSU), United States

Citation

Heath, S. C., Wu, R., eds. (2023). *Insights in statistical genetics and methodology: 2022*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-3645-2

Table of contents

ORIGINAL RESEARCH
published: 15 July 2022
doi: 10.3389/fgene.2022.938259

Cuproptosis-Associated lncRNA Establishes New Prognostic Profile and Predicts Immunotherapy Response in Clear Cell Renal Cell Carcinoma

*Shengxian Xu[1†], Dongze Liu[1†], Taihao Chang[1†], Xiaodong Wen[1†], Shenfei Ma[1], Guangyu Sun[1], Longbin Wang[2], Shuaiqi Chen[1], Yong Xu[1] and Hongtuan Zhang[1]**

[1]Department of Urology, National Key Specialty of Urology Second Hospital of Tianjin Medical University Tianjin Key Institute of Urology Tianjin Medical University, Tianjin, China, [2]Department of Family Planning, The Second Hospital of Tianjin Medical University, Tianjin, China

OPEN ACCESS

Edited by:
Rongling Wu,
The Pennsylvania State University (PSU), United States

Reviewed by:
Dalong Cao,
Fudan University, China
Yao Lin,
Fujian University of Traditional Chinese Medicine, China

Correspondence:
Hongtuan Zhang
zhtlml@163.com

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

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

Received: *07 May 2022*
Accepted: *09 June 2022*
Published: *15 July 2022*

Citation:
Xu S, Liu D, Chang T, Wen X, Ma S, Sun G, Wang L, Chen S, Xu Y and Zhang H (2022) Cuproptosis-Associated lncRNA Establishes New Prognostic Profile and Predicts Immunotherapy Response in Clear Cell Renal Cell Carcinoma. Front. Genet. 13:938259. doi: 10.3389/fgene.2022.938259

Background: Clear cell renal cell carcinoma (ccRCC) accounts for 80% of all kidney cancers and has a poor prognosis. Recent studies have shown that copper-dependent, regulated cell death differs from previously known death mechanisms (apoptosis, ferroptosis, and necroptosis) and is dependent on mitochondrial respiration (Tsvetkov et al., Science, 2022, 375 (6586), 1254–1261). Studies also suggested that targeting cuproptosis may be a novel therapeutic strategy for cancer therapy. In ccRCC, both cuproptosis and lncRNA were critical, but the mechanisms were not fully understood. The aim of our study was to construct a prognostic profile based on cuproptosis-associated lncRNAs to predict the prognosis of ccRCC and to study the immune profile of clear cell renal cell carcinoma (ccRCC).

Methods: We downloaded the transcriptional profile and clinical information of ccRCC from The Cancer Genome Atlas (TCGA). Co-expression network analysis, Cox regression method, and least absolute shrinkage and selection operator (LASSO) method were used to identify cuproptosis-associated lncRNAs and to construct a risk prognostic model. In addition, the predictive performance of the model was validated and recognized by an integrated approach. We then also constructed a nomogram to predict the prognosis of ccRCC patients. Differences in biological function were investigated by GO, KEGG, and immunoassay. Immunotherapy response was measured using tumor mutational burden (TMB) and tumor immune dysfunction and rejection (TIDE) scores.

Results: We constructed a panel of 10 cuproptosis-associated lncRNAs (HHLA3, H1-10-AS1, PICSAR, LINC02027, SNHG15, SNHG8, LINC00471, EIF1B-AS1, LINC02154, and MINCR) to construct a prognostic prediction model. The Kaplan–Meier and ROC curves showed that the feature had acceptable predictive validity in the TCGA training, test, and complete groups. The cuproptosis-associated lncRNA model had higher diagnostic efficiency compared to other clinical features. The analysis of Immune cell infiltration and ssGSEA further confirmed that predictive features were significantly associated with the immune status of ccRCC patients. Notably, the superimposed effect of patients in the high-risk group and high TMB resulted in shorter survival. In addition, the higher TIDE scores in the high-risk group suggested a poorer outcome for immune checkpoint blockade response in these patients.

Conclusion: The ten cuproptosis-related risk profiles for lncRNA may help assess the prognosis and molecular profile of ccRCC patients and improve treatment options, which can be further applied in the clinic.

Keywords: cuproptosis, lncRNA, ccRCC, prognostic model, bioinformatics

INTRODUCTION

Renal cell carcinoma is a common genitourinary malignancy that causes nearly 170,000 deaths each year. The most common histologic type of renal cell carcinoma is clear cell renal cell carcinoma (ccRCC), which accounts for approximately 80% of cases (Delman, 2020). Due to the asymptomatic nature of renal clear cell carcinoma, metastases are usually already present at the time of diagnosis. Surgery is also difficult to remove renal cell carcinoma metastases, and recurrence is common after nephrectomy. Also, ccRCC differs from other urologic tumors in that it is insensitive to both radiotherapy and chemotherapy (Ljungberg et al., 2015). As a highly immunogenic tumor, ccRCC may benefit from immunotherapy. Although immunotherapy has indeed made considerable breakthroughs in ccRCC, treatment outcomes still vary from individual to individual (Motzer et al., 2019). Therefore, there is an urgent need to better understand the heterogeneity of ccRCC patients and establish an accurate and comprehensive risk model to stratify patients to design personalized treatment plans in terms of prognosis prediction and drug selection.

Long non-coding RNA (lncRNA) refers to RNAs that are longer than 200bp and do not have protein-coding functions, which play an important regulatory role in immune response processes, such as immune cell infiltration, antigen recognition, antigen exposure, and tumor clearance (Quinn and Chang, 2016).

LncRNAs play specific roles in carcinogenesis and metastasis by transcription and post-transcriptional modifications of genes (Du et al., 2020; Gao et al., 2020; Liu et al., 2021). Lv pointed out that lncRNAs were associated with tumor autophagy in ccRCC(9). At the same time, a number of studies have shown that lncRNAs can influence the expression of target genes by acting as competing RNAs (Liu and Lei, 2021; Shan et al., 2022a; Zhang et al., 2022). LncRNAs are also connected to drug resistance in tumors (Barik et al., 2021). However, studies on the role of cuproptosis-associated lncRNAs in ccRCC prognosis and tumor immunity (TIME) are still unclear.

Copper is an indispensable cofactor for all organisms to maintain life activities, as it plays an important role in biological processes such as mitochondrial respiration, antioxidant/detoxification, and iron uptake (Ruiz et al., 2021). However, it can become harmful if the concentration of copper in the body exceeds the threshold that can be maintained by homeostatic mechanisms. Recent studies have indicated that copper-regulated cell death occurs in a manner that is different from previously known death mechanisms (apoptosis, ferroptosis, and necroptosis) and that it is closely linked to mitochondrial respiration. Specifically, cuproptosis occurs through direct binding of copper to the lipidated components of the tricarboxylic acid (TCA) cycle. The combination of the two will lead to lipid-acylated protein aggregation and subsequent loss of iron-sulfur cluster proteins, further leading to proteotoxic stress and ultimately cell death (Tsvetkov et al., 2022). Several links have been observed between copper and cancer. Copper accumulation is closely associated with tumor cell development, angiogenesis, and metastasis (Lelièvre et al., 2020; Li, 2020; Ruiz et al., 2021; Ge et al., 2022). Currently, the mechanism of copper-mediated death regulation in tumors is unclear, and studies on the role of copper-death-associated lncRNAs in ccRCC are inconclusive. Therefore, our study aims to explore the role of cuproptosis-related lncRNAs in ccRCC using bioinformatics.

MATERIALS AND METHODS

Data Collection

RNA sequencing data and clinical characterization data for ccRCC were obtained on 9 April 2022 by downloading from the TCGA database (https://portal.gdc.cancer.gov/repository), which included a dataset of 539 tumor samples and a dataset of 72 normal tissue samples (Liu et al., 2018). Using the Perl programming language (version Strawberry-Perl-5.30.0; https://www.perl.org), the RNA-seq data were extracted in the fragment per kilobase million (FPKM) format that has been normalized (Conesa et al., 2016). At the same time, the clinical data were preprocessed with Pearl to obtain the complete pathological information of the clinical samples.

Screening and Differential Expression Analysis of Cuproptosis-Associated lncRNAs

Using the packages “limma,” “dplyr,” “ggalluvial,” and “ggplot2,” we plotted the Sankey relationship between cuproptosis genes and cuproptosis-associated lncRNAs (Ritchie et al., 2015). Our team was filtered using Pearson’s correlation analysis with the criteria of |Pearson R| > 0.4 and $p < 0.001$.

Modeling and Validation of Prognostic Risk Assessment

The KIRC dataset from TCGA was randomly divided into a training risk set and a test risk set using the caret R package in a 1:

1 ratio. The train set was utilized to construct cuproptosis-related lncRNA signatures, and the test set and the whole set were applied to validate the signature.

Univariate Cox regression analysis was applied to identify prognosis-associated lncRNAs among those cuproptosis-associated lncRNAs ($p < 0.05$), and forest plots were drawn. Also, we mapped these lncRNAs by the "limma," "pheatmap," "reshape2,″ and "ggpubr" packages. Then, by performing LASSO Cox regression algorithm analysis (using the penalty parameter estimated by 10-fold cross-validation) on the obtained prognostic lncRNAs, we determined the best group of prognostic lncRNAs and established the risk model. This approach minimizes overfitting in the modeling process. Finally, we developed a prognostic risk model based on optimal lncRNA using multivariate Cox regression and calculated the risk score for each patient with ccRCC according to the following equation:

$$\text{risk score} = \sum \text{i} = \text{1nCoef}\,(\text{i})\times\text{Expr}\,(\text{i}).$$

Coef (i) and Expr (i) in the formula denote the regression coefficient of the multiple Cox regression analysis for each lncRNA and normalized expression level for each lncRNA, respectively. The median of the training set was used as a cut-off point to classify all samples containing KIRC as low- or high-risk subsets. Kaplan–Meier (KM) curves were adopted to explore whether there is a difference in the overall survival and progression-free survival of ccRCC patients between the high-risk and low-risk subsets in the training and testing sets using the "survival" R package. The chi-square test was utilized by us to evaluate the correlation between the model and the clinical characteristics. Based on survival, caret, glmnet, rms, survminer, and timeROC packages, we generated ROC curves and calculated the area under the curve (AUC) and applied the consistency index (C-index) together to measure the accuracy of the model.

Nomogram and Calibration

Combining risk scores with various clinical pathological factors, the rms package was applied to create line graphs for 1-, 3-, and 5-years OS for ccRCC patients. The calibration curve based on the Hosmer–Lemeshow test was used to show the predictive power of the nomogram models developed.

PCA, GO, and KEGG Analysis

The expression patterns of cuproptosis-related lncRNAs for ccRCC samples were classified using principal component analysis to visualize the spatial distribution of high- and low-risk samples. In addition, for the differential genes in the low- and high-risk groups, we used Gene Ontology (GO) analysis, which consisted of three components: biological process (BP), cellular component (CC), and molecular function (MF). Also, differentially expressed KEGG pathways in the two groups were analyzed using the Hs. eg.db, clusterProfiler, and enrichplot packages. $p < 0.05$ and FDR <0.05 were considered as significantly enriched biological processes and pathways.

Tumor Immune Analysis

In order to explore the relationship between this model and immune infiltration status, our team calculated the immune infiltration profile of the TCGA-KIRC dataset using seven algorithms (XCELL, TIMER, QUANTISEQ, MCPCOUNTER, EPIC, CIBERSORT-ABS, and CIBERSORT) (Aran et al., 2017; Li et al., 2017; Racle et al., 2017; Chen et al., 2018; Dienstmann et al., 2019; Finotello et al., 2019; Li et al., 2020; Tamminga et al., 2020). Wilcoxon signed-rank test, limma, tidyverse, scales, ggplot2, and ggtext R packages were used to perform the analysis of the differences in the content of immune infiltrating cells in the different risk groups explored and the outcomes were shown in the bubble plots.

Then, based on the ESTIMATE algorithm, we explored the abundance of immune and stromal cells between different groups and calculated the StromalScore, ImmuneScore, and ESTIMATEScore (StromalScore + ImmuneScore) for each group (Chen et al., 2018). In addition, we investigated the differential expression of immune checkpoints in high- and low-risk populations and showed them in box plots. Subsequently, single-sample GSEA (ssGSEA) scoring of infiltrating immune cells and immune-related functions in ccRCC was performed by the "limma," "GSVA," and "GSEABase" packages and presented as a heat map.

Tumor Mutation Burden and Tumor Immune Dysfunction and Exclusion Score

After downloading the somatic mutation data from the TCGA website, we applied the Pearl programming language to extract the mutation data. Then, we examined and integrated TCGA data using the "maftools" package and analyzed the differences in TMB and survival rates between the high-risk and low-risk groups. The tumor immune dysfunction and exclusion (TIDE) scoring file was retrieved from the TIDE website (http://tide.dfci.harvard.edu) (Jiang et al., 2018). We then assessed potential differences in immune checkpoint blockade (ICB) responses between the low- and high-risk groups using the "ggpubr" package. Finally, our team used the R package pRRophetic to predict the IC50 values of drugs available for the treatment of ccRCC in the high- and low-risk groups.

Validation of the Expression Level of Screened Hub Cuproptosis-Associated lncRNAs in KIRC by qRT-PCR

Cancer and adjacent normal tissues were collected from six patients with renal clear cell carcinoma admitted to the Second Hospital of Tianjin Medical University. Each patient was informed and signed the consent form. The study was approved by the Institutional Review Board of the Second Hospital of Tianjin Medical University. All tissues were rapidly stored in liquid nitrogen after excision. After tissue grinding, total RNA was extracted from ccRCC tissue using TRIzol reagent (Invitrogen, China) according to the manufacturer's protocol. Finally, we performed a quantitative reverse transcription-polymerase chain reaction (qRT-PCR) on cDNA using FastStart Universal SYBR Green Master (ROX,

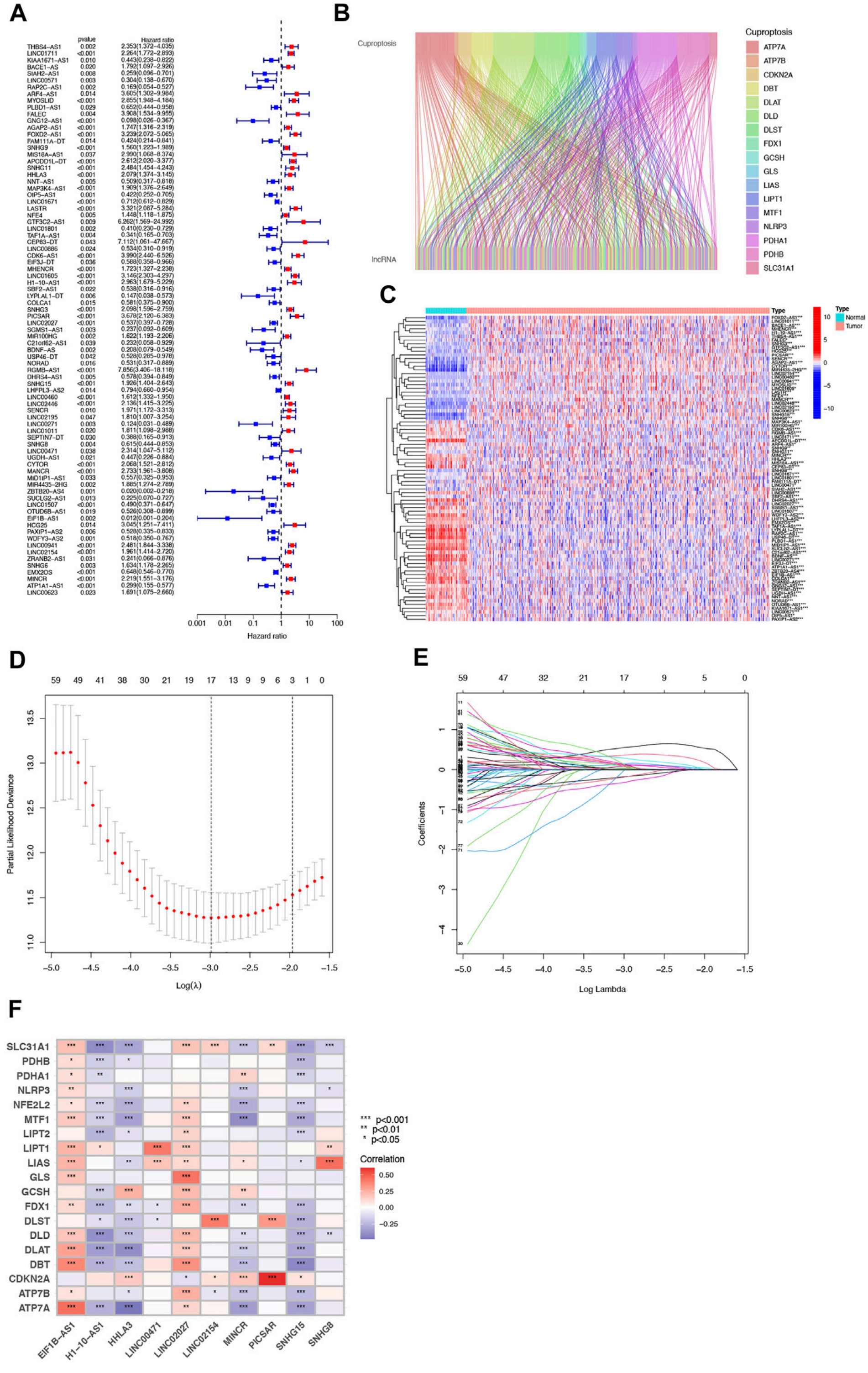


FIGURE 1 | Identification of Cuproptosis-associated lncRNA prognostic features in ccRCC. The forest plot shows prognosis-related genes for cuproptosis-associated lncRNAs **(A)**. Sankey relationship diagram of cuproptosis genes and cuproptosis-associated lncRNAs **(B)**. Differential expression of 81 cuproptosis-associated lncRNAs associated with survival between ccRCC and normal samples **(C)**. Distribution of the LASSO coefficients of cuproptosis-associated lncRNAs **(D)**. The 10-fold cross-validation of variable selection in the least absolute shrinkage and selection operator (LASSO) algorithm **(E)**. Correlation of lncRNAs with cuproptosis-related genes in risk models **(F)**.

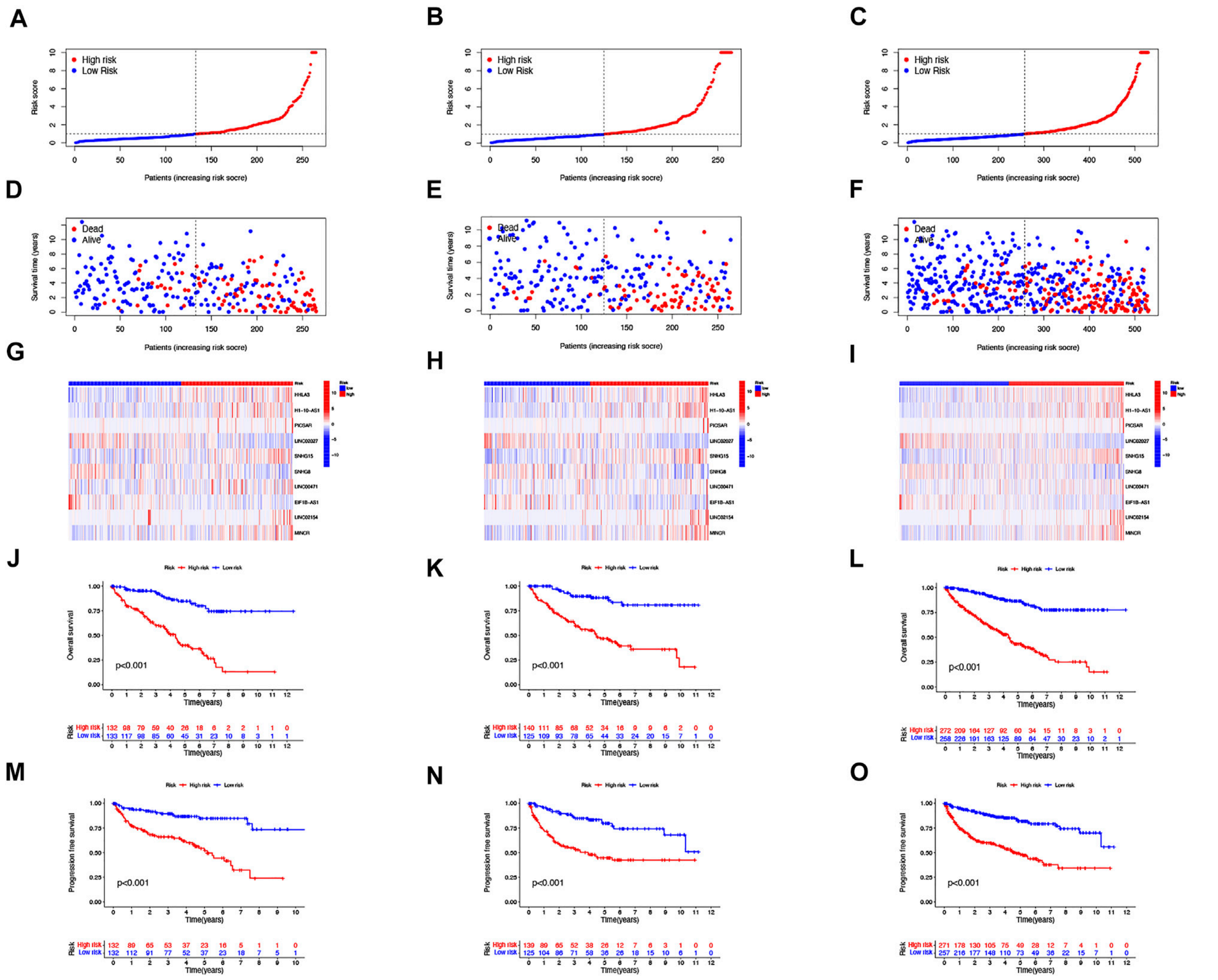


FIGURE 2 | Prognosis of the risk model in different groups. The distribution of overall survival risk scores **(A–C)**, survival time and survival status **(D–F)**, heat maps of 10 lncRNA expressions **(G–I)**, Kaplan–Meier survival curves of overall survival of ccRCC patients **(J–L)**, and Kaplan–Meier survival curves of progression-free survival of ccRCC patients **(M–O)** between low- and high-risk groups in the train, test, and entire sets, respectively.

Roche; United States). GAPDH was used as a reference. The following primer sequences were used: GAPDH-F: GGAAGGTGAAGGTCGGAGTCA, GAPDH-R: GTCATTGATGGCAACAATATATCCACT; SNHG15-F: TGGCAGACCTGTACTCCGTA, SNHG15-R: CCTGGGCTCAGGAATGGTCA; LINC00471-F: TATCACCAAGCAGGAGGGGA, LINC00471-R: ATCGGGAACCCCCTACAGAA.

RESULTS

Prognosis-Related lncRNAs With Coexpression of Cuproptosis

Our team identified 434 lncRNAs with co-expression relationships in ccRCC (|Pearson R| > 0.4 and p < 0.001) (**Figure 1B**). Univariate Cox analysis (p < 0.05) was utilized to choose 81 differentially expressed prognostic-related lncRNAs: THBS4-AS1, LINC01711, MACORIS, KIAA1671-AS1, BACE1-AS, SIAH2-AS1, LINC00571, RAP2C-AS1, ARF4-AS1, MYOSLID, PLBD1-AS1, FALEC, GNG12-AS1, AGAP2-AS1, OXCT1-AS1, FOXD2-AS1, SNHG9, LINC00882, APCDD1L-DT, SNHG11, OXCT1-AS1, CTBP1-DT, HHLA3, NNT-AS1, MAP3K4-AS1, OIP5-AS1, LINC01671, LASTR, NFE4, GTF3C2-AS1, LINC01801, LINC00886, CDK6-AS1, EIF3J-DT, MHENCR, LINC01605, H1-10-AS1, SBF2-AS1, PCCA-DT, LYPLAL1-DT, COLCA1, SNHG3, GAS6-DT, LINC02027, SGMS1-AS1, BDNF-AS, KLHL7-DT, NORAD, DHRS4-AS1, SNHG15, LHFPL3-AS2, LINC00460, LINC02446, LINC02195, LINC00271, GATA2-AS1, LINC01011, SEPTIN7-DT, SNHG8, UGDH-AS1, CYTOR, MANCR, MIR4435-2HG, ITGA9-AS1, ZBTB20-AS4, SUCLG2-AS1, LINC01507, OTUD6B-AS1, EIF1B-AS1, HCG25, PAXIP1-AS2, WDFY3-AS2, TGFB2-AS1, BAALC-AS1, LINC00941, LINC02154, SNHG6, EMS2OS,

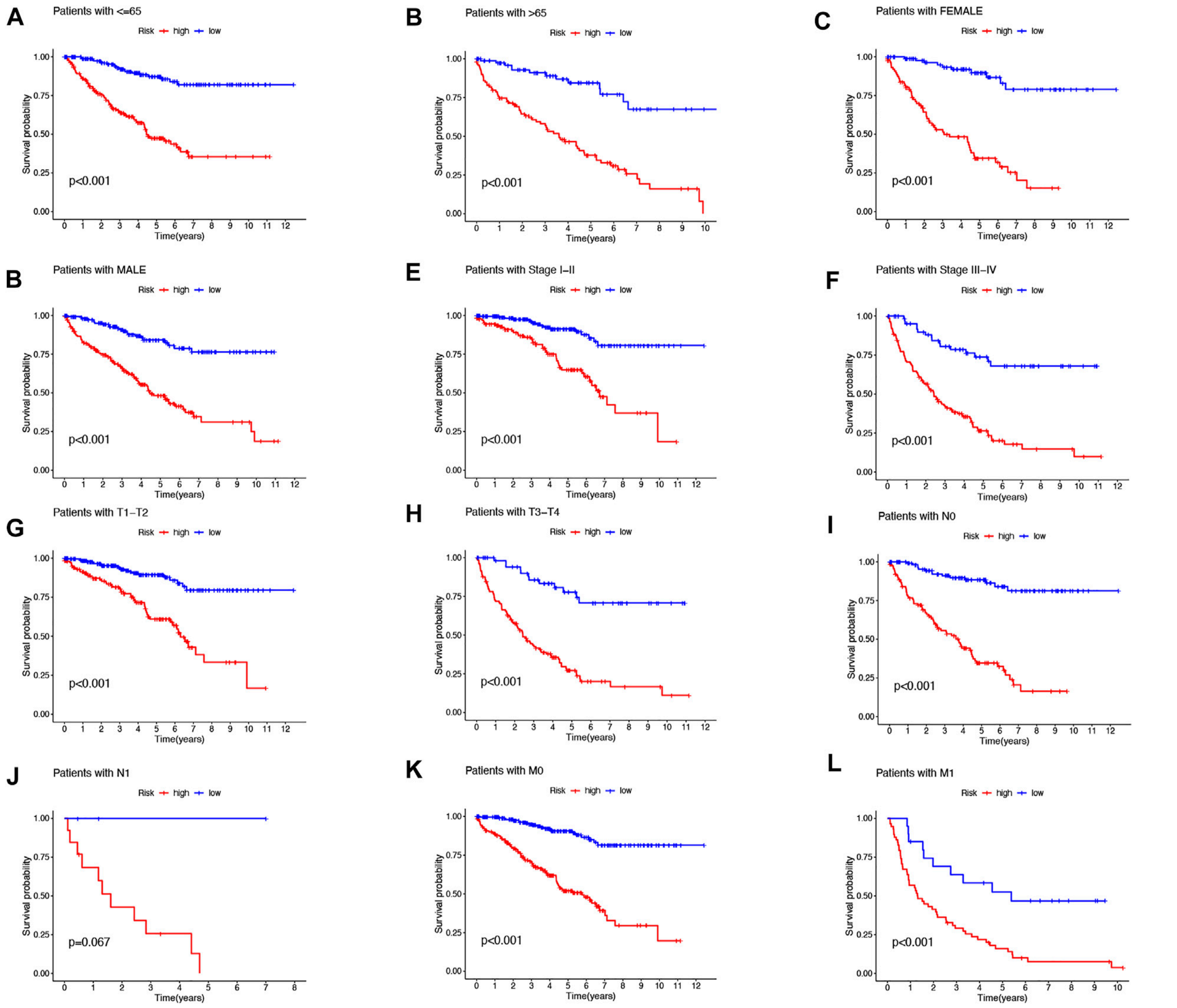


FIGURE 3 | Kaplan–Meier survival curves for low- and high-risk populations by different clinical variables. Age **(A,B)**, sex **(C,D)**, stage **(E,F)**, T stage **(G,H)**, N stage **(I,J)**, and M stage **(K,L)**.

MINCR, ATP1A1-AS1, LINC00623, and LINC01415 (**Figure 1A and C**).

Construction of the Cuproptosis-Related LncRNA Predictive Signature

Then, we performed a LASSO Cox regression analysis using the training set and obtained the lncRNAs with the highest prognostic values using the "glmnet" package of R software (**Figure 1D–F**). Finally, we obtained 17 lncRNAs, 10 of which were introduced into the multi-Cox proportional risk model. The risk score was obtained using the multivariate Cox regression formula: risk score = HHLA3 × (0.4223) + H1-10-AS1 × (0.5960) + PICSAR × (0.9702) + LINC02027 × (−0.5392) + SNHG15 × (0.3602) + SNHG8 × (−0.6352) + LINC00471 × (1.2766) + EIF1B-AS1 × (−3.8776) + LINC02154 × (0.7232) + MINCR × (0.3724). Overall survival was significantly shorter for all patients in the high-risk group in the complete set and training and validation partitions (**Figure 2A–L**). Similarly, the progression-free survival was significantly lower in the high-risk group compared to the low-risk group (**Figure 2M–O**). Meanwhile, ccRCC patients were grouped by age, sex, stage, T-stage, N-stage, and M-stage to investigate the correlation between survival probability and risk score in generic clinicopathological characteristics. The results showed that for different classifications, except for stage N1 (**Figure 3J**), the overall survival rate was much higher in the low-risk group (**Figures 3A–I**, **Figure 3K-L**). A possible interpretation of the N1 stage was the limited number of patients because of the bad prognosis of advanced ccRCC. The results suggested that the model

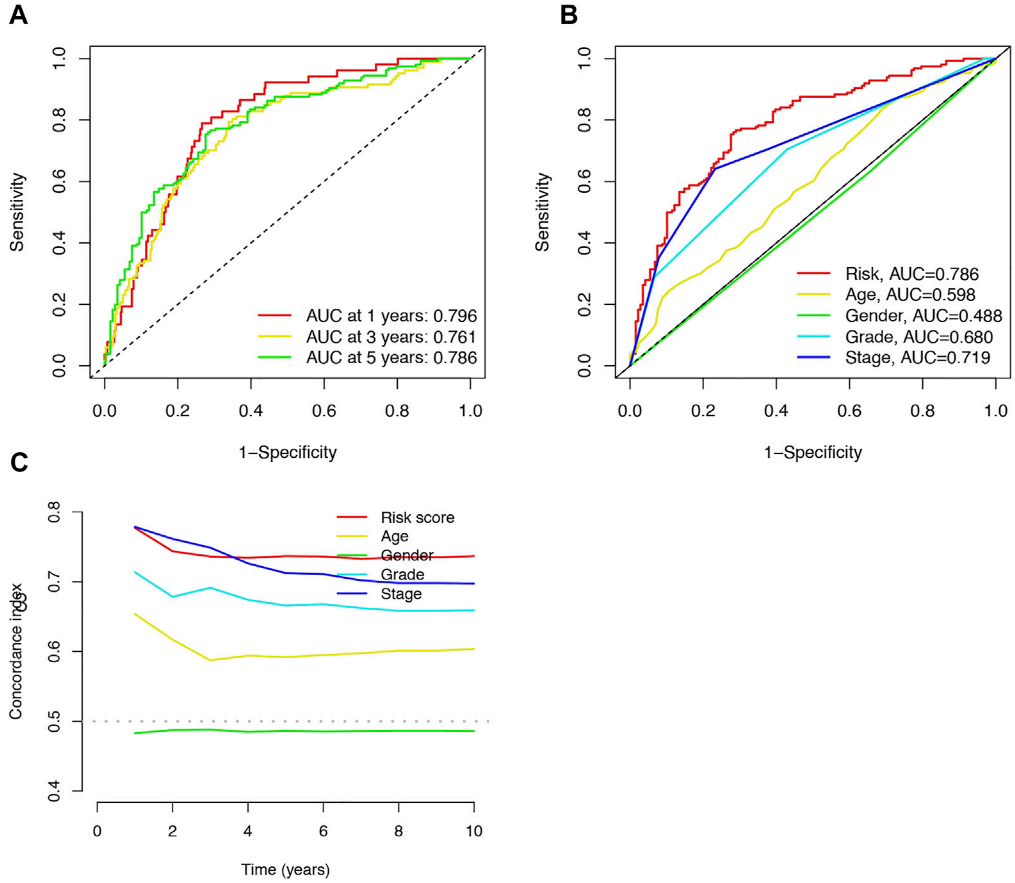


FIGURE 4 | Accuracy of the risk characteristic based on a whole-group prediction of 1-, 3-, and 5-years receiver operating characteristic curves **(A)**. Predictive accuracy of the risk model compared with clinicopathologic characteristics such as age, sex, and stage **(B)**. C-index curve of the risk model **(C)**.

can be used to help predict the prognosis of patients with ccRCC with different clinicopathological variables.

An Independent Prognostic Indicator of ccRCC of the Cuproptosis-Related lncRNA Signature

The area under the curve (AUC) was 0.796, 0.761, and 0.786 for the 1-, 3-, and 5-years ROCs, respectively (**Figure 4A**). The AUC of the risk score was 0.786 in the 5-years ROC of the model, showing extremely strong predictive power compared to other clinicopathological characteristics (**Figure 4B**). The 10-years C-index in the risk model was also higher than the other clinical features (**Figure 4C**).

Construction and Validation of the lncRNA-Based Nomogram

Our team predicted the prognosis of ccRCC patients at 1, 3, and 5 years by constructing a nomogram that included clinical characteristics and risk scores (**Figure 5A**). The calibration curves showed good agreement between the nomogram and the predicted results (**Figure 5B**).

The Principal Component Analysis and Biological Pathways Analyses

We then utilized PCA to explore the differences between the high- and low-risk groups in four expression profiles (total gene expression profiles, cuproptosis genes, cuproptosis-associated lncRNAs, and risk models classified by the expression profiles of 10 cuproptosis-associated lncRNAs) (**Figure 6A–D**). The outcomes indicated that the 10 cuproptosis-associated lncRNAs were of best discriminatory capacity to distinguish well between low- and high-risk populations. GO analysis showed that cuproptosis-associated lncRNAs were strongly associated with the development of immune responses (**Figure 7A and B**). KEGG analysis resulted mainly in cytokine–cytokine receptor interaction and PI3K-AKT signaling pathway (**Figures 7C and D**).

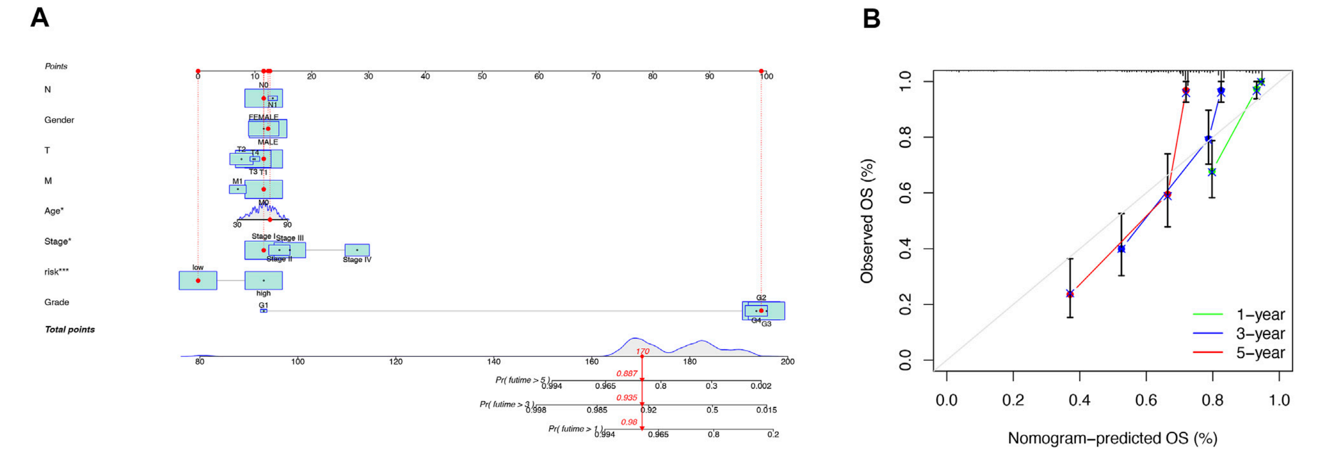


FIGURE 5 | Construction and validation of the nomogram. A nomogram combining clinicopathological variables and risk scores predicts 1-, 3-, and 5-years overall survival in patients with ccRCC **(A)**. Calibration curves test the agreement between actual and predicted outcomes at 1, 3, and 5 years **(B)**.

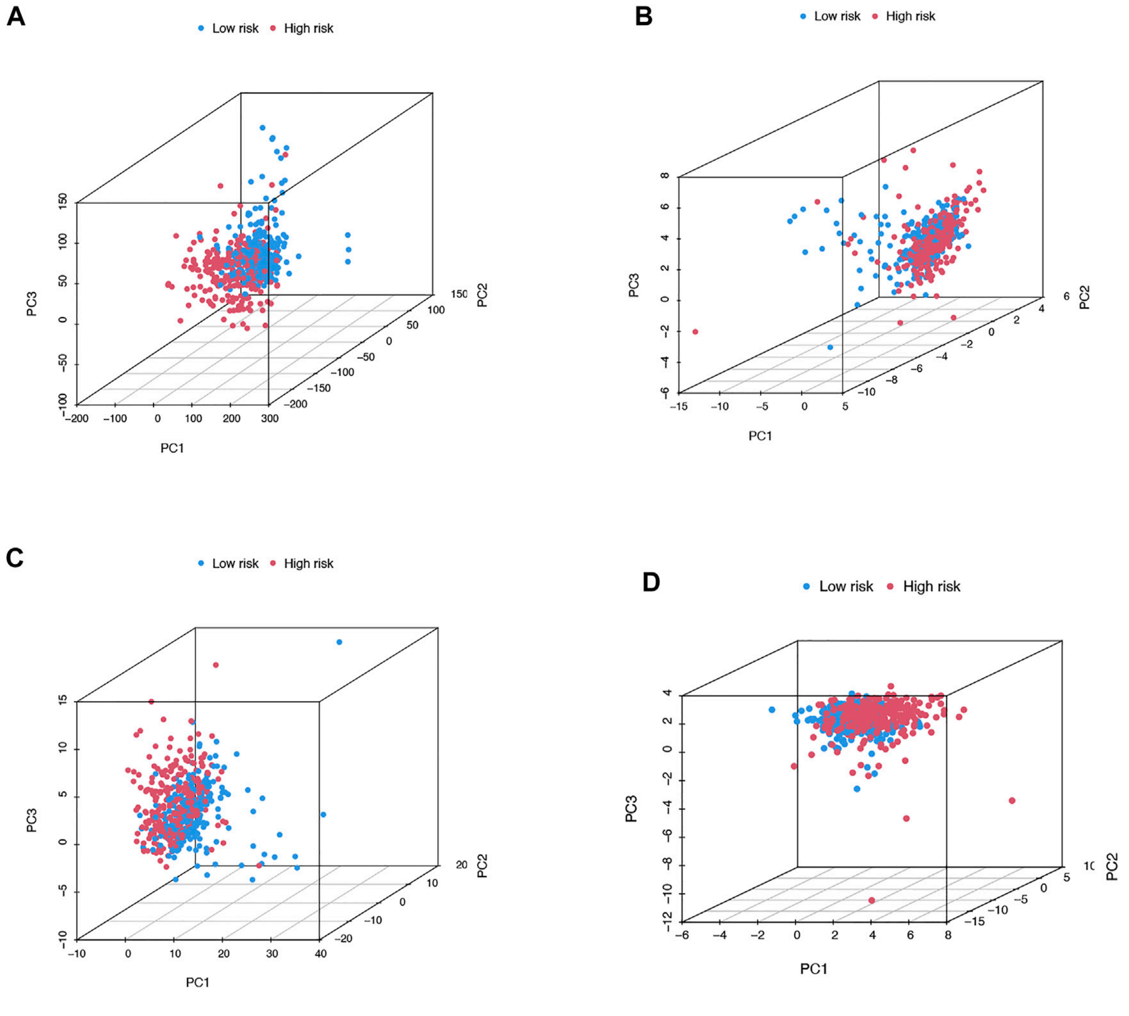


FIGURE 6 | PCA in both groups of patients. PCA of all genes **(A)**. PCA of cuproptosis genes **(B)**. PCA of cuproptosis-related lncRNAs **(C)**. PCA of risk lncRNAs **(D)**.

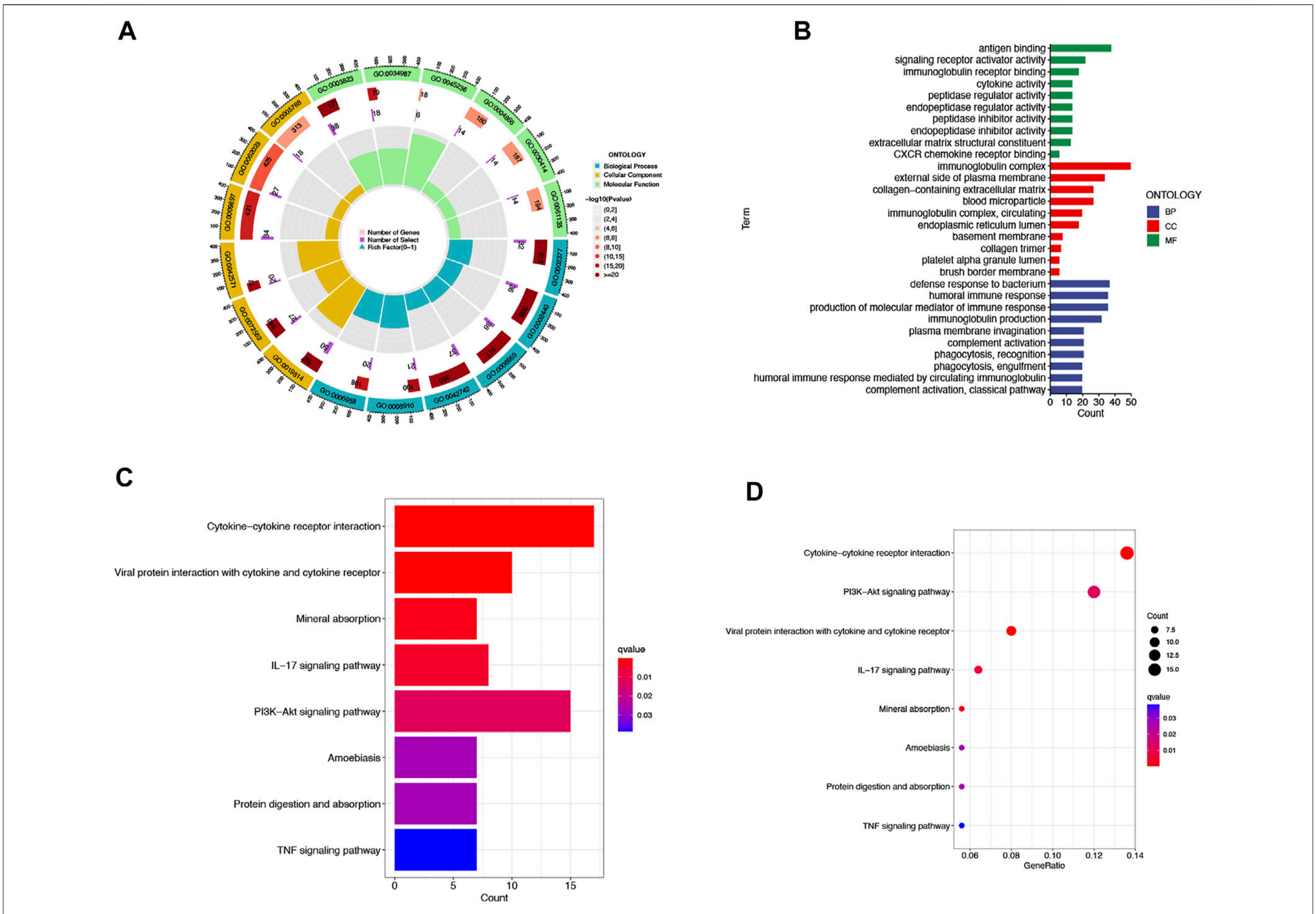


FIGURE 7 | GO and KEGG analysis. Gene Ontology (GO) analysis demonstrated the richness of molecular biological processes (BP), cellular components (CC), and molecular functions (MF) **(A,B)**. KEGG pathway analysis showed the significantly enriched pathways **(C,D)**.

Examination of Immune Characteristics in High- and Low-Risk Groups

In immune cell bubble graphs, our team found that samples from the high-risk group were significantly positively correlated with infiltration of regulatory T cells, B cell memory, NK cells, and T cell follicular helper and negatively correlated with neutrophil infiltration (all $p < 0.05$) (**Figure 8A**). Details of the infiltration of the aforementioned cells are shown in **Supplementary Figure S1**. In addition, we analyzed the differences in immune checkpoints between the high-risk and low-risk groups (**Figure 8B**). Interestingly, most of the immune checkpoints had higher expression in the high-risk patients, which may explain the poorer OS in the high-risk group. Subsequently, our team investigated the connection between risk scores and immune-related activities in ccRCC. The box plots of the results indicated that type II IFN response, Type I IFN response, cytolytic activity, inflammation-promoting, check point, T-cell co-stimulation, CCR, and parainflammation were dramatically different in the risk scores (**Figure 8C**). In terms of TME scores, immune scores and ESTIMATE scores were higher in high-risk patients than in low-risk patients, with no difference in stromal scores between them (**Figure 8D–F**).

TMB, TIDE, and Therapeutic Drug Sensitivity

We then downloaded the somatic mutation data from the TGCA database and analyzed the changes in somatic mutations in the high- and low-risk groups. The 10 most highly mutated genes were VHL, PBRM1, TTN, SETD2, BAP1, MTOR, MUC16, DNAH9, KDM5C, and LRP2. (**Figure 9A and B**). Among these genes, VHL, PBRM1, SETD2, BAP1, KDM5C, and MTOR were the most frequently mutated genes in ccRCC. However, in general, there was no significant difference in TMB between the two groups (**Figure 9C**). In addition, patients in the high TMB and high-risk cohorts had the worst prognosis than the other groups (**Figures 9D and E**). Compared to the low-risk group, the TIDE scores were dramatically higher in the high-risk group (**Figure 9F**). By comparing drug sensitivity, we found significant differences in IC50 values between the low- and high-risk groups for multiple drugs. Drugs sensitive to the high-risk group and drugs sensitive to the low-risk group are shown in **Supplementary Figures S2 and S3**, respectively. Of these drugs, sorafenib was more effective in

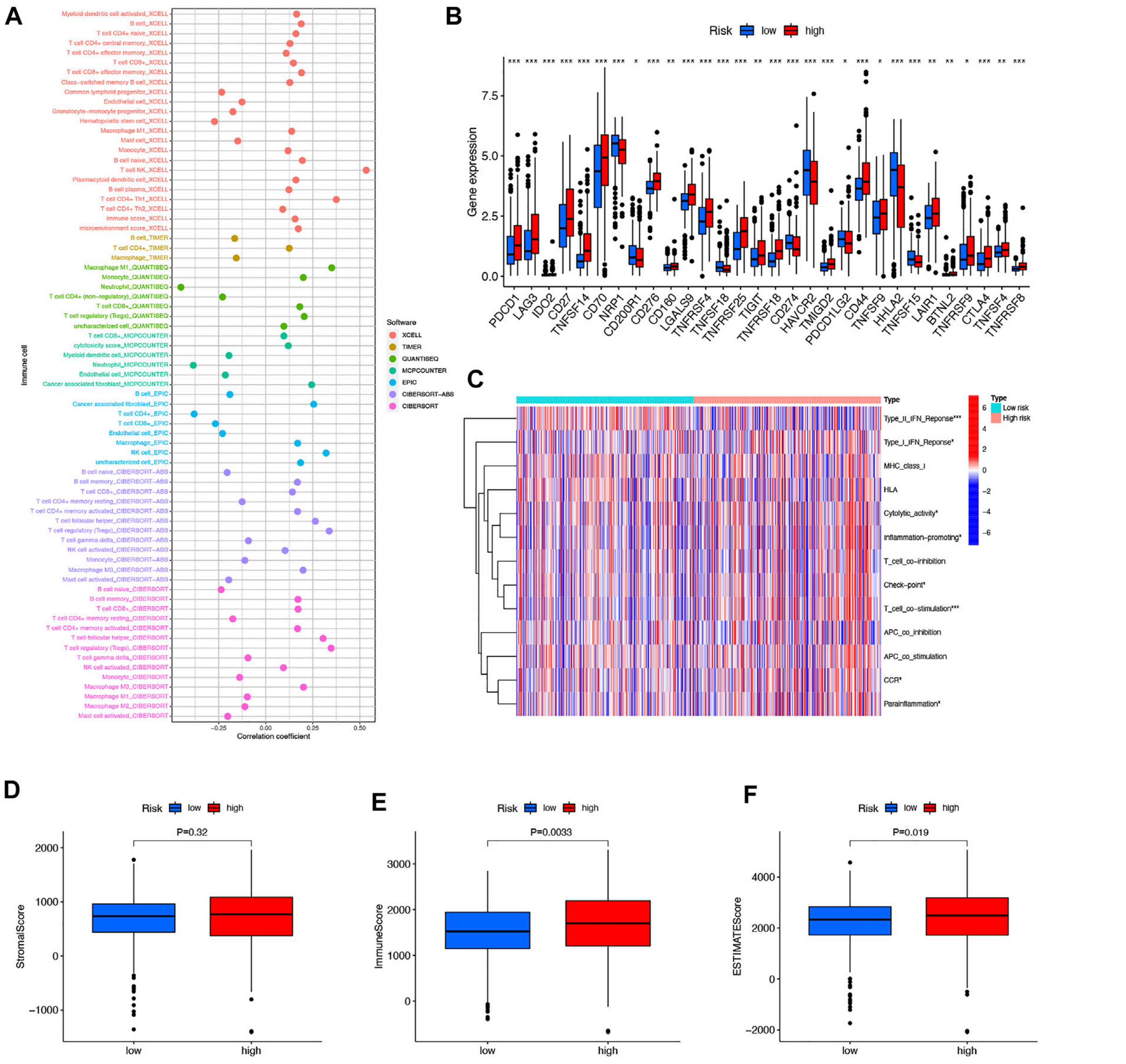


FIGURE 8 | Differences in the tumor immune microenvironment between the low- and high-risk groups. Immune cell bubble of risk groups **(A)**. Differences in expression of common immune checkpoints in the risk groups **(B)**. ssGSEA scores of immune cells and immune function in the risk group **(C)**. Box plots comparing StromalScore, ImmuneScore and ESTIMATEScore between the low- and high-risk groups, respectively **(D–F)**. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

the high-risk group and, conversely, pazopanib was more effective in the low-risk group (**Figure 10A and B**).

External Validation of Cuproptosis-Related lncRNAs as a Potential Biomarker

Then, the KM survival analysis was utilized to verify the prognostic value of SNHG15 and LINC00471 in the external Kaplan–Meier Plotter database. The results showed that SNHG15, as a poor prognostic factor, was dramatically correlated with OS (HR = 2.46 (1.79–3.39), Log-rank p = 1.1e-08) (**Figure 11A**). LINC00471, an indicator of bad prognosis, was also significantly associated with OS (HR = 1.6 (1.18–2.15), Log-rank p = 0.002) (**Figure 11B**). The results of the survival analysis of external datasets were consistent with our outcomes.

In Vitro Experimental Validation of Cuproptosis-Related lncRNAs as a Potential Biomarker

To further validate the prognostic value of this cuproptosis death-associated lncRNA model, our team performed *in vitro* experiments to illustrate the expression trends of hub differentially expressed cuproptosis-associated lncRNAs. RT-qPCR results indicated an overall trend of increased SNHG15 and LINC00471 expression levels in ccRCC tissues compared to adjacent paired normal tissues,

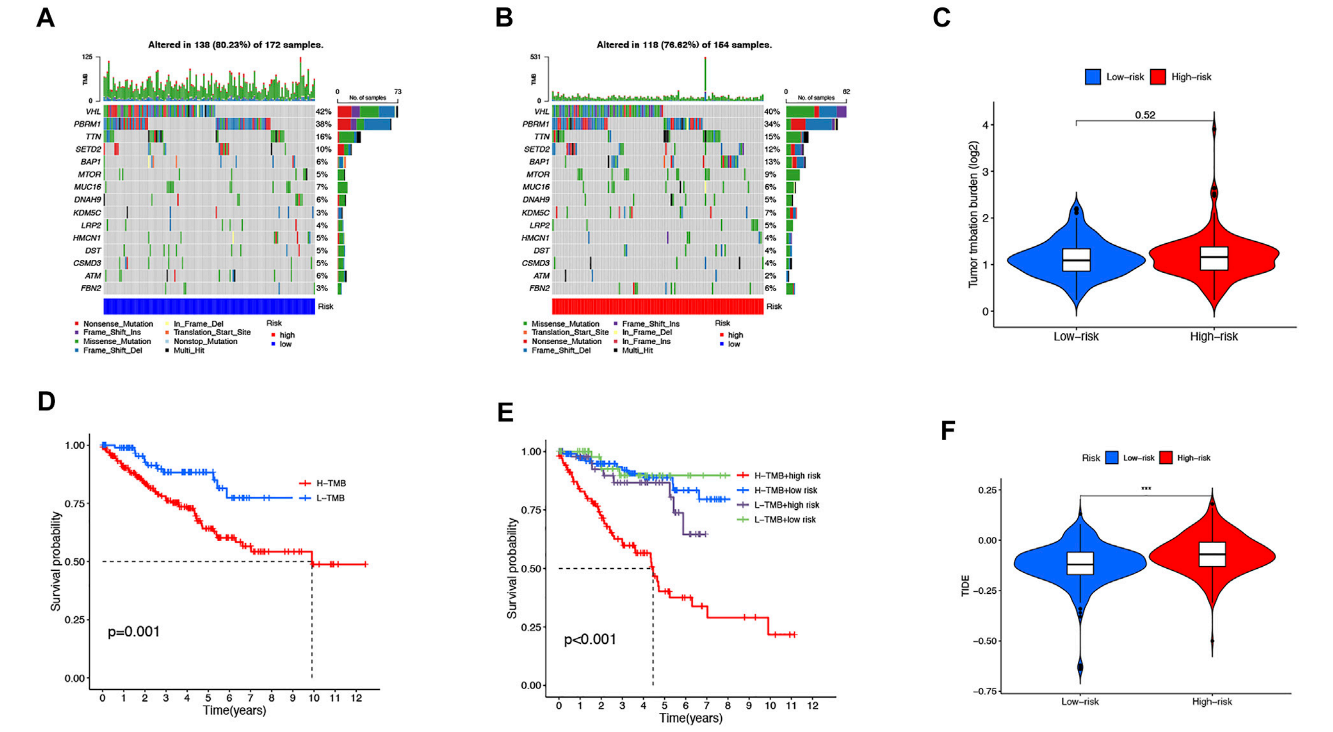


FIGURE 9 | TMB, TIDE, and Chemotherapeutic Sensitivity. Waterfall plots of somatic mutation characteristics in the two groups **(A-B)**. TMB between the low-risk and high-risk groups **(C)**. K–M survival curves between the high- and low-TMB groups **(D)**. K–M survival curves between the four groups **(E)**. TIDE scores between the two groups **(F)**.

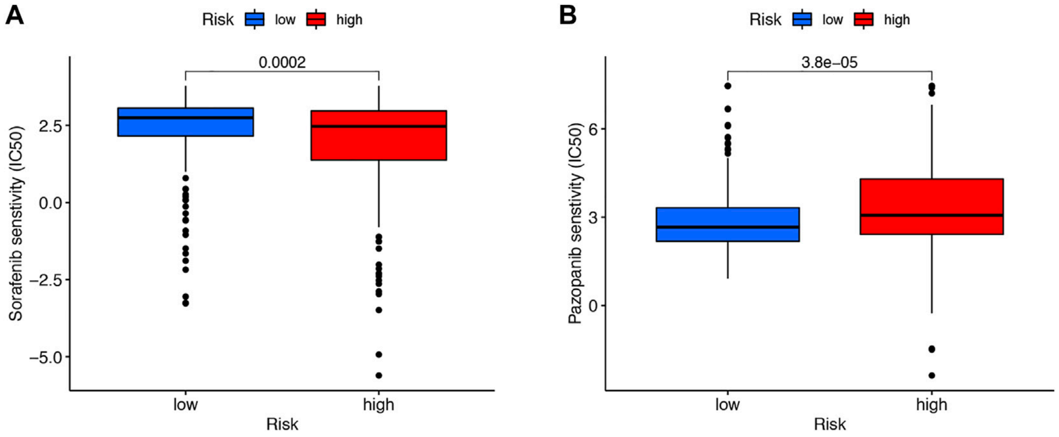


FIGURE 10 | Drug sensitivity. Sorafenib was more effective in the high-risk group **(A)**. Pazopanib was more effective in the low-risk group **(B)**.

which matched the results of our previous bioinformatics analysis based on public databases (**Figure 12A and B**).

DISCUSSION

CCRCC, as the most aggressive subtype, is the predominant histological type of renal cancer.

Although surgical resection is a moderate treatment option for localized ccRCC, the outcome of advanced or metastatic ccRCC remains dissatisfactory. Therefore, the identification of prospective prognostic and molecular signatures specific to patients with ccRCC is essential to improve the patient's prognosis.

Recent studies have shown that intracellular copper accumulation triggers the aggregation of mitochondrial

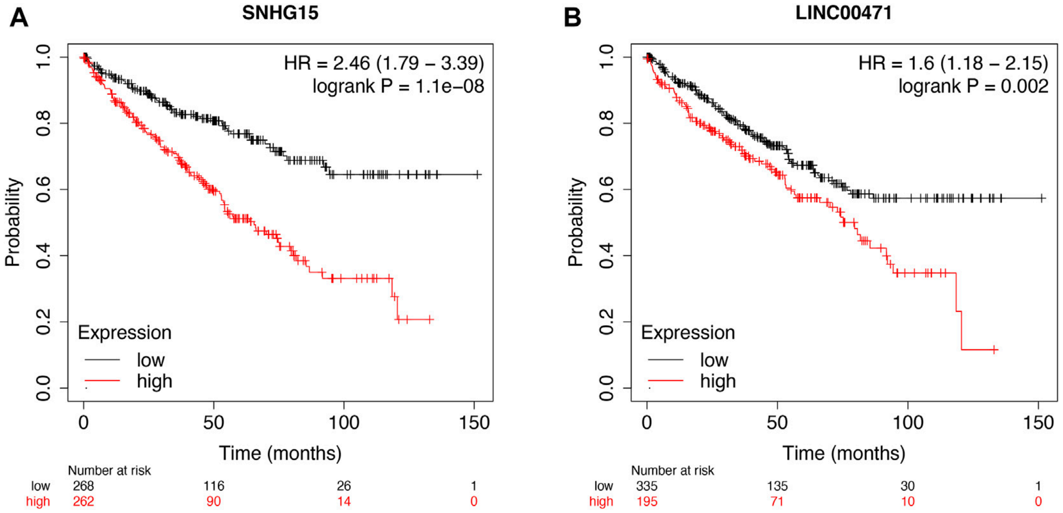


FIGURE 11 | External validation of cuproptosis-associated lncRNAs as potential biomarkers. OS analysis of SNHG15 and LINC00471 in the Kaplan–Meier Plotter datasets **(A** and **B)**.

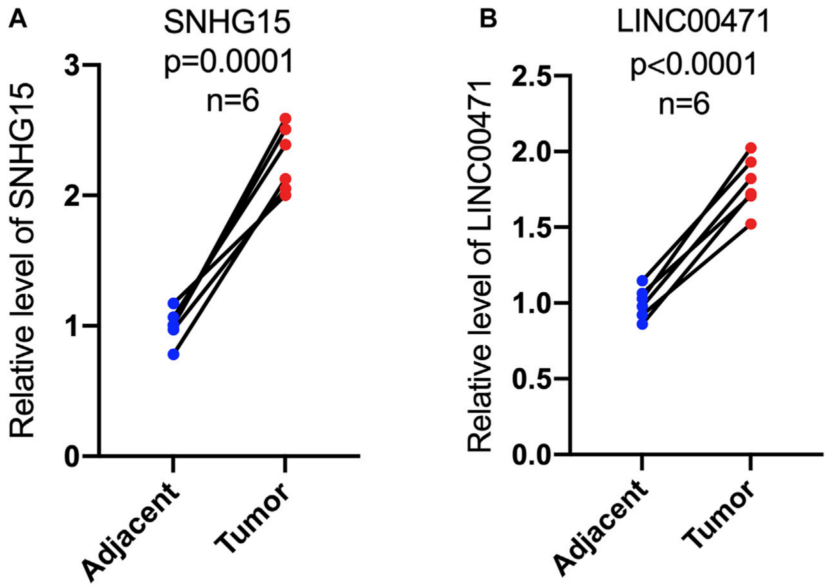


FIGURE 12 | Expression levels of cuproptosis-associated lncRNAs in paired tumor tissues. RT-qPCR was used to measure the expression of SNHG15 and LINC00471 in paired tumor tissues **(A,B)**.

lipid acylated proteins and the loss of Fe–S cluster proteins, resulting in a proteotoxic stress-induced death called cuproptosis (Tsvetkov et al., 2022). Significantly, the accumulation of intracellular copper is dependent on the transport of copper ionophores. Therefore, copper ionophores are a powerful tool for studying copper toxicity (Hunsaker and Franz, 2019). Traditional cancer treatments usually harm normal cells, so novel therapeutic agents are being developed with the aim of improving selectivity and, thus, reducing side effects. In addition, these agents should target cancer stem cells, thus, overcoming the resistance of cancer cells. Cancer cells are usually preferentially induced by cuproptosis compared to normal cells, and some copper ionophores have shown promise in this direction (Li, 2020; Steinbrueck et al., 2020; Babak and Ahn, 2021; Michniewicz et al., 2021; Shanbhag et al., 2021). Therefore, cuproptosis-related studies are urgently needed for a deeper understanding.

Previous studies have shown that lncRNAs play an important regulatory role in the development and progression of ccRCC.

Professor Liu confirmed that LINC01232 promotes clear cell renal cell carcinoma by binding miR-204-5p to upregulate RAB22A (Liu et al., 2021). Lv noted that the long non-coding RNA TUG1 promotes cell proliferation through the MIR-31-5p/FLOT1 axis in clear cell renal cell carcinoma and inhibits apoptosis and autophagy (Lv et al., 2020). However, lncRNAs associated with cuproptosis have never been studied in ccRCC. Here, our team constructed a cuproptosis-associated lncRNA signature to predict the prognostic status of ccRCC patients. In our research, we obtained 81 cuproptosis-related lncRNAs associated with prognosis by analysis. We screened and identified 10 cuproptosis-related lncRNAs significantly associated with OS by univariate, LASSO, and multivariate Cox regression analysis (HHLA3, H1-10-AS1, PICSAR, LINC02027, SNHG15, SNHG8, LINC00471, EIF1B-AS1, LINC02154, and MINCR). With the aforementioned lncRNAs, we constructed cuproptosis-related lncRNA features to predict the prognosis of ccRCC patients. Among these lncRNAs, the lncRNA PICSAR was reported to be highly expressed in tumors and could promote proliferation and migration and inhibit apoptosis in cutaneous squamous cell carcinoma and hepatocellular carcinoma (Liu et al., 2020; Lu et al., 2021). LINC02027 was an important member of the ccRCC prognostic model (Chen et al., 2022). LncRNA SNHG15 was a novel lncRNA identified as a tumor promoter in various human cancers, including hepatocellular carcinoma (HCC), colorectal cancer (CRC), breast cancer (BRCA), pancreatic cancer (PC), gastric cancer (GC), and clear cell carcinoma (ccRCC) (Guo et al., 2018; Jin et al., 2018; Kong and Qiu, 2018; Huang et al., 2019; Yang et al., 2020; Chen et al., 2021). Studies in ccRCC have shown that increased expression of lncRNA SNHG15 was an independent predictor of shorter RFS. In addition, SNHG15 expression levels were significantly regulated by DNA methylation in ccRCC (Yang et al., 2020). All findings suggested that SNHG15 was promising as a biomarker and therapeutic target for cancer patients. Similarly, SNHG8 was considered to be an oncogenic factor and was upregulated in various types of cancer (Yuan et al., 2021), such as gastric cancer, melanoma, nasopharyngeal cancer, and esophageal cancer (Shan et al., 2022b; Luan et al., 2022; Wu et al., 2022; Zhu et al., 2022). LINC00471 was an essential member of the prognostic model of childhood acute myeloid leukemia and esophageal squamous cell carcinoma (Zhang et al., 2019a; Yu et al., 2019). LINC02154 was involved in the construction of a prognostic model for laryngeal squamous cell carcinoma (Zhang et al., 2019b; Gong et al., 2020). MINCR was highly expressed in nasopharyngeal, colon, non-small cell lung cancers, and hepatocellular carcinoma and promotes cancer development (Cao et al., 2018; Chen et al., 2019; Yu et al., 2020; Zhong et al., 2020). The remaining three lncRNAs (HHLA3, H1-10-AS1, and EIF1B-AS1) are the first publicly available. In particular, these newly discovered cuproptosis-related lncRNAs can help us better understand ccRCC and find new targets for cancer therapy. We then divided patients with ccRCC into low-risk and high-risk cohorts according to median values. The Roc and c-index curves were used to validate the prognostic accuracy of the risk score. We could find that the risk score could be used as a criterion to predict the prognosis. Then, we constructed a nomogram to predict the prognosis of patients with ccRCC. The calibration curves showed excellent agreement between actual results and predictions. Then the PCA results showed that the 10 cuproptosis-associated lncRNAs had the best ability to discriminate well between low- and high-risk populations. GO analysis suggested that immune responses were strongly associated with lncRNAs associated with cuproptosis. KEGG analysis showed that cytokine–cytokine receptor interactions and the PI3K-AKT signaling pathway were most active in cuproptosis-associated lncRNAs. The PI3K-Akt signaling pathway was widely present in a variety of cells and can be involved in cell proliferation, apoptosis, invasion, metastasis, and angiogenesis by altering the activation status of downstream signaling molecules, which had been regarded by scientists as the primary pathway for cancer cell survival (Polivka and Janku, 2014). Normally, immune cell infiltration in the tumor microenvironment varies with tumor progression. Sierra et al. (2021) found *in vitro* experiments that an increase in NK cells suppressed the proliferation of CD8$^+$ T cells and suggested that infiltration of NK cells impairs the immune regulatory function of the body. A study showed that T cell follicular helper cells, T cell regulation, and B cell memory were associated with adverse outcomes of ccRCC (Yu et al., 2020). The characteristics of the high-risk group we established were highly consistent with the aforementioned study and predicted a poorer prognosis for the high-risk group. Furthermore, the results of ssGSEA pointed to an immune profile of type II inactivation of the IFN response and activation of T cell co-stimulation in high-risk populations. These results suggested that our features may be involved in the tumor immune microenvironment of ccRCC, acting by blocking the immune response, and may be a factor in the progression of ccRCC. We also performed immune scores, stromal scores, and ESTIMATE scores on different subgroups of the population, resulting in higher-risk groups having higher immune scores and lower tumor purity. As previously reported, the TIDE algorithm was used to assess the clinical response of patients to ICI therapy; the higher the TIDE score, the greater the likelihood of immune escape, which may imply a limited response and shorter survival time for patients treated with ICI. Compared to the low-risk group, patients in the high-risk group had higher TIDE scores, suggesting that patients in the high-risk group may have a more limited response to ICI therapy. Previous clinical trials have confirmed that the benefits of pazopanib are more prominent in the low-risk group, which is consistent with our study (Méndez-Vidal et al., 2018). However, in the case of sorafenib, there is no evidence in the literature that it is more beneficial in the high-risk group, and the exact mechanism remains to be confirmed by more studies. Our team constructed 10 copper death-associated lncRNAs to predict the prognosis of patients with ccRCC through adequate bioinformatics analysis. However, our study still had some drawbacks and shortcomings. First, we could not get validation from the GEO and ICGC databases. Even though we tried the GEO and ICGC databases, we still could not obtain proper lncRNA information due to the bias and limitation of commercial microarray data compared with GTEx and TCGA. Therefore, we validated the potential ability of two of these lncRNAs as biomarkers by PCR together with the external database Kaplan–Meier Plotter database. In addition, the immune cell bubble plots showed the results of immune infiltration from multiple platforms, which in a sense can be considered as external validation. In addition, our team will subsequently collect additional clinical datasets to validate the value of cuproptosis-associated lncRNAs.

CONCLUSION

The 10 cuproptosis-related-associated lncRNA risk profiles may help to assess the prognosis and molecular profile of ccRCC patients and improve treatment options, which may be further applied in the clinic.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

SX, DL, TC, XW, and SM studied and designed all bioinformatics analysis. SX, SM, LW, GS, and SC critically reviewed the manuscript. SX, DL, TC, and XW worked together on the manuscript. Administrative, technical, and material support were provided by YX and HZ. All authors approved the final version of the manuscript. SX, DL, TC, and XW contributed equally to our research.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (81972412 and 81772758) and the Science and Technology Foundation of Tianjin Municipal Health Care Commission (KJ20051).

ACKNOWLEDGMENTS

The authors are grateful for the financial support from HZ. The authors also thank TCGA databases for the availability of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.938259/full#supplementary-material

REFERENCES

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18 (1), 220. doi:10.1186/s13059-017-1349-1

Babak, M. V., and Ahn, D. (2021). Modulation of Intracellular Copper Levels as the Mechanism of Action of Anticancer Copper Complexes: Clinical Relevance. *Biomedicines* 9 (8). doi:10.3390/biomedicines9080852

Barik, G. K., Sahay, O., Behera, A., Naik, D., and Kalita, B. (2021). Keep Your Eyes Peeled for Long Noncoding RNAs: Explaining Their Boundless Role in Cancer Metastasis, Drug Resistance, and Clinical Application. *Biochimica Biophysica Acta (BBA) - Rev. Cancer* 1876 (2), 188612. doi:10.1016/j.bbcan.2021.188612

Cao, J., Zhang, D., Zeng, L., and Liu, F. (2018). Long Noncoding RNA MINCR Regulates Cellular Proliferation, Migration, and Invasion in Hepatocellular Carcinoma. *Biomed. Pharmacother.* 102, 102–106. doi:10.1016/j.biopha.2018.03.041

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* 1711, 243–259. doi:10.1007/978-1-4939-7493-1_12

Chen, S., Gu, T., Lu, Z., Qiu, L., Xiao, G., Zhu, X., et al. (2019). Roles of MYC-Targeting Long Non-coding RNA MINCR in Cell Cycle Regulation and Apoptosis in Non-small Cell Lung Cancer. *Respir. Res.* 20 (1), 202. doi:10.1186/s12931-019-1174-z

Chen, X., Tu, J., Ma, L., Huang, Y., Yang, C., and Yuan, X. (2022). Analysis of Ferroptosis-Related LncRNAs Signatures Associated with Tumor Immune Infiltration and Experimental Validation in Clear Cell Renal Cell Carcinoma. *Ijgm* 15, 3215–3235. doi:10.2147/ijgm.s354682

Chen, Z., Zhong, T., Li, T., Zhong, J., Tang, Y., Liu, Z., et al. (2021). LncRNA SNHG15 Modulates Gastric Cancer Tumorigenesis by Impairing miR-506-5p Expression. *Biosci. Rep.* 41 (7). doi:10.1042/bsr20204177

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A Survey of Best Practices for RNA-Seq Data Analysis. *Genome Biol.* 17, 13. doi:10.1186/s13059-016-0881-8

Delman, K. A. (2020). Introducing the "Virtual Tumor Board" Series in CA: A Cancer Journal for Clinicians. *CA A Cancer J. Clin.* 70 (2), 77. doi:10.3322/caac.21598

Dienstmann, R., Villacampa, G., Sveen, A., Mason, M. J., Niedzwiecki, D., Nesbakken, A., et al. (2019). Relative Contribution of Clinicopathological Variables, Genomic Markers, Transcriptomic Subtyping and Microenvironment Features for Outcome Prediction in Stage II/III Colorectal Cancer. *Ann. Oncol.* 30 (10), 1622–1629. doi:10.1093/annonc/mdz287

Du, X.-h., Wei, H., Qu, G.-x., Tian, Z.-c., Yao, W.-t., and Cai, Q.-q. (2020). Gene Expression Regulations by Long Noncoding RNAs and Their Roles in Cancer. *Pathology - Res. Pract.* 216 (6), 152983. doi:10.1016/j.prp.2020.152983

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., et al. (2019). Molecular and Pharmacological Modulators of the Tumor Immune Contexture Revealed by Deconvolution of RNA-Seq Data. *Genome Med.* 11 (1), 34. doi:10.1186/s13073-019-0638-6

Gao, J., Wang, F., Wu, P., Chen, Y., and Jia, Y. (2020). Aberrant LncRNA Expression in Leukemia. *J. Cancer* 11 (14), 4284–4296. doi:10.7150/jca.42093

Ge, E. J., Bush, A. I., Casini, A., Cobine, P. A., Cross, J. R., DeNicola, G. M., et al. (2022). Connecting Copper and Cancer: from Transition Metal Signalling to Metalloplasia. *Nat. Rev. Cancer* 22 (2), 102–113. doi:10.1038/s41568-021-00417-2

Gong, S., Xu, M., Zhang, Y., Shan, Y., and Zhang, H. (2020). The Prognostic Signature and Potential Target Genes of Six Long Non-coding RNA in

Laryngeal Squamous Cell Carcinoma. *Front. Genet.* 11, 413. doi:10.3389/fgene.2020.00413

Guo, X. B., Yin, H. S., and Wang, J. Y. (2018). Evaluating the Diagnostic and Prognostic Value of Long Non-coding RNA SNHG15 in Pancreatic Ductal Adenocarcinoma. *Eur. Rev. Med. Pharmacol. Sci.* 22 (18), 5892–5898. doi:10.26355/eurrev_201809_15917

Huang, L., Lin, H., Kang, L., Huang, P., Huang, J., Cai, J., et al. (2019). Aberrant Expression of Long Noncoding RNA SNHG15 Correlates with Liver Metastasis and Poor Survival in Colorectal Cancer. *J. Cell. Physiology* 234 (5), 7032–7039. doi:10.1002/jcp.27456

Hunsaker, E. W., and Franz, K. J. (2019). Emerging Opportunities to Manipulate Metal Trafficking for Therapeutic Benefit. *Inorg. Chem.* 58 (20), 13528–13545. doi:10.1021/acs.inorgchem.9b01029

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T Cell Dysfunction and Exclusion Predict Cancer Immunotherapy Response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Jin, B., Jin, H., Wu, H. B., Xu, J. J., and Li, B. (2018). Long Non-coding RNA SNHG15 Promotes CDK14 Expression via miR-486 to Accelerate Non-small Cell Lung Cancer Cells Progression and Metastasis. *J. Cell Physiol.* 233 (9), 7164–7172. doi:10.1002/jcp.26543

Kong, Q., and Qiu, M. (2018). Long Noncoding RNA SNHG15 Promotes Human Breast Cancer Proliferation, Migration and Invasion by Sponging miR-211-3p. *Biochem. Biophysical Res. Commun.* 495 (2), 1594–1600. doi:10.1016/j.bbrc.2017.12.013

Lelièvre, P., Sancey, L., Coll, J. L., Deniaud, A., and Busser, B. (2020). The Multifaceted Roles of Copper in Cancer: A Trace Metal Element with Dysregulated Metabolism, but Also a Target or a Bullet for Therapy. *Cancers (Basel)* 12 (12). doi:10.3390/cancers12123594

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77 (21), e108–e110. doi:10.1158/0008-5472.can-17-0307

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for Analysis of Tumor-Infiltrating Immune Cells. *Nucleic Acids Res.* 48 (W1), W509–W514. doi:10.1093/nar/gkaa407

Li, Y. (2020). Copper Homeostasis: Emerging Target for Cancer Treatment. *IUBMB Life* 72 (9), 1900–1908. doi:10.1002/iub.2341

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 173 (2), 400–e11. doi:10.1016/j.cell.2018.02.052

Liu, Q., and Lei, C. (2021). LINC01232 Serves as a Novel Biomarker and Promotes Tumour Progression by Sponging miR-204-5p and Upregulating RAB22A in Clear Cell Renal Cell Carcinoma. *Ann. Med.* 53 (1), 2153–2164. doi:10.1080/07853890.2021.2001563

Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y., and Maher, C. A. (2021). Long Noncoding RNAs in Cancer Metastasis. *Nat. Rev. Cancer* 21 (7), 446–460. doi:10.1038/s41568-021-00353-1

Liu, Z., Mo, H., Sun, L., Wang, L., Chen, T., Yao, B., et al. (2020). Long Noncoding RNA PICSAR/miR-588/EIF6 axis Regulates Tumorigenesis of Hepatocellular Carcinoma by Activating PI3K/AKT/mTOR Signaling Pathway. *Cancer Sci.* 111 (11), 4118–4128. doi:10.1111/cas.14631

Ljungberg, B., Bensalah, K., Canfield, S., Dabestani, S., Hofmann, F., Hora, M., et al. (2015). EAU Guidelines on Renal Cell Carcinoma: 2014 Update. *Eur. Urol.* 67 (5), 913–924. doi:10.1016/j.eururo.2015.01.005

Lu, X., Gan, Q., Gan, C., Zheng, Y., Cai, B., Li, X., et al. (2021). Long Non-coding RNA PICSAR Knockdown Inhibits the Progression of Cutaneous Squamous Cell Carcinoma by Regulating miR-125b/YAP1 axis. *Life Sci.* 274, 118303. doi:10.1016/j.lfs.2020.118303

Luan, Q., Yang, R., Lin, L., and Li, X. (2022). SNHG8 Promotes Cell Proliferation, Migration, and Invasion of Nasopharyngeal Carcinoma Cells as an Oncogene through miR-588/HMGA2 axis. *Can. J. Physiol. Pharmacol.* 100 (2), 158–166. doi:10.1139/cjpp-2021-0149

Lv, D., Xiang, Y., Yang, Q., Yao, J., and Dong, Q. (2020). Long Non-coding RNA TUG1 Promotes Cell Proliferation and Inhibits Cell Apoptosis, Autophagy in Clear Cell Renal Cell Carcinoma via MiR-31-5p/FLOT1 Axis. *Ott* 13, 5857–5868. doi:10.2147/ott.s254634

Méndez-Vidal, M. J., Molina, Á., Anido, U., Chirivella, I., Etxaniz, O., Fernández-Parra, E., et al. (2018). Pazopanib: Evidence Review and Clinical Practice in the Management of Advanced Renal Cell Carcinoma. *BMC Pharmacol. Toxicol.* 19 (1), 77. doi:10.1186/s40360-018-0264-8

Michniewicz, F., Saletta, F., Rouaen, J. R. C., Hewavisenti, R. V., Mercatelli, D., Cirillo, G., et al. (2021). Copper: An Intracellular Achilles' Heel Allowing the Targeting of Epigenetics, Kinase Pathways, and Cell Metabolism in Cancer Therapeutics. *ChemMedChem* 16 (15), 2315–2329. doi:10.1002/cmdc.202100172

Motzer, R. J., Penkov, K., Haanen, J., Rini, B., Albiges, L., Campbell, M. T., et al. (2019). Avelumab Plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* 380 (12), 1103–1115. doi:10.1056/nejmoa1816047

Polivka, J., Jr., and Janku, F. (2014). Molecular Targets for Cancer Therapy in the PI3K/AKT/mTOR Pathway. *Pharmacol. Ther.* 142 (2), 164–175. doi:10.1016/j.pharmthera.2013.12.004

Quinn, J. J., and Chang, H. Y. (2016). Unique Features of Long Non-coding RNA Biogenesis and Function. *Nat. Rev. Genet.* 17 (1), 47–62. doi:10.1038/nrg.2015.10

Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data. *Elife* 6. doi:10.7554/eLife.26476

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 (7), e47. doi:10.1093/nar/gkv007

Ruiz, L. M., Libedinsky, A., and Elorza, A. A. (2021). Role of Copper on Mitochondrial Function and Metabolism. *Front. Mol. Biosci.* 8, 711227. doi:10.3389/fmolb.2021.711227

Shan, B., Qu, S., Lv, S., Fan, D., and Wang, S. (2022). YY1-induced Long Non-coding RNA Small Nucleolar RNA Host Gene 8 Promotes the Tumorigenesis of Melanoma via the microRNA-656-3p/SERPINE1 mRNA Binding Protein 1 axis. *Bioengineered* 13 (3), 4832–4843. doi:10.1080/21655979.2022.2034586

Shan, G., Huang, T., and Tang, T. (2022). Long Non-coding RNA MEG8 Induced by PLAG1 Promotes Clear Cell Renal Cell Carcinoma through the miR-495-3p/G3BP1 axis. *Pathology - Res. Pract.* 229, 153734. doi:10.1016/j.prp.2021.153734

Shanbhag, V. C., Gudekar, N., Jasmer, K., Papageorgiou, C., Singh, K., and Petris, M. J. (2021). Copper Metabolism as a Unique Vulnerability in Cancer. *Biochimica Biophysica Acta (BBA) - Mol. Cell Res.* 1868 (2), 118893. doi:10.1016/j.bbamcr.2020.118893

Sierra, J. M., Secchiari, F., Nuñez, S. Y., Iraolagoitia, X. L. R., Ziblat, A., Friedrich, A. D., et al. (2021). Tumor-Experienced Human NK Cells Express High Levels of PD-L1 and Inhibit CD8+ T Cell Proliferation. *Front. Immunol.* 12, 745939. doi:10.3389/fimmu.2021.745939

Steinbrueck, A., Sedgwick, A. C., Brewster, J. T., 2nd, Yan, K.-C., Shang, Y., Knoll, D. M., et al. (2020). Transition Metal Chelators, Pro-chelators, and Ionophores as Small Molecule Cancer Chemotherapeutic Agents. *Chem. Soc. Rev.* 49 (12), 3726–3747. doi:10.1039/c9cs00373h

Tamminga, M., Hiltermann, T. J. N., Schuuring, E., Timens, W., Fehrmann, R. S., and Groen, H. J. (2020). Immune Microenvironment Composition in Non-small Cell Lung Cancer and its Association with Survival. *Clin. Transl. Immunol.* 9 (6), e1142. doi:10.1002/cti2.1142

Tsvetkov, P., Coy, S., Petrova, B., Dreishpoon, M., Verma, A., Abdusamad, M., et al. (2022). Copper Induces Cell Death by Targeting Lipoylated TCA Cycle Proteins. *Science* 375 (6586), 1254–1261. doi:10.1126/science.abf0529

Wu, Y., Liang, Y., Li, M., and Zhang, H. (2022). Knockdown of Long Non-coding RNA SNHG8 Suppresses the Progression of Esophageal Cancer by Regulating miR-1270/BACH1 axis. *Bioengineered* 13 (2), 3384–3394. doi:10.1080/21655979.2021.2021064

Yang, W., Zhang, K., Li, L., Ma, K., Hong, B., Gong, Y., et al. (2020). Discovery and Validation of the Prognostic Value of the lncRNAs Encoding snoRNAs in Patients with Clear Cell Renal Cell Carcinoma. *Aging* 12 (5), 4424–4444. doi:10.18632/aging.102894

Yu, J., Wu, X., Huang, K., Zhu, M., Zhang, X., Zhang, Y., et al. (2019). Bioinformatics Identification of lncRNA Biomarkers Associated with the Progression of Esophageal Squamous Cell Carcinoma. *Mol. Med. Rep.* 19 (6), 5309–5320. doi:10.3892/mmr.2019.10213

Yu, Y., Chang, Z., Han, C., Zhuang, L., Zhou, C., Qi, X., et al. (2020). Long Noncoding RNA MINCR Aggravates Colon Cancer via Regulating miR-708-5p-Mediated Wnt/β-Catenin Pathway. *Biomed. Pharmacother.* 129, 110292. doi:10.1016/j.biopha.2020.110292

Yuan, X., Yan, Y., and Xue, M. (2021). Small Nucleolar RNA Host Gene 8: A Rising Star in the Targets for Cancer Therapy. *Biomed. Pharmacother.* 139, 111622. doi:10.1016/j.biopha.2021.111622

Zhang, G., Fan, E., Zhong, Q., Feng, G., Shuai, Y., Wu, M., et al. (2019). Identification and Potential Mechanisms of a 4-lncRNA Signature that Predicts Prognosis in Patients with Laryngeal Cancer. *Hum. Genomics* 13 (1), 36. doi:10.1186/s40246-019-0230-6

Zhang, N., Chen, Y., Shen, Y., Lou, S., and Deng, J. (2019). Comprehensive Analysis the Potential Biomarkers for the High-Risk of Childhood Acute Myeloid Leukemia Based on a Competing Endogenous RNA Network. *Blood Cells, Mol. Dis.* 79, 102352. doi:10.1016/j.bcmd.2019.102352

Zhang, Z., Fu, X., Gao, Y., and Nie, Z. (2022). LINC01535 Attenuates ccRCC Progression through Regulation of the miR-146b-5p/TRIM2 Axis and Inactivation of the PI3K/Akt Pathway. *J. Oncol.* 2022, 2153337. doi:10.1155/2022/2153337

Zhong, Q., Chen, Y., and Chen, Z. (2020). RETRACTED ARTICLE: LncRNA MINCR Regulates Irradiation Resistance in Nasopharyngeal Carcinoma Cells via the microRNA-223/ZEB1 axis. *Cell Cycle* 19 (1), 53–66. doi:10.1080/15384101.2019.1692176

Zhu, W., Tan, L., Ma, T., Yin, Z., and Gao, J. (2022). Long Noncoding RNA SNHG8 Promotes Chemoresistance in Gastric Cancer via Binding with hnRNPA1 and Stabilizing TROY Expression. *Dig. Liver Dis.*, S1590-S8658(22)00202-X. doi:10.1016/j.dld.2022.02.011

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 September 2022
DOI 10.3389/fgene.2022.970933

Causal relationship between bipolar disorder and inflammatory bowel disease: A bidirectional two-sample mendelian randomization study

Zhe Wang[1], Xinyu Wang[1], Xushi Zhao[1], Zhaoliang Hu[1], Dongwei Sun[2], Donglei Wu[1] and Yanan Xing[1]*

[1]Department of Surgical Oncology, Department of General Surgery, First Affiliated Hospital, China Medical University, Shenyang, China, [2]Department of International Special Medical Center, First Affiliated Hospital, China Medical University, Shenyang, China

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Lan Luo,
Merck, United States
Dalin Li,
Cedars Sinai Medical Center,
United States

*CORRESPONDENCE
Yanan Xing,
xingyanan8124@163.com,
orcid.org/0000-0002-9944-7675

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

RECEIVED 16 June 2022
ACCEPTED 29 August 2022
PUBLISHED 20 September 2022

CITATION
Wang Z, Wang X, Zhao X, Hu Z, Sun D, Wu D and Xing Y (2022), Causal relationship between bipolar disorder and inflammatory bowel disease: A bidirectional two-sample mendelian randomization study.
Front. Genet. 13:970933.
doi: 10.3389/fgene.2022.970933

Background: Growing evidence suggests a bidirectional association between bipolar disorder (BD) and inflammatory bowel disease (IBD); however, observational studies are prone to confounding, making causal inference and directional determination of these associations difficult.

Methods: We performed bidirectional two-sample Mendelian randomization (MR) and selected single nucleotide polymorphisms (SNPs) associated with BD and IBD as instrumental variables (IV). SNPs and genetic associations with BD and IBD were obtained from the latest genome-wide association studies (GWAS) in Europeans (BD: cases/controls: 20352/31358; IBD: 12882/21770; Crohn's disease (CD): 5,956/14927; ulcerative colitis (UC): 6968/20464). The inverse-variance-weighted method was the major method used in MR analyses. MR-Egger, weight mode, simple mode, and weighted median were used for quality control.

Results: Genetically predicted BD (per log-odds ratio increase) was significantly positively associated with risk of IBD (OR: 1.18, 95% CI: 1.04–1.33), and UC (OR = 1.19, 95% CI: 1.05–1.35), but not CD (OR = 1.18, 95% CI: 0.95–1.48). The validation analysis found that combined OR of IBD, CD, and UC increased per log-OR of BD were 1.16(95% CI: 1.02–1.31), 1.20(95% CI: 0.98–1.48) 1.17(95% CI: 1.02–1.35), respectively. In contrast, no causal relationship was identified between genetically influenced IBD and BD.

Conclusion: Our results confirm a causal relationship between BD and IBD, which may influence clinical decisions on the management of BD patients with intestinal symptoms. Although the reverse MR results did not support a causal effect of IBD on BD, the effect of the IBD active period on BD remains to be further investigated.

KEYWORDS

inflammatory bowel disease, mendelian randomization (MR), bipolar disorder, causal relationship, crohn's disease, ulcerative colitis

Introduction

Inflammatory bowel disease (IBD) causes a high disease burden worldwide and comprises of two major diseases, Crohn's disease (CD) and ulcerative colitis (UC), both characterized by visible chronic and progressive intestinal inflammation, weight loss, diarrhea, and gastrointestinal bleeding (Ng et al. 2017, Le Berre et al. 2021). Interactions between genetic predisposition and environmental risk factors including poor dietary habits, antibiotic exposure, smoking, major social stressors, and unfavorable lifestyle are thought to be the main pathogenesis of IBD, as they may contribute to improper intestinal immune activation and disrupt the proinflammatory microbiome (Ananthakrishnan 2015, Piovani et al. 2019, Ramos and Papadakis 2019). However, studies have shown that psycho-neuro-endocrine-immune regulation via the brain-gut axis may also lead to abnormal activation of gut immunity and alter pro-inflammatory flora, suggesting a role in IBD pathogenesis (Bonaz and Bernstein 2013, Gracie et al. 2019). Several recent studies have indicated that individuals with mood disorders may be affected by inflammatory changes in the gut, particularly during the manic and psychotic phases of the disease (Severance et al. 2010, Bernstein et al. 2019, Marrie et al. 2019).

Bipolar disorder (BD), a chronic psychiatric disorder characterized by intermittent mania, depression, or mixed mood states, is an important manifestation of mood disorders (Ferrari et al. 2016). Data from a serological and gene expression study suggests that inflammation may be an important pathology in BD patients (Severance et al. 2014). Another study found that sTNF-R1, IL-1Ra, OPG, and IL-6 were significantly altered in the affective state and that they correlated with the severity of affective symptoms in BD patients (Hope et al. 2011). BD has a high heritability (about 70%), and nonpsychiatric comorbidities (including IBD) are prevalent in patients with BD (Vieta et al. 2018, McIntyre et al. 2020). The increased prevalence of BD in patients with IBD has led to a growing number of studies investigating potential associations between IBD and BD (Severance et al. 2014, Bernstein et al. 2019, Nikolova et al. 2022). For instance, in a cross-sectional study involving over 1.5 million people in the UK, people with BD were nearly twice as likely to develop IBD as those without a BD diagnosis (Smith et al. 2013). This was consistent with the Kao et al. observational study of 3590 IBD patients and 14360 controls from a population survey database in Taiwan (Kao et al. 2019). However, another population-based study from Canada found that patients with IBD had lower BD incidence than the general population (Walker et al. 2008). Conclusions from previous observational studies are controversial, and previously described associations may be affected by reverse causality and residual confounders. Therefore, the directional and causal relationships between BD and IBD remain unclear.

Mendelian randomization (MR) is a more convincing causal reasoning method which minimizes the limitations of observational studies (Emdin et al. 2017, Davey Smith et al. 2020). MR uses genetic variations identified through genome-wide association studies (GWAS) as instrumental variables (IVs) to infer causality between outcome and lifetime exposure, which may effectively avoid confounding factors and reverse causality (Emdin et al. 2017, Porcu et al. 2019, Zhao et al. 2019). Thus, the aim of this study was to investigate the potential bidirectional causal relationship between genetically predicted BD and IBD using the latest and most comprehensive GWAS meta-analysis on IBD and BD, implementing a two-sample MR study design.

Materials and methods

Study design

A schematic overview of the bidirectional two-sample MR study design and data sources is detailed in Figure 1. The causal relationship of BD with IBD, including UC and CD, was explored using summary-level statistics including the most comprehensive current IBD GWAS of 59957 individuals of European ancestry, then validated using another comprehensive GWAS study from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). In reverse MR analysis, summary-level data was extracted from the most extensive current BD-related GWAS, which included 198,882 individuals from 14 countries, to test the association between IBD (including CD and UC) and BD risk. MR depends on three key assumptions: ① IVs should significantly relate to exposure; ② IVs should not connect to any confounding factors of the exposure-outcome association; ③ IVs affect the outcome only via exposure. Our analysis was limited to participants of mostly European ancestry to reduce racial mismatches. Details of the data can be found in the Supplementary Tables.

Data sources and SNP selection for BD

The latest and most comprehensive published GWAS for BD was used (Stahl et al. 2019), which included 198882 individuals from 14 countries in 32 cohorts including Europe, North America, and Australia. In this study, BD was diagnosed via international consensus criteria DSM-IV or ICD-10 and assessed by trained interviewers, clinically managed checklists, or medical record review using structured diagnostic tools for a lifetime diagnosis of BD. Only patients of European ancestry were including in the present study (20352 cases versus 31358 controls) to reduce bias due to racial mismatch. In most cohorts, controls underwent lifetime psychiatric screening and were randomly selected from the population. GWAS cohort analysis using Plink 'clumping' to identify a set of linkage disequilibrium (LD) trimmed found GW AS meta-analyses BD-associated variants ($p < 0.0001$, distance > 500 kilobases (kb) or LD

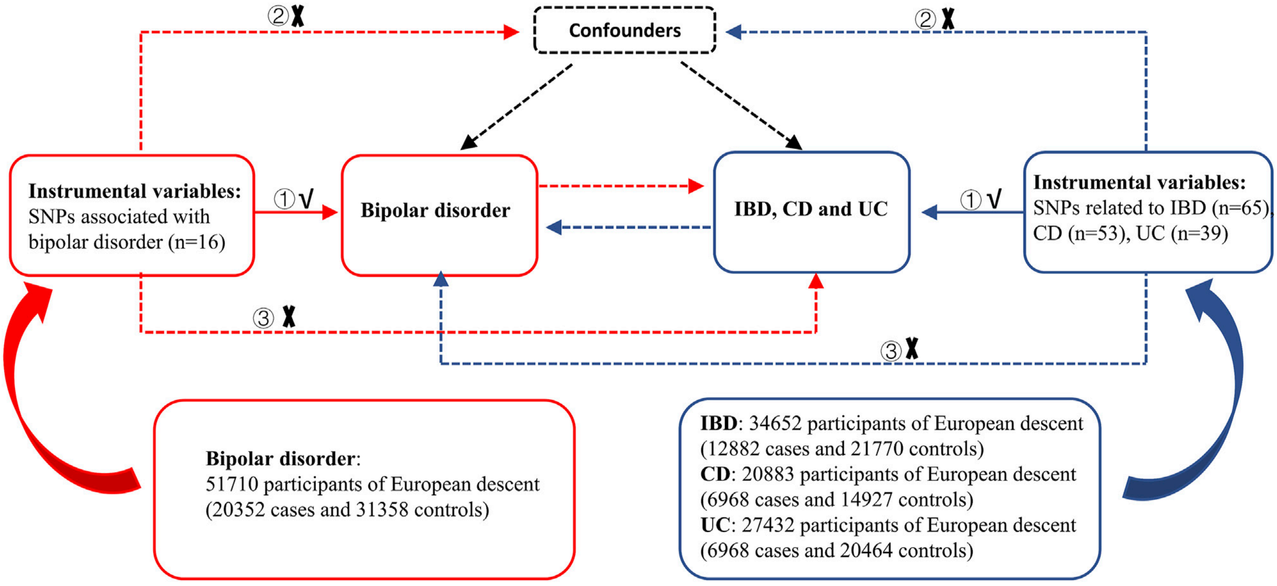

Confounders
②✗
②✗
Instrumental variables: SNPs associated with bipolar disorder (n=16)
①√
Bipolar disorder
IBD, CD and UC
①√
Instrumental variables: SNPs related to IBD (n=65), CD (n=53), UC (n=39)
③✗
③✗
Bipolar disorder: 51710 participants of European descent (20352 cases and 31358 controls)
IBD: 34652 participants of European descent (12882 cases and 21770 controls)
CD: 20883 participants of European descent (6968 cases and 14927 controls)
UC: 27432 participants of European descent (6968 cases and 20464 controls)

FIGURE 1
Overview of the study design in this bidirectional MR study. MR analysis depends on three key assumptions: ① IVs should be significantly related to exposure; ② IVs should not be connected to any confounding factors of the exposure-outcome association; ③ IVs affect the outcome only *via* exposure. IBD: inflammatory bowel disease, CD: Crohn's disease, UC: ulcerative colitis, SNPs: single nucleotide polymorphisms.

r2 < 0.1) for use in subsequent cohorts analysis. The summary GWAS has undergone strict quality control in cohort analysis, follow-up cohort analysis, genome-wide polygenic risk scores (PRS) analysis, and LD score regression analysis. Logistic regression association tests were performed on BD in each cohort, adjusting for covariates of the seven principal components including age, sex, and genetic ancestry.

We excluded 8 SNPs associated with more than one phenotype (e.g., some SNPs are also associated with schizophrenia) to avoid any potential pleiotropic IVs. After removing pleiotropic SNPs, sixteen independent BD-related loci were identified in this GWAS with genome-wide significance thresholds (p < 5 × 10-8), and satisfactory variants were selected to construct instrumental variables.

Data sources and SNP selection for IBD

A GWAS meta-analysis on IBD was recently conducted by de Lange et al., consisting of 59957 individuals (25042 cases and 34915 controls) of predominantly European ancestry (UC: 12366 cases/33609 controls; CD:12194 cases/28072 controls) (de Lange et al. 2017). All included cases were diagnosed by recognized radiological, endoscopic, and histopathological assessments and met clinical diagnostic criteria for IBD. The results of the fixed-effects meta-analysis were further filtered, and sites with strong evidence of heterogeneity (I2>0.90) were discarded. Only sites where all cohorts passed our quality control filters were included in the analysis. Another large GWAS summary of IBD data from Liu et al. was used as a validation analysis (Liu et al. 2015), with a study population arising from Europe, Iran, India or East Asia. To reduce bias caused by racial mismatch, only the population with European ancestry was used as the research object. Mark QC, sample QC, population correlation analysis QC and Genomic inflation factor QC were performed respectively, and all SNPs that did not conform to Hardy-Weinberg equilibrium were eliminated. After quality control (QC) and 1000-genome estimates, adjusted for covariates including smoking status, race, sex, family history, age of disease onset, extraintestinal manifestations and surgery, the number of cases and controls were 12882/21770 for IBD, 5956/14927 for CD, and 6968/20464 for UC, respectively.

For reverse MR analyses, after removing 44 pleiotropic SNPs with more than one phenotype, 65 independent genetic SNPs with p values less than 5 × 10−8 were selected from the summary-level GWAS of Liu et al. to construct the IBD genetic instruments (Liu et al. 2015). For CD and UC, 53 and 39 independent genetic SNPs were selected by the same method. At the same time, we also tested the potential causal relationship of de Lange et al IBD GWAS data to BD, details of the significant IBD SNPs of de Lange et al were in Supplementary Table S3.

Selection of instrumental variables

SNPs were identified at a threshold of genome-wide significance (p < 5 × 10-8). Stringent clumping criteria were set to further filter SNPs with low LD (r2 = 0.001 in 10000 Kb windows) and high minor allele frequency (MAF > 0.01). R2 and F statistics were calculated to represent the variance ratio of exposure factors explained by IVs and the association between IVs and risk exposures of interest (Burgess et al. 2011). F-statistics was calculated by the formula: $(N-2)\times\frac{R^2}{1-R^2}$ to check for bias due to weak IVs, and it is generally recommended to use an F-statistic threshold >10 for MR analysis (Burgess et al. 2011).

Given the selection of SNPs from a very large GWAS, IVs may have effects on traits other than exposure, such as directly affecting outcomes. If those variants that are more strongly associated with outcome than exposure cannot be excluded from the MR analysis, the MR analysis results may be inaccurate due to the reverse causality between exposure and outcome. Therefore, it is necessary to determine whether the SNP is primarily associated with the exposure of interest rather than the outcome. To clarify the direction of causality for each IV with respect to exposure and outcome, we applied MR Steiger filtering to remove those SNPs that were strongly associated with outcomes (Hemani et al. 2017). Steiger filtering assumes that the IV should explain more exposure variation than the outcome; the direction of the instrument is “TRUE” if the IV meets the criteria, and “FALSE” otherwise. After removing those SNPs with the “FALSE” orientation using Steiger filtering, we proceeded to the next MR analysis.

Statistical analyses

Inverse variance weighted (IVW) MR was the main method used to estimate the potential bidirectional relationship between BD and IBD since it avoids confounding factors in the absence of horizontal pleiotropy and produces unbiased estimates (Burgess et al. 2013). At the same time, weighted mode, simple mode, weighted median, and MR Egger methods were used for supplementary and substitution analysis (Bowden et al. 2016, Bowden et al. 2018). In MR-Egger regression, the MR-Egger intercept was used to test for directional horizontal pleiotropy effect (Bowden et al. 2015). Cochran’s Q statistic and funnel plots were then used to verify the heterogeneity of the IVW methods and MR-Egger regression. Cochran’s Q-test statistic was used to examine heterogeneity among all SNPs in each database. Finally, the leave-one-out method was used for sensitivity analysis to verify the stability of the results. Leave-one-out analysis was performed by excluding each SNP in turn and applying the IVW method to the remaining SNPs to assess the potential effect of specific variants on the estimates. When Cochran’s Q test suggests that there is heterogeneity in SNPs, leave-one-out analysis can be a good way to verify the stability of MR analysis. All statistical analyses in this study were performed using the TwoSampleMR packages (https://mrcieu.github.io/TwoSampleMR, version 0.5.6) in R (version 4.1.3, www.r-project.org/).

Results

The causal effect of BD on IBD

Among the 16 BD-associated variants, one SNP was unavailable in the summary-level GWAS of IBD, UC, and CD. In addition, we excluded five SNPs for IBD, CD, and UC due to ambiguous palindromes. We ultimately included 10 SNPs in the MR analysis as genetic instruments for IBD, CD, and UC. The R2 and the (minimal - maxima) F-statistics (112.62–258.55) indicated that all IVs were suitable for MR analysis. (Supplementary Table S1).

The result of the MR analysis showed that genetically predicted BD significantly positively correlated with IBD. (Table 1). The odds ratio (OR) for IBD with 95% confidence interval (CI) per log-OR increment in BD liability was 1.18 (95% CI: 1.04–1.33; p = 0.008) in the IVW model, consistent with the trend of the median weight model, although the median weight model did not reach statistical significance. The scatter diagram is shown in Figure 2 and the forest diagram is shown in Supplementary Figure S1. MR-Egger regression did not reveal a potential horizontal pleiotropy for BD on IBD (egger-intercept = 0.03, p = 0.47), which was similar to the conclusion for BD on CD and UC. Cochran’s Q value suggested no notable heterogeneity (Q = 12.82, p = 0.12), consistent with the conclusion of the MR analysis shown in the funnel plot (Supplementary Figure S2). Furthermore, as shown in the leave-one-out analysis, no significant association changes were observed after removing any individual variant (Supplementary Figure S3).

Genetic susceptibility of BD had a significant positive correlation with UC (OR = 1.19, 95% CI: 1.05–1.35; p = 0.005), but no obvious association was identified for CD (OR = 1.18, 95% CI: 0.95–1.48; p = 0.14). (Table 1). Scatter and forest diagrams are shown in Figure 2 and Supplementary Figure S1, respectively. Cochran’s Q test and funnel plot (Supplementary Figure S2) suggested heterogeneity in the CD database (p values of Cochran’s Q = 0.005) but not in the UC database (p values of Cochran’s Q = 0.73). In addition, the leave-one-out analysis also indicated that the results were stable. (Supplementary Figure S3).

The validation analysis (Table 1) used scatter plots (Figure 2) and forest plots (Supplementary Figure S1) to find MR results consistent with the initial analysis. IVW estimates were analyzed to genetically predict that the combined OR of IBD, CD, and UC increases per log-OR of BD were 1.16(95% CI: 1.02–1.31), 1.20(95% CI: 0.98–1.48), and 1.17 (95% CI: 1.02–1.35), respectively. (Table 1). The Egger’s test showed no potential horizontal pleiotropy except for the relationship between BD and risk of CD. Cochran’s Q test and funnel plot analysis (Supplementary Figure S2) showed significant heterogeneity between BD and CD risk, but no heterogeneity was shown in

TABLE 1 Effects of genetically predicted BD on the risk of IBD in the MR analysis.

Exposure	Outcome	No. SNP	Methods	OR (95% CI)	pval	Egger_intercept	p-Egger_intercept
BD	IBD*	10	MR Egger	0.88 (0.41–1.89)	0.749	0.03	0.47
			Weighted median	1.12 (0.97–1.28)	0.119		
			IVW	1.18 (1.04–1.33)	0.008		
			Simple mode	1.16 (0.91–1.48)	0.253		
			Weighted mode	1.16 (0.96–1.40)	0.164		
BD	UC*	10	MR Egger	1.71 (0.79–3.69)	0.211	-0.03	0.38
			Weighted median	1.16 (0.97–1.37)	0.081		
			IVW	1.19 (1.05–1.35)	0.005		
			Simple mode	1.13 (0.90–1.43)	0.312		
			Weighted mode	1.15 (0.90–1.46)	0.289		
BD	CD*	10	MR Egger	0.44 (0.12–1.61)	0.251	0.09	0.17
			Weighted median	1.12 (0.92–1.35)	0.257		
			IVW	1.18 (0.95–1.48)	0.142		
			Simple mode	1.21 (0.90–1.63)	0.243		
			Weighted mode	1.18 (0.86–1.62)	0.319		
BD	IBD#	14	MR Egger	0.58 (0.29–1.17)	0.149	0.06	0.08
			Weighted median	1.12 (0.97–1.30)	0.125		
			IVW	1.16 (1.02–1.30)	0.024		
			Simple mode	1.10 (0.88–1.37)	0.415		
			Weighted mode	1.11 (0.88–1.40)	0.375		
BD	UC#	14	MR Egger	0.99 (0.51–2.39)	0.978	0.02	0.71
			Weighted median	1.19 (0.99–1.43)	0.057		
			IVW	1.17 (1.02–1.35)	0.029		
			Simple mode	1.18 (0.85–1.62)	0.321		
			Weighted mode	1.20 (0.88–1.62)	0.258		
BD	CD#	14	MR Egger	0.30 (0.10–0.89)	0.051	0.13	0.03
			Weighted median	1.13 (0.92–1.39)	0.221		
			IVW	1.20 (0.98–1.48)	0.079		
			Simple mode	1.20 (0.90–1.61)	0.231		
			Weighted mode	1.19 (0.88–1.61)	0.289		

*Data from de Lange et al.

#Data from Liu et al.

BD, on IBD* MR, Egger (Q = 12.82, p = 0.12), BD on UC* MR Egger (Q = 5.27, p = 0.73), BD on CD* MR Egger (Q = 21.96, p = 0.005), BD on IBD# MR, Egger (Q = 11.52, p = 0.48), BD on UC# MR Egger (Q = 6.02, p = 0.91), BD on CD# MR Egger (Q = 15.23, p = 0.22), Q: Cochran's Q statistics.

BD, bipolar disorder; IBD, inflammatory bowel disease; CD, crohn's disease; UC, ulcerative colitis; IVW, inverse variance weighted.

IBD and UC. The leave-one-out analysis demonstrated the stability of the results (Supplementary Figure S3).

The causal effect of IBD on BD

In the reverse MR analysis, we utilized 61 variants for IBD, 48 variants for CD, and 35 variants for UC as genetic instruments. A summary and detailed information about the variants for each exposure are presented in Supplementary Table S2.

As shown in Table 2, we observed no causal relationship between genetically determined IBD (including both CD and UC) and BD in the outcome database, with ORs close to 1. The Scatter diagram and forest diagram are shown in Supplementary Figure S4 and Supplementary Figure S5, respectively. The Egger's test showed no potential horizontal pleiotropy in reverse MR analysis. Cochran's Q test and funnel plot analysis (Supplementary Figure S6) suggested notable heterogeneity. Therefore, we used a random-effects IVW model to estimate the MR effect size and found no causal relationship between IBD (including CD and UC) and BD. After individual SNPs were deleted, the results remained consistent in the leave-one-out analyses (Supplementary Figure S7). Steiger filtering showed that all genetic IVs used for IBD explained more variance in IBD than in BD in any database (Supplementary Table S2). In addition, we also validated the effect of IBD data on BD from de

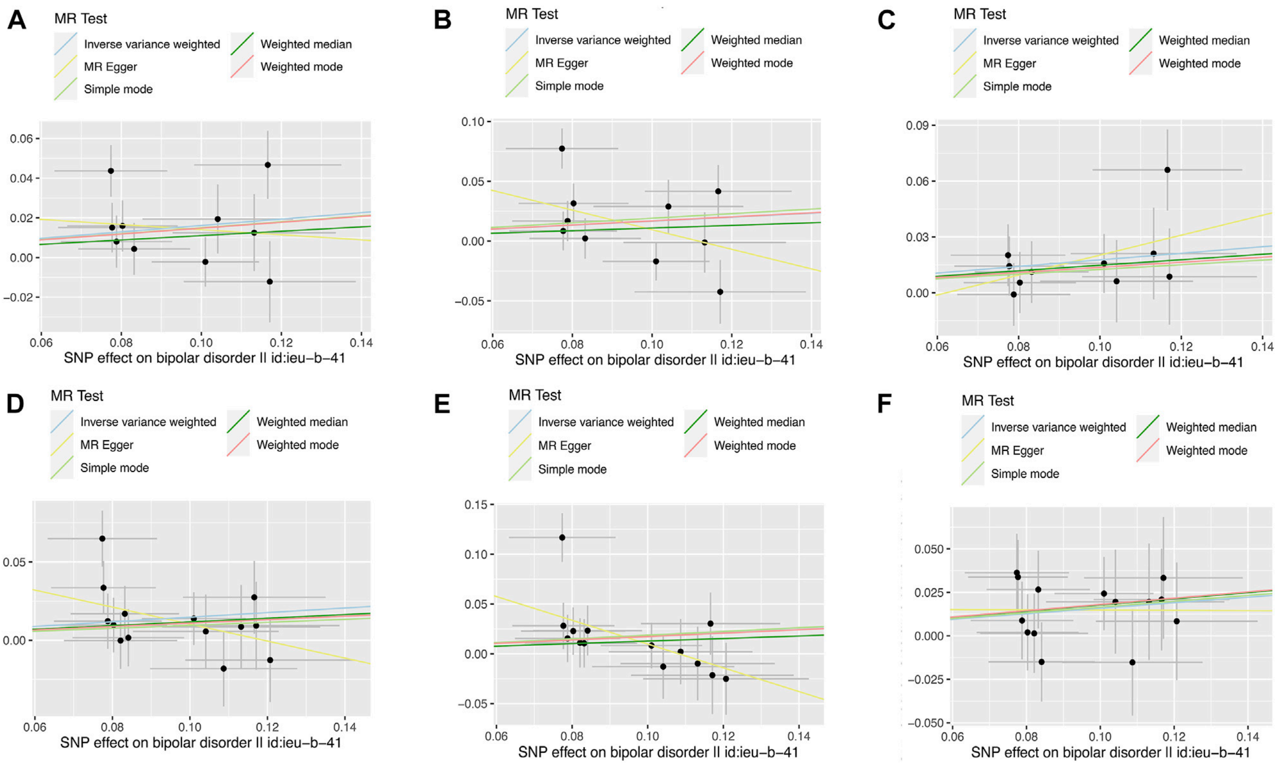


FIGURE 2
Scatter plots of the relationship between genetically predicted BD on IBD, CD and UC. The x-axes represent the genetic instrument−BD associations and y-axes represent genetic instrument−IBD associations from different outcome databases. Black dots denote the genetic instruments included in the primary MR analyses. The colored lines represent the MR fitting results. The line at each point actually reflects the 95% confidence interval. **(A)** BD on IBD*; **(B)** BD on CD*; **(C)** BD on UC*; **(D)** BD on IBD#; **(E)** BD on CD#; **(F)** BD on UC#. BD: bipolar disorder, IBD: inflammatory bowel disease, CD: Crohn's disease, UC: ulcerative colitis. * Data from de Lange et al. # Data from Liu et al.

Lange et al., and the results suggest that there is no casual relationship between IBD and BD (Supplementary Table S4).

Discussion

We tested the potential bidirectional association between BD and IBD and found evidence that genetically predicted BD associates with an increased risk of IBD and UC, with a non-significant trend towards increased risk with CD (Supplementary Table S5). The reverse MR analyses implicated that genetic liability to IBD or any subtype does not significantly associate with BD.

Previous observational studies have shown that BD is positively associated with IBD risk, consistent with our results (Eaton et al. 2010, Smith et al. 2013). However, observational studies on the effect of IBD on BD risk remain controversial (Nikolova et al. 2022). One Canadian study of IBD patients found a lower BD prevalence in IBD patients than in controls, while another cross-sectional study showed that patients with IBD were 2.1 times more likely to develop BD than control subjects (Walker et al. 2008, Kao et al. 2019). These controversial findings may result from methodological limitations and small patient sample sizes. In addition, racial differences and confounding factors may influence the association between BD and IBD. Our MR study was based on the largest available set of GWAS data and restricted the population to those with European ancestry to avoid bias due to small sample size or ethnic differences. Our MR analysis suggests that genetic prediction of IBD is not significantly associated with BD risk, suggesting that the previously observed association may be due to confounding factors or ethnic differences.

Although the biological link between IBD and BD remains unclear, several proposed hypotheses are worth investigating. Recent evidence has shown that patients with BD have significantly higher serological anti-Saccharomyces cerevisiae antibodies (ASCA) levels than non-psychotic patients (Severance et al. 2014). ASCA is commonly used as a predictor of IBD and has significant disease associations with immune reactivity to wheat gluten and bovine casein. However, IBD may accelerate exposure to food antigens to the systemic circulation, which may help explain the elevated levels of gluten and casein antibodies seen in patients with BD (Severance et al. 2014). Another possible hypothesis is that the digestive byproducts of these foods are exorphins which may

TABLE 2 Effect of genetically predicted IBD on the risk of BD in the MR analysis.

Exposure	Outcome	No. SNP	Methods	OR (95% CI)	pval	Egger_intercept	p-Egger_intercept
IBD#	BD	61	MR Egger	0.99 (0.90–1.07)	0.74	0.002	0.772
			Weighted median	1.01 (0.97–1.05)	0.64		
			IVW	1.00 (0.97–1.03)	0.88		
			Simple mode	0.98 (0.90–1.07)	0.69		
			Weighted mode	0.99 (0.93–1.05)	0.81		
CD#	BD	48	MR Egger	1.01 (0.94–1.07)	0.96	0.003	0.681
			Weighted median	1.01 (0.97–1.04)	0.55		
			IVW	1.01 (0.99–1.04)	0.34		
			Simple mode	1.04 (0.98–1.10)	0.23		
			Weighted mode	1.02 (0.98–1.06)	0.41		
UC#	BD	35	MR Egger	0.95 (0.87–1.04)	0.29	0.009	0.309
			Weighted median	0.99 (0.95–1.03)	0.66		
			IVW	0.99 (0.96–1.03)	0.73		
			Simple mode	1.01 (0.94–1.10)	0.72		
			Weighted mode	0.99 (0.94–1.05)	0.91		

#Data from Liu et al.
IBD on BD MR Egger (Q = 93.92, p = 0.003), CD on BD MR Egger (Q = 87.42, p = 0.0002), UC on BD MR Egger (Q = 49.92, p = 0.03), Q: Cochran's Q statistics.
BD, bipolar disorder; IBD, inflammatory bowel disease; CD, Crohn's disease; UC, ulcerative colitis; IVW, inverse variance weighted.

directly interact with tight junction proteins or undergo epithelial cell transcytosis to potentially affect brain physiology by acting on opioid receptors (Peeters et al. 2001, Lammers et al. 2008, Tripathi et al. 2009). Secondly, many previous studies have suggested that inflammatory cytokines may play a key role in IBD pathogenesis (Neurath 2014, Friedrich et al. 2019), and psychiatric diseases promote intestinal inflammation by regulating the microbiota-gut-brain axis (Osadchiy et al. 2019). Altered mood increases gut permeability, enabling gut bacteria to translocate to peripheral lymphoid organs and trigger innate immune responses (Peppas et al. 2021). Affective disorders can activate the hypothalamic-pituitary-adrenal axis, thereby aggravating chronic inflammation and promoting immune response, consistent with previous observations that BD patients have elevated levels of inflammatory cytokines (Modabbernia et al. 2013, Munkholm et al. 2013, Kostic et al. 2014, Gracie et al. 2019). Activation of the hypothalamic-pituitary-adrenal axis stimulates secretion of corticotropin-releasing factor (CRF), followed by release of adrenocorticotropic hormone (ACTH) from the anterior pituitary. CRF and ACTH increase intestinal permeability by inducing mast cell degranulation and cytokine secretion (Santos et al. 1999, Hill et al. 2013). In addition, BD also stimulates activation of the sympathetic nervous system through the stress response, mediating changes in the autonomic nervous system and increasing catecholamine secretion to exert a pro-inflammatory effect (Farhadi et al. 2005, Luo et al. 2021). A series of inflammatory reactions caused by BD increases intestinal permeability and damages the epithelial barrier, further activating the immune response to disrupt gastrointestinal homeostasis and ultimately lead to IBD.

Two main advantages of our study are worth noting. Observational studies suggest a bidirectional relationship between BD and IBD, but studies present opposing conclusions due to potential confounding factors. A major advantage of this MR study is that we explored the results from a genetic susceptibility perspective, avoiding reverse causality and minimizing residual confounding. Second, we used the largest available resource of exposure GWAS data and the broadest summary-level IBD and BD data from different samples and validated our results across different datasets. Although potential sample overlap cannot be completely avoided, two-sample MR greatly reduces bias due to potential sample overlap between exposures and outcomes. The consistency of the two analyses suggests our results are accurate.

However, some limitations should be acknowledged. First, our study subjects were primarily individuals of European ancestry, which may limit the generality of our findings to other ethnic groups. However, selecting populations of the same ancestry for studies helps avoid genetic differences between races, making our conclusions more convincing. Second, previous studies have shown that patients with active IBD have significantly higher rates of affective disorders than patients with inactive disease (Walker et al. 2008, Kao et al. 2019). While our conclusions do not support a causal effect of IBD on BD, the GWAS data we included only considered the dichotomous diagnosis of IBD, i.e., incidence, but not the course of IBD. Because IBD is characterized by alternating remissions and relapses and its onset is difficult to predict, dissecting the genetic makeup associated with IBD activity remains a challenge. Therefore, due to the lack of GWAS data on the active phase of IBD, we were unable to explore the causal relationship between active IBD and BD using MR methods. Third, in partial negative MR results, Cochran's Q value suggested significant heterogeneity of IVs. Therefore, we performed further random effect IVW analysis and leave-one-out analysis to support the stability of the results.

Despite the well-established bidirectional relationship between IBD and mental illness, psychotherapy for patients with IBD is currently rarely recommended as an adjunctive treatment to improve quality of life (Lamb et al. 2019). The causal relationship of BD to IBD observed in our study should draw attention to intestinal symptoms in BD patients for more accurate clinical treatment. Affective disorders can lead to chronic inflammation and stress response intensification, so clinicians should improve the IBD suspicion index for BD patients (Marrie et al. 2019). Persistent gastrointestinal symptoms should not be ignored, and antidepressant treatments may need to be tailored for their different effects on bowel habits. Multiple studies have also demonstrated that dietary intervention and probiotic therapy can have a positive impact on BD (Liu et al. 2019, Nikolova et al. 2021). Therefore, people with BD and concomitant lower gastrointestinal disorders should consider using this therapy to obtain maximal benefit.

Conclusion

Our results confirm a causal relationship between BD and IBD, which may influence clinical decisions on the management of BD patients with intestinal symptoms. Although the reverse MR results did not support a causal effect of IBD on BD, the effect of active IBD on BD remains to be further investigated.

Data availability statement

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

Author contributions

ZW contributed the conception and design of the study, and drafted the manuscript; YX contributed to analyze and interpretation of data and revised the manuscript. XZ, XW, and ZH participated in data acquisition. DS and DW participated in literature research. All authors read and approved the final manuscript.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.970933/full#supplementary-material

References

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* 12 (4), 205–217. doi:10.1038/nrgastro.2015.34

Bernstein, C. N., Hitchon, C. A., Walld, R., Bolton, J. M., Sareen, J., Walker, J. R., et al. (2019). Increased burden of psychiatric disorders in inflammatory bowel disease. *Inflamm. Bowel Dis.* 25 (2), 360–368. doi:10.1093/ibd/izy235

Bonaz, B. L., and Bernstein, C. N. (2013). Brain-gut interactions in inflammatory bowel disease. *Gastroenterology* 144 (1), 36–49. doi:10.1053/j.gastro.2012.10.003

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Bowden, J., Spiller, W., Del Greco, M. F., Sheehan, N., Thompson, J., Minelli, C., et al. (2018). Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *Int. J. Epidemiol.* 47 (4), 1264–1278. doi:10.1093/ije/dyy101

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37 (7), 658–665. doi:10.1002/gepi.21758

Burgess, S., Thompson, S. G., and Collaboration, C. C. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidemiol.* 40 (3), 755–764. doi:10.1093/ije/dyr036

Davey Smith, G., Holmes, M. V., Davies, N. M., and Ebrahim, S. (2020). Mendel's laws, mendelian randomization and causal inference in observational data: Substantive and nomenclatural issues. *Eur. J. Epidemiol.* 35 (2), 99–111. doi:10.1007/s10654-020-00622-7

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 49 (2), 256–261. doi:10.1038/ng.3760

Eaton, W. W., Pedersen, M. G., Nielsen, P. R., and Mortensen, P. B. (2010). Autoimmune diseases, bipolar disorder, and non-affective psychosis. *Bipolar Disord.* 12 (6), 638–646. doi:10.1111/j.1399-5618.2010.00853.x

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *JAMA* 318 (19), 1925–1926. doi:10.1001/jama.2017.17219

Farhadi, A., Keshavarzian, A., Van de Kar, L. D., Jakate, S., Domm, A., Zhang, L., et al. (2005). Heightened responses to stressors in patients with inflammatory bowel disease. *Am. J. Gastroenterol.* 100 (8), 1796–1804. doi:10.1111/j.1572-0241.2005.50071.x

Ferrari, A. J., Stockings, E., Khoo, J. P., Erskine, H. E., Degenhardt, L., Vos, T., et al. (2016). The prevalence and burden of bipolar disorder: Findings from the global burden of disease study 2013. *Bipolar Disord.* 18 (5), 440–450. doi:10.1111/bdi.12423

Friedrich, M., Pohin, M., and Powrie, F. (2019). Cytokine networks in the pathophysiology of inflammatory bowel disease. *Immunity* 50 (4), 992–1006. doi:10.1016/j.immuni.2019.03.017

Gracie, D. J., Hamlin, P. J., and Ford, A. C. (2019). The influence of the brain-gut axis in inflammatory bowel disease and possible implications for treatment. *Lancet. Gastroenterol. Hepatol.* 4 (8), 632–642. doi:10.1016/S2468-1253(19)30089-5

Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* 13 (11), e1007081. doi:10.1371/journal.pgen.1007081

Hill, L. T., Kidson, S. H., and Michell, W. L. (2013). Corticotropin-releasing factor: A possible key to gut dysfunction in the critically ill. *Nutrition* 29 (7-8), 948–952. doi:10.1016/j.nut.2012.12.023

Hope, S., Dieset, I., Agartz, I., Steen, N. E., Ueland, T., Melle, I., et al. (2011). Affective symptoms are associated with markers of inflammation and immune activation in bipolar disorders but not in schizophrenia. *J. Psychiatr. Res.* 45 (12), 1608–1616. doi:10.1016/j.jpsychires.2011.08.003

Kao, L. T., Lin, H. C., and Lee, H. C. (2019). Inflammatory bowel disease and bipolar disorder: A population-based cross-sectional study. *J. Affect. Disord.* 247, 120–124. doi:10.1016/j.jad.2019.01.014

Kostic, A. D., Xavier, R. J., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology* 146 (6), 1489–1499. doi:10.1053/j.gastro.2014.02.009

Lamb, C. A., Kennedy, N. A., Raine, T., Hendy, P. A., Smith, P. J., Limdi, J. K., et al. (2019). British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. *Gut* 68, s1–s106. doi:10.1136/gutjnl-2019-318484

Lammers, K. M., Lu, R., Brownley, J., Lu, B., Gerard, C., Thomas, K., et al. (2008). Gliadin induces an increase in intestinal permeability and zonulin release by binding to the chemokine receptor CXCR3. *Gastroenterology* 135 (1), 194–204. e193. doi:10.1053/j.gastro.2008.03.023

Le Berre, C., and Peyrin-Biroulet, L.S.-I. s. group (2021). Selecting end points for disease-modification trials in inflammatory bowel disease: The SPIRIT consensus from the IOIBD. *Gastroenterology* 160 (5), 1452–1460.e21. doi:10.1053/j.gastro.2020.10.065

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47 (9), 979–986. doi:10.1038/ng.3359

Liu, R. T., Walsh, R. F. L., and Sheehan, A. E. (2019). Prebiotics and probiotics for depression and anxiety: A systematic review and meta-analysis of controlled clinical trials. *Neurosci. Biobehav. Rev.* 102, 13–23. doi:10.1016/j.neubiorev.2019.03.023

Luo, J., Xu, Z., Noordam, R., van Heemst, D., and Li-Gao, R. (2021). Depression and inflammatory bowel disease: A bidirectional two-sample mendelian randomization study. *J. Crohns Colitis.* doi:10.1093/ecco-jcc/jjab191

Marrie, R. A., Walld, R., Bolton, J. M., Sareen, J., Walker, J. R., Patten, S. B., et al. (2019). Rising incidence of psychiatric disorders before diagnosis of immune-mediated inflammatory disease. *Epidemiol. Psychiatr. Sci.* 28 (3), 333–342. doi:10.1017/S2045796017000579

McIntyre, R. S., Berk, M., Brietzke, E., Goldstein, B. I., Lopez-Jaramillo, C., Kessing, L. V., et al. (2020). Bipolar disorders. *Lancet* 396 (10265), 1841–1856. doi:10.1016/S0140-6736(20)31544-0

Modabbernia, A., Taslimi, S., Brietzke, E., and Ashrafi, M. (2013). Cytokine alterations in bipolar disorder: A meta-analysis of 30 studies. *Biol. Psychiatry* 74 (1), 15–25. doi:10.1016/j.biopsych.2013.01.007

Munkholm, K., Brauner, J. V., Kessing, L. V., and Vinberg, M. (2013). Cytokines in bipolar disorder vs. healthy control subjects: A systematic review and meta-analysis. *J. Psychiatr. Res.* 47 (9), 1119–1133. doi:10.1016/j.jpsychires.2013.05.018

Neurath, M. F. (2014). Cytokines in inflammatory bowel disease. *Nat. Rev. Immunol.* 14 (5), 329–342. doi:10.1038/nri3661

Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., et al. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: A systematic review of population-based studies. *Lancet* 390 (10114), 2769–2778. doi:10.1016/S0140-6736(17)32448-0

Nikolova, V. L., Cleare, A. J., Young, A. H., and Stone, J. M. (2021). Updated review and meta-analysis of probiotics for the treatment of clinical depression: Adjunctive vs. Stand-alone treatment. *J. Clin. Med.* 10 (4), 647. doi:10.3390/jcm10040647

Nikolova, V. L., Pelton, L., Moulton, C. D., Zorzato, D., Cleare, A. J., Young, A. H., et al. (2022). The prevalence and incidence of irritable bowel syndrome and inflammatory bowel disease in depression and bipolar disorder: A systematic review and meta-analysis. *Psychosom. Med.* 84 (3), 313–324. doi:10.1097/PSY.0000000000001046

Osadchiy, V., Martin, C. R., and Mayer, E. A. (2019). The gut-brain Axis and the microbiome: Mechanisms and clinical implications. *Clin. Gastroenterol. Hepatol.* 17 (2), 322–332. doi:10.1016/j.cgh.2018.10.002

Peeters, M., Joossens, S., Vermeire, S., Vlietinck, R., Bossuyt, X., and Rutgeerts, P. (2001). Diagnostic value of anti-Saccharomyces cerevisiae and antineutrophil cytoplasmic autoantibodies in inflammatory bowel disease. *Am. J. Gastroenterol.* 96 (3), 730–734. doi:10.1111/j.1572-0241.2001.03613.x

Peppas, S., Pansieri, C., Piovani, D., Danese, S., Peyrin-Biroulet, L., Tsantes, A. G., et al. (2021). The brain-gut Axis: Psychological functioning and inflammatory bowel diseases. *J. Clin. Med.* 10 (3), 377. doi:10.3390/jcm10030377

Piovani, D., Danese, S., Peyrin-Biroulet, L., Nikolopoulos, G. K., Lytras, T., and Bonovas, S. (2019). Environmental risk factors for inflammatory bowel diseases: An umbrella review of meta-analyses. *Gastroenterology* 157 (3), 647–659. e644. doi:10.1053/j.gastro.2019.04.016

Porcu, E., Rueger, S., Lepik, K., e, Q. C., Consortium, B., Santoni, F. A., et al. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* 10 (1), 3300. doi:10.1038/s41467-019-10936-0

Ramos, G. P., and Papadakis, K. A. (2019). Mechanisms of disease: Inflammatory bowel diseases. *Mayo Clin. Proc.* 94 (1), 155–165. doi:10.1016/j.mayocp.2018.09.013

Santos, J., Saunders, P. R., Hanssen, N. P., Yang, P. C., Yates, D., Groot, J. A., et al. (1999). Corticotropin-releasing hormone mimics stress-induced colonic epithelial pathophysiology in the rat. *Am. J. Physiol.* 277 (2), G391–G399. doi:10.1152/ajpgi.1999.277.2.G391

Severance, E. G., Dupont, D., Dickerson, F. B., Stallings, C. R., Origoni, A. E., Krivogorsky, B., et al. (2010). Immune activation by casein dietary antigens in bipolar disorder. *Bipolar Disord.* 12 (8), 834–842. doi:10.1111/j.1399-5618.2010.00879.x

Severance, E. G., Gressitt, K. L., Yang, S., Stallings, C. R., Origoni, A. E., Vaughan, C., et al. (2014). Seroreactive marker for inflammatory bowel disease and associations with antibodies to dietary proteins in bipolar disorder. *Bipolar Disord.* 16 (3), 230–240. doi:10.1111/bdi.12159

Smith, D. J., Martin, D., McLean, G., Langan, J., Guthrie, B., and Mercer, S. W. (2013). Multimorbidity in bipolar disorder and undertreatment of cardiovascular disease: A cross sectional study. *BMC Med.* 11, 263. doi:10.1186/1741-7015-11-263

Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., et al. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* 51 (5), 793–803. doi:10.1038/s41588-019-0397-8

Tripathi, A., Lammers, K. M., Goldblum, S., Shea-Donohue, T., Netzel-Arnett, S., Buzza, M. S., et al. (2009). Identification of human zonulin, a physiological modulator of tight junctions, as prehaptoglobin-2. *Proc. Natl. Acad. Sci. U. S. A.* 106 (39), 16799–16804. doi:10.1073/pnas.0906773106

Vieta, E., Berk, M., Schulze, T. G., Carvalho, A. F., Suppes, T., Calabrese, J. R., et al. (2018). Bipolar disorders. *Nat. Rev. Dis. Prim.* 4, 18008. doi:10.1038/nrdp.2018.8

Walker, J. R., Ediger, J. P., Graff, L. A., Greenfeld, J. M., Clara, I., Lix, L., et al. (2008). The manitoba IBD cohort study: A population-based study of the prevalence of lifetime and 12-month anxiety and mood disorders. *Am. J. Gastroenterol.* 103 (8), 1989–1997. doi:10.1111/j.1572-0241.2008.01980.x

Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *Int. J. Epidemiol.* 48 (5), 1478–1492. doi:10.1093/ije/dyz142

TYPE Original Research
PUBLISHED 05 October 2022
DOI 10.3389/fgene.2022.1021178

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University,
United States

REVIEWED BY
Zhaoming Qi,
Northeast Agricultural University, China
Xuehai Zhang,
Henan Agricultural University, China

*CORRESPONDENCE
Kaiyi Wang,
wangky@nercita.org.cn
Jiankang Wang,
wangjiankang@caas.cn

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

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

RECEIVED 17 August 2022
ACCEPTED 20 September 2022
PUBLISHED 05 October 2022

CITATION
Zhang L, Wang X, Wang K and Wang J (2022), GAHP: An integrated software package on genetic analysis with bi-parental immortalized heterozygous populations.
Front. Genet. 13:1021178.
doi: 10.3389/fgene.2022.1021178

GAHP: An integrated software package on genetic analysis with bi-parental immortalized heterozygous populations

Luyan Zhang[1†], Xinhui Wang[2†], Kaiyi Wang[2*] and Jiankang Wang[1,3*]

[1]National Key Facility for Crop Gene Resources and Genetic Improvement, and Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China, [2]Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China, [3]National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences (CAAS), Hainan, China

GAHP is a freely available software package for genetic analysis with bi-parental immortalized heterozygous and pure-line populations. The package is project-based and integrated with multiple functions. All operations and running results are properly saved in a project, which can be recovered when the project is re-open by the package. Four functionalities have been implemented in the current version of GAHP, i.e., 1) MHP: visualization of genetic linkage maps; 2) VHP: analysis of variance (ANOVA) and estimation of heritability on phenotypic data; 3) QHP: quantitative trait locus (QTL) mapping on both genotypic and phenotypic data; 4) SHP: simulation of bi-parental immortalized heterozygous and pure-line populations, and power analysis of QTL mapping. VHP and QHP can be conducted in individual populations, as well as in multiple populations by the combined analysis. Input files are arranged either in the plain text format with an extension name same as the functionality or in the MS Excel formats. Output files have the same prefix name as the input file, but with different extensions to indicate their contents. Three characters before the extension names stand for the types of populations used in analysis. In the interface of the software package, input files are grouped by functionality, and output files are grouped by individual or combined mapping populations. In addition to the text-format outputs, the constructed linkage map can be visualized per chromosome or for a number of selected chromosomes; line plots and bi-plots can be drawn from QTL mapping results and phenotypic data. Functionalities and analysis methods available in GAHP help the investigation of genetic architectures of complex traits and the mechanism of heterosis in plants.

KEYWORDS

bi-parental population, immortalized heterozygous population, analysis of variance, QTL mapping, genetic simulation

1 Introduction

In past decades, the methodology on quantitative trait locus (QTL) mapping has been extensively applied in genetic studies to dissect the individual genes of complex traits in both animals and plants. Bi-parental segregating populations, such as backcross (BC), doubled haploids (DH), recombinant inbred lines (RIL), and F_2, are commonly developed and then used for QTL mapping studies in plants. A number of mapping methods have been proposed, such as interval mapping (IM; Lander and Botstein, 1989), composite interval mapping (CIM; Zeng, 1994), multiple interval mapping (MIM; Kao et al., 1999), inclusive composite interval mapping (ICIM; Li et al., 2007; Zhang et al., 2008), and multiple QTL model (MQM; van Ooijen, 2009). Some frequently used software packages for bi-parental populations are R/qtl (Broman et al., 2003), QTL Cartographer (Wang et al., 2007), QTLNetwork (Yang et al., 2008), MAPQTL (van Ooijen, 2009), and QTL IciMapping (Meng et al., 2015).

By comparison with the other mapping methods, ICIM is more efficient in background control *via* a two-step mapping strategy. In the first step of ICIM, stepwise regression is applied to identify the most-significant regression variables representing the marker genotypes. In the second step, interval mapping is performed on phenotypic values adjusted by marker variables identified in the first step (Li et al., 2007; Zhang et al., 2008; Meng et al., 2015). In recent years, the ICIM algorithm has been extended to epistatic mapping (Li et al., 2008a), QTL by environment interaction analysis (Li et al., 2015), hybrid F_1 populations derived from two heterozygous parents, double cross F_1 populations derived from four homozygous parents (Zhang et al., 2015a), and pure-line populations derived from four to eight homozygous parents (Zhang et al., 2017; Shi et al., 2019). The ICIM-based algorithms have been implemented in three integrated software packages, i.e. QTL IciMapping for bi-parental populations (Meng et al., 2015), GACD for hybrid F_1 from two heterozygous parents and double cross F_1 from four homozygous parents (Zhang et al., 2015b), and GAPL for multi-parental pure-line populations (Zhang et al., 2019).

Conventional heterozygous populations, such as BC, F_2, and F_3, may be used to estimate the dominance-related effects and investigate the genetic mechanism of heterosis. However, these populations cannot be phenotyped in multi-environmental trials, and thus the analysis for QTL stability and QTL by environment interaction cannot be conducted. To avoid these problems, the concept of immortalized F_2 and BC has been proposed by using the bi-parental pure lines. For example, Hua et al. (2003) investigated the genetic basis of an elite rice hybrid using an immortalized F_2 population by randomly permutated inter-mating of 240 bi-parental RILs. Liu et al. (2017) started from one RIL population of two maize inbred lines S-951 and Qi319, and developed one immortalized F_2 population for QTL detection on leaf width. Yi et al. (2019) investigated the genetic bases of yield-related traits and heterosis in maize using immortalized F_2 and RIL populations. Li et al. (2008b) reported two immortalized BC populations in rice and used them to identify the main-effect QTLs and digenic epistatic loci underlying the heterosis of agronomic and economic traits. Aakanksha et al. (2021) investigated the heterosis on yield in *Brassica juncea* by using a DH and two-directional immortalized BC populations. Li et al. (2018) developed two-directional immortalized BC populations and one immortalized F_2 population, and used them to detect QTLs affecting fiber quality traits in upland cotton.

In studies mentioned above, immortalized heterozygous populations were treated as a kind of bi-parental populations in genetic analysis, and analyzed separately from their pure-line populations. The joint analysis of pure lines and their derived immortalized heterozygous populations provides more genetic information, and improves the mapping accuracy. In addition, no software package has been developed when heterozygous and pure-line populations are both available. In this study, we report an integrated software package which is called GAHP, i.e. genetic analysis with bi-parental immortalized heterozygous populations. By using this package, the phenotypic and genetic analysis can be performed in bi-parental immortalized populations and their pure lines either separately or jointly.

2 Materials and methods

2.1 Genetic mapping populations

Four kinds of populations, which are essentially derived from the same two homozygous parents, can be handled in GAHP for both phenotypic and genetic analysis. These populations are called by bi-parental pure-inbred lines (PIL), immortalized backcross population with the first parent (IB1), immortalized backcross population with the second parent (IB2), and immortalized F_2 population (IF2). It should be noted that the pure-inbred lines (or pure lines in short) can be either DHs or RILs derived from two inbred homozygous parents. Relationship between the four populations is shown in Figure 1. Genotype of each line in population PIL can be maintained by selfing, which is the reason to be called 'permanent'. IB1 is generated by the hybridization between the PIL population and the first inbred parent, similar to the backcrossing of F_1 hybrid with the first inbred parent. IB2 is generated by the hybridization between the PIL population and the second inbred parent, similar to the backcrossing of F_1 hybrid with the second inbred parent. IF2 is generated by the hybridization between two lines in the PIL population, similar to selfing of the F_1 hybrid. As each line in population PIL can be maintained by selfing, IB1, IB2 and IF2 can be repeatedly produced like the typical F_1 hybrids whenever needed, which is the reason to be called 'immortalized'. Due to their repeatability, each of the four

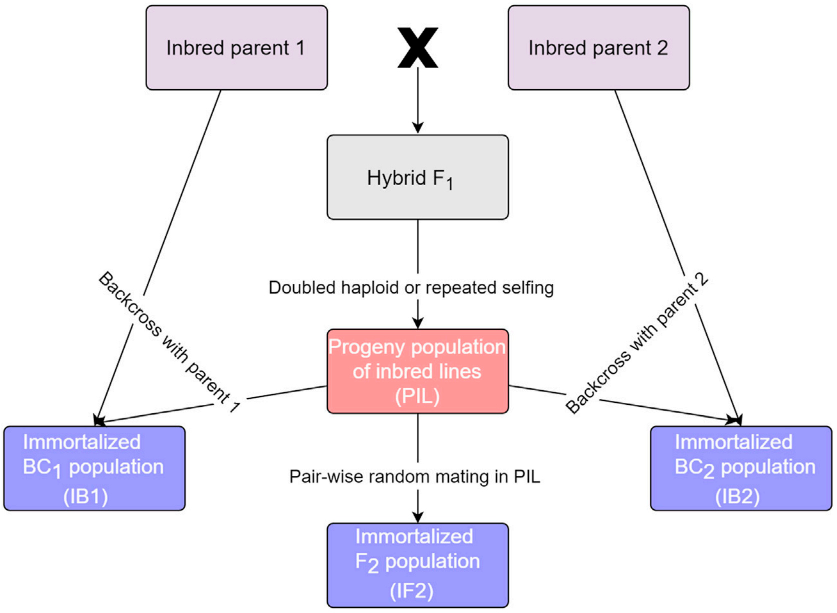

Inbred parent 1
X
Inbred parent 2
Hybrid F_1
Doubled haploid or repeated selfing
Backcross with parent 1
Backcross with parent 2
Progeny population of inbred lines (PIL)
Immortalized BC_1 population (IB1)
Immortalized BC_2 population (IB2)
Pair-wise random mating in PIL
Immortalized F_2 population (IF2)

FIGURE 1
Relationship between populations that can be handled in GAHP.

kinds of populations can be evaluated in multi-environmental trials with replications.

2.2 Coding criteria of marker types and phenotypic values

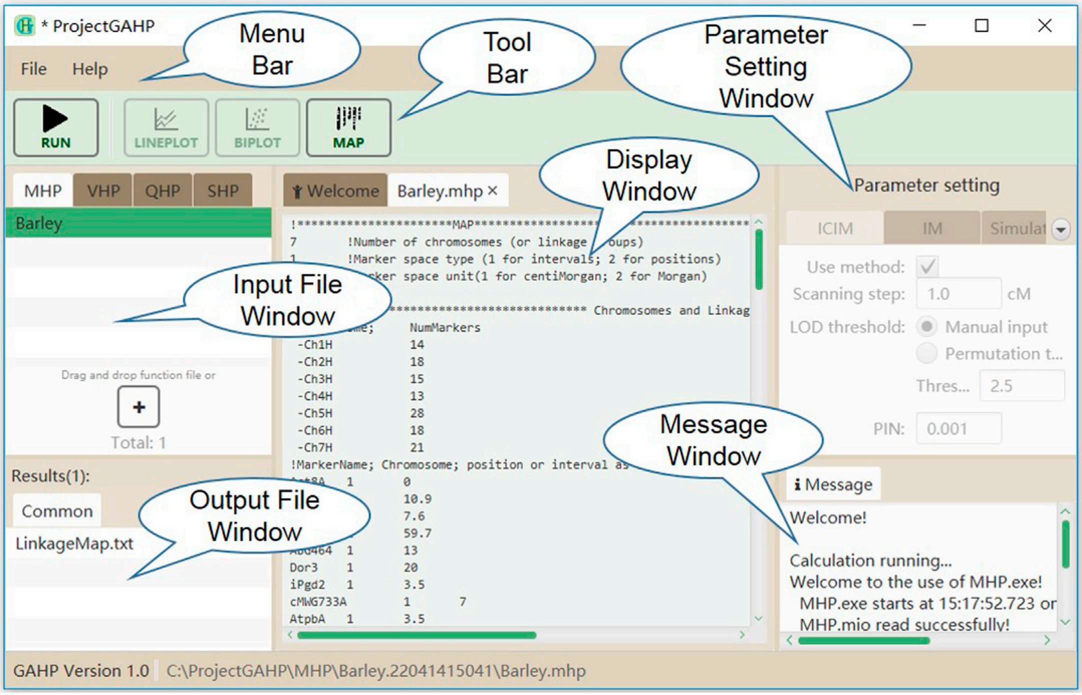
Both independent population and combined analysis can be conducted in GAHP. For genetic analysis, the genotypic data is only needed for population PIL. Genotypes of heterozygous lines in populations IB1, IB2, or IF2 can be deduced from the genotypes of homozygous lines in PIL. Assuming there are two homozygous parents P_1 (or Parent A) and P_2 (or Parent B), two bands can be observed in the two parents at one polymorphic marker locus. Markers having no polymorphism or heterozygous in either parent cannot be used. Assuming *AA* is the genotype of P_1, *BB* is the genotype of P_2, and *AB* is the genotype of their F_1 hybrid. Marker types could be coded by numbers, letters, or the mixed numbers and letters. As individual lines in PIL are assumed to be homozygous, only homozygous genotypes in PIL are useful in genetic analysis. Heterozygous genotypes in PIL are treated as missing values. When numbers are used in coding, the two parental bands are coded as 2 and 0, respectively. When letters are used, Parent A is coded as A or AA; Parent B is coded as B or BB. Codes 1, H and AB are acceptable for heterozygotes, and missing values of marker types are coded as -1, X, XX, *, or **. Mixed coding with numbers and capital letters is allowed in the software, but it is not recommended. Missing phenotypic values are represented by “NA”, “na”, “*“, “.“, or “-100”, which will be replaced by population mean in QTL mapping.

2.3 Development of the GAHP software

In GAHP, core modules for phenotypic data analysis, QTL mapping, genetic population simulation, and power analysis were written in Intel Fortran 90/95. The interface and core modules for setting parameters, viewing results and drawing figures were written in JAVA. The software runs on Microsoft Windows XP/Vista/7/10/11. GAHP is an integrated and project-based software package. When the software is initiated, the first thing to do is to create a new project or open an existing project. The use of project will assure that all operations and running results are properly saved when the software is closed. When the project is open the next time by the software, previous operations and results can be recovered. Introduced below are the four functionalities implemented in the current version of GAHP.

2.4 The MHP functionality

Functionality MHP displays the completed linkage maps in a format (or style) which can be easily modified by users.



```
!*********************MAP*****************        ****************
7       !Number of chromosomes (or linkage    oups)
1       !Marker space type (1 for intervals; 2 for positions)
          rker space unit(1 for centiMorgan; 2 for Morgan)

          ***************************** Chromosomes and Linkag
         me;      NumMarkers
 -Ch1H            14
 -Ch2H            18
 -Ch3H            15
 -Ch4H            13
 -Ch5H            28
 -Ch6H            18
 -Ch7H            21
!MarkerName; Chromosome; position or interval as
        1        0
                 10.9
                 7.6
                 59.7
        1        13
Dor3    1        20
iPgd2   1        3.5
cMWG733A         1       7
AtpbA   1        3.5
```

Use method: ☑
Scanning step: 1.0 cM
LOD threshold: ● Manual input ○ Permutation t...
Thres... 2.5
PIN: 0.001

Welcome!

Calculation running...
Welcome to the use of MHP.exe!
  MHP.exe starts at 15:17:52.723 or
  MHP.mio read successfully!

C:\ProjectGAHP\MHP\Barley.22041415041\Barley.mhp

FIGURE 2
The interface of functionality MHP.

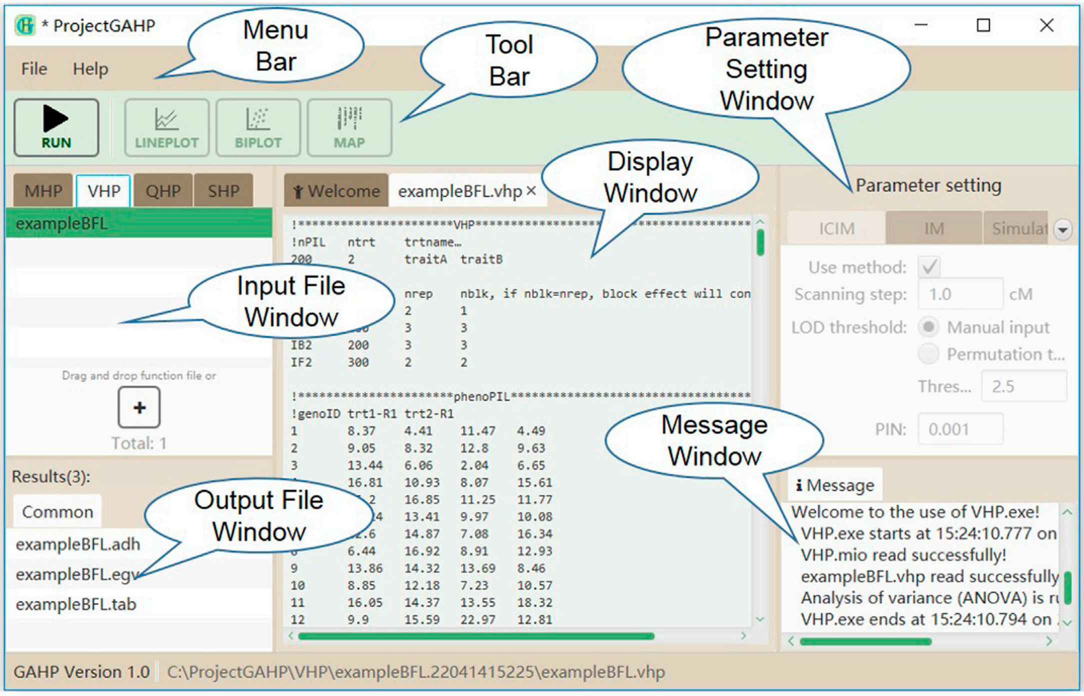
Linkage maps should have been built by other software packages. Chromosome information and marker positions have to be provided. The input file for MHP consists of three parts: 1) general information on linkage maps, 2) marker number information, and 3) linkage map information. The example given in Supplementary Figure S1 represents a linkage map with seven chromosomes. Markers on their chromosomes were defined by marker interval, i.e. distance between adjacent markers in cM (Supplementary Figure S1A). Marker number on each chromosome and linkage map information are given in Supplementary Figures S1B and S1C, respectively.

Figure 2 shows the interface of functionality MHP. The menu and tool bars are located on the top of the interface. The input and output file windows are located on the left side, showing names of the loaded input files and associated output files. In the input file window, files are grouped by functionalities, i.e. MHP, VHP, QHP, and SHP. In the output file window, files are grouped by population names, i.e. PIL, IB1, IB2 and IF2 *etc.* In the middle is the display window, which shows the detailed information of input or output files. At the right side are the parameter setting and running message windows. No parameter is needed to run functionality MHP. While the input file is properly loaded, the users may click “Run” on the tool bar to run the functionality.

2.5 The VHP functionality

Heritability may be the most important concept in quantitative genetics, which quantifies the proportion of genetic variation included in phenotypic values. Analysis of variance (ANOVA) can be used to estimate the variance components, based on which the broad-sense heritability can be estimated in genetic populations. Here the mapping populations can be some or all of the four populations as shown in Figure 1. Combined ANOVA will be applied if more than one population is included in the input file. The input file for VHP consists of five parts: 1) general information of the genetic populations, 2) phenotype of PIL, 3) phenotype of IB1, 4) phenotype of IB2, and 5) phenotype of IF2. If one population has no phenotypic data, the corresponding part in the input file is left to be empty. Supplementary Figure S2 represents an example of input file for VHP, where all the four populations have phenotypic values. Population sizes of PIL, IB1, IB2 and IF2 are equal to 200, 200, 200, and 300, respectively (Supplementary Figure S2A). Phenotypic values of the four populations were defined in Supplementary Figures S2B–S2E, respectively. It should be noted that populations IB1 and IB2 must have the same size as PIL, if included.

Figure 3 shows the interface of functionality VHP. Input files for this functionality are grouped on the VHP tab in the input file



```
!*********************VHP*****************                 ****************
!nPIL    ntrt     trtname…
200      2        traitA  traitB

                  nrep    nblk, if nblk=nrep, block effect will con
                  2       1
                  3       3
IB2      200      3       3
IF2      300      2       2

!*********************phenoPIL*********************************
!genoID trt1-R1 trt2-R1
1        8.37     4.41     11.47    4.49
2        9.05     8.32     12.8     9.63
3        13.44    6.06     2.04     6.65
         16.81    10.93    8.07     15.61
          2       16.85    11.25    11.77
          4       13.41    9.97     10.08
          .6      14.87    7.08     16.34
         6.44     16.92    8.91     12.93
9        13.86    14.32    13.69    8.46
10       8.85     12.18    7.23     10.57
11       16.05    14.37    13.55    18.32
12       9.9      15.59    22.97    12.81
```

exampleBFL.adh
exampleBFL.egv
exampleBFL.tab

Use method: ☑
Scanning step: 1.0 cM
LOD threshold: ● Manual input
○ Permutation t...
Thres... 2.5
PIN: 0.001

ℹ Message

```
Welcome to the use of VHP.exe!
 VHP.exe starts at 15:24:10.777 on
 VHP.mio read successfully!
 exampleBFL.vhp read successfully
 Analysis of variance (ANOVA) is ru
 VHP.exe ends at 15:24:10.794 on
```

C:\ProjectGAHP\VHP\exampleBFL.22041415225\exampleBFL.vhp

FIGURE 3
The interface of functionality VHP.

TABLE 1 Naming of the combined QTL mapping in functionalities QHP and SHP.

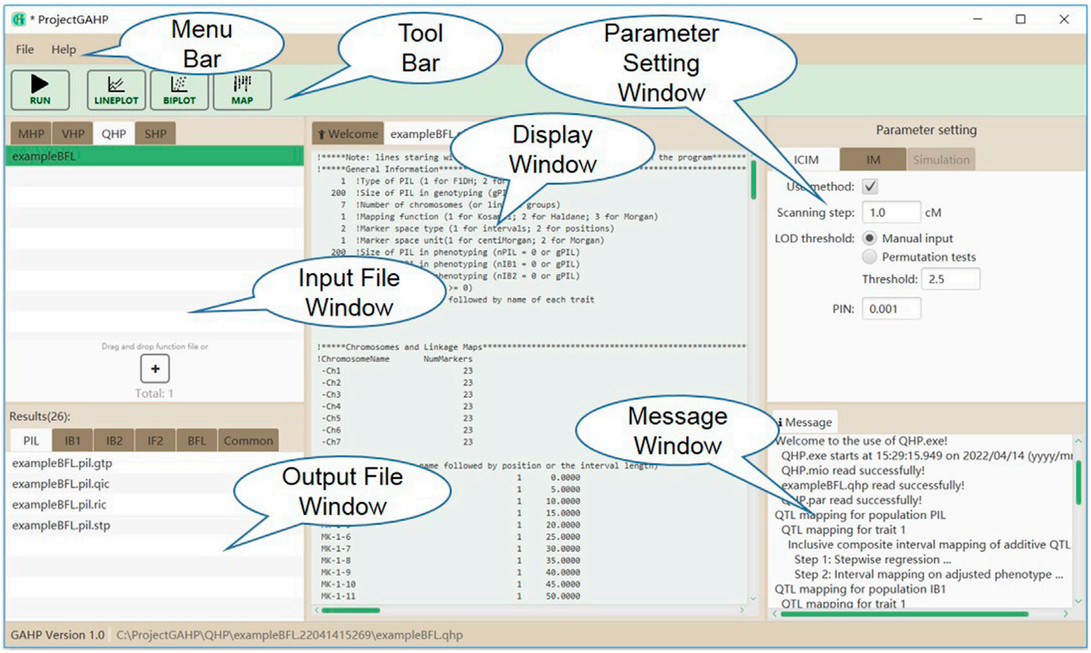
Combined analysis	Populations needed
IBC	IB1 and IB2
IFL	IF2 and PIL
IBL	IB1, IB2 and PIL
IBF	IB1, IB2 and IF2
BFL	IB1, IB2, IF2 and PIL

window. No parameter is needed to run this functionality. While the input file is properly loaded, the users may click “Run” on the tool bar to run the functionality.

2.6 The QHP functionality

As many as four populations, i.e. PIL, IB1, IB2, and IF2, can be used in QTL mapping either independently or together in functionality QHP, depending on the populations available. Firstly, the included populations are analyzed independently. Independent analysis is named by the respective population. Secondly, combined analysis is conducted using the included populations as many as possible. Names of the combined analysis are given in Table 1. Combined analysis using populations IB1 and IB2 is named by IBC; using populations IF2 and PIL is named by IFL; using populations IB1, IB2, and PIL is named by IBL; using populations IB1, IB2, and IF2 is named by IBF; and using populations IB1, IB2, IF2 and PIL is named by BFL (Table 1). The input file for QHP is composed of eight parts: 1) general information of mapping populations, 2) marker number information, 3) linkage map information, 4) marker types of PIL, 5) phenotype of PIL, 6) phenotype of IB1, 7) phenotype of IB2, and 8) phenotype of IF2. If one population has no phenotypic data, the corresponding part in the input file is left to be empty.

Supplementary Figure S3 represents an example of input file for QHP, where all the four populations have phenotypic values. Eleven parameters are included in general information (Supplementary Figure S3A): (1) type of pure lines in PIL, i.e. 1 for DHs, and 2 for RILs; (2) size of PIL in genotyping, i.e. number of genotyped pure lines in PIL (denoted as gPIL); (3) number of chromosomes or linkage groups; (4) mapping function, i.e. 1 for Kosambi’s function, 2 for Haldane’s function, and 3 for Morgan’s function; (5) marker space type, i.e. 1 for marker positions, and 2 for marker intervals; (6) marker space unit, i.e. 1 for centi-Morgan, and 2 for Morgan; (7) size of PIL in phenotyping; (8) size of IB1 in phenotyping; (9) size of IB2 in phenotyping; (10) size of IF2 in phenotyping; and (11) number of traits, followed by name of each trait. Population sizes



exampleBFL

Drag and drop function file or

+

Total: 1

Results(26):

PIL | IB1 | IB2 | IF2 | BFL | Common

exampleBFL.pil.gtp
exampleBFL.pil.qic
exampleBFL.pil.ric
exampleBFL.pil.stp

```
!*****Note: lines staring wi                          the program*******
!*****General Information*****                        *******************
     1  !Type of PIL (1 for F1DH; 2 for
   200  !Size of PIL in genotyping (gPI
     7  !Number of chromosomes (or lin     groups)
     1  !Mapping function (1 for Kosam  ; 2 for Haldane; 3 for Morgan)
     2  !Marker space type (1 for intervals; 2 for positions)
     1  !Marker space unit(1 for centiMorgan; 2 for Morgan)
   200  !Size of PIL in phenotyping (nPIL = 0 or gPIL)
                  in phenotyping (nIB1 = 0 or gPIL)
                   henotyping (nIB2 = 0 or gPIL)
                   >= 0)
                   followed by name of each trait


!*****Chromosomes and Linkage Maps******************************************
!ChromosomeName       NumMarkers
 -Ch1                         23
 -Ch2                         23
 -Ch3                         23
 -Ch4                         23
 -Ch5                         23
 -Ch6                         23
 -Ch7                         23

                  name followed by position or the interval length)
                                        1      0.0000
                                        1      5.0000
                                        1     10.0000
                                        1     15.0000
                                        1     20.0000
 MK-1-6                                 1     25.0000
 MK-1-7                                 1     30.0000
 MK-1-8                                 1     35.0000
 MK-1-9                                 1     40.0000
 MK-1-10                                1     45.0000
 MK-1-11                                1     50.0000
```

Parameter setting

Use method: ☑

Scanning step: 1.0 cM

LOD threshold: ● Manual input ○ Permutation tests

Threshold: 2.5

PIN: 0.001

```
Welcome to the use of QHP.exe!
  QHP.exe starts at 15:29:15.949 on 2022/04/14 (yyyy/m
  QHP.mio read successfully!
  exampleBFL.qhp read successfully!
   P.par read successfully!
QTL mapping for population PIL
  QTL mapping for trait 1
    Inclusive composite interval mapping of additive QTL
      Step 1: Stepwise regression ...
      Step 2: Interval mapping on adjusted phenotype ...
QTL mapping for population IB1
  QTL mapping for trait 1
```

GAHP Version 1.0   C:\ProjectGAHP\QHP\exampleBFL.22041415269\exampleBFL.qhp

FIGURE 4
The interface of functionality QHP.

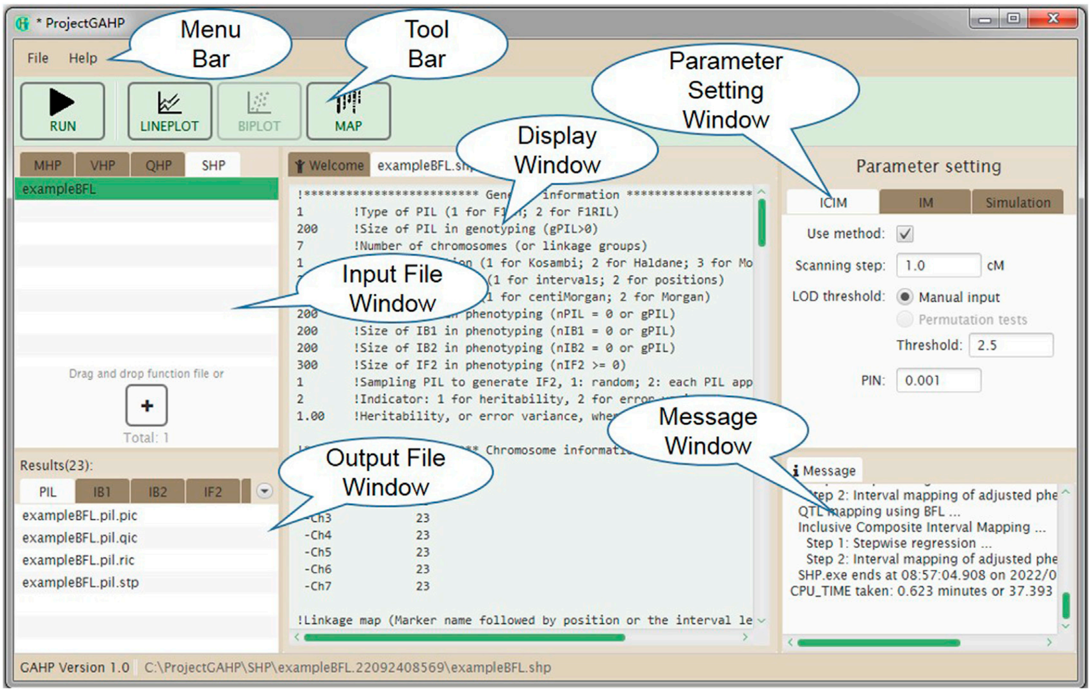
of PIL, IB1, IB2 and IF2 in the example as given in Supplementary Figure S3A were equal to 200, 200, 200, and 300, respectively. Kosambi's mapping function was used to convert recombination frequency to marker distance. Markers on the seven chromosomes were defined by positions. The unit of marker space was cM, and the number of phenotypic traits was equal to 1, named by simuTait. Marker number and linkage map information were given in Supplementary Figures S3B and S3C, respectively. Genotypic data at all polymorphic markers for all pure lines in PIL was given in Supplementary Figure S3D. Phenotypic values of the four populations were given in Supplementary Figures 3E–3H, respectively. As for functionality QHP, sizes of populations PIL, IB1, and IB2 have to be equal, if included.

Figure 4 shows the interface of functionality QHP. Input files are grouped on the QHP tab in the input file window. Mapping parameters can be set in the parameter setting window, located at the right side of the interface. Two mapping methods are available in QHP, i.e., 1) IM: the conventional interval mapping for additive and dominant QTLs (Lander and Botstein, 1989); 2) ICIM: inclusive composite interval mapping for additive and dominant QTLs (Li et al., 2007; Zhang et al., 2008). After the mapping method selection and parameter setting, the users may click the "Run" button in the tool bar to run the functionality. Mapping results will be listed in the output file window, when the functionality is completed successfully.

2.7 The SHP functionality

In functionality SHP, populations PIL, IB1, IB2 and IF2 are generated for a set of predefined QTLs, and then power analysis is conducted on the simulated populations. Similar to functionality QHP, mapping methods IM and ICIM are provided in SHP. QTL mapping can be conducted in individual populations, as well as in multiple populations by combined analysis. Only one trait can be defined and simulated in one input file. The input file for SHP is composed of five parts: 1) general information of mapping populations, 2) marker number information, 3) linkage map information, 4) gene or QTL information, and 5) genotypic values of the predefined QTLs.

Supplementary Figure S4 represents an example input file to run functionality SHP, where all the four populations are simulated for power analysis. Thirteen parameters are included in general information of populations (Supplementary Figure S4A). The first ten parameters are same as those in functionality QHP. The other parameters are: (11) sampling PIL to generate IF2, i.e. 1 for random sampling, and 2 for sampling method that each line in PIL appears the same times in IF2; (12) indicator to define the content of the next parameter, i.e. 1 for heritability, and 2 for error variance; (13) heritability or error variance depending on the previous indicator, where F_2 is used as the reference population to convert between heritability and error variance. Name of each chromosome and number of markers on the chromosome are specified first



```
!************************ Gen   information *****************
1       !Type of PIL (1 for F1 ; 2 for F1RIL)
200     !Size of PIL in genotyping (gPIL>0)
7       !Number of chromosomes (or linkage groups)
1                    ion (1 for Kosambi; 2 for Haldane; 3 for Mo
                       (1 for intervals; 2 for positions)
                       1 for centiMorgan; 2 for Morgan)
200            phenotyping (nPIL = 0 or gPIL)
200     !Size of IB1 in phenotyping (nIB1 = 0 or gPIL)
200     !Size of IB2 in phenotyping (nIB2 = 0 or gPIL)
300     !Size of IF2 in phenotyping (nIF2 >= 0)
1       !Sampling PIL to generate IF2, 1: random; 2: each PIL app
2       !Indicator: 1 for heritability, 2 for error
1.00    !Heritability, or error variance, whe

                    ** Chromosome informati

 -Ch3          23
 -Ch4          23
 -Ch5          23
 -Ch6          23
 -Ch7          23

!Linkage map (Marker name followed by position or the interval le
```

Use method: ☑
Scanning step: 1.0 cM
LOD threshold: ● Manual input
○ Permutation tests
Threshold: 2.5
PIN: 0.001

```
  tep 2: Interval mapping of adjusted phe
 QTL mapping using BFL ...
 Inclusive Composite Interval Mapping ...
   Step 1: Stepwise regression ...
   Step 2: Interval mapping of adjusted phe
 SHP.exe ends at 08:57:04.908 on 2022/0
CPU_TIME taken: 0.623 minutes or 37.393
```

FIGURE 5
The interface of functionality SHP.

(Supplementary Figure S4B), followed by the definition of each chromosome (Supplementary Figure S4C). Each chromosome is defined by all markers located on, and the marker positions. The fourth part provides the number of QTLs or genes and their positions on each chromosome (Supplementary Figure S4D), and the fifth part provides the genotypic values of additive-dominant QTLs and epistatic networks (Supplementary Figure S4E).

Figure 5 shows the interface of functionality SHP. Input files are grouped on the SHP tab in the input file window. In addition to the parameters for mapping methods (similar to functionality QHP), those for the simulation purpose also need to be specified in the parameter setting window, including random seed, number of runs, indicator whether or not to output the simulated populations, and support interval in cM for counting the true and false QTLs detected in simulated populations. After mapping method selection and parameter setting, the users may click the "Run" button in the tool bar to conduct the population simulation and QTL detection power analysis.

3 Results

3.1 Outputs of the MHP functionality

For the four functionalities implemented in the current version of GAHP, most output files have the same prefix name as the input file but with different extension names. Output file with extension name '*.txt' is pure-text, providing the connection between interface and calculation kernel. There is only one output file after running MHP, named by 'LinkageMap.txt' (see the "common" tab in output file window in Figure 2), which contains the information of linkage maps given in the input file. GAHP provides the user-friendly interface to draw the linkage maps for individual chromosomes (Supplementary Figure S5A), or all chromosomes simultaneously (Supplementary Figure S5B). Options are provided for users to change the style of map drawing, including the position label, marker name, separator line, chromosome height, number of chromosomes per row, and gradient color.

3.2 Outputs of the VHP functionality

Three output files are generated after running the VHP functionality (see the "common" tab in output file window in Figure 3). Output with extension name '*.adh' contains the estimates of variance components and heritability (Supplementary Figure S6). The first part provides the estimates of genotypic variance (Vgeno), error variance (Verror), phenotypic variance (Vpheno), heritability in the broad sense (Hbroad), and degree of freedom of random error (DFerror) for each trait in each population. The second part provides the estimates of additive variance (Vadd_F2), dominant

TABLE 2 Description of output files from the QHP functionality.

Group	Extension name	Description of contents
Results related to individual population or combined QTL mapping	STP	Selected marker variables and their effects from the first step of stepwise regression in inclusive composite interval mapping (ICIM)
	QIM, QIC	QTL identified from interval mapping (IM), and ICIM
	RIM, RIC	Results at every one-dimensional scanning position from IM and ICIM
	TIM, TIC	LOD score from permutation tests for IM and ICIM
	GTP	Bayesian classification of genotypes at QTLs identified from ICIM
Common, i.e. results not related to QTL mapping	COE	Lower triangular matrix of pairwise correlation coefficient between markers in population PIL
	MTP	Frequency of marker types, Chi-square test for segregation distortion, and missing-imputed marker types
	STA	Descriptive statistics of phenotypes
	TXT	Three text files, i.e. 'LinkageMap.txt', 'Phenotype.txt' and 'Threshold.txt', are used for the connection between interface and QTL mapping kernel

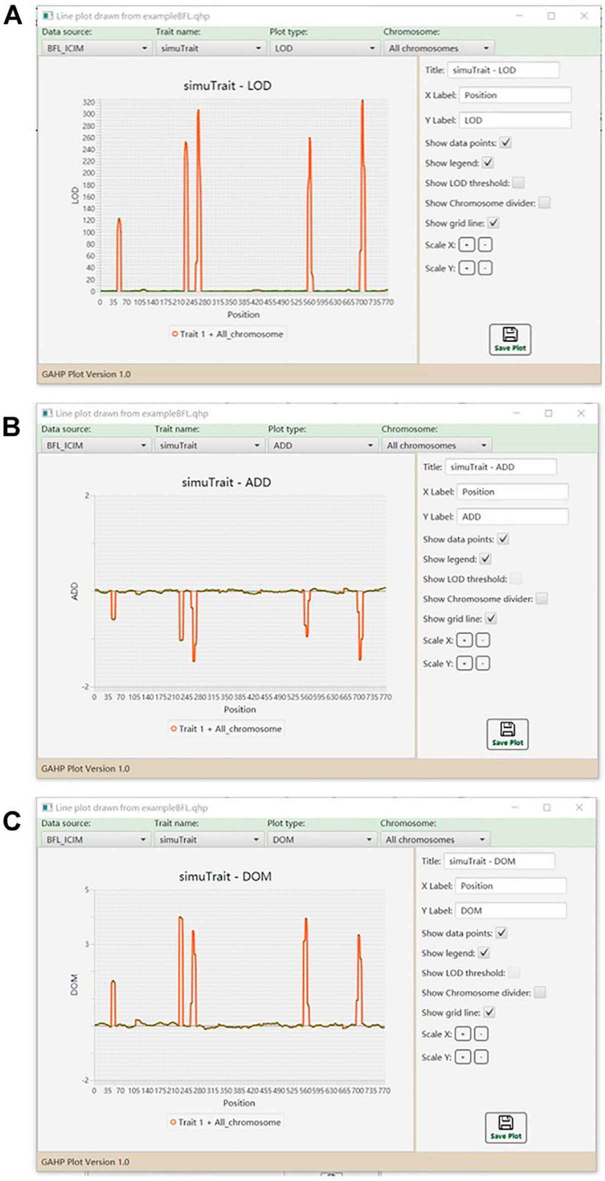
variance (Vdom_F2), error variance (Verror_F2), heritability in the narrow sense (Hnarrow_F2), and degree of freedom of random error (DFerror) from the combined ANOVA using all populations, where F_2 is assumed to be the reference population. Output with extension name '*.egv' contains the estimated genotypic value of each line in population PIL or each hybrid in populations IB1, IB2 and IF2 for each trait (Supplementary Figure S7). Output with extension name '*.tab' contains the conventional ANOVA table for each trait. As an example, Supplementary Figure S8 shows ANOVA tables of two traits in population PIL. All populations included in input files have their corresponding ANOVA tables in this output file.

3.3 Outputs of the QHP functionality

QHP is the key functionality in GAHP. Outputting results are grouped by names of individual population (i.e. PIL, IB1, IB2, or IF2) and combined QTL mapping (i.e. IBC, IFL, IBL, IBF, or BFL; see the lower left window in Figure 4). For output files arranged in each group, three lower case characters after the prefix indicate the group name, i.e. '*.pil', '*.ib1', '*.ib2', '*.if2', '*.ibc', '*.ifl', '*.ibl', '*.ibf', or '*.bfl'. The last three lower case characters are the extension name, indicating contents in each output. Each mapping method (i.e. IM, and ICIM) has three kinds of outputting information, which are labeled by Q for detected QTLs, R for results at every scanning position, and T for permutation tests (Table 2). For ICIM, two additional output files with extension names '*.stp' and '*.gtp' are provided, containing the results from stepwise regression, and the predicted genotypes at each detected QTL and genotypic values, respectively. As many as four mapping populations can be included, and thus there may be at most five groups of '*.stp', '*.gtp', Q, R and T output files, four for independent population mapping, and one for combined QTL mapping. As an example, Supplementary Figure S9 gives part of the content in output '*.bfl.ric' from simulated populations, i.e., mapping results from ICIM in combined mapping BFL (denoted as BFL-ICIM) at each scanning position; Supplementary Figure S10 gives the content in output '*.bfl.qic' from ICIM, i.e., information of the detected QTLs. For each QTL, the chromosomal position, nearest left marker, nearest right marker, total LOD score, LOD score for additive effect, LOD score for dominant effect, total phenotypic variance explained (PVE), additive PVE, dominant PVE, additive effect, dominant effect, and one-LOD confidence interval are reported.

Outputs not related to QTL mapping are listed under the 'Common' group (see the lower left window in Figure 4). There are six such output files recording the relevant information in mapping populations (Table 2). Output with extension name '*.coe' contains the pair-wise correlation coefficients between markers in population PIL, which may be used to check the quality of linkage maps. Output with extension name '*.mtp' contains marker summary, and marker types after the imputation of missing values. Output with extension name '*.sta' contains the descriptive statistics of phenotypic values in each population. Three text files, i.e. 'LinkageMap.txt', 'Phenotype.txt' and 'Threshold.txt' contain information of the linkage map, phenotypic values, and threshold LOD score, respectively, which are used for the connection between interface and QTL mapping kernel.

Graphs of LOD score and genetic effects on each chromosome or on all chromosomes are available in the QHP functionality. Figure 6 shows the one-dimensional profile of LOD score, additive and dominant effects on one trait in simulated populations from BFL-ICIM. Tool bars are provided for the users to select the source of data, and modify the parameters so as to change the style of graphs. Bi-plot graphs for phenotypic data are also available. For example, Supplementary Figure S11 shows the bi-plot for phenotypic data of individuals in population IF2 together with their mid-parental values.



Line plot drawn from exampleBFL.qhp

Data source: BFL_ICIM; Trait name: simuTrait; Plot type: LOD; Chromosome: All chromosomes

Title: simuTrait - LOD; X Label: Position; Y Label: LOD

Show data points: ☑ Show legend: ☑ Show LOD threshold: □ Show Chromosome divider: □ Show grid line: ☑ Scale X: + - Scale Y: + -

Line plot drawn from exampleBFL.qhp

Data source: BFL_ICIM; Trait name: simuTrait; Plot type: ADD; Chromosome: All chromosomes

Title: simuTrait - ADD; X Label: Position; Y Label: ADD

Show data points: ☑ Show legend: ☑ Show LOD threshold: □ Show Chromosome divider: □ Show grid line: ☑ Scale X: + - Scale Y: + -

Line plot drawn from exampleBFL.qhp

Data source: BFL_ICIM; Trait name: simuTrait; Plot type: DOM; Chromosome: All chromosomes

Title: simuTrait - DOM; X Label: Position; Y Label: DOM

Show data points: ☑ Show legend: ☑ Show LOD threshold: □ Show Chromosome divider: □ Show grid line: ☑ Scale X: + - Scale Y: + -

FIGURE 6
Line plots for QTL mapping results. **(A)** LOD score. **(B)** Additive effect. **(C)** Dominant effect.

3.4 Outputs of the SHP functionality

Similar to QHP, outputting results from functionality SHP are also grouped by names of individual population and combined QTL mapping (see the lower left window in Figure 5). For output files arranged in each group, three lower case characters after the prefix indicate the group name. The last three lower case characters are the extension name, indicating contents in each output. Each mapping method (i.e. IM, and ICIM) generates three kinds of output files, labeled by Q for detected QTLs, R for results at all scanning positions, and P for power analysis (Table 3). Output file '*.stp' is generated only for ICIM. There may be at most five groups of '*.stp', Q, R and P files, four for individual population mapping, and one for combined QTL mapping. By looking into the P output files, the users can compare the QTL detection power from different mapping methods. Formats of the Q and R outputs are similar to those from the QHP functionality, but the Q output files in SHP contain the detected QTLs from each simulation run, and the R output files in SHP contain the average LOD score and effects across all simulation runs. Supplementary Figure S12 gives part of the content in output file '*.bfl.pic' from an example input file. The first part contains the detection power, LOD score and estimated effects from ICIM for each QTL in simulation, and the second part contains the corresponding information for each marker interval.

Outputs not related to QTL mapping are listed under the 'Common' group (see the lower left window in Figure 5). One output has the name 'SHP.gmd', which is arranged in a format that can be directly used as the input of the Blib platform of genetics and breeding simulation, i.e., genetic model of the simulated trait (Table 3). Two text files, i.e. 'LinkageMap.txt' and 'Threshold.txt' contain information of the linkage map and threshold LOD score. If the check box "Outputting population" in the parameter setting window is clicked, the simulated populations are arranged in the format that can be directly used as input files for the QHP functionality.

SHP also provides the graphic option of LOD scores and genetic effects on one chromosome or on all chromosomes, averaged from all simulation runs, which are similar to functionality QHP.

4 Discussion

4.1 Applications of the GAHP software package in genetic studies

Heterozygous populations are needed in order to investigate the dominance-related genetic effects, which are critical to understanding the genetic mechanism of heterosis in plants. Conventional bi-parental F_2 are such populations, but have the disadvantage in conducting the multi-environmental and replicated phenotyping trials. As one replacement, immortalized F_2 populations can overcome the disadvantage and provide the estimates of additive, dominant and epistatic effects. In addition, genotyping is only needed on pure lines in population PIL, which are the direct parents of F_1 hybrids consisting of the immortalized population (Hua et al., 2003; Liu et al., 2020). Immortalized BC population with one parental line has only two genotypes at each locus, and therefore cannot provide the full information to estimate the dominant effect. However, when used together, immortalized BC populations at both directions to the original two parental lines can also be used in investigating the genetic basis of heterosis (Li et al., 2008b; Aakanksha et al., 2021).

TABLE 3 Description of output files from the SHP functionality.

Group	Extension name	Description of contents
Results related to individual population or combined QTL mapping	STP	Selected marker variables and their effects from the first step of stepwise regression in inclusive composite interval mapping (ICIM) for each simulation run
	QIM, QIC	QTL identified from interval mapping (IM), and ICIM
	RIM, RIC	Results at all one-dimensional scanning positions from IM and ICIM
	PIM, PIC	Power of predefined QTLs together with false positives from IM and ICIM
Common, i.e. results not related to QTL mapping	TXT	Two text files, i.e. 'LinkageMap.txt', and 'Threshold.txt', are used for the connection between interface and QTL mapping kernels
	GMD	Input file for the Blib simulation platform, which defines the genetic model on the simulated trait
	QHP (optional)	Simulated populations in the format that can be directly loaded to functionality QHP

GAHP is freely available from https://isbreeding.caas.cn. Users' manual and sample datasets are automatically included when the package is properly installed in local personal computers. A video tutorial is provided on the software webpage. GAHP can conduct the phenotypic data analysis, and QTL mapping on pure-line populations and their derived immortalized BC and F_2 populations, either separately or in combination. Both additive and dominant variances can be estimated by the combined ANOVA in the SHP functionality, by which the broad-sense and narrow-sense heritabilities can be calculated. Both additive and dominant effects of QTLs can be estimated by the combined QTL mapping on immortalized BC and F_2 populations in the QHP functionality. Combined mapping utilizes more populations, and improves the estimation accuracy of genetic variances, heritabilities, and positions and effects of QTLs. In addition, GAHP can simulate the four kinds of mapping populations (Figure 1), based on the user-defined information on linkage map, QTL locations and effects, and error variance (or heritability). Mapping results from the simulated populations allow the users to investigate of efficiency of genetic studies on immortalized populations. Furthermore, the SHP functionality in GAHP allows a perspective comparison of mapping methods through power analysis. QTL detection power is affected by many factors, such as population size, heritability of phenotypic trait, QTL locations and effects, marker density, and the linkage relationship between QTLs (Li et al., 2010). Evaluation of mapping methods can be based on QTL detection power and false discovery rate (FDR). A better mapping method in the sense of statistics should have higher detection power and lower FDR (Li et al., 2010). The SHP functionality provides an approach to comparing the mapping methods in immortalized populations by considering the factors affecting mapping efficiency. SHP can also be used to investigate the efficiency of combined analysis using different populations, effect of population size on QTL detection, and various crossing schemes in PIL to generate the IF2 population *etc*. When new mapping methods are developed, the simulated populations generated by SHP can be used to evaluate their efficiency.

4.2 Features of the GAHP integrated package

In most QTL mapping packages, only the independent population analysis is provided, such as QTL IciMapping (Meng et al., 2015), GACD (Zhang et al., 2015b) and GAPL (Zhang et al., 2019). The four kinds of populations that can be handled in GAHP are highly related (Figure 1), which provides the opportunity for combined analysis. Mapping accuracy of independent population in the QHP functionality is actually the same as the BIP functionality in QTL IciMapping (Li et al., 2007; Zhang et al., 2008; Meng et al., 2015). It is expected that the combined QTL mapping in QHP on multiple populations should provide more accurate estimation on QTL positions and effects. Functionality AOV in QTL IciMapping (Meng et al., 2015) and VHP in GAHP are both developed for phenotypic ANOVA and heritability estimation. AOV in QTL IciMapping is suitable for individual populations phenotyped in single-environmental or multi-environmental trials, by which only the broad-sense heritability can be estimated. VHP in GAHP is specifically designed for the four related populations as shown in Figure 1, by which both broad-sense and narrow-sense heritabilities can be estimated, since the additive and dominant variances can be separated by the combined ANOVA across populations. It should be noted that only the phenotypic values from single-environmental trials are acceptable in the current version of GAHP.

Linkage map used in functionality QHP is based on genotypes of pure lines in population PIL, which should be constructed by other software packages, such as QTL IciMapping (Meng et al., 2015; Zhang et al., 2020). There is no need to rebuild the linkage maps in immortalized BC or F_2 populations. Therefore, map construction is not considered in GAHP. Instead, functionality MHP is developed in GAHP to display the completed linkage maps. MHP can handle larger number of markers and make higher quality of linkage maps, in comparison with QTL IciMapping. In input files of functionality QHP, genotypes are only needed for population PIL; genotypes of hybrids in immortalized BC and F_2 populations can be derived from pure lines and two original inbred parents. When using functionalities VHP and QHP, it is expected that the phenotypic values of different populations are collected in the same environment so as to avoid the effect of environments and genotype by environment interactions.

Time spent in QTL mapping should be taken into consideration when a large number of markers are included. When populations PIL, IB1, IB2 and IF2 are fixed at a size of 1000, the time spent for SHP to complete one simulation run was around 1, 12 and 55 min for marker numbers 200, 2000 and 20000, respectively. The time spent in one run was to complete four independent population analysis, and one combined analysis. The time spent for independent population analysis was close to that in QTL IciMapping for the same dataset. The time spent for combined analysis is slightly longer than that for independent population. The current version of GAHP can handle a number of markers as much as 50000. In most bi-parental populations, number of polymorphic markers may be much smaller than 50000. When more markers are included, binning analysis can be conducted to reduce the marker number and running time.

4.3 Further refinement of the GAHP package

At present, only one-dimensional QTL mapping is available in GAHP. In addition to additive and dominant effects, epistasis is also an important source of variation of complex traits, which maintains the additive variance and assures the long-term genetic gain in breeding (Zhang et al., 2012). Epistasis plays an important role in genetic basis of heterosis as well (Hua et al., 2003). QTL by environment interaction (QEI) widely exists in plants. Studies on epistasis and QEI contribute to the better understanding of genetic architecture of quantitative traits and heterosis (Li et al., 2015; Liu et al., 2020). It can be imagined that the algorithms of epistatic and QEI mapping would be more complicated than that of additive and dominant mapping in one environment. Nevertheless, ICIM has been extended to epistatic and QEI mapping in bi-parental populations (Zhang et al., 2012; Li et al., 2015). In the future, we may consider the extension of ICIM to epistatic and QEI mapping using multiple immortalized populations, and implement the mapping algorithms in GAHP. In addition, heterosis can also be studied by diversity inbred lines and their F_1 hybrids obtained by suitable crossing designs. The hybrid population derived from a diversity of inbred lines has different structure from population IF2 as discussed in this study, which may require further studies on genetic analysis method. Once developed and validated, the analysis method can be added as a separate functionality to extend the applications of GAHP in genetic studies.

Data availability statement

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

Author contributions

LZ wrote the Fortran codes for ANONA, QTL mapping and simulation, and wrote the manuscript draft. XW wrote the JAVA scripts for project management, interfaces, menus, tool bars, and various visualization tools. KW designed the structure of GAHP and tested the package. JW designed the structure of GAHP, and wrote the Fortran codes for ANONA, QTL mapping and simulation. All authors read and revised the manuscript draft.

Funding

This work was supported by the National Key Research and Development Program of China (No. 2019YFD1101100), the National Natural Science Foundation of China (Project No. 31861143003), and the Agricultural Science and Technology Innovation Program of CAAS.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1021178/full#supplementary-material

References

Aakanksha, Yadava, S. K., Yadav, B. G., Gupta, V., Mukhopadhyay, A., Pental, D., et al. (2021). Genetic analysis of heterosis for yield influencing traits in *Brassica juncea* using a doubled haploid population and its backcross progenies. *Front. Plant Sci.* 12, 721631. doi:10.3389/fpls.2021.721631

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19 (7), 889–890. doi:10.1093/bioinformatics/btg112

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S., et al. (2003). Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. U. S. A.* 100 (5), 2574–2579. doi:10.1073/pnas.0437907100

Kao, C. H., Zeng, Z. B., and Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* 152 (3), 1203–1216. doi:10.1093/genetics/152.3.1203

Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121 (1), 185–199. doi:10.1093/genetics/121.1.185

Li, C., Yu, H., Li, C., Zhao, T., Dong, Y., Deng, X., et al. (2018). QTL mapping and heterosis analysis for fiber quality traits across multiple genetic populations and environments in upland cotton. *Front. Plant Sci.* 9, 1364. doi:10.3389/fpls.2018.01364

Li, H., Hearne, S., Bänziger, M., Li, Z., and Wang, J. (2010). Statistical properties of QTL linkage mapping in biparental genetic populations. *Heredity* 105 (3), 257–267. doi:10.1038/hdy.2010.56

Li, H., Ribaut, J.-M., Li, Z., and Wang, J. (2008a). Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theor. Appl. Genet.* 116 (2), 243–260. doi:10.1007/s00122-007-0663-5

Li, H., Ye, G., and Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics* 175, 361–374. doi:10.1534/genetics.106.066811

Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., and Li, X. (2008b). Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. *Genetics* 180 (3), 1725–1742. doi:10.1534/genetics.108.091942

Li, S., Wang, J., and Zhang, L. (2015). Inclusive composite interval mapping of QTL by environment interactions in biparental populations. *PLOS ONE* 10 (7), e0132414. doi:10.1371/journal.pone.0132414

Liu, J., Li, M., Zhang, Q., Wei, X., and Huang, X. (2020). Exploring the molecular basis of heterosis for plant breeding. *J. Integr. Plant Biol.* 62 (3), 287–298. doi:10.1111/jipb.12804

Liu, R., Meng, Q., Zheng, F., Kong, L., Yuan, J., and Lübberstedt, T. (2017). Genetic mapping of QTL for maize leaf width combining RIL and IF_2 populations. *PLOS ONE* 12, e0189441. doi:10.1371/journal.pone.0189441

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3 (3), 269–283. doi:10.1016/j.cj.2015.01.001

Shi, J., Wang, J., and Zhang, L. (2019). Genetic mapping with background control for quantitative trait locus (QTL) in 8-parental pure-line populations. *J. Hered.* 110 (7), 880–891. doi:10.1093/jhered/esz050

van Ooijen, J. W. (2009). *MapQTL®6, Software for the mapping of quantitative trait loci in experimental populations of diploid species*. Wageningen, Netherlands: Kyazma B.V.

Wang, S., Basten, C. J., and Zeng, Z. B. (2007). *Windows QTL cartographer 2.5*. Raleigh, NC: Department of Statistics, North Carolina State University.

Yang, J., Hu, C., Hu, H., Yu, R., Xia, Z., Ye, X., et al. (2008). QTLNetwork: Mapping and visualizing genetic architecture of complex traits in experimental populations. *Bioinformatics* 24 (5), 721–723. doi:10.1093/bioinformatics/btm494

Yi, Q., Liu, Y., Hou, X., Zhang, X., Li, H., Zhang, J., et al. (2019). Genetic dissection of yield-related traits and mid-parent heterosis for those traits in maize (*Zea mays* L.). *BMC Plant Biol.* 19 (1), 392. doi:10.1186/s12870-019-2009-2

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136 (4), 1457–1468. doi:10.1093/genetics/136.4.1457

Zhang, L., Li, H., Ding, J., Wu, J., and Wang, J. (2015a). Quantitative trait locus mapping with background control in genetic populations of clonal F_1 and double cross. *J. Integr. Plant Biol.* 57 (12), 1046–1062. doi:10.1111/jipb.12361

Zhang, L., Li, H., Li, Z., and Wang, J. (2008). Interactions between markers can be caused by the dominance effect of quantitative trait loci. *Genetics* 180 (2), 1177–1190. doi:10.1534/genetics.108.092122

Zhang, L., Li, H., Meng, L., and Wang, J. (2020). Ordering of high-density markers by the k-Optimal algorithm for the traveling-salesman problem. *Crop J.* 8 (5), 701–712. doi:10.1016/j.cj.2020.03.005

Zhang, L., Li, H., and Wang, J. (2012). The statistical power of inclusive composite interval mapping in detecting digenic epistasis showing common F_2 segregation ratios. *J. Integr. Plant Biol.* 54 (4), 270–279. doi:10.1111/j.1744-7909.2012.01110.x

Zhang, L., Meng, L., and Wang, J. (2019). Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses. *Crop J.* 7 (3), 283–293. doi:10.1016/j.cj.2018.10.006

Zhang, L., Meng, L., Wu, W., and Wang, J. (2015b). Gacd: Integrated software for genetic analysis in clonal F_1 and double cross populations. *J. Hered.* 106 (6), 741–744. doi:10.1093/jhered/esv080

Zhang, S., Meng, L., Wang, J., and Zhang, L. (2017). Background controlled QTL mapping in pure-line genetic populations derived from four-way crosses. *Heredity* 119 (4), 256–264. doi:10.1038/hdy.2017.42

TYPE Review
PUBLISHED 06 October 2022
DOI 10.3389/fgene.2022.1014947

Recent advances and challenges of rare variant association analysis in the biobank sequencing era

Wenan Chen[1]*†, Brandon J. Coombes[2]* and Nicholas B. Larson[2]*

[1]Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, United States, [2]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Zheyang Wu,
Worcester Polytechnic Institute,
United States
Ningbo Chen,
Northwest A&F University, China

*CORRESPONDENCE
Wenan Chen,
wenan.chen@stjude.org
Brandon J. Coombes,
coombes.brandon@mayo.edu
Nicholas B. Larson,
Larson.Nicholas@mayo.edu

†These authors have contributed equally to this work

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

RECEIVED 09 August 2022
ACCEPTED 22 September 2022
PUBLISHED 06 October 2022

CITATION
Chen W, Coombes BJ and Larson NB (2022), Recent advances and challenges of rare variant association analysis in the biobank sequencing era.
Front. Genet. 13:1014947.
doi: 10.3389/fgene.2022.1014947

Causal variants for rare genetic diseases are often rare in the general population. Rare variants may also contribute to common complex traits and can have much larger per-allele effect sizes than common variants, although power to detect these associations can be limited. Sequencing costs have steadily declined with technological advancements, making it feasible to adopt whole-exome and whole-genome profiling for large biobank-scale sample sizes. These large amounts of sequencing data provide both opportunities and challenges for rare-variant association analysis. Herein, we review the basic concepts of rare-variant analysis methods, the current state-of-the-art methods in utilizing variant annotations or external controls to improve the statistical power, and particular challenges facing rare variant analysis such as accounting for population structure, extremely unbalanced case-control design. We also review recent advances and challenges in rare variant analysis for familial sequencing data and for more complex phenotypes such as survival data. Finally, we discuss other potential directions for further methodology investigation.

KEYWORDS

rare variant, sequencing data, variant annotations, population structure, external controls, family-based design, complex phenotypes, case-control

Introduction

High-throughput next-generation sequencing (NGS) technologies, including whole-exome sequencing (WES) and whole-genome sequencing (WGS), are increasingly being applied in studies of both rare diseases and common complex traits. In contrast to the array-based genotyping commonly applied in genome-wide association studies (GWAS), WES/WGS can directly capture relevant variation not interrogated by common genotyping platform designs, including rare variants (RVs). Identifying rare variants is important because pathogenic rare germline mutations can cause many human diseases. For example, many SOD1 mutations can cause amyotrophic lateral sclerosis (ALS) (Sau et al., 2007), NF1 mutations can cause pediatric brain tumors (Campian and

TABLE 1 Comparison between CV and RV association analysis.

Considerations	CV association analysis	RV association analysis
Assays	Typically captured using inexpensive genotyping microarrays	Often requires NGS, especially for detecting extremely rare/novel variants
Number of variants tested	Often single variant based (e.g., GWAS)	Often multiple variants based due to low power of single-variant methods
Population structure	Confounding can be adequately controlled using PCA or mixed models	Rare variants are likely more recent and reflect finer subpopulations. May need either more PCs or specifically designed methods
Null distributions of test statistics	Ordinary asymptotic distributions work well	Null distributions are often complex mixtures and more sophisticated methods may be necessary
Use of annotations	Statistical test for each variant is often performed without relying on annotations	Due to the large number of rare variants in a region, annotations are often used to filter rare variants
Interpretation	Due to potential LD, single-variant associations may be tag-SNPs	May be unclear which RVs are "driving" a significant RV association result using aggregative testing, especially those considering both directions

Gutmann, 2017), RB1 mutations can cause retinoblastoma (Yun et al., 2011), and ETV6 mutations can cause pediatric acute lymphoblastic leukemia (Hock and Shimamura, 2017). For adult cancers, mutations in BRCA1/BRCA2 can cause breast and ovarian cancer (Mavaddat et al., 2013), mutations in TP53 are responsible for many pediatric or adult cancers or syndromes (Olivier et al., 2010). Mutations in APP, PSEN1, PSNE2 can increase the risk of early onset Alzheimer disease (Lanoiselee et al., 2017). Therefore, sequencing technologies have often been prioritized for studying both somatic and germline DNA mutations in cancers (Consortium, 2020), and germline pathogenic mutations in rare Mendelian diseases (Gilissen et al., 2011).

There is also increasing interest in exploring the contributions of RVs to variability in common complex traits, driven in large part by the phenomenon of "missing heritability" (Manolio et al., 2009). This missing heritability is defined by the commonly observed gap between complex trait heritability estimates from family-based studies and trait variation explained by common single-nucleotide polymorphisms (SNPs) discovered by large-scale GWAS, leading to the common-disease/rare-variant (CD-RV) hypothesis (Schork et al., 2009). The CD-RV genetic model postulates that common complex traits may be the result of multiple RVs that impact one or multiple genes that would not be tagged by conventional GWAS SNPs. RVs have also largely remained unexplored in the GWAS era of genetic association analysis, and the vast majority of human genetic variation is rare. Technology and sample sizes have started to bear this hypothesis out, as RVs have recently been shown to account for unexplained heritability in highly polygenic traits, such as height and BMI (Wainschtein et al., 2022). Given the increasing empirical evidence that RVs play a role in various complex traits, cancers and rare diseases, such as results from WES profiling of the United Kingdom Biobank (Wang et al., 2021), NGS is increasingly being used to investigate RV associations in risk of human disease.

Unlike common variants (CVs), application of traditional single-variant analysis methods on RVs is often underpowered for typical NGS study sample sizes due to low minor allele frequencies (MAFs). The multiple testing burden for single RV analysis also increases as a function of sample size due to the fact that more unique RV positions will be detected. Consequently, adequate power for single-variant RV analyses requires extremely large sample sizes that often are practically and/or economically unfeasible. Moreover, it is possible *via* allelic heterogeneity that multiple RVs within a gene may affect the same trait. Therefore, RV analysis using NGS data is typically performed using "aggregative" testing, whereby identified variants are tested collectively in some fashion based on physical overlap with pre-defined genomic regions. Table 1 shows a comparison between CV and RV association analysis.

In this review, we discuss emerging challenges and methodological advancements in RV association analysis, covering topics related to variant filtering and annotation, population structure, implications of study design and use of externally-sequenced control samples, and adaptation of existing methods to different phenotypes. With the growing availability of DNA sequencing datasets with sufficiently large sample sizes for well-powered RV association analyses, the content of this review is particularly topical as investigators focus their attention on the role of RVs in human traits.

Background on RV association testing methodology

While many RV testing methods have been available for over a decade, they may still largely be considered niche even among genetic epidemiologists given the only recent emergence of DNA sequencing datasets with sufficiently large sample sizes. In this section, we briefly review a basic background of RV association analysis, orienting the reader to core concepts that contextualize modern methodological challenges and advancements.

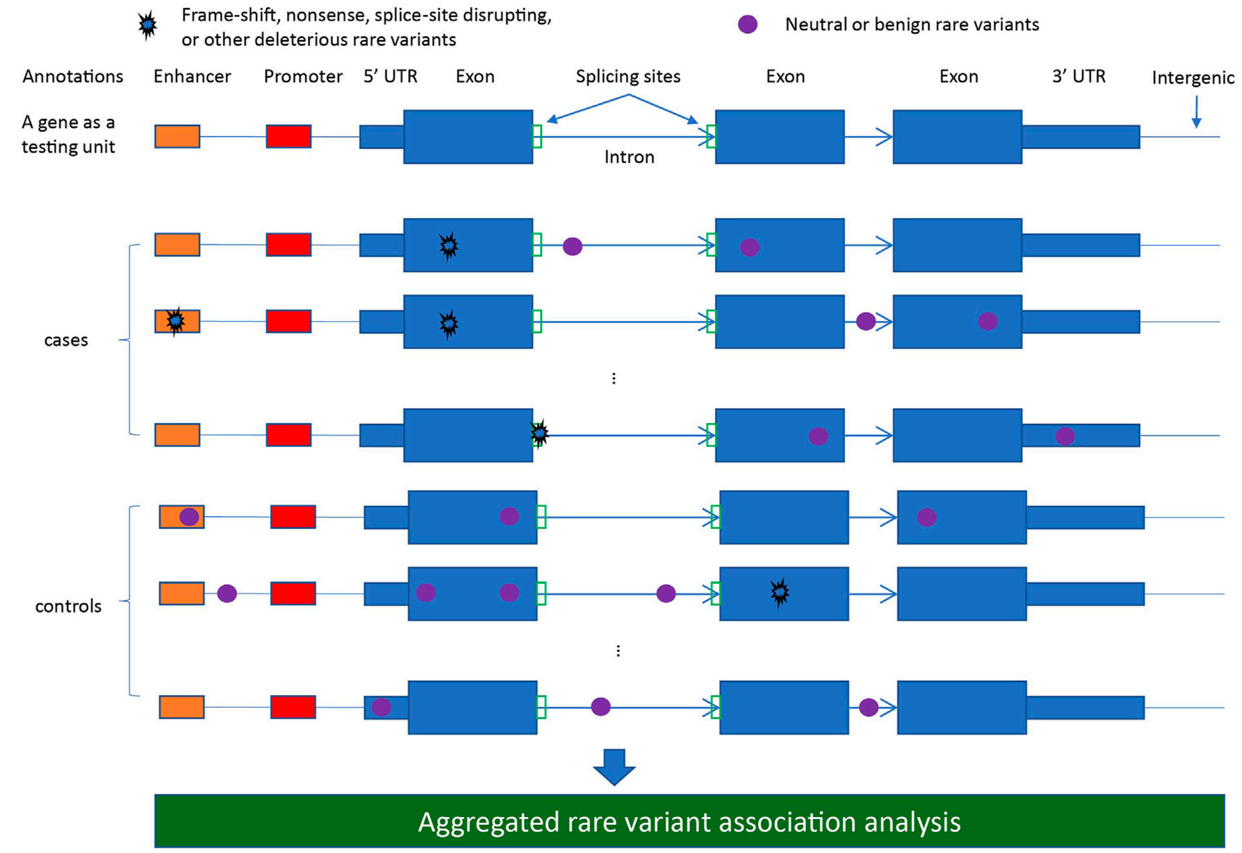

Frame-shift, nonsense, splice-site disrupting, or other deleterious rare variants
Neutral or benign rare variants
Annotations
Enhancer
Promoter
5′ UTR
Exon
Splicing sites
Exon
Exon
3′ UTR
Intergenic
A gene as a testing unit
Intron
cases
⋮
controls
⋮
Aggregated rare variant association analysis

FIGURE 1
A diagram illustrating different rare variant types defined from annotations for aggregated rare variant association analysis.

What is "rare"?

No formal threshold is defined for what qualifies a variant as an RV. For GWAS, minimum MAF thresholds are often applied to exclude SNPs that are underpowered for single-variant association analysis - typically in the range of 0.5%–5.0%, depending on available sample size. Current convention partitions variants into ultra-rare, rare, low-frequency, and common, with respective population MAF thresholds of 0.05%, 1% and 5% often observed in the literature. For RV association testing, this definition is more readily important, as it defines which variants are eligible for analysis. While this threshold is left to the investigator, 1% and 5% thresholds are commonly applied in practice for common complex traits, while even lower MAF (e.g., 0.1% or 0.05%) have been used for cancer predisposition variants or rare Mendelian diseases.

Defining variant sets

Conducting aggregative testing naturally requires defining eligible variant sets for analysis, which generally is akin to defining genomic region(s) by which overlapping RVs are grouped. Such regions should be defined *a priori*, as they 1) enumerate the anticipated multiple testing burden and 2) prevent overfitting *via* selection of genomic regions that correspond to chance RV enrichment. The most commonly applied region-based testing unit is a gene (Figure 1), particularly for large-scale agnostic scans (e.g., WES/WGS). More focused candidate gene studies may examine a finer regional granularity, such as individual exons or protein functional domains. Alternative approaches to standard region-based testing include scan-type statistics (Ionita-Laza et al., 2012; Schaid et al., 2013b), where the testing unit is a sliding genomic window, and pathway/gene-set testing (Wu and Zhi, 2013), where gene-level results may be further combined across biologically-related sets of genes.

Types of RV tests

Many aggregative RV analysis methods have been proposed in the literature, with the majority falling into two broad classes: 1) burden tests and 2) variance-component, or "kernel", tests. For the latter, the set-based Sequence Kernel Association Test (SKAT) (Wu et al., 2011) and its variations (e.g., SKAT-O (Lee et al., 2012)) are most widely applied, although other competing approaches and modifications have been developed.

First, we must define some relevant notation for RV testing. Specifically, let us consider a sequencing-based genetic association study of N samples on some phenotype of interest, defined by vector $Y_{N\times 1}$. For our purposes, we assume Y to be continuous or binary in nature, as these phenotype classes are broadly supported by most statistical methods for RV association analysis. We define available genotype allelic dosage data on M identified variants, $G_{N\times M}$, such that $G_{ij} \in \{0, 1, 2\}$. Some methods also allow for covariate adjustment (e.g., age and sex), and we define the set of P additional adjusting covariates by the matrix $X_{N\times P}$.

The first class of RV tests is a burden test. Generally, a burden test generates a test statistic based upon a (potentially weighted) sum of observed RVs, which implicitly assumes that causal variants share effect directionality (e.g., benign vs deleterious). The RV burden for subject i may then be calculated as $B_i = \sum_{j=1}^{M} w_j G_{i,j}$, where optional variant weights are defined as $w = (w_1, \ldots, w_M)'$. These weights should be defined to reflect relative confidence in causal status and/or anticipated magnitude of effect on the phenotype of interest, and well-informed weight definitions can substantially impact analysis results. One of the simplest burden testing procedures is the collapsing and sum test (CAST) (Morgenthaler and Thilly, 2007), which is a 2 × 2 Fisher's Exact Test for a binary RV carrier status for case-control studies. In this test, burden is further reduced to an indicator variable $B_i^* = I(B_i > 0)$ and samples are classified in the contingency table by burden status. The concept of variant burden has been generalized to a large number of testing frameworks where a univariate exposure is compared to an outcome of interest (e.g., combined multivariate and collapsing test, weighted-sum statistic test). Burden measures can also be used as predictors in regression analysis if additional covariate adjustment is desired. Adaptive burden testing approaches were developed to incorporate data-driven approaches to weighting and filtering of variants, including the variable threshold test (Price et al., 2010) (Han and Pan, 2010). Many of these adaptive burden tests employed permutations to compute p-values, which can be computationally burdensome. The burden test can also be framed as a score test to derive analytical p-values, such that the statistic $Q_B = (\sum_{i=1}^{N} (Y_i - \hat{Y}_i) B_i)^2$ follows a scaled χ_1^2 distribution, where $\hat{Y}_i$ is the predicted value of Y_i under a null model $Y_i = \beta_0 + X_i\beta + \epsilon_i$, X represents non-genetic variables with effects β, and the genetic effects corresponding to G are all fixed at zero.

In contrast to burden tests, kernel tests are robust to the presence of non-causal variation and heterogeneity of effect directionality. These tests are based upon measures of genetic similarity in the form of a kernel matrix $K_{N\times N}$, where $K_{i,j} = \kappa(G_i, G_j)$ for some kernel function $\kappa(\cdot,\cdot)$ describing the similarity between genotype vector G_i of subject i and genotype vector G_j of subject j. A common kernel function is the weighted linear kernel, such that $K = GWWG'$, where $W = diag(\sqrt{w_1}, \ldots, \sqrt{w_M})$ for the vector of marker weights w. The score test statistic is then given by the equation $Q = (Y - \hat{Y})' K (Y - \hat{Y})$ where $\hat{Y}$ is the predicted value of Y under the null model. The null distribution for Q then follows a mixture of χ^2 distributions, which can be well-approximated by a variety of methods or exactly computed.

SKAT was later extended to a generalized framework that includes formulation of a kernel function for the burden test score statistic, Q_B. SKAT-O, aka "Optimal SKAT" (Lee et al., 2012), is a type of hybrid approach to RV testing that optimally combines both burden and kernel statistics, Q_B and Q_S, respectively, into a weighted average, such that $Q_\rho = \rho Q_B + (1-\rho) Q_S$. Selection of ρ is conducted by SKAT-O using a simple grid search over the unit interval. Also known as omnibus tests, methods like SKAT-O are data-adaptive and consider a broad spectrum of potential genetic architectures rather than selecting one over the other. In general, there is no uniformly most powerful test across all potential conditions, since factors such as magnitude and direction of effect sizes, relationships between effect size and MAF, and proportion of causal variation all influence the relative power for a given test. While the robust property of kernel tests has great appeal, a burden test will be more powerful under conditions of high causal variant proportion. For large agnostic scans (e.g., WES/WGS studies), flexible omnibus tests like SKAT-O are often recommended.

Many other RV methods have been proposed that are neither burden tests nor variance component tests, such as the replication-based test (RBT) or p-value combination methods. The RBT instead tests for enrichment of rare alleles in cases and controls (Ionita-Laza et al., 2011). Alternatively, p-value combination methods combine the group of RV p-values in a given gene using either Fisher-like Method such as Fisher (Derkach et al., 2013), TFisher (Zhang et al., 2020), GFisher (Zhang and Wu, 2022), or some other transformation. These methods include the Aggregated Cauchy Association Test (ACAT) (Liu et al., 2019) which transforms p-values to the Cauchy distribution, the Higher Criticism (HC) or generalized HC test which combines ordered p-values using the HC statistic (Xuan et al., 2014) (Barnett et al., 2017), and the Generalized Berk-Jones (GBJ) test (Liu et al., 2021).

TABLE 2 Outline of advances and challenges of RV association analysis.

Topic	Motivation/Challenges
Incorporating variant annotations	There is growing knowledge available on potential variant impact on protein structure and function, and annotations may provide useful information in selecting functional variants. However, relevant annotation may vary by gene and phenotype, and annotation-informed filtering/weighting of variants may lead to improved or decreased statistical power
Accounting for population structure	Population structure is a primary confounding factor in genetic association analysis, and properly controlling for these confounding effects may differ relative to common variants
Accounting for extremely unbalanced case-control designs	Large biobanks with rare outcomes have led to extremely unbalanced case-control designs. This inflates type I error of standard RV methods relying on large sample theory based asymptotic distributions
Increasing power using external controls	To reduce the sequencing cost, often only cases and few controls are sequenced. In order to perform RV association analysis, external controls are used. One main challenge of this design is the potential confounding batch effect from different sequencing and processing platforms between cases and controls
Analysis of familial sequencing data	Family based design has the advantage of being robust from population structure, it is also the standard way for heritability estimation. It is important that RV association analysis methods can accommodate studies using the family based design
Allowing for more complex phenotypes	While case-control studies and analyses of quantitative traits are most common in RV analysis, RV methods have also been developed for multivariate phenotypes and time-to-event outcomes

Recent advances and challenges in RV association analysis

RV association studies present a number of unique challenges that have driven methodological development in the last decade; however, many challenges remain outstanding. We summarize our review of recent advances and challenges of RV association analysis in Table 2. The essential themes in these topics align with fundamentals of hypothesis testing: type I error control, maximizing the statistical power, and how to model different data types in a statistical test. For example, accounting for population structure and extremely unbalanced case-control designs address the challenge of inflated type I error in RV association tests. Incorporating variant annotations and using external controls aim to increase the statistical power of RV association tests. Analysis of familial sequencing data needs to model the inheritance patterns of genotypes and genotype correlations among family members, treating related samples as unrelated will lead to inflated type I error. Analysis with more complex phenotypes requires modeling the additional complexities in phenotypes in order to achieve well controlled type I error and powerful test results. We provide a more detailed review of each topic in the following sections.

Incorporating variant annotations in RV analysis

The statistical power of most aggregative RV testing methods is highly dependent on the proportion of truly causal variants included in the RV set. Given that the functional relevance status of individual variants is generally not known *a priori*, variant filtering and/or weighting is common practice to leverage biological knowledge and improve power, and many RV testing methods are designed to flexibly accommodate variant weights in the testing procedure. For burden tests, it has been shown that the optimal weights will be proportional to the true absolute variant effect sizes (King and Nicolae, 2014). Absent any relevant functional annotation, weighting schemes based on MAF, such as the Madsen-Browning weights (Madsen and Browning, 2009) or beta density function weights (Wu et al., 2011), are commonly employed. This is motivated by an assumed inverse relationship between allele frequency and functional impact imposed by strong purifying selection pressure on highly damaging variants.

For gene-based RV testing, the simplest strategies incorporating annotation involve variant filtering based on the likely functional impact on the resulting protein product. Standard bioinformatics annotation tools (Wang et al., 2010) (Cingolani et al., 2012) (Mclaren et al., 2016) can rapidly assign basic qualitative functional variant effects based on the open reading frame of protein-coding gene transcript(s), and prioritization of loss-of-function variants (i.e., nonsense, splice-site disrupting, frame-shift indels) is commonly applied given the severity of the effects on the resultant protein structure. Variants that impose more modest changes to the amino acid sequence (i.e., missense, in-frame indels) may be more likely tolerated in relation to protein function, and a vast array of functional impact prediction tools have been developed to provide quantitative functional prediction scores to reflect the likelihood of deleteriousness (Livesey and Marsh, 2020). Synonymous and non-coding RVs may also impact a given gene through other mechanisms beyond direct alteration of the amino acid sequence, including disruption of regulatory sequences as well as epigenomic impacts. Many such annotations may also be cell-type specific, requiring consideration for the phenotype under study. Appropriate consideration for variant filtering and weighting may substantially improve statistical power for RV association discovery (Byrnes et al., 2013); conversely, misspecification of variant weights could lead to loss of power by inadvertently removing and/or down-weighting key disease-related functional RVs (Minica et al., 2017).

Given the number and heterogeneity of available variant annotations along with the uncertainty as to which annotations are most relevant to a particular gene-phenotype relationship, various methods have recently been proposed to dynamically accommodate and combine multiple annotations. For example, Wu et al. (Wu et al., 2013) proposed a multi-kernel approach using perturbation to perform kernel-based testing while simultaneously considering multiple candidate kernels, which could be defined by various competing weighting schemes. Due to the computational considerations of permutation/perturbation-based strategies, He et al. (He et al., 2017a) proposed the functional score test (FST), which similarly accommodates multiple candidate variant weighting schemes by partitioning the overall genetic effect attributable to the various annotation sources. The authors then apply a minP approach for combining test results across weight sets, and derive a computationally efficient resampling-based procedure for p-value calculation. More recently, Li et al. (2020) developed STAAR, which applies principal components analysis to matrices of various candidate annotation classes in order to reduce the annotation dimensionality. For gene-based testing, STAAR also considers testing stratified by variant classes, and all tests are then combined under an omnibus using the ACAT method.

Accounting for population structure in RV analysis

The primary confounding factor in genetic association analysis of both common SNPs and RVs is population stratification, which is the systematic difference in allele frequencies across sub-populations due to non-random mating and genetic drift. Various statistical methods have been successfully developed to address confounding by population stratification for common SNP association testing in genome-wide association studies. The most popular of these approaches include principal component analysis (PCA) (Price et al., 2006) and (generalized) linear mixed models (GLMMs) (Kang et al., 2010). PCA-based methods often address population stratification by adjusting for the leading PCs derived from the genotype-dosage matrix as covariates in a regression-based analysis. In contrast, GLMMs can simultaneously account for population stratification and cryptic relatedness by modeling a random effect whose covariance structure is defined by an estimated genetic relatedness matrix (GRM).

Since most modern RV association testing methods are also regression-based, both PC adjustment and GLMM-based strategies can be readily accommodated to address population stratification in RV analyses. However, it has been less clear whether the same methods applied for common SNPs can be similarly effective for RV association testing. From a population genetics perspective, it has been argued that RV associations are more prone to confounding effects of population stratification, as RVs are likely to be more recent and thus will reflect finer population substructure (e.g., regional geographic differences) (Mcclellan and King, 2010) (O'connor et al., 2015). To this end, a larger number of leading PCs could be required when performing RV testing to account for more nuanced population stratification (Mathieson and Mcvean, 2012). However, it has been shown that this may not be sufficient, as additional PCs derived from common SNPs may not capture fine-scale population stratification (Persyn et al., 2018). This is commensurate with other findings that demonstrate that common and RVs can reflect systematically different patterns of structure (Mathieson and Mcvean, 2012; Ma and Shi, 2020). Similarly, substantially different PCs may be obtained when derived from genotype matrices that are composed of common variants, RVs, and both (Liu et al., 2013; Ma and Shi, 2020).

Given the uncertainty as to how to properly account for population stratification in a regression-based analysis framework for RVs, alternative strategies based on sample matching have also been proposed. Matching based on genetic ancestry typically involves the use of leading PCs and makes less assumptions about the functional relationship of the PCs confounding the association between RV genotypes and outcome. Cheng et al. (2022) proposed a family of RV tests based on conditional logistic regression (CLoMAT), along with a matching algorithm based on PCA output. Another recently developed method used local permutations (LocPerm) to account for the population structure in the association test (Bouaziz et al., 2021; Mullaert et al., 2021). LocPerm first defines the K-nearest neighborhoods of each sample based on top PCs calculated from common variants. Then it selects permutations such that each phenotype is drawn from the K-nearest neighbors. Simulation results by the authors showed that LocPerm can control type I error rates under a variety of study conditions. However, the permutation procedure may require high computation cost when the sample size becomes large.

Accounting for extremely unbalanced case control design in RV analysis

The decrease in sequencing costs and the increase in large biobanks established around the world now enable researchers to identify the role of RVs in complex and sometimes rare outcomes (Backman et al., 2021). Many of these samples contain rich phenotypic data through surveys and questionnaires as well as linking to the electronic health record, which allows for investigation of RV associations phenome-wide. Barring any concerns of selection bias, it is generally optimal under these study conditions to include all genotyped samples in an association analysis. Since most diseases have a low prevalence in these biobanks, this leads to association tests with extremely unbalanced case-control samples. Many of the single RV and multiple RV tests mentioned above, such as SKAT and weighted versions of SKAT, take advantage of the score test framework to dramatically increase computational efficiency of RV tests by avoiding calculation of the likelihood or maximum-likelihood estimator under the full model. In the case of severe imbalance, violation of the large sample theory assumptions used to derive the asymptotic distribution leads to inflated type I error rates of the score test (Zhang et al., 2019). Recent methods have addressed this by applying either Firth regression (Wang, 2014) or a saddle-point approximation (SPA) (Zhou et al., 2018) to both single RV and multiple RV tests.

Firth regression uses a penalized likelihood approach to remove bias from the maximum-likelihood estimates. As the sample size increases, this penalization shrinks to zero; however, in the instance of extreme imbalance, this term helps maintain control of the type I error rate (Wang, 2014). A limitation of this approach involves requiring the calculation of the maximum likelihood under both the null and the full model for a likelihood ratio test, which is computationally expensive in large biobank-scale datasets and becomes impractical when considering RV testing across the genome. Alternatively, instead of assuming a normal approximation for the score test, application of SPA estimates the null distribution using all the cumulants hence all the moments in the case of severe imbalance and controls the type I error rates well (Dey et al., 2017).

The SPA approach is implemented in SAIGE (Zhou et al., 2018) and in REGENIE (Mbatchou et al., 2021) for testing single-variant association across the genome in the case of extreme imbalance. The SPA approach has also been used to extend SKAT and SKAT-O testing of multiple RVs and avoid the inflated type I error of those tests in the case of severe case-control imbalance (Zhao et al., 2020). REGENIE also alternatively implements approximate Firth regression to allow for usable SNP effect sizes because the SPA approach can sometimes fail to produce good estimates of SNP effect sizes and standard errors. A comparison of these methods in the United Kingdom Biobank testing for association in rare diseases found that SAIGE and REGENIE (SPA and Firth) appropriately controlled the type I error, but the SAIGE and REGENIE-SPA had inflated effect-size estimates (Mbatchou et al., 2021). Furthermore, REGENIE was 4.4 times faster than SAIGE in terms of CPU time (Mbatchou et al., 2021). Finally, the SPA approach has also been implemented in SPAGE to allow for scalable genome-wide single-variant gene-environment interaction analyses, which are well calibrated for severe case-control imbalance (Bi et al., 2019).

Using external controls in RV analysis

Because RV analysis often requires tens of thousands of samples to reach adequate statistical power, using available external sequencing data as a source of controls is a cost-effective approach for case-control RV association studies (Wojcik et al., 2022). One major challenge of using external controls is the potential confounding batch effect due to different sequencing platforms and genotype calling bioinformatics pipelines. For example, the sequencing depth between cases and controls can vary considerably if cases are WES samples (average depth 80x) and controls are low read depth WGS samples from the 1,000 Genomes Project (average depth 7x) (Genomes Project et al., 2015).

Several computational methods have been developed to address these challenges (Table 3). When individual sequencing data are available, statistical models have been developed to incorporate the read depth or genotype likelihood into the association test. Derkach et al. (2014) developed a score statistic that uses the expected genotype instead of the called genotype to account for the differences in read depth. Hu et al. (2016) developed a likelihood-based approach incorporating the sequencing reads depth directly without calling the genotypes; however, due to the direct use of raw sequencing reads, the computational cost might be high. Chen and Lin (2020) proposed regression calibration (RC)-based and maximum likelihood (ML)-based methods to incorporate the genotype likelihood in the association test and also allow inclusion of covariates to adjust for confounding, such as population structure. When internal controls are available, Li and Lee (2021) developed a weighted sum of score statistics to allow inclusion of both the internal and external controls by assessing the existence of batch effects between the internal and external controls for each variant.

Methods have also been developed using publicly available summary genotype counts of external controls, such as gnomAD (Karczewski et al., 2020). Since summary counts have less information than individual sequencing data, it is even more challenging to correct for batch effects between cases and external controls. When both internal controls and external summary counts are available, Lee et al. developed a method iECAT-O (Lee et al., 2017) that can use external summary counts when batch effects between internal and external controls cannot be detected. There are other methods developed that do not assume the existence of internal controls and aim to adjust for the batch effects between cases and external controls. ProxECAT (Hendricks et al., 2018) assumes the non-functional variants within a gene can be used as a proxy of how the variants are

TABLE 3 Summary of methods using external controls for improvement of statistical power.

Method	External control data	Require internal control?	Require sequencing depth for cases and controls?	Method correcting for batch differences between case controls	Can the method adjust for covariates?	Test
RVS (Derkach et al., 2014)	Individual genotype likelihood	N	N	Modeling the effect of sequencing depth	N	Single variant based test, burden test and variance component based test
TASER (Hu et al., 2016)	Individual Bam files	N	N	Modeling the effect of sequencing depth	N	Burden test
Chen and Lin (Chen and Lin, 2020)	Individual genotype likelihood	N	N	Modeling the effect of sequencing depth	Y	Single common variant based test
iECAT-Score (Li and Lee, 2021)	Individual genotypes	Y	N	Only use the external control if no batch effect exists	Y	Single variant based test for common and rare
iECAT-O (Lee et al., 2017)	Summary counts	Y	N	Only use the external control if no batch effect exists	N	A combination of burden test and variance component based test
ProxECAT (Hendricks et al., 2018)	Summary counts	N	N	Use non-functional variants as a baseline in the test	N	Burden test based on rare allele counts
TRAPD (Guo et al., 2018)	Summary counts	N	≥ 10 in 90% of samples	Adjusting filtering criteria	N	Burden test based on sample counts
RV- EXCALIBER (Lali et al., 2021)	Summary counts	Preferred	≥ 20 in 90% of samples	Adjust the expected counts sample-wise and gene-wise	N	Burden test based on rare allele counts
CoCoRV (Chen et al., 2022)	Summary counts	N	≥10 in 90% of samples	Consistent filtering to keep high quality variants	N	Burden test based on sample counts

sequenced and called. The total number of rare alleles from functional variants and non-functional variants are then compared between cases and controls. TRAPD (Guo et al., 2018) uses coverage summary statistics to keep high quality positions and then uses synonymous variants to tune variant filtering parameters between cases and controls. A burden test is used assuming RVs are independent from each other and thus can be pooled together from summary counts of individual variants. RV-EXCALIBER (Lali et al., 2021) also uses coverage summary statistics to keep high quality positions, instead of using the raw summary counts from public controls, it adjusts them using gene-wise and sample-wise correction factors and then compares the corrected values from public controls with observed values in cases. In addition to using coverage summary statistics to filter variants, a recently developed method CoCoRV (Chen et al., 2022) can provide consistent filtering between cases and controls. It also uses a blacklist to filter out potential problematic variants that show large discrepancies between the WES and WGS cohort. CoCoRV also provides a way to handle RVs in high linkage disequilibrium (LD) and can perform ethnicity-stratified association analysis which ameliorates potential confounding due to population structure.

A notable limitation of methods using summary counts is that they cannot adjust for covariates, given that only the summary information is available for controls. Therefore, adjusting for the confounding due to population structure in these methods remains challenging. Careful matching of race/ethnicity between cases and controls is critical in these analyses. Given that high-coverage WES (~80x) and WGS (~30x) external control data are becoming more and more common, evaluating the performance of methods modeling sequencing depth directly or using simple read-depth based filtering criterion would provide guidance on how to combine sequencing data sets in association tests.

RV analysis of familial sequencing data

Familial or pedigree-based design has the advantage of being robust to population stratification when using proper analysis methods. It is also indispensable if the interest is to study the effect of pathogenic *de novo* variation on risk of the disease. In addition, pedigree data from previous linkage mapping efforts might be sequenced for additional analysis (Ott et al., 2015).

Recent advances in RV association analysis for pedigree data in general can be summarized into two categories. The first category includes methods developed to analyze RVs based on the transmission disequilibrium test (TDT) or family-based association test (FBAT) (Laird and Lange, 2006). The second category includes the association test methods that adjust for relatedness and population structure using mixed models.

RV association analysis for unrelated individuals has been introduced to FBAT, which is robust to the presence of population structure. For example, the burden test was introduced to FBAT by De et al. (2013). Ionita-Laza later introduced the SKAT-type test to FBAT (Ionita-Laza et al., 2013) and showed that the statistical power for dichotomous traits was comparable between a family-based study for 500 trios and population-based study of 500 cases and 500 controls. Hecker et al. (2020) recently proposed a general framework for RV association tests including the burden test, SKAT-type test, and higher criticism based test, which was more powerful when the signal was sparse. By combining the p-values from different RV association tests using ACAT (Liu et al., 2019), Hecker et al. (2020) demonstrated the proposed method had robust and more powerful performance than other TDT extensions, such as RV-TDT (He et al., 2014), RV-GDT (He et al., 2017b), and gTDT (Chen et al., 2015). Under the FBAT model, the phenotype is treated as fixed and the genotypes as random variables. Because FBAT conditions on the phenotype, it is robust to different ascertainment schemes based on phenotypes, such as selecting pedigrees enriched with cases (Schaid et al., 2013a; Hecker et al., 2019). One disadvantage of FBAT is that it conditions on the parental genotypes and does not use between-family information (Schaid et al., 2013a; Ionita-Laza et al., 2013), which can result in loss of power compared with the association tests adjusting for relatedness using regression models.

The second category of association methods account for the relatedness in a regression model. Schifano et al. (2012) and Chen et al. (2013) developed similar RV association tests for a quantitative trait using a linear mixed model. These methods extend the SKAT method to handle pedigree data by including a random variable to account for the correlation between individuals within the same pedigree. The correlation matrix between individuals within a pedigree can be defined using twice the kinship coefficient (Sinnwell et al., 2014). If the pedigree information is not explicitly available, often the GRM estimated using genome-wide common variants is used. For binary traits, the logistic mixed model approach GMMAT was developed by Chen et al. (2016). To account for unbalanced case-control ratios using the saddlepoint approximation and efficient resampling as used in SAIGE (Zhou et al., 2018), Zhou et al. developed SAIGE-GENE (Zhou et al., 2020) using the generalized linear mixed model which can handle both binary and quantitative traits. For the mixed model methods, they regard the genotype as fixed and the phenotype as random. The relatedness within each pedigree is then included in the covariance matrix of the phenotype. Besides the mixed models, two similar retrospective likelihood-based methods, PedGene (Schaid et al., 2013a) and FARVAT (Choi et al., 2014) were also developed. As in FBAT, both methods treat the phenotype as fixed, and the genotype as random variables. The covariance matrix of genotypes incorporates both the LD information and the pedigree information, and a score statistic is derived. Power evaluations have shown that for quantitative traits, based on a recent review (Choi et al., 2014) (Larson et al., 2019), PedGene had similar power to the mixed model based methods developed by Schifano et al. (2012) and Chen et al. (2013). In addition to burden and SKAT-like tests, a robust SKAT-O-like method was also developed in FARVAT. FARVAT was written in C++ and has a speed advantage over PedGene. Evaluations (Wang et al., 2016; Fernandez et al., 2018) have shown that PedGene and FARVAT are usually more powerful than TDT based methods such as RV-TDT (He et al., 2014) or RV-GDT (He et al., 2017b). Even though the regression model based methods that account for the relatedness are likely more powerful than TDT based methods, how well they can account for the population structure might need further investigation (Mathieson and Mcvean, 2012).

For RV association analysis using pedigree data, because the two categories of methods have their own advantages and potential disadvantages, it might be a good idea to try methods in both categories and summarize their results for a robust interpretation of the data.

Allowing for more complex phenotypes in RV analysis

Many RV tests were developed to accommodate single binary and/or continuous outcomes. However, a given study may collect multiple and potentially highly related outcome measures. One extension of the above described methods is to consider these multiple correlated outcomes in order to increase statistical power and reveal potential pleiotropy. As is the case for testing association of multiple RVs with a single phenotype, testing for association of RVs with a multivariate outcome primarily uses either burden-like (Zhao and Thalamuthu, 2011; Zhu et al., 2015; Kaakinen et al., 2017) or SKAT-like (Ray et al., 2016; Liu and Lin, 2018; Dutta et al., 2019; Liu and Lin, 2019; Luo et al., 2020) approaches. Additional methods used are a standard MANOVA approach (Ferreira and Purcell, 2009) and a regression approach that flips the outcomes and RV predictor using proportional odds regression (MultiPhen) to test for association of a group of phenotypes with the RV as an outcome (O'reilly et al., 2012). However, no test among these is uniformly most powerful and many of these methods are sensitive to deviations from normality in the case of multivariate quantitative phenotypes (Ray and Chatterjee, 2020).

Another type of outcome that is especially common to biobanks is time-to-event data. Cox proportional hazards (PH) regression models are heavily used in this context, but fitting the maximum partial likelihood for these models is often not scalable to large GWAS. For that reason, kernel statistics using martingale residuals in place of residuals from a generalized linear model (e.g. SKAT) have been initially proposed for gene- or region-based RV testing across the genome (Chen et al., 2014; Larson et al., 2019), such as the method implemented in rareSurvival software (Syed et al., 2021). In the case of extremely unbalanced case-control designs, SPACox has been proposed to correct the inflated type I error rates in GWAS of RVs (Bi et al., 2020). This approach scales well by first fitting a Cox PH regression model only once across the genome-wide analysis and then using the SPA approach to calibrate the score statistics.

Discussion

In this review, we have covered the basic background on RV association testing using sequencing data, and outlined leading areas of methodological development in RV association analysis. The growth in availability of large datasets with RVs measured will finally allow researchers to assess the impact that RVs have on rare and common diseases. This growing availability of large sequencing data not only makes RV analyses feasible, but may yield novel analytical issues. For example, many analytical issues may occur when trying to coordinate RV analyses across multi-site/biobank studies where incorporating all datasets into one conglomerated analysis is near impossible due to data sharing concerns and patient privacy. This means that RV analyses will likely require federated analyses with each site performing the analysis at their respective site for which results are combined afterward. Given the large number of potential rare variants that may be involved in a significant result, questions also remain as to how to optimally validate rare variant findings and how to design large-scale functional validation assays of the findings. Regardless of these potential challenges, the methodological advancements we have highlighted in this review demonstrate a very active scientific community dedicated to tackling these issues.

Author contributions

WC, BC, and NL discussed and decided the scope of the review, selected topics, reviewed each topic and wrote the manuscript collectively. All authors approved the final version of the manuscript. WC, BC, and NL contributed equally to the review.

Funding

This study was supported in part by American Lebanese Syrian Associated Charities (ALSAC), the National Cancer Institute grant P30 CA021765 and 5U54NS092091-08.

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.

References

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., et al. (2021). Exome sequencing and analysis of 454, 787 UK Biobank participants. *Nature* 599, 628–634. doi:10.1038/s41586-021-04103-z

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

Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S., and Lee, S. (2020). A fast and accurate method for genome-wide time-to-event data analysis and its application to UK biobank. *Am. J. Hum. Genet.* 107, 222–233. doi:10.1016/j.ajhg.2020.06.003

Bi, W., Zhao, Z., Dey, R., Fritsche, L. G., Mukherjee, B., and Lee, S. (2019). A fast and accurate method for genome-wide scale phenome-wide G × E analysis and its application to UK biobank. *Am. J. Hum. Genet.* 105, 1182–1192. doi:10.1016/j.ajhg.2019.10.008

Bouaziz, M., Mullaert, J., Bigio, B., Seeleuthner, Y., Casanova, J. L., Alcais, A., et al. (2021). Controlling for human population stratification in rare variant association studies. *Sci. Rep.* 11, 19015. doi:10.1038/s41598-021-98370-5

Byrnes, A. E., Wu, M. C., Wright, F. A., Li, M., and Li, Y. (2013). The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet. Epidemiol.* 37, 666–674. doi:10.1002/gepi.21747

Campian, J., and Gutmann, D. H. (2017). CNS tumors in neurofibromatosis. *J. Clin. Oncol.* 35, 2378–2385. doi:10.1200/JCO.2016.71.7199

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A., et al. (2014). Sequence kernel association test for survival traits. *Genet. Epidemiol.* 38, 191–197. doi:10.1002/gepi.21791

Chen, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* 37, 196–204. doi:10.1002/gepi.21703

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. doi:10.1016/j.ajhg.2016.02.012

Chen, R., Wei, Q., Zhan, X., Zhong, X., Sutcliffe, J. S., Cox, N. J., et al. (2015). A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis. *Bioinformatics* 31, 1452–1459. doi:10.1093/bioinformatics/btu860

Chen, S., and Lin, X. (2020). Analysis in case-control sequencing association studies with different sequencing depths. *Biostatistics* 21, 577–593. doi:10.1093/biostatistics/kxy073

Chen, W., Wang, S., Tithi, S. S., Ellison, D. W., Schaid, D. J., and Wu, G. (2022). A rare variant analysis framework using public genotype summary counts to prioritize disease-predisposition genes. *Nat. Commun.* 13, 2592. doi:10.1038/s41467-022-30248-0

Cheng, S., Lyu, J., Shi, X., Wang, K., Wang, Z., Deng, M., et al. (2022). Rare variant association tests for ancestry-matched case-control data based on conditional logistic regression. *Brief. Bioinform.* 23, bbab572. doi:10.1093/bib/bbab572

Choi, S., Lee, S., Cichon, S., Nothen, M. M., Lange, C., Park, T., et al. (2014). Farvat: A family-based rare variant association test. *Bioinformatics* 30, 3197–3205. doi:10.1093/bioinformatics/btu496

Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly. (Austin)* 6, 80–92. doi:10.4161/fly.19695

Consortium, I. T. P.-C. a. O. W. G. (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi:10.1038/s41586-020-1969-6

De, G., Yip, W. K., Ionita-Laza, I., and Laird, N. (2013). Rare variant analysis for family-based design. *PLoS One* 8, e48495. doi:10.1371/journal.pone.0048495

Derkach, A., Chiang, T., Gong, J., Addis, L., Dobbins, S., Tomlinson, I., et al. (2014). Association analysis using next-generation sequence data from publicly available control groups: The robust variance score statistic. *Bioinformatics* 30, 2179–2188. doi:10.1093/bioinformatics/btu196

Derkach, A., Lawless, J. F., and Sun, L. (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet. Epidemiol.* 37, 110–121. doi:10.1002/gepi.21689

Dey, R., Schmidt, E. M., Abecasis, G. R., and Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* 101, 37–49. doi:10.1016/j.ajhg.2017.05.014

Dutta, D., Scott, L., Boehnke, M., and Lee, S. (2019). Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genet. Epidemiol.* 43, 4–23. doi:10.1002/gepi.22156

Fernandez, M. V., Budde, J., Del-Aguila, J. L., Ibanez, L., Deming, Y., Harari, O., et al. (2018). Evaluation of gene-based family-based methods to detect novel genes associated with familial late onset alzheimer disease. *Front. Neurosci.* 12, 209. doi:10.3389/fnins.2018.00209

Ferreira, M. A., and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics* 25, 132–133. doi:10.1093/bioinformatics/btn563

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Gilissen, C., Hoischen, A., Brunner, H. G., and Veltman, J. A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biol.* 12, 228. doi:10.1186/gb-2011-12-9-228

Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N., and Lippincott, M. F. (2018). Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am. J. Hum. Genet.* 103, 522–534. doi:10.1016/j.ajhg.2018.08.016

Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* 70, 42–54. doi:10.1159/000288704

He, Z., O'roak, B. J., Smith, J. D., Wang, G., Hooker, S., Santos-Cortez, R. L., et al. (2014). Rare-variant extensions of the transmission disequilibrium test: Application to autism exome sequence data. *Am. J. Hum. Genet.* 94, 33–46. doi:10.1016/j.ajhg.2013.11.021

He, Z., Xu, B., Lee, S., and Ionita-Laza, I. (2017a). Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am. J. Hum. Genet.* 101, 340–352. doi:10.1016/j.ajhg.2017.07.011

He, Z., Zhang, D., Renton, A. E., Li, B., Zhao, L., Wang, G. T., et al. (2017b). The rare-variant generalized disequilibrium test for association analysis of nuclear and extended pedigrees with application to alzheimer disease WGS data. *Am. J. Hum. Genet.* 100, 193–204. doi:10.1016/j.ajhg.2016.12.001

Hecker, J., Laird, N., and Lange, C. (2019). A comparison of popular TDT-generalizations for family-based association analysis. *Genet. Epidemiol.* 43, 300–317. doi:10.1002/gepi.22181

Hecker, J., William Townes, F., Kachroo, P., Laurie, C., Lasky-Su, J., Ziniti, J., et al. (2020). A unifying framework for rare variant association testing in family-based designs, including higher criticism approaches, SKATs, and burden tests. *Bioinformatics* 36, 5432–5438. doi:10.1093/bioinformatics/btaa1055

Hendricks, A. E., Billups, S. C., Pike, H. N. C., Farooqi, I. S., Zeggini, E., Santorico, S. A., et al. (2018). ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. *PLoS Genet.* 14, e1007591. doi:10.1371/journal.pgen.1007591

Hock, H., and Shimamura, A. (2017). ETV6 in hematopoiesis and leukemia predisposition. *Semin. Hematol.* 54, 98–104. doi:10.1053/j.seminhematol.2017.04.005

Hu, Y. J., Liao, P., Johnston, H. R., Allen, A. S., and Satten, G. A. (2016). Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. *PLoS Genet.* 12, e1006040. doi:10.1371/journal.pgen.1006040

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* 7, e1001289. doi:10.1371/journal.pgen.1001289

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet.* 21, 1158–1162. doi:10.1038/ejhg.2012.308

Ionita-Laza, I., Makarov, V., Consortium, A. a. S., and Buxbaum, J. D. (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am. J. Hum. Genet.* 90, 1002–1013. doi:10.1016/j.ajhg.2012.04.010

Kaakinen, M., Magi, R., Fischer, K., Heikkinen, J., Jarvelin, M. R., Morris, A. P., et al. (2017). A rare-variant test for high-dimensional data. *Eur. J. Hum. Genet.* 25, 988–994. doi:10.1038/ejhg.2017.90

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

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7

King, C. R., and Nicolae, D. L. (2014). GWAS to sequencing: Divergence in study design and analysis. *Genes (Basel)* 5, 460–476. doi:10.3390/genes5020460

Laird, N. M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7, 385–394. doi:10.1038/nrg1839

Lali, R., Chong, M., Omidi, A., Mohammadi-Shemirani, P., Le, A., Cui, E., et al. (2021). Calibrated rare variant genetic risk scores for complex disease prediction using large exome sequence repositories. *Nat. Commun.* 12, 5852. doi:10.1038/s41467-021-26114-0

Lanoiselee, H. M., Nicolas, G., Wallon, D., Rovelet-Lecrux, A., Lacour, M., Rousseau, S., et al. (2017). APP, PSEN1, and PSEN2 mutations in early-onset alzheimer disease: A genetic screening study of familial and sporadic cases. *PLoS Med.* 14, e1002270. doi:10.1371/journal.pmed.1002270

Larson, N. B., Chen, J., and Schaid, D. J. (2019). A review of kernel methods for genetic association studies. *Genet. Epidemiol.* 43, 122–136. doi:10.1002/gepi.22180

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237. doi:10.1016/j.ajhg.2012.06.007

Lee, S., Kim, S., and Fuchsberger, C. (2017). Improving power for rare-variant tests by integrating external controls. *Genet. Epidemiol.* 41, 610–619. doi:10.1002/gepi.22057

Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., et al. (2020). Dynamic incorporation of multiple *in silico* functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* 52, 969–983. doi:10.1038/s41588-020-0676-4

Li, Y., and Lee, S. (2021). Novel score test to increase power in association test by integrating external controls. *Genet. Epidemiol.* 45, 293–304. doi:10.1002/gepi.22370

Liu, Q., Nicolae, D. L., and Chen, L. S. (2013). Marbled inflation from population structure in gene-based association studies with rare variants. *Genet. Epidemiol.* 37, 286–292. doi:10.1002/gepi.21714

Liu, W., Guo, Y., and Liu, Z. (2021). An omnibus test for detecting multiple phenotype Associations based on GWAS summary level data. *Front. Genet.* 12, 644419. doi:10.3389/fgene.2021.644419

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, Z., and Lin, X. (2019). A geometric perspective on the power of principal component association tests in multiple phenotype studies. *J. Am. Stat. Assoc.* 114, 975–990. doi:10.1080/01621459.2018.1513363

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

Livesey, B. J., and Marsh, J. A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* 16, e9380. doi:10.15252/msb.20199380

Luo, L., Shen, J., Zhang, H., Chhibber, A., Mehrotra, D. V., and Tang, Z. Z. (2020). Multi-trait analysis of rare-variant association summary statistics using MTAR. *Nat. Commun.* 11, 2850. doi:10.1038/s41467-020-16591-0

Ma, S., and Shi, G. (2020). On rare variants in principal component analysis of population stratification. *BMC Genet.* 21, 34. doi:10.1186/s12863-020-0833-x

Madsen, B. E., and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5, e1000384. doi:10.1371/journal.pgen.1000384

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

Mathieson, I., and Mcvean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* 44, 243–246. doi:10.1038/ng.1074

Mavaddat, N., Peock, S., Frost, D., Ellis, S., Platte, R., Fineberg, E., et al. (2013). Cancer risks for BRCA1 and BRCA2 mutation carriers: Results from prospective analysis of EMBRACE. *J. Natl. Cancer Inst.* 105, 812–822. doi:10.1093/jnci/djt095

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* 53, 1097–1103. doi:10.1038/s41588-021-00870-7

Mcclellan, J., and King, M. C. (2010). Genetic heterogeneity in human disease. *Cell* 141, 210–217. doi:10.1016/j.cell.2010.03.032

Mclaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17, 122. doi:10.1186/s13059-016-0974-4

Minica, C. C., Genovese, G., Hultman, C. M., Pool, R., Vink, J. M., Neale, M. C., et al. (2017). The weighting is the hardest part: On the behavior of the likelihood ratio test and the score test under a data-driven weighting scheme in sequenced samples. *Twin Res. Hum. Genet.* 20, 108–118. doi:10.1017/thg.2017.7

Morgenthaler, S., and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat. Res.* 615, 28–56. doi:10.1016/j.mrfmmm.2006.09.003

Mullaert, J., Bouaziz, M., Seeleuthner, Y., Bigio, B., Casanova, J. L., Alcais, A., et al. (2021). Taking population stratification into account by local permutations in rare-variant association studies on small samples. *Genet. Epidemiol.* 45, 821–829. doi:10.1002/gepi.22426

O'connor, T. D., Fu, W., Project, N. G. E. S., Genetics, E. S. P. P., Mychaleckyj, J. C., Logsdon, B., et al. (2015). Rare variation facilitates inferences of fine-scale population structure in humans. *Mol. Biol. Evol.* 32, 653–660.

O'reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., 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

Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: Origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2, a001008. doi:10.1101/cshperspect.a001008

Ott, J., Wang, J., and Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* 16, 275–284. doi:10.1038/nrg3908

Persyn, E., Redon, R., Bellanger, L., and Dina, C. (2018). The impact of a fine-scale population stratification on rare variant association test results. *PLoS One* 13, e0207677. doi:10.1371/journal.pone.0207677

Price, A. L., Kryukov, G. V., De Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838. doi:10.1016/j.ajhg.2010.04.005

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847

Ray, D., and Chatterjee, N. (2020). Effect of non-normality and low count variants on cross-phenotype association tests in GWAS. *Eur. J. Hum. Genet.* 28, 300–312. doi:10.1038/s41431-019-0514-2

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

Sau, D., De Biasi, S., Vitellaro-Zuccarello, L., Riso, P., Guarnieri, S., Porrini, M., et al. (2007). Mutation of SOD1 in ALS: A gain of a loss of function. *Hum. Mol. Genet.* 16, 1604–1618. doi:10.1093/hmg/ddm110

Schaid, D. J., Mcdonnell, S. K., Sinnwell, J. P., and Thibodeau, S. N. (2013a). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet. Epidemiol.* 37, 409–418. doi:10.1002/gepi.21727

Schaid, D. J., Sinnwell, J. P., Mcdonnell, S. K., and Thibodeau, S. N. (2013b). Detecting genomic clustering of risk variants from sequence data: Cases versus controls. *Hum. Genet.* 132, 1301–1309. doi:10.1007/s00439-013-1335-y

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L., Peyser, P. A., et al. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* 36, 797–810. doi:10.1002/gepi.21676

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 19, 212–219. doi:10.1016/j.gde.2009.04.010

Sinnwell, J. P., Therneau, T. M., and Schaid, D. J. (2014). The kinship2 R package for pedigree data. *Hum. Hered.* 78, 91–93. doi:10.1159/000363105

Syed, H., Jorgensen, A. L., and Morris, A. P. (2021). *rareSurvival: rare variant association analysis for "time-to-event" outcomes*. doi:10.1101/2021.12.19.473338

Wainschtein, P., Jain, D., Zheng, Z., Group, T. O. a. W., Consortium, N. T.-O. F. P. M., Cupples, L. A., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* 54, 263–273. doi:10.1038/s41588-021-00997-7

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Wang, L., Choi, S., Lee, S., Park, T., and Won, S. (2016). Comparing family-based rare variant association tests for dichotomous phenotypes. *BMC Proc.* 10, 181–186. doi:10.1186/s12919-016-0027-8

Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* 597, 527–532. doi:10.1038/s41586-021-03855-y(

Wang, X. (2014). Firth logistic regression for rare variant association tests. *Front. Genet.* 5, 187. doi:10.3389/fgene.2014.00187

Wojcik, G. L., Murphy, J., Edelson, J. L., Gignoux, C. R., Ioannidis, A. G., Manning, A., et al. (2022). Opportunities and challenges for the use of common controls in sequencing studies. *Nat. Rev. Genet.* doi:10.1038/s41576-022-00487-4

Wu, G., and Zhi, D. (2013). Pathway-based approaches for sequencing-based genome-wide association studies. *Genet. Epidemiol.* 37, 478–494. doi:10.1002/gepi.21728

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi:10.1016/j.ajhg.2011.05.029

Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., et al. (2013). Kernel machine SNP-set testing under multiple candidate kernels. *Genet. Epidemiol.* 37, 267–275. doi:10.1002/gepi.21715

Xuan, J., Yang, L., and Wu, Z. (2014). Higher criticism approach to detect rare variants using whole genome sequencing data. *BMC Proc.* 8, S14. doi:10.1186/1753-6561-8-S1-S14

Yun, J., Li, Y., Xu, C. T., and Pan, B. R. (2011). Epidemiology and Rb1 gene of retinoblastoma. *Int. J. Ophthalmol.* 4, 103–109. doi:10.3980/j.issn.2222-3959.2011.01.24

Zhang, H., Tong, T., Landers, J., and Wu, Z. (2020). TFisher: A powerful truncation and weighting procedure for combining p-values. *Ann. Appl. Stat.* 14, 178–201. doi:10.1214/19-aoas1302

Zhang, H., and Wu, Z. (2022). The generalized Fisher's combination and accurate p-value calculation under dependence. *Biometrics*. doi:10.1111/biom.13634

Zhang, X., Basile, A. O., Pendergrass, S. A., and Ritchie, M. D. (2019). Real world scenarios in rare variant association analysis: The impact of imbalance and sample size on the power *in silico*. *BMC Bioinforma.* 20, 46. doi:10.1186/s12859-018-2591-6

Zhao, J., and Thalamuthu, A. (2011). Gene-based multiple trait analysis for exome sequencing data. *BMC Proc.* 5 (9), S75. doi:10.1186/1753-6561-5-S9-S75

Zhao, Z., Bi, W., Zhou, W., Vandehaar, P., Fritsche, L. G., and Lee, S. (2020). UK biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *Am. J. Hum. Genet.* 106, 3–12. doi:10.1016/j.ajhg.2019.11.012

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

Zhou, W., Zhao, Z., Nielsen, J. B., Fritsche, L. G., Lefaive, J., Gagliano Taliun, S. A., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* 52, 634–639. doi:10.1038/s41588-020-0621-6

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

Nomenclature

GLMM: Generalized linear mixed model

GRM: Genetic relatedness matrix

GWAS: Genome-wide association study

PCA: Principal components analysis

RV: Rare variant

CV: Common variant

SNP: Single-nucleotide polymorphism

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fgene.2022.989639

Polygenic power calculator: Statistical power and polygenic prediction accuracy of genome-wide association studies of complex traits

Tian Wu[1], Zipeng Liu[1,2,3], Timothy Shin Heng Mak[3,4] and Pak Chung Sham[1,2,3]*

[1]Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China, [2]State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China, [3]Centre for PanorOmic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China, [4]Fano Labs, Hong Kong, Hong Kong SAR, China

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Christoph Lippert,
Hasso Plattner Institute, Germany
Dongjun Chung,
The Ohio State University, United States

*CORRESPONDENCE
Pak Chung Sham,
pcsham@hku.hk

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

RECEIVED 08 July 2022
ACCEPTED 02 September 2022
PUBLISHED 10 October 2022

CITATION
Wu T, Liu Z, Mak TSH and Sham PC (2022), Polygenic power calculator: Statistical power and polygenic prediction accuracy of genome-wide association studies of complex traits. *Front. Genet.* 13:989639.
doi: 10.3389/fgene.2022.989639

Power calculation is a necessary step when planning genome-wide association studies (GWAS) to ensure meaningful findings. Statistical power of GWAS depends on the genetic architecture of phenotype, sample size, and study design. While several computer programs have been developed to perform power calculation for single SNP association testing, it might be more appropriate for GWAS power calculation to address the probability of detecting any number of associated SNPs. In this paper, we derive the statistical power distribution across causal SNPs under the assumption of a point-normal effect size distribution. We demonstrate how key outcome indices of GWAS are related to the genetic architecture (heritability and polygenicity) of the phenotype through the power distribution. We also provide a fast, flexible and interactive power calculation tool which generates predictions for key GWAS outcomes including the number of independent significant SNPs, the phenotypic variance explained by these SNPs, and the predictive accuracy of resulting polygenic scores. These results could also be used to explore the future behaviour of GWAS as sample sizes increase further. Moreover, we present results from simulation studies to validate our derivation and evaluate the agreement between our predictions and reported GWAS results.

KEYWORDS

GWAS, polygenic model, power calculation, online tool, statistical method

Introduction

Genome-wide association studies (GWAS) aim to systematically identify single-nucleotide polymorphisms (SNPs) associated with complex phenotypes. Though not necessarily causal, associated SNPs are good starting points for elucidating biological mechanisms of diseases and related phenotypes. GWAS on a wide range of phenotypes have confirmed the polygenic nature of most common traits, with thousands of SNPs each making a small contribution to individual differences in the population (Visscher et al., 2017). The recent increase in the sample size of GWAS and meta-GWAS has resulted in more of these SNPs to be identified, leading not only to more comprehensive understanding of disease etiology (Cano-Gamez and Trynka, 2020), but also greater accuracy in the calculation of polygenic scores to predict individual genetic liability to develop disease (Vilhjalmsson et al., 2015; Mak et al., 2017; Torkamani et al., 2018).

Adequate statistical power is necessary to both detect enough SNPs to inform etiology and to obtain accurate effect size estimate for polygenic score calculations (Dudbridge, 2013). Several computer programs have been developed to perform power calculation for single SNP association testing. For example, Genetic Power Calculator (GPC) (Purcell et al., 2003) used closed-form analytic results (Sham and Purcell, 2014) to perform power calculations for linkage and association studies. Genetic Association Study Power Calculator (GAS) (Johnson and Abecasis, 2017) performs power calculation for genetic association studies under case-control design. However, these tools perform power calculation for single SNPs, ignoring the polygenic nature of complex diseases, and the simultaneous testing of millions of SNPs that is now standard in GWAS (Sham and Purcell, 2014). Meta-GWAS Accuracy and Power (MetaGAP) (de Vlaming et al., 2017) performs GWAS power calculations and introduces genetic correlation parameters to account for effect size heterogeneity between studies. However, it is restricted to quantitative phenotype and random samples.

Since the goal of GWAS is to detect any truly associated SNPs, power calculation might more appropriately address the probability of detecting any number of associated SNPs, than the probability of detecting a specific associated SNP. Such a calculation would require specification of the entire distribution of effect size of all analysed SNPs, rather than the effect size of a single SNP. Several methods have been proposed to infer the underlying genetic effect size distribution based on significant GWAS hits or GWAS summary statistics (Park et al., 2010; So et al., 2010; Chatterjee et al., 2013; Moser et al., 2015; Zhang et al., 2018). Evidence shows that a point-normal distribution is adequate to fit the distribution of true effects of common variants for some complex traits (Zhang et al., 2018) and it is more practical than the infinitesimal model (Visscher et al., 2017).

This report describes a fast, flexible and interactive power calculation tool for GWAS under the assumption of a point-normal distribution of standardized effect sizes. The program generates predictions for the key outcomes of GWAS, including the distribution of statistical power across all independent causal SNPs, the expected number of independent genome-wide significant SNPs, total phenotypic variance explained by these SNPs, and the predictive accuracy of optimally weighted polygenic scores (PGS). It

TABLE 1 Key input parameters and output indices.

General parameters

n	GWAS sample size
m	Number of nearly independent SNPs, after removing SNPs in strong LD
h^2	SNP heritability of quantitative phenotype or of liability to disease
π_0	Proportion of SNPs that do not contribute to SNP heritability
Parameter in qualitative phenotype model	
K	Population disease prevalence
Study design parameters	
T_L	Lower threshold for extreme sample selection
T_U	Upper threshold for extreme sample selection
P_L	Proportion of samples below T_L, in extreme sample selection
ω	Proportion of cases in case-control design
Output indices	
$E(S)$	Expected number of independent significant SNPs
$E(C)$	Expected number of detected causal SNPs
$\sum_{j\in\Omega}\hat{\beta}_j^2$	Apparent phenotypic variance explained by independent significant SNPs
$\sum_{j\in\Omega}[E(\beta_j^2\vert\hat{\beta}_j)]^2$	Corrected phenotypic variance explained by the independent significant SNPs

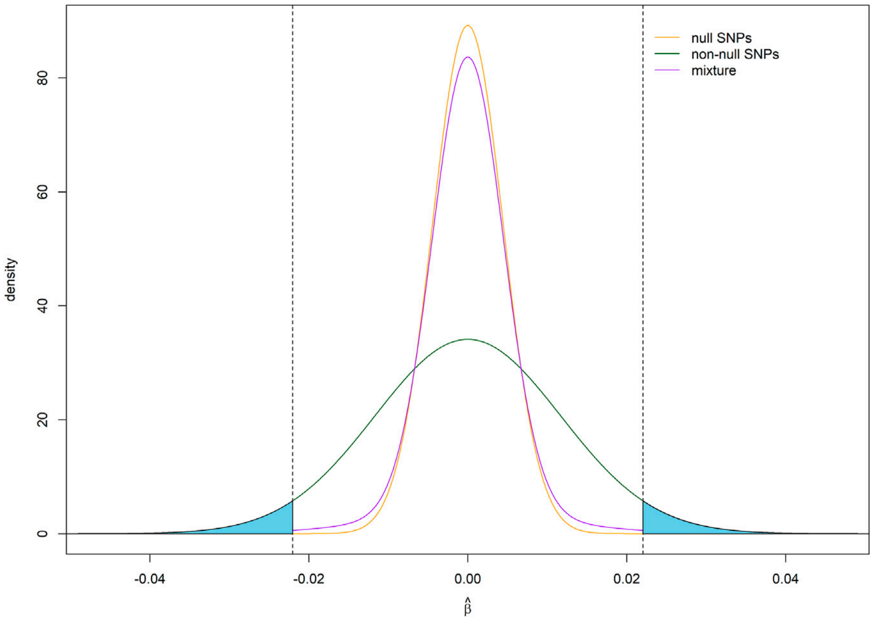


FIGURE 1
Assumed distribution of effect size estimates under a point-normal model. For illustration, the critical values for statistical significance are shown as vertical dotted lines, while average statistical power for detecting non-null SNPs is given by the shaded areas under density curve for non-null SNPs. Parameter values $h^2 = 0.7$, $m = 60{,}000$, $\pi_0 = 0.9$, $n = 50{,}000$, $\alpha = 5 \times 10^{-8}$.

allows the user to specify the nature of the phenotype under consideration (quantitative or dichotomous), its epidemiological features (e.g., disease prevalence) and genetic architecture (e.g., SNP-heritability), and the study design (e.g., case-control).

Material and methods

The input parameters and the output indices of the program are summarized in Table 1.

Model description

The phenotype is either an observed quantitative trait or a disease determined by a latent continuous liability (Falconer, 1965). For simplicity, SNPs are assumed to have been made nearly independent by clumping or pruning; the total number of SNPs (m) is the effective number of independent SNPs in the entire genome. A proportion π_0 of independent SNPs do not contribute to phenotypic variance (i.e., the null SNPs), while the remaining $(1 - \pi_0) \times m$ SNPs are causally associated with the phenotype (i.e., the non-null SNPs), explaining a proportion h^2 of the phenotypic variance, known as the SNP heritability. The effect size of a SNP j on phenotype (observed or latent), β_j, is defined as the regression coefficient of the standardized quantitative phenotype on the standardized genotype. The effect sizes of causal SNPs are assumed to be drawn from a normal distribution with mean zero and variance $\frac{h^2}{m(1-\pi_0)}$. Overall, the distribution of effect sizes of all SNPs follow a point-normal distribution:

$$\beta \sim \pi_0\delta_0 + (1 - \pi_0)N\left(0, \frac{h^2}{m(1 - \pi_0)}\right)$$

where δ_0 denotes a point mass at zero. When π_0 is zero, effect sizes become normally distributed, corresponding to the infinitesimal model (Falconer, 1996).

For disease phenotypes, standardised log-odds ratios (γ_j) from the logistic regression model can be transformed approximately to effect size on the liability scale (β_j), assuming knowledge of disease prevalence K in the population (Wu and Sham, 2021).

$$\beta_j \approx \frac{K(1 - K)}{\phi(\Phi^{-1}(K))}\gamma_j$$

where ϕ is the standard normal probability density function and $\Phi^{-1}$is the inverse of the standard normal cumulative distribution function.

Distribution of effect size estimates

For quantitative traits, the regression coefficient estimate $\hat{\beta}_j$ for a SNP with a true effect size β_j is normally distributed with mean β_j and variance approximately $\frac{1}{n}$, where n is the sample size

(Dudbridge, 2013). Thus, the overall distribution of $\hat{\beta}$ is a mixture of two normal distributions (Figure 1):

$$\hat{\beta} \sim \pi_0 N\left(0, \frac{1}{n}\right) + (1 - \pi_0) N\left(0, \frac{h^2}{m(1-\pi_0)} + \frac{1}{n}\right)$$

For binary traits, the sampling variance of the per-standard deviation effect estimate on the liability scale depends on the disease prevalence (K) in the population and the proportion of cases (w) in the sample, as well as the total (case and control) sample size, as follows (Wu and Sham, 2021):

$$Var\left(\hat{\beta}\right) \approx Var(\beta) + \frac{1}{n} \frac{K^2 (1-K)^2}{w(1-w)} \frac{1}{\phi^2 \left(\Phi^{-1}(K)\right)}$$

The sample size n can be rescaled by a factor $\frac{w(1-w)}{K^2(1-K)^2}\phi^2(\Phi^{-1}(K))$ to obtain the size of a random sample with equivalent sampling variance for $\hat{\beta}$, for an observed quantitative trait with the same parameters (π_0 ,m,andh^2) as the disease liability.

Distribution of statistical power across causal single-nucleotide polymorphisms

The statistical power for an individual SNP is determined by its effect size, the sample size, and the desired significance level. In a random sample of size n, the test statistic for the association between a quantitative phenotype and a SNP is $\hat{\beta}\sqrt{n}$, which approximately follows a non-central chi-squared distribution with non-centrality parameter (NCP) $n\beta^2$. The statistical power of detecting a SNP is given by the tail area of this distribution beyond the critical value for the desired significance level. Thus, given an assumed distribution of β across all non-null SNPs, we can obtain the distribution of statistical power, for any sample size and desired level of statistical significance. This was done by partitioning possible β values, for example, $[-10\,sd, 10\,sd]$ of the assumed effect size distribution, into narrow intervals, and calculating the probability of the effect size to be within intervals and the statistical power for an effect size at the mid-point of the intervals. This method provides increasingly more accurate approximations to the probability density function of statistical power as the intervals become narrower. Based on this approximate probability density function of statistical power, we calculated the average and variance of statistical power across causal SNPs ($E(p)$ and $Var(p)$).

Distribution of the number of and variance explained by independent significant single-nucleotide polymorphisms

From the expectation and variance of statistical power, we derived formulae for the expectation and variance of the number of independent significant SNPs, as well as the proportion of phenotypic variance explained by these SNPs. These formulae were validated by simulation studies. For SNP j, $(j = 1, 2, \ldots m)$, we generated minor allele frequency $f_j \sim Uniform\,(0.01, 0.5)$ and independent genotype value $X_j \sim Binomial\,(2, f_j)$ (subsequently standardised to have mean zero and variance one). We randomly selected $m(1-\pi_0)$ SNPs to be causal, with standardised effect size $\beta_j \sim N(0, \frac{h^2}{m(1-\pi_0)})$; the remaining $m\pi_0$ SNPs were assigned effect size zero. We also generated error term $\varepsilon \sim N(0, 1-h^2)$, which was added to the total effect of the causal SNPs to calculate the phenotypic value of each individual. We then performed association analysis for SNPs to obtain the estimated effect sizes $\hat{\beta}_j$ and associated p-values. This procedure was repeated 100 times using LDAK (Speed et al., 2017), and the results were checked for consistency with the theoretical number of significant SNPs and its 95% probability interval calculated by our formulae.

Polygenic score predictive accuracy

The polygenic model specifies that the phenotypic value is related to SNP genotypes by $y_i = G_i + \varepsilon_i$, where $G_i = \sum_{j=1}^{m} \beta_j x_{ij}$ is defined as the true additive genetic value of individual i, and m is the number of SNPs. In practice, the true effect size β_j are unknown, and we calculate individual PGS using estimates of $\beta_j$as weights, i.e., $\tilde{G}_i = \sum_{j=1}^{m} \tilde{\beta}_j x_{ij}$.

A number of different methods to determine the weights $\tilde{\beta}_j$ have been proposed. The simplest method is to use the regression coefficient estimates ($\hat{\beta}_j$) from simple linear or logistic regression of the phenotype, on each SNP separately. When the SNPs are independent and both phenotype and genotype data are standardised to have mean 0 and variance 1, the sampling variance of the regression coefficient estimate for a quantitative phenotype is $Var(\hat{\beta}_j) = \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_{ij}-\bar{x})^2} = \frac{\sigma_e^2}{(n-1)s^2} \approx \frac{\sigma_e^2}{n} \approx \frac{1}{n}$ and the efficacy of PGS relative to the true additive genetic value is $r^2(\hat{G}_i, G_i) = \frac{1}{1+\frac{m}{nh^2}}$, $i = 1, 2, \ldots n$ (Daetwyler et al., 2008), where $\hat{G}_i$ denotes the PGS constructed by $\hat{\beta}_j$. The prediction accuracy of PGS on phenotype, i.e., $r^2(\hat{G}_i, y_i)$,is then given by $r^2(\hat{G}_i, G_i)h^2$ (Wray et al., 2013). Furthermore, the prediction accuracy of PGS for binary phenotypes on the liability scale can be easily obtained based on the aforementioned effect size transformation. Once the variance explained on the liability scale is obtained, it can be easily transformed to the area under the curve (AUC) of receiver-operator characteristic (ROC) or Nagelgerke's pseudo-R^2 following Lee et al. (2012). However, the marginal effect estimates are poor proxies of true SNP effect sizes. Also, not all SNPs contribute to the phenotypic variance, so only a number of SNPs should be included in the PGS. To address these issues, shrinkage methods to construct PGS have been proposed (Purcell et al., 2009; Vilhjalmsson et al., 2015; Bigdeli et al., 2016; Mak et al., 2016; So and Sham, 2017; Qian et al., 2020; Song et al., 2020). A classic way of selecting SNPs contributing to PGS is p-value thresholding (Euesden et al., 2015), where only SNPs with GWAS p-value less than a certain threshold are retained, in effect shrinking the regression coefficient estimates of SNPs with

p-value above the threshold to zero. The threshold is usually determined by optimizing the PGS prediction accuracy of the target phenotype by split-sample or out-sample validation. Another, more sophisticated, shrinkage method is to replace the regression coefficient by the posterior expectation $E(\beta_j|\hat{\beta}_j)$, assuming a certain prior distribution for β_j (Vilhjalmsson et al., 2015; Lloyd-Jones et al., 2019; Song et al., 2020). Thus the magnitude of shrinkage depends on the value of $\hat{\beta}_j$ non-linearly, with small values being shrunk to zero while large values are relatively unchanged. The efficacy of PGS constructed by various shrinkage methods can be calculated by $r^2(G_i, \tilde{G}_i) = \frac{Cov^2(\beta_j,\hat{\beta}_j)}{Var(\beta_j)Var(\hat{\beta}_j)}$, where $\tilde{G}_i$ denotes the estimated PGS constructed by shrunk estimators of $\hat{\beta}_j$. Numeric method is adopted to calculate this efficacy index given the parameters in the genetic effect-size distribution.

Other study designs and meta-genome wide association studies analysis

We enabled the above framework to be used for power calculation in other study designs, including phenotypic selection of continuous traits (e.g., extreme phenotype design), and case-control studies of binary traits, by deriving the equivalent sample size n^*, defined as the sample size that would give the same power to detect associated SNPs as a population study of a continuous phenotype with sample size n. For meta-analysis of case-control studies of a binary trait, we first calculate the equivalent sample sizes of the component studies (which may have different case-control ratios) and then combine them to give a total equivalent sample size.

Application to real data

We applied our method to four phenotypes including height, body mass index (BMI), major depressive disorder (MDD) and schizophrenia (SCZ) to evaluate how well the predicted GWAS outcomes match up with the reported GWAS outcomes (Wray et al., 2018; Yengo et al., 2018; The Schizophrenia Working Group of the Psychiatric Genomics Consortium Ripke et al., 2020). We selected these four phenotypes because at least three sizeable GWAS or meta-GWAS had been conducted, so that earlier GWAS outcomes could be used to set a reasonable range for π_0. For example, given Wood et al. (2014) (Wood et al., 2014) reported 623 independent genome-wide significant SNPs detected by meta-analysis for height, we searched for π_0 such that the 95% probability interval of the predicted number of significant SNPs covered 623. As a result, the range of π_0 is estimated as [0.6505, 0.6800]. Similarly, we used Locke et al. (2015), Hyde et al. (2016), and Ripke et al. (2014) to estimate the range of π_0 for BMI, MDD, and SCZ, respectively (Supplementary Table S1).

For SNP heritability, we assumed the latest estimated value reported in literature; when several SNP heritability estimates were reported at about the same time, their average value was used. Specifically, we assumed the SNP heritabilities of height, BMI, MDD, and SCZ were 0.483 (Yengo et al., 2018), 0.249 (see *Web resources*), 0.089 (Howard et al., 2019) and 0.23 (Lam et al., 2019; Lee et al., 2019), respectively. In all of our applications, we set m as 60,000 (Wray et al., 2013), assuming meta-analysis samples are from European ancestry. For quantitative trait GWAS using a population cohort, the parameter n was simply the sample size of GWAS or meta-GWAS, whereas for binary phenotypes, we used the equivalent sample size described above. If earlier study was a meta-analysis, we calculated the equivalent sample size for each cohort in the meta-analysis, and used the sum of equivalent sample sizes as our model parameter n (Supplementary Tables S2, S3). We set the genome-wide significant level α as 5×10^{-8} except when predicting GWAS key outcomes for height and BMI. For these two studies, α was set as 1×10^{-8} to be consistent with the literature.

Results

Distribution of statistical power across causal single-nucleotide polymorphisms

Our model is based on the assumption that the effect size follows a point-normal distribution. Accordingly, the effect size estimate follows a normal mixture distribution (Figure 1). Figure 2A shows the relationship between statistical power and sample size for different effect sizes for a single SNP. We define SNP explaining 0.01%, 0.1%, and 1% of SNP heritability as having small, moderate and large effect, respectively. When the effect size is large, power curve increased rapidly and saturated soon. The proportion of SNPs with at least that level of statistical power on the x-axis is shown in Figure 2B. This proportion is equivalent to one minus the cumulative probability of power. With the increase of sample size, larger proportions of SNPs remain high statistical power. The expectation and variance of power, given different levels of heritability, π_0, and sample sizes, are shown in Table 2.

Distribution of number of independent significant single-nucleotide polymorphisms

The number of independent significant SNPs is a function of statistical power across all causal SNPs. Testing the significance of each independent SNP could be regarded as a Bernoulli trial X_j, which is either 0 or 1, with probability of success rate $s_j = \pi_0\alpha + (1-\pi_0)p_j, \quad j = 1, 2, \ldots m$,where α is the Type 1 error rate and p_j is the statistical power of detecting SNP j. Hence, the total number of significant SNPs $S = \sum_{j=1}^{m} X_j$ and its

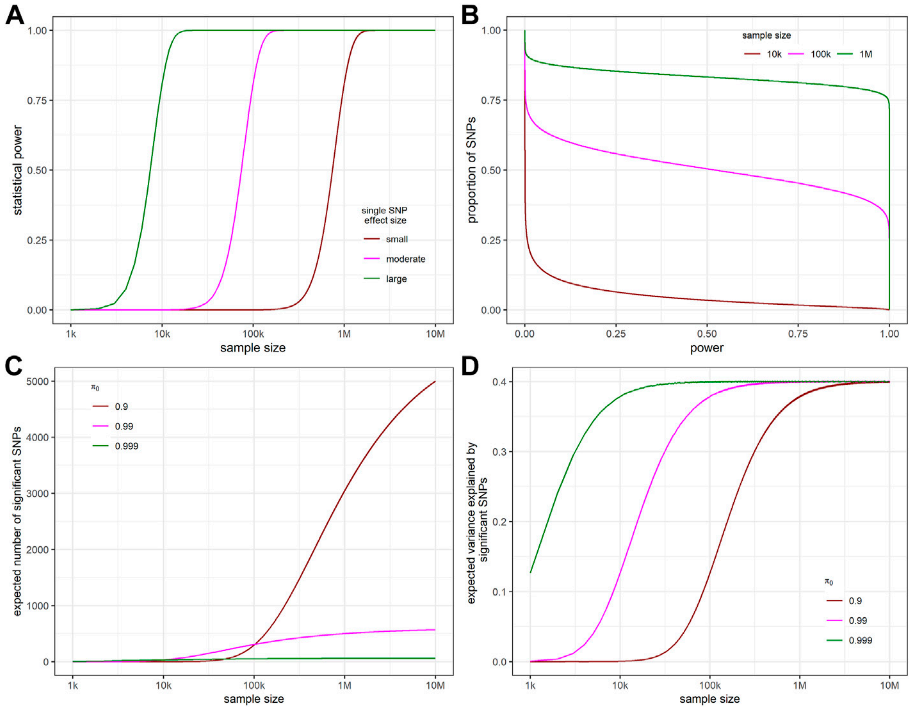


FIGURE 2
The relationship between statistical power, sample size, expected number of significant SNPs, and apparent variance explained by significant SNPs. **(A)** The relationship between sample size and the statistical power to detect a single SNP with different effect sizes "small", "moderate", and "large" representing SNPs that explain 0.01%, 0.1%, and 1% of SNP heritability. **(B)** Proportion of SNPs with at least that level of statistical power on the x-axis for different sample sizes. **(C)** Relationship between expected number of significant SNPs and sample sizes. **(D)** Relationship between the expected variance explained by the significant SNPs and sample sizes. For all figures, h^2 = 0.4, m = 60,000, α = 5 × 10^{-8}. For B, π_0 = 0.99.

expectation is $E(S) = m\pi_0\alpha + (1-\pi_0)E(\sum_{j=1}^{m} p_j) = m[\pi_0\alpha + (1-\pi_0)\ E(p)]$, where $E(p)$ is the average power of causal SNPs. The expected number of detected causal SNPs $E(C) = (1-\pi_0)E(\sum_{j=1}^{m} p_j) = m(1-\pi_0)\ E(p)$.

When calculating the variance of the number of significant SNPs, null and non-null SNPs are also considered separately. For null SNPs, the number of significant SNPs is binomial with mean $m\pi_0\alpha$ and variance $m\pi_0\alpha(1-\alpha)$. As α is often small in GWAS, the variance is approximately $m\pi_0\alpha$ thus the distribution is approximately a Poisson. For non-null SNPs, the number of significant SNPs is a convolution of $m(1-\pi_0)$ Bernoulli trials with different success rates p_j, i.e., a Poisson binomial distribution. The variance of the number of significant SNPs is therefore $m(1-\pi_0)[E(p)(1-E(p)) - Var(p)]$, where $Var(p)$ is the variance of power across causal SNPs. Hence, $Var(S) = m\pi_0\alpha(1-\alpha) + m(1-\pi_0)[E(p)(1-E(p)) - Var(p)]$. This variance is used to construct the 95% probability interval of the number of significant SNPs.

In our model, sample size and π_0 of phenotype are two factors that would affect the number of independent significant SNPs. Specifically, the more polygenic a phenotype is, the smaller the averaged effect size. With the increase of sample size, the smaller the averaged effect size, the slower the expected number of significant SNPs curve plateaus out (Figure 2C).

Distribution of variance explained by independent significant single-nucleotide polymorphisms

The phenotypic variance explained by independent significant SNPs in a GWAS is $\frac{Var(\sum_{j\in\Omega}\beta_j x_{ij})}{Var(y_i)} = \sum_{j\in\Omega}\beta_j^2$, $i = 1, 2, \ldots n$, where Ω denotes the set of such SNPs. However, since the true effect size is unknown, an approximation of the variance explained is $\sum_{j\in\Omega}\hat{\beta}_j^2$. This is referred as the apparent variance explained, because substituting β by $\hat{\beta}$ would inflate the result due to Winner's curse (Palmer and Pe'er, 2017). To correct this overestimation, we use $E(\beta_j^2|\hat{\beta}_j)$, the possibly best estimator of β_j^2, to replace $\hat{\beta}_j^2$, i.e., $\sum_{j\in\Omega}[E(\beta_j^2|\hat{\beta}_j)]^2$ .This is referred as the corrected variance explained.

When effect size estimates are calculated in different samples, the number of significant SNPs and $\hat{\beta}_j$would vary due to

TABLE 2 The expectation and variance of statistical power across causal SNPs for different SNP heritability, polygenicity, and sample sizes. m = 60,000, $\alpha = 5 \times 10^{-8}$.

h^2	π_0	Sample size	Expected power	Variance of power
0.1	0.9	10^3	6.43×10^{-8}	5.13×10^{-16}
0.1	0.9	10^5	8.43×10^{-4}	7.17×10^{-5}
0.1	0.9	10^7	0.67	0.19
0.1	0.99	10^3	4.49×10^{-7}	2.96×10^{-12}
0.1	0.99	10^5	0.19	0.11
0.1	0.99	10^7	0.89	0.08
0.1	0.999	10^3	8.43×10^{-4}	7.17×10^{-5}
0.1	0.999	10^5	0.67	0.19
0.1	0.999	10^7	0.97	0.03
0.4	0.9	10^3	1.31×10^{-7}	3.34×10^{-14}
0.4	0.9	10^5	0.05	0.02
0.4	0.9	10^7	0.83	0.12
0.4	0.99	10^3	2.42×10^{-5}	9.44×10^{-8}
0.4	0.99	10^5	0.51	0.21
0.4	0.99	10^7	0.95	0.04
0.4	0.999	10^3	0.05	0.02
0.4	0.999	10^5	0.83	0.12
0.4	0.999	10^7	0.98	0.01

sampling error. In other words, both the number of significant SNPs S and $\hat{\beta}_j$ are random variables. The expected variance explained by the significant SNPs is

$$E\left(\sum_{j\in\Omega}\hat{\beta}_j^2\right) = E\left(\sum_{j=1}^{m} X_j\right)E\left(\hat{\beta}_j^2 \middle| \hat{\beta}_j > T\right)$$

T is the critical value given the significance level.

The variance of variance explained by the significant SNPs is obtained using the law of total variance.

$$Var\left(\sum_{j\in\Omega}\hat{\beta}_j^2\right) = E\left(Var\left(\sum_{j\in\Omega}\hat{\beta}_j^2 \middle| S\right)\right) + Var\left(E\left(\sum_{j\in\Omega}\hat{\beta}_j^2 \middle| S\right)\right)$$

Similarly, the variance of corrected variance explained by significant SNPs can also be calculated.

The relationship between the expected apparent variance explained and sample size shows consistent pattern with that of expected number of significant SNPs and sample size (Figure 2D).

Simulation results

To validate the derived formula, we performed simulation studies using specific genetic architecture parameters (Figure 3). For both continuous and binary phenotypes, the 95% probability intervals of the theoretical number of significant SNPs and variance explained covers the mean of 100-time simulation results, which supports our analytic derivation. In addition, In Table 3, we listed necessary sample sizes to detect 5%, 50%, and 95% of causal SNPs for traits with different levels of π_0 and SNP heritability. It shows that we need disproportional increase of sample size to detect more significant SNPs.

Application to other study designs

For study design with phenotypic selection of continuous traits, we first consider the extreme phenotype (EP) study design (Barnett et al., 2013), which recruits subjects with extreme phenotypic values from both tail regions of truncated normal distribution (Y_S). This sampling strategy is shown to be effective for detecting rare variants that contribute to complex traits (Amanat et al., 2020). This is because rare variants are assumed to be enriched in individuals with extreme phenotypic values, and the statistical power to detect these variants is thus increased.

The relationship between sample regression coefficient $\hat{\beta}_{j_S}$ and regression coefficient without phenotypic value selection is $\hat{\beta}_j = \frac{\hat{\beta}_{j_S}}{var(Y_S)}$. Under this study design, the equivalent sample size $n^* = nVar(Y_S)^2$, where $Var(Y_S)$ can be calculated by the law of total variance:

$$\begin{aligned} Var(Y_S) = {} & Var(Y|A_1)P(A_1) + Var(Y|A_2)P(A_2) + E(Y|A_1)^2(1 \\ & - P(A_1))P(A_1) + E(Y|A_2)^2(1 - P(A_2))P(A_2) \\ & - 2E(Y|A_1)(Y|A_2)P(A_1)P(A_2). \end{aligned}$$

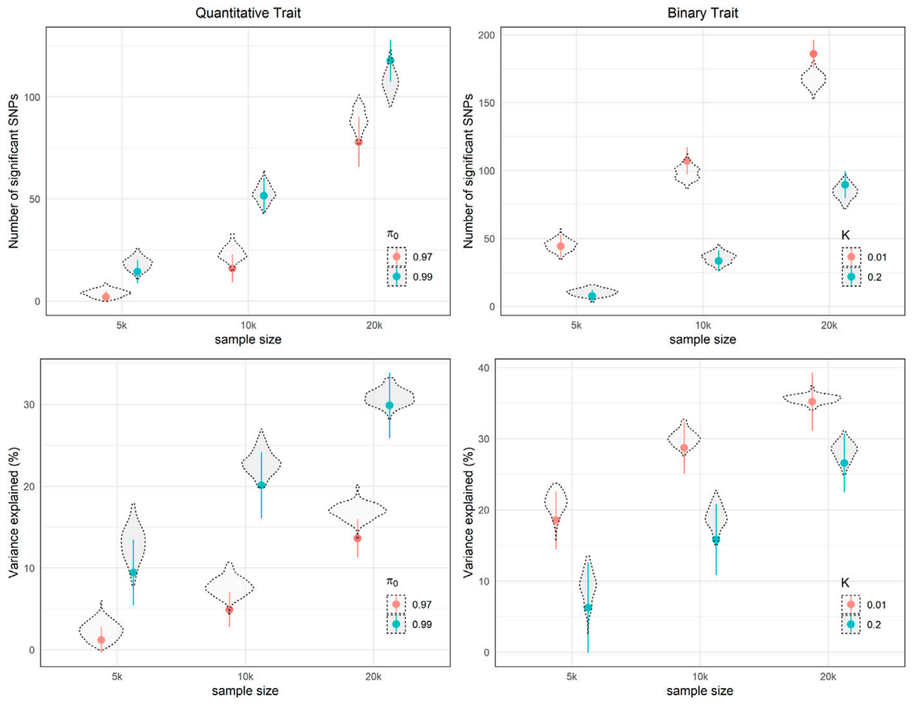


FIGURE 3
Theoretical expected number of independent significant SNPs and variance explained with 95% probability intervals i.e., dots and whiskers, with different parameters settings in 100 simulations. h^2 = 0.4, m = 50,000, α = 10^{-6}. For binary trait, π_0 = 0.99, w = 0.5.

$P(A_1)$is the proportion of samples with extreme small phenotypic values whereas $P(A_2)$ is the proportion of extreme large samples. In fact, this method applies to any method of selection based on Y, not just the truncated normal selection.

Similarly, to calculate the equivalent sample size for case-control study, the key is to build up the relationship between the estimated log odds ratio based on standardised genotype, i.e., γ, and the per-standard deviation effect on the liability scale. The equivalent sample size for a case-control study is $\frac{K^2(1-K)^2}{w(1-w)}\frac{1}{\phi(\Phi^{-1}(K))^2}n$ as mentioned in the Material and methods section.

Efficacy of polygenic scores is improved using shrinkage method

Under the assumption of point-normal genetic effect distribution, we also compared the efficacy of PGS constructed by the ordinary least square estimate (OLSE), p-value thresholding method and the aforementioned posterior expectation shrinkage relative to the true additive genetic value (Figure 4). In this figure, the p-value threshold is chosen to maximize the $r^2(\hat{G},G)$. When PGS is constructed by OLSE, $\pi_0$would not affect the PGS efficacy. When sample size is large enough, PGS constructed by p-value thresholding method can provide efficacious polygenic prediction. However, when the proportion of causal SNPs is high and effect sizes are small, shrinkage method can greatly improve polygenic score efficacy.

Real data results

We compared the predicted results with the reported meta-GWAS outcomes (Table 4). The predicted number of independent significant SNPs, the apparent and corrected variance explained are calculated based on $\pi_0$such that 95% probability interval of the predicted number of significant SNPs would cover the number reported in earlier GWAS.

For BMI and MDD, the predicted key GWAS outcomes are close to the reported values. However, our model over-estimated the results for height and SCZ. For height, one of the possible reasons is that the effect size distribution is not as simple as a point-normal, which is supported by other reference (Zhang et al., 2018). For schizophrenia, mixed population in discovery samples, for example, Asian samples are included in Ripke et al. (2014) and PGC3—SCZ (The Schizophrenia Working Group of the Psychiatric Genomics Consortium Ripke et al., 2020), may lead to the phenomenon that the reported number of significant SNPs is less than expected and it is out of the scope of our model. For different populations, m would be different, but how exactly

TABLE 3 The sample sizes needed to detect 5%, 50%, and 95% of independent significant SNPs for phenotypes with different levels of polygenicity, assuming the effect size following point-normal distribution, m = 60,000. m_1 is the total number of causal SNPs.

h^2	π_0	Total number of independent significant SNPs (m_1)	Sample size needed to detect 5% of m_1	Sample size needed to detect 50% of m_1	Sample size needed to detect 95% of m_1
0.1	0.95	3,000	2.02×10^5	1.93×10^6	2.27×10^8
	0.98	1,200	8.08×10^4	7.72×10^5	9.07×10^7
	0.99	600	4.04×10^4	3.86×10^5	4.53×10^7
0.3	0.95	3,000	6.74×10^4	6.43×10^5	7.56×10^7
	0.98	1,200	2.69×10^4	2.57×10^5	3.02×10^7
	0.99	600	1.35×10^4	1.29×10^5	1.51×10^7
0.5	0.95	3,000	4.04×10^4	3.86×10^5	4.53×10^7
	0.98	1,200	1.62×10^4	1.54×10^5	1.81×10^7
	0.99	600	8.08×10^3	7.72×10^5	9.07×10^6

the mixed population in discovery sample would affect the detected number of significant SNPs needs further study.

Discussion

In this paper, we derived theoretical results and provided computational algorithms for predicting the key outcomes of GWAS or meta-GWAS using parameters regarding the genetic architecture of phenotype and sample size, under the assumption that the standardised effect sizes of all SNPs in the genome follow a point-normal distribution. We conducted simulation studies to validate our theoretical results, and applied our model to GWAS data on four example complex traits.

Our results show that the density function of statistical power across causal SNPs under the assumed effect size distribution is bimodal with peaks near 0 and 1 (a variation of Figure 2B; Supplementary Figure S1). In other words, most causal SNPs have statistical power close to either zero or one, because of "floor" and "ceiling" effects. The relative heights of the two peaks are influenced by sample size; increasing sample size will increase the statistical power of all causal SNPs and thus reduce the height of the peak near zero and increase the height near one. From the distribution of statistical power, the expectations and variances of key GWAS outcomes, such as the number of independent genome-wide significant SNPs and the phenotypic variance explained by these SNPs, can be calculated. These calculations have been implemented in an online interactive tool named Polygenic Power Calculator.

For many phenotypes, meta-GWAS sample sizes have not reached the halfway point of the desired level to detect most of the contributing SNPs. Taking MDD as an example, we estimate that 7.36 × 10 (de Vlaming et al., 2017) equivalent total samples are needed to detect 95% of all causal SNPs when MDD prevalence is 15% whereas the existing equivalent sample size only reaches 3.05 × 10 (Torkamani et al., 2018). On the other hand, it takes a much smaller sample size to capture most of the genetic variance. Figures 2C,D shows that when π_0 is 0.9, i.e., there are 6,000 causal SNPs, it takes ~10 million samples to detect ~80% causal SNPs but only takes ~400 thousand samples to capture ~80% of SNP heritability. This is because under the assumed normal distribution of causal effects, detecting the SNPs with very small effects requires a very large sample size but does not add very much to variance explained. In practice, with the increase of global collaboration in studying genetics of complex traits, meta-GWAS sample sizes for many phenotypes are steadily increasing. As a result, we would expect to be increasingly able to identify more trait-associated SNPs with small effect sizes. However, we will eventually see a diminishing marginal return in terms of the variance explained and polygenic score prediction accuracy.

In genetic association studies, the most common definition of effect size is the per-allele effect b, estimated by regressing phenotypic value on allele count. However, we adopted the per-standard deviation effect $\beta = \sqrt{2f(1-f)}b$, where f is the allele frequency. Our assumption that the distribution of β is independent of allele frequency implies that per-allele effect sizes are inversely related to SNP variance. Although the per-allele effect has more explicit biological meaning, adopting per-standard deviation effect and assuming this to be independent of allele frequency simplifies power calculation. Indeed, theoretical models and analytical methods of complex trait genetics have widely adopted standardised effect sizes (Yang et al., 2010; Bulik-Sullivan et al., 2015; Privé et al., 2020). It is possible to relax the assumption of independence between standardized effect size and allele frequency; this would then require the allele frequency distribution in the population to be specified. Since the relationship between effect size and allele frequency depends on selective pressure on the phenotype, it is expected to be different for different phenotypes.

The parameter π_0 in this paper is not equivalent to polygenicity in the usual sense, which usually refers to the proportion of all SNPs that directly influence the phenotypes, and can be estimated by tools such as GENESIS (Zhang et al., 2018) and MiXeR (Holland et al., 2020).

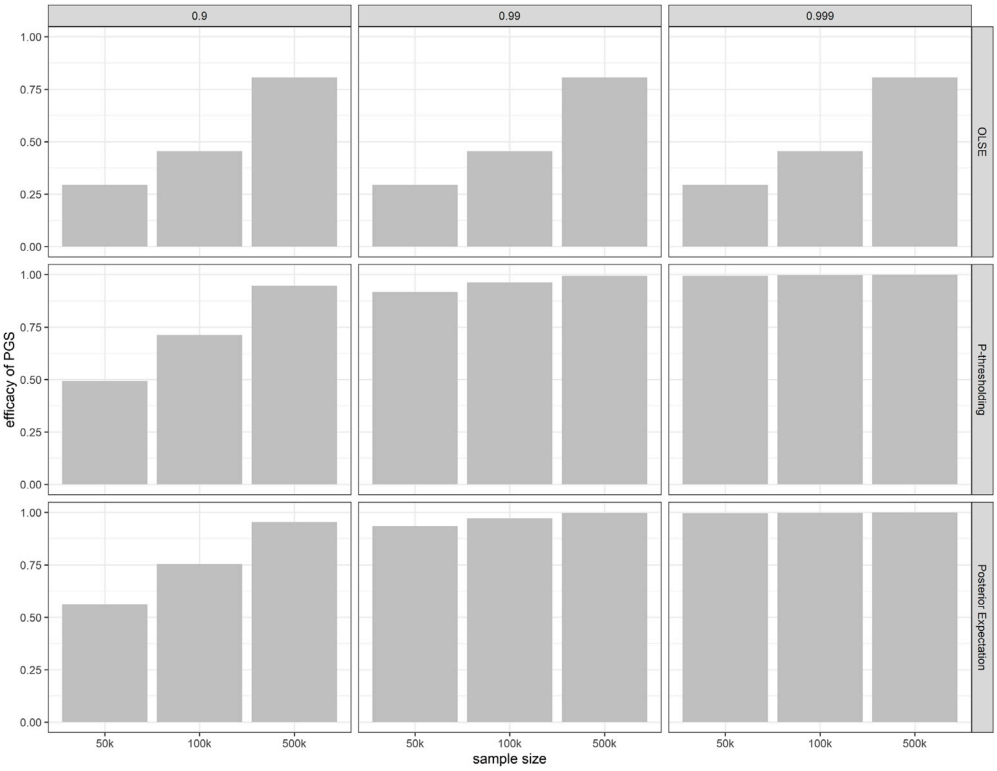


FIGURE 4
Efficacy of PGS constructed under different π_0 by different methods relative to the true additive genetic value, against sample size. OLSE: ordinary least square estimate. p-value threshold is chosen to maximize r^2. m = 60,000. h^2 = 0.5.

TABLE 4 Predicted versus reported numbers of independent significant SNPs and variance explained by these SNPs with 95% probability intervals (PIs) based on the range of estimated π_0 for height, body mass index (BMI), major depressive disorder (MDD), and schizophrenia (SCZ).

Phenotype (SNP heritability[a])	Estimated π_0[b]	Sample size	Number of significant SNPs		Variance explained by significant SNPs (%)		
			Predicted	Reported	Apparent	Corrected	Reported
Height (0.483)	0.66 [0.65, 0.68]	693,529	3466.91 [3380.98, 3547.29]	2388	30.67 [29.24, 32.14]	27.27 [25.87, 28.72]	24.6[d]
BMI (0.249)	0.36 [0.30, 0.40]	681,275	523.47 [419.5, 637.36]	656	3.27 [2.58, 4.05]	2.17 [1.66, 2.76]	6.0[d]
MDD (0.089)	0.85 [0.83, 0.88] (K = 0.15)	305,431[c]	62.2 [31.74,101.84]	44	0.76 [0.37, 1.28]	0.45 [0.19, 0.8]	0.51[e]
	0.88 [0.86, 0.90] (K = 0.25)	262,344[c]	60.34 [31.07, 97.97]		0.86 [0.43, 1.45]	0.53 [0.23, 0.93]	
SCZ (0.23)	0.86 [0.85, 0.87]	265,238[c]	494.71 [430.06, 556.73]	294	8.26 [6.98, 9.55]	6.57 [5.3, 7.55]	2.6

[a]SNP heritability is on the liability scale for MDD and SCZ.
[b]π_0 was estimated based on earlier GWAS. Details of calculations are listed in Supplementary Table S3.
[c]Equivalent sample sizes. Details of calculations are listed in Supplementary Tables S1, S2.
[d]The reported variance explained included nearly independent SNPs detected using GCTA-COJO, i.e., 3,290 and 941 nearly independent SNPs.
[e]This value is the average of liability variance explained by SNPs with p-value less than 5×10^{-8} in row 29, Supplementary Table S4 of Wray et al. (2018).

Instead, our model makes the simplification of considering only independent SNPs (obtained via linkage disequilibrium pruning or clumping), so that $1 - \pi_0$ is the proportion of causal SNPs in ~60,000 nearly independent SNPs. Taking the total number of SNPs in the genome to be approximately 4.5 million (Genomes Project Consortium Auton et al., 2015), each independent SNP on average represents approximately 75 SNPs in the genome. We have assumed that the testing of an equivalent number of independent SNP will have similar properties to the testing of all genotyped and imputable SNPs in current GWAS.

In the early days of GWAS, only a few independent significant SNPs were observed from GWAS and meta-GWAS due to limited sample size. Visscher et al. (Visscher et al., 2012) made the empirical observation of a roughly linear relationship between discovery sample size and the number of genome-wide significant hits, once the sample size reached a level sufficient to detect a few SNPs. This pattern matches the linear part of the S-shape in Figure 2C. In this study, we further extended the range of sample size to that needed to detect nearly all $m\pi_0$ independent SNPs, and obtained the predicted relationship in the entire range.

Our method has some limitations. First, we assumed the SNPs to be independent, on the basis that GWAS or meta-GWAS usually report independent SNPs after pruning or clumping. This assumption simplifies the model and bridges the relationship between genetic architecture parameters and key GWAS outcomes directly in a concise manner. We adopted 60,000 as the number of independent SNPs, but the appropriate number may depend on the population, minor allele frequency cutoff, and sample size. A more satisfactory approach in the future may be to explicitly take LD into account, expressing marginal SNP effects by weighted sums of joint effects, while making reasonable assumptions for the joint effect size distribution. Second, we adopted the per standard deviation allele effect as effect size and ignored possible differences in the relationships between allele frequency to effect size distribution for different phenotypes. Although this definition has been widely adopted (Daetwyler et al., 2008; Dudbridge, 2013), models taking allele frequency into account in effect size distribution are not uncommon (Park et al., 2010; So et al., 2010). Third, we assumed the standardised effect sizes followed a point-normal distribution but several other effect size distributions have been proposed (Zhou et al., 2013). Thus, it would be interesting to investigate how these other distributions would alter the predicted behaviour of GWAS outcomes. Fourth, our model ignores the contribution of rare variants (allele frequency < 1%). As GWAS are increasing in both sample size and number of genotyped or imputed SNPs, more rare variants with large effect size are being detected. The observed discrepancies between the predicted values from our model and the reported empirical results for height and schizophrenia also suggest possible inadequacies in our model, including misspecification of effect size distribution, inaccurate estimates of parameters such as π_0 and m, the ignoring of rare variants, and the failure to account for cross-study phenotypic or population heterogeneity in the meta-GWAS.

Web resources

Heritability of BMI can be found here: http://www.nealelab.is/uk-biobank/. The online power calculator is available at https://twexperiment.shinyapps.io/PPC_v2_1/.

Data availability statement

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

Author contributions

PS conceived of the presented idea. TW, ZL, and TM developed the theory. TW performed the computations and drafted the article. ZL and PS made revision of the article. All authors discussed the results, contributed to, and approved the final manuscript.

Funding

This work was supported by Hong Kong Research Grants Council Collaborative Research Grant C7044-19G, Theme-based Research Scheme Grant T12-712/21-R, Hong Kong Innovation and Technology Bureau funding for the State Key Laboratory of Brain and Cognitive Sciences, and National Natural Science Foundation of China (32170637).

Conflict of interest

Author TM was employed by Fano Labs.

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.989639/full#supplementary-material

References

Amanat, S., Requena, T., and Lopez-Escamez, J. A. (2020). A systematic review of extreme phenotype strategies to search for rare variants in genetic studies of complex disorders. *Genes* 11, 987. doi:10.3390/genes11090987

Barnett, I. J., Lee, S., and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet. Epidemiol.* 37, 142–151. doi:10.1002/gepi.21699

Bigdeli, T. B., Lee, D., Webb, B. T., Riley, B. P., Vladimirov, V. I., Fanous, A. H., et al. (2016). A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. *Bioinformatics* 32, 2598–2603. doi:10.1093/bioinformatics/btw303

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, 291–295. doi:10.1038/ng.3211

Cano-Gamez, E., and Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* 11, 424. doi:10.3389/fgene.2020.00424

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

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3, e3395. doi:10.1371/journal.pone.0003395

de Vlaming, R., Okbay, A., Rietveld, C. A., Johannesson, M., Magnusson, P. K. E., Uitterlinden, A. G., et al. (2017). Meta-GWAS accuracy and power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLoS Genet.* 13, e1006495. doi:10.1371/journal.pgen.1006495

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9, e1003348. doi:10.1371/journal.pgen.1003348

Euesden, J., Lewis, C. M., and O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics* 31, 1466–1468. doi:10.1093/bioinformatics/btu848

Falconer, D. S. (1996). *Introduction to quantitative genetics*. Harlow, United Kingdom: Prentice-Hall.

Falconer, D. S. (1965). The inheritance of liability to certain diseases estimated from the incidence among relatives. *Ann. Hum. Genet.* 29, 51–76. doi:10.1111/j.1469-1809.1965.tb00500.x

Genomes Project ConsortiumAuton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Holland, D., Frei, O., Desikan, R., Fan, C. C., Shadrin, A. A., Smeland, O. B., et al. (2020). Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.* 16, e1008612. doi:10.1371/journal.pgen.1008612

Howard, D. M., Adams, M. J., Clarke, T. K., Hafferty, J. D., Gibson, J., Shirali, M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* 22, 343–352. doi:10.1038/s41593-018-0326-7

Hyde, C. L., Nagle, M. W., Tian, C., Chen, X., Paciga, S. A., Wendland, J. R., et al. (2016). Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* 48, 1031–1036. doi:10.1038/ng.3623

Johnson, J. L., and Abecasis, G. (2017). GAS power calculator: Web-based power calculator for genetic association studies. bioRxiv.

Lam, M., Chen, C. Y., Li, Z. Q., Martin, A. R., Bryois, J., Ma, X. X., et al. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* 51, 1670–1678. doi:10.1038/s41588-019-0512-x

Lee, P. H., Anttila, V., Won, H., Feng, Y. A., Rosenthal, J., Zhu, Z., et al. (2019). Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* 179, 1469–1482. e1411. doi:10.1016/j.cell.2019.11.020

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genet. Epidemiol.* 36, 214–224. doi:10.1002/gepi.21614

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* 10, 5086. doi:10.1038/s41467-019-12653-0

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Felix, R., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206. doi:10.1038/nature14177

Mak, T. S. H., Kwan, J. S., Campbell, D. D., and Sham, P. C. (2016). Local true discovery rate weighted polygenic scores using GWAS summary data. *Behav. Genet.* 46, 573–582. doi:10.1007/s10519-015-9770-2

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* 41, 469–480. doi:10.1002/gepi.22050

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

Palmer, C., and Pe'er, I. (2017). Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.* 13, e1006916. doi:10.1371/journal.pgen.1006916

Park, J. H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* 42, 570–575. doi:10.1038/ng.610

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2020). LDpred2: Better, faster, stronger. *Bioinformatics* 36, 5424–5431. doi:10.1093/bioinformatics/btaa1029

Purcell, S., Cherny, S. S., and Sham, P. C. (2003). Genetic power calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19, 149–150. doi:10.1093/bioinformatics/19.1.149

Purcell, S., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., et al.International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi:10.1038/nature08185

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., et al. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* 16, e1009141. doi:10.1371/journal.pgen.1009141

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K., Holmans, P. A., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427. doi:10.1038/nature13595

The Schizophrenia Working Group of the Psychiatric Genomics ConsortiumRipke, S., Walters, J. T. R., and O'Donovan, M. C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. medRxiv. doi:10.1101/2020.09.12.20192922

Sham, P. C., and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* 15, 335–346. doi:10.1038/nrg3706

So, H. C., and Sham, P. C. (2017). Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Sci. Rep.* 7, 41262. doi:10.1038/srep41262

So, H. C., Yip, B. H., and Sham, P. C. (2010). Estimating the total number of susceptibility variants underlying complex diseases from genome-wide association studies. *PLoS One* 5, e13898. doi:10.1371/journal.pone.0013898

Song, S., Jiang, W., Hou, L., and Zhao, H. (2020). Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Comput. Biol.* 16, e1007565. doi:10.1371/journal.pcbi.1007565

Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., Balding, D. J., and Consortium, U. (2017). Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* 49, 986–992. doi:10.1038/ng.3865

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19, 581–590. doi:10.1038/s41576-018-0018-x

Vilhjalmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindstrom, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592. doi:10.1016/j.ajhg.2015.09.001

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

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

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. doi:10.1038/ng.3097

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

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

Wu, T., and Sham, P. C. (2021). On the transformation of genetic effect size from logit to liability scale. *Behav. Genet.* 51, 215–222. doi:10.1007/s10519-021-10042-2

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi:10.1038/ng.608

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649. doi:10.1093/hmg/ddy271

Zhang, Y., Qi, G., Park, J. H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* 50, 1318–1326. doi:10.1038/s41588-018-0193-x

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

TYPE Original Research
PUBLISHED 18 October 2022
DOI 10.3389/fgene.2022.976579

Causal association between rheumatoid arthritis and celiac disease: A bidirectional two-sample mendelian randomization study

Lijiangshan Hua[1†], Shate Xiang[2†], Rixiang Xu[3], Xiao Xu[1], Ting Liu[1], Yanan Shi[1], Lingyun Wu[1], Rongyun Wang[1,2]* and Qiuhua Sun[1]*

[1]School of Nursing, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China, [2]School of Basic Medical Sciences, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China, [3]School of Humanities and Management, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China

Objectives: Rheumatoid Arthritis (RA) has been associated with Celiac Disease (CD) in previous observational epidemiological studies. However, evidence for this association is limited and inconsistent, and it remains uncertain whether the association is causal or due to confounding or reverse causality. This study aimed to assess the bidirectional causal relationship between RA and CD.

Methods: In this two-sample Mendelian randomization (MR) study, instrumental variables (IVs) for RA were derived from a genome-wide association studies (GWAS) meta-analysis including 58,284 subjects. Summary statistics for CD originated from a GWAS meta-analysis with 15,283 subjects. The inverse-variance weighted (IVW) method was used as the primary analysis. Four complementary methods were applied, including the weighted-median, weighted mode, MR pleiotropy residual sum and outlier (MR-PRESSO) test and MR-Egger regression, to strengthen the effect estimates.

Results: Positive causal effects of genetically increased RA risk on CD were derived [IVW odds ratio (OR): 1.46, 95% confidence interval (CI): 1.19–1.79, p = 3.21E-04]. The results of reverse MR analysis demonstrated no significant causal effect of CD on RA (IVW OR: 1.05, 95% CI: 0.91–1.21, p = 0.499). According to the sensitivity analysis, horizontal pleiotropy was unlikely to distort the causal estimates.

Conclusion: This study reveals a causality of RA on CD but not CD on RA among patients of European descent. This outcome suggests that the features and indicators of CD should regularly be assessed for RA patients.

KEYWORDS

mendelian randomization, rheumatoid arthritis, celiac disease, bidirectional, causality

OPEN ACCESS

EDITED BY
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

REVIEWED BY
Mohammad Rostami-Nejad,
Shahid Beheshti University of Medical Sciences, Iran
Mahyar Heydarpour,
Brigham and Women's Hospital and Harvard Medical School, United States

*CORRESPONDENCE
Rongyun Wang,
wry1113@hotmail.com
Qiuhua Sun,
sunqiuhua@zcmu.edu.cn

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

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

RECEIVED 23 June 2022
ACCEPTED 06 October 2022
PUBLISHED 18 October 2022

CITATION
Hua L, Xiang S, Xu R, Xu X, Liu T, Shi Y, Wu L, Wang R and Sun Q (2022), Causal association between rheumatoid arthritis and celiac disease: A bidirectional two-sample mendelian randomization study.
Front. Genet. 13:976579.
doi: 10.3389/fgene.2022.976579

Introduction

Rheumatoid arthritis (RA) is a multi-systemic inflammatory autoimmune disease characterized by synovitis and joint damage, with a prevalence of 0.5%–1% (Smolen et al., 2016). Numerous studies demonstrated that RA causes a heavy burden on both individuals and society (Cutolo et al., 2014; Safiri et al., 2019; Hsieh et al., 2020). Celiac disease (CD) is an autoimmune disorder that occurs in genetically predisposed individuals who develop an immune reaction to gluten, with a worldwide prevalence of 1–2% (Gnodi et al., 2022). It is often accompanied by either, or both, intestinal and non-intestinal symptoms, such as diarrhea, steatorrhea, constipation, weight loss, anemia, hypo-proteinemia, and osteoporosis (Rej and Sanders, 2021). Genetic and some environmental factors, such as alteration of the gut microbiome and inflammation are believed responsible for the development of RA and CD (Smolen et al., 2016; Lebwohl et al., 2018). Genetically, the human leukocyte antigen (HLA) risk alleles play an essential role in the susceptibility of RA and CD. Individuals carrying HLA-DR shared epitope alleles have an increased risk of developing RA, whereas those carrying HLA-DQ2.5 and/or HLA-DQ8 alleles are more likely to develop celiac disease (Koning et al., 2015).

The association between RA and CD has recently received much attention (Lerner and Matthias, 2015). It has been estimated that the prevalence rate of CD in RA patients is approximately 3%, which is triple the healthy population (Elhami et al., 2018). Results from several cross-sectional and retrospective studies highlight that CD is associated with a high frequency of rheumatoid factor-IgA (RF-IgA), implying the prevalence of RA in CD patients might be higher than in healthy controls (Fayyaz et al., 2019; Ghozzi et al., 2022). Moreover, a recent epidemiological study clarifies that children with one multiple chronic inflammatory diseases (CIDs) affected parent are at a higher risk of developing the same CIDs as their parents as well as other specific CIDs reliant on the parents' CIDs (Andersen et al., 2021). Given that, children of patients comorbid with both RA and CD are considered at an increased risk for developing RA and CD in the future compared to children with no diseased parents (Andersen et al., 2021). In addition, patients accompanied by both RA and CD have a higher risk of osteoporosis and fractures, which would largely decrease the life quality and increase the risk of mortality of patients (Choi et al., 2018; Ganji et al., 2019).

Even though the exact mechanisms of the relationship between RA and CD observed in the epidemiological and observational studies are not fully understood, the gut-joint axis hypothesis was proposed as an indispensable explanation of the pathogenic link (Lerner and Matthias, 2015). Abnormal intestinal barrier permeability occurs not only in patients with CD (Lerner and Matthias, 2015) but also in RA patients (Zaiss et al., 2021). The primary mechanism of barrier disruption in the gut is potentially *via* increased zonulin production, an essential regulator of the integrity of the tight junctions in the intestinal epithelium (Fasano, 2020). Notably, identified triggers for zonulin release from intestinal epithelial cells include gluten (Drago et al., 2006), a protein that causes CD, and dysbiotic microbiota (El Asmar et al., 2002; Ciccia et al., 2017). Furthermore, autoantibodies related to RA could be generated within the inflamed intestine. Pro-inflammatory immune cells primed in intestinal tissues could traffic to the joints and systemic sites, exacerbating inflammation in genetically susceptible individuals and contributing to RA and CD occurrence (Teng et al., 2016; Zaiss et al., 2021). In addition, a moderate inflammation of the small bowel mucosa has been reported with an increased number of intraepithelial lymphocytes (IELs) in patients with RA (Molberg and Sollid, 2006). IELs have been observed to migrate from joints to the gut mucosa and *vice versa*. Notably, $CD4^+$ T lymphocytes detected in synovial fluid of RA patients have been demonstrated to express NKG2D, one of the NK-cell family receptors and a typical IEL marker of CD patients. However, observational studies might be confounded by potential confounding factors and reverse causation. Whether the observed relationships between RA and CD reflect causality requires more investigation.

Mendelian randomization (MR) is used to determine any association between risk factors and disease outcomes by employing genetic variations as instrumental variables (IVs), per the law of independent assortment, where genetic variants are allocated randomly at conception (Davey Smith and Hemani, 2014; Yarmolinsky et al., 2018; Mukamal et al., 2020). This statistical approach avoids confusion and the bias associated with reverse causation since genotypes precede the disease process and are usually unaffected by postnatal lifestyle or environmental influences (Ebrahim and Davey Smith, 2008; Lawlor et al., 2008). Based on the current genetic databank, genetic variants controlling RA could be utilized as IVs to investigate the effect of RA on the risk of developing CD, thus removing confounding variables from the data.

No MR analysis has been reported investigating a possible causal relationship between RA and CD. Investigating the causal relationship between these two diseases is of great significance since it will consolidate existing knowledge of RA and CD pathogeneses and improve treatments. This study is the first MR analysis to examine the potential causal relationships of genetically predicted RA with the risk for CD. We also undertook reverse MR to investigate the causal effect of CD on RA.

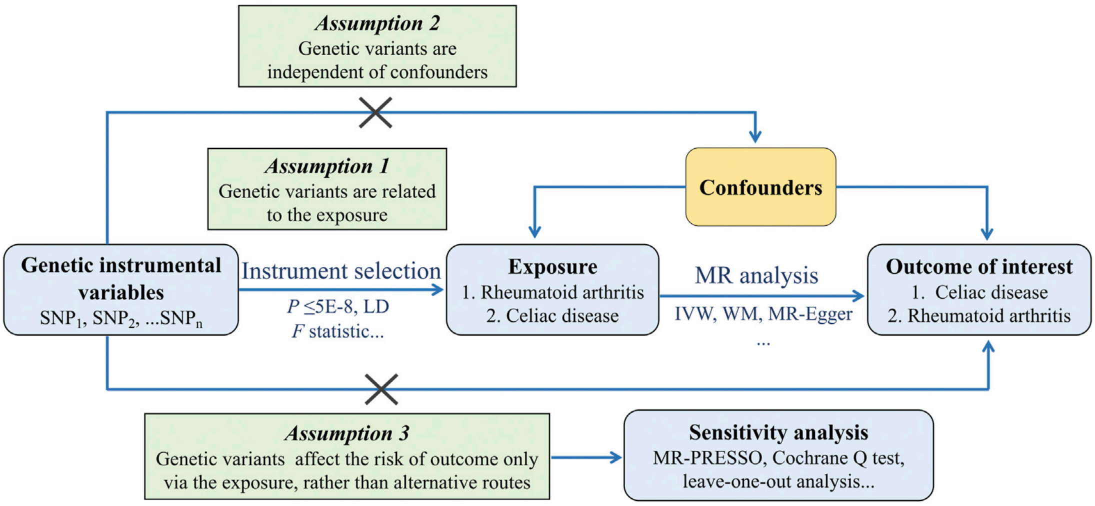

Assumption 2
Genetic variants are independent of confounders
Assumption 1
Genetic variants are related to the exposure
Confounders
Genetic instrumental variables
SNP$_1$, SNP$_2$, ...SNP$_n$
Instrument selection
$P \leq$5E-8, LD
F statistic...
Exposure
1. Rheumatoid arthritis
2. Celiac disease
MR analysis
IVW, WM, MR-Egger
...
Outcome of interest
1. Celiac disease
2. Rheumatoid arthritis
Assumption 3
Genetic variants affect the risk of outcome only via the exposure, rather than alternative routes
Sensitivity analysis
MR-PRESSO, Cochrane Q test, leave-one-out analysis...

FIGURE 1
Schematics for the bidirectional MR design. Abbreviation: MR, Mendelian randomization; SNP, single nucleotide polymorphism; LD, linkage disequilibrium; IVW, Inverse Variance Weighted; WM, Weighted Median; PRESSO, Pleiotropy REsidual Sum and Outlier.

Materials and methods

Ethics/consent statement

No further ethical approval or participation consent was required, as this study drew on published articles and public databases.

The framework of the two-sample MR study is shown in Figure 1. Genetic variations were used to investigate the causal relationship of RA on CD and the reverse causation separately. To obtain reliable results, selected IVs must meet three essential assumptions: 1) the IVs are strongly related to the exposure, 2) the IVs have no relationship to any confounders affecting both exposure and outcome, and 3) the IVs influence outcome only through the exposure. As for each inference direction, the MR analysis includes three key procedures: extracting single-nucleotide polymorphisms (SNPs) associated with interested exposure as IVs, performing primary MR analysis, and for significant associations, a series of sensitivity analysis procedures were undertaken.

Data source

In this MR study, a crucial step was to choose appropriate genetic variants from the publicly available genome-wide association studies (GWAS) database. The selected SNPs as IVs were chosen for exposures and outcomes from the IEU GWAS database (https://gwas.mrcieu.ac.uk/datasets/).

Summary statistics for RA originated from a large-scale GWAS meta-analysis involving 58,284 subjects of European ancestry (14,361 RA cases and 43,923 controls) (Okada et al., 2014). SNPs associated with CD were derived from a GWAS meta-analysis including 15,283 subjects of European ancestry (4,533 CD cases and 10,750 controls) (Dubois et al., 2010) (Supplementary Table S1).

Potentially, population stratification may introduce bias into MR analysis. Since the allele frequencies differ, a single SNP could be associated with ancestry, whereas it may be related to disease risk. SNPs and their corresponding summary statistics in the MR analysis were restricted to European descent for the exposures and outcomes to mitigate this bias.

Selection of instrumental variables

A series of quality control steps were performed to select eligible SNPs. Firstly, SNPs associated with exposures were extracted with genome-wide significance (P < 5E–08), which were the potential IVs. Secondly, independent SNPs were selected *via* setting the linkage disequilibrium (LD) threshold for clumping to $r^2 < 0.01$, and the clumping window size was 5,000 kb. The independent SNPs could not have an overlap with the reported fourteen shared loci between RA and CD (Zhernakova et al., 2011). Moreover, if the r^2 of these independent SNPs and the fourteen shared loci were greater than 0.01, the independent SNP would also be excluded from the IVs. Thirdly, to satisfy the assumptions of eligible IVs, SNPs associated with traits of outcomes were excluded by manually searching in the PhenoScanner GWAS database (http://phenoscanner.medschl.cam.ac.uk). Fourthly, SNPs with a minor allele frequency (MAF) less than 0.01 were also eliminated. Finally, the effect alleles of genetic instruments were harmonized across the exposure and outcome GWAS.

The *F* statistics were calculated to assess the strength of the selected IVs. If the *F* statistic is much greater than 10 for the instrument-exposure association, the possibility of weak instrumental variable bias is slight (Pierce et al., 2011).

Statistical analysis

This study applied multiple complementary methods, including the inverse variance weighted (IVW) method, the MR-Egger regression, the weighted median (WM) approach, and the weighted mode regression, to investigate the causal relationship between exposures and outcomes.

Specifically, the fixed-effects or random-effects IVW method was performed as the primary analysis of causal estimates, which would provide the most precise results when all the IVs were valid (Burgess et al., 2020). The WM approach uses the median MR estimate as the causative estimate (Bowden et al., 2016), and the MR Egger regression allows the intercept to indicate average pleiotropic bias (Bowden et al., 2015). These two methods are relatively robust to horizontal pleiotropy at the sacrifice of statistical power. Moreover, the weighted mode method could assess the causal association of the subset with the largest number of SNPs *via* clustering the SNPs into subsets resting on the resemblance of causal effects (Hartwig et al., 2017).

Additionally, the MR Pleiotropy RESidual Sum and Outlier (MR-PRESSO) test was applied to detect potential horizontal pleiotropy and correct it by removing outliers. The Cochrane Q test was used to evaluate heterogeneity between SNPs in the IVW method. When heterogeneity exists ($p < 0.05$), the random-effects IVW test was utilized to provide a more conservative yet robust estimate. At last, the leave-one-out analysis was performed to guarantee the reliability of the affiliation between the SNPs and exposures, evaluating whether any SNP was responsible for the significant results.

All the bidirectional MR analyses were undertaken using R (version 4.1.3) with the "*TwoSampleMR*" and the "*MRPRESSO*" packages.

Results

Effects of rheumatoid arthritis on celiac disease

After a series of approaches selecting eligible IVs and excluding potential pleiotropic SNPs, five SNPs strongly related to RA were identified as IVs in the MR analysis (Supplementary Table S2). These 5 SNPs explain 3% of the variance in RA across the population. The *F* statistic of these SNPs ranged from 210 to 528, indicating the instrument was sufficiently robust to eliminate the potential of null association due to instrument bias (Pierce et al., 2011).

The primary analysis indicated a significant causal relationship between an increased risk of RA and changes in CD risk (IVW OR: 1.46, 95% CI: 1.19–1.79, p = 3.21E-04) (Figure 2). The WM method yielded the same pattern of effects (OR: 1.39, 95% CI: 1.08–1.79, p = 0.012). Moreover, the MR-PRESSO test and the MR-Egger regression did not detect any horizontal pleiotropy among the instrumental SNPs (Table 1). No heterogeneity was observed in the Cochrane Q test (Table 1). The result of the leave-one-out analysis demonstrated that the risk estimate of genetically predicted RA on CD was remarkably stable after leaving out one SNP at a time (Supplementary Figure S1). The scatter plots and forest plots are presented in Figure 3.

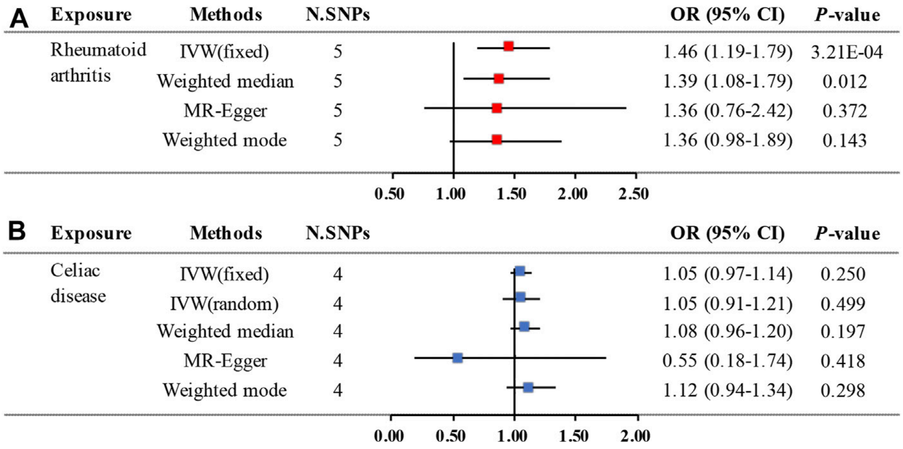
Effects of celiac disease on rheumatoid arthritis

In the reverse MR analysis, four significant (P < 5E−08) and independent SNPs ($r^2 < 0.01$) were incorporated as IVs for CD and explained 6.9% of the phenotypic variation. All the *F* statistics are greater than 10 (ranging from 205 to 301), indicating no evidence of weak instrument bias (Supplementary Table S3).

The MR analysis demonstrated that genetic liability to RA is not significantly associated with CD diagnosis. To be specific, the corresponding effect estimate is 1.05 (95% CI: 0.97–1.14, p = 0.250) in the IVW (fixed effects) method and remained consistent in the WM method (OR: 1.08, 95% CI: 0.96–1.20, p = 0.197) (Figure 2). The MR-PRESSO test results indicate no outlier, and the MR-Egger intercept did not identify any pleiotropic SNPs. However, the Cochrane Q test evidences the existence of slight heterogeneity (p = 0.034) (Table 1). Then, the IVW method based on the multiplicative random effects was performed, indicating that the onset of RA was not causally associated with suffering from CD (OR: 1.05, 95% CI: 0.91–1.21, p = 0.499) (Figure 2). Finally, the leave-one-out analysis demonstrated that the observed relationship was not driven by a single SNP (Supplementary Figure S2). Scatter plots and forest plots are shown in Figure 4.

Discussion

This study is the first MR analysis to investigate the bidirectional causal association between RA and CD, using large-scale GWAS data by conducting multiple MR approaches. The results suggest that genetically predicted RA is causally related to CD in individuals of European descent. Conversely, the current study did not observe evidence



A

Exposure	Methods	N.SNPs	OR (95% CI)	*P*-value
Rheumatoid arthritis	IVW(fixed)	5	1.46 (1.19-1.79)	3.21E-04
	Weighted median	5	1.39 (1.08-1.79)	0.012
	MR-Egger	5	1.36 (0.76-2.42)	0.372
	Weighted mode	5	1.36 (0.98-1.89)	0.143

B

Exposure	Methods	N.SNPs	OR (95% CI)	*P*-value
Celiac disease	IVW(fixed)	4	1.05 (0.97-1.14)	0.250
	IVW(random)	4	1.05 (0.91-1.21)	0.499
	Weighted median	4	1.08 (0.96-1.20)	0.197
	MR-Egger	4	0.55 (0.18-1.74)	0.418
	Weighted mode	4	1.12 (0.94-1.34)	0.298

FIGURE 2
Two-sample MR estimates results of causal associations between genetically predicted RA and CD. **(A)** Causal estimates result for RA on CD. **(B)** Causal estimates result for CD on RA. Abbreviation: MR, Mendelian randomization; RA, rheumatoid arthritis; CD, celiac disease; N.SNPs is the number of SNPs being used as IVs; SNPs, single nucleotide polymorphisms; OR, odds ratio; CI, confidence interval; IVW, Inverse Variance Weighted.

TABLE 1 Heterogeneity and horizontal pleiotropy analyses between RA and CD.

Exposure	Outcome	MR-PRESSO global test		MR-Egger		IVW	
		RSSobs	*p*-value	Intercept	*p*-intercept	Q statistic	Q-pval
RA	CD	2.34	0.877	0.009	0.819	1.43	0.839
CD	RA	16.84	0.12	0.175	0.384	8.66	0.034

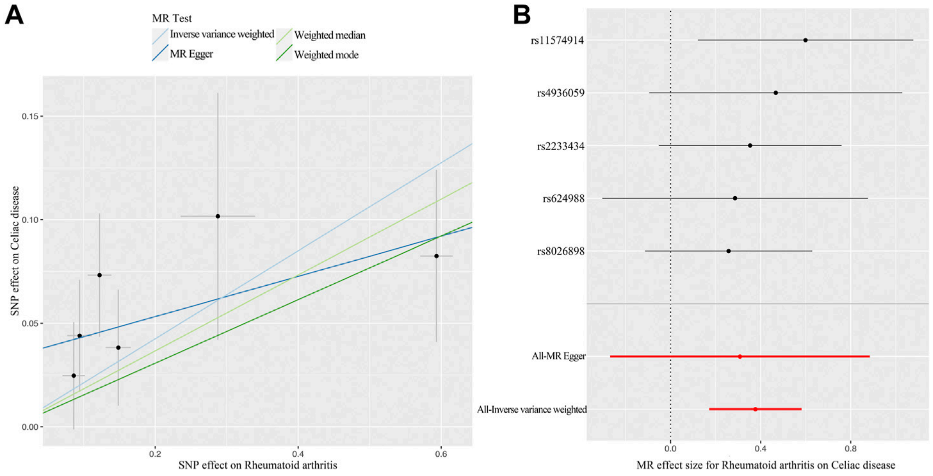


FIGURE 3
MR plots for the causal effect of RA on CD. **(A)** Scatter plot for the causal relationship of RA on CD. **(B)** Forest plot for the causal relationship of RA on CD. Abbreviation: MR, Mendelian randomization; RA, rheumatoid arthritis; CD, celiac disease; SNP, single nucleotide polymorphism.

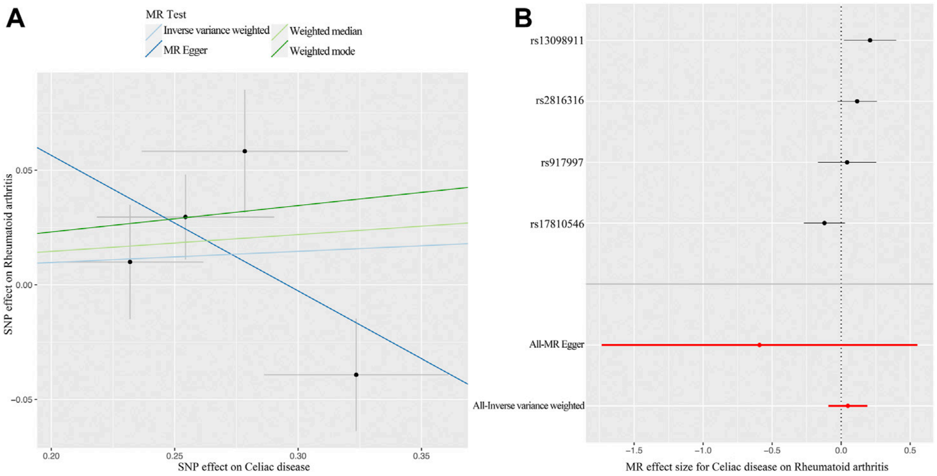


FIGURE 4
MR plots for the causal effect of CD on RA. **(A)** Scatter plot for the causal relationship of CD on RA. **(B)** Forest plot for the causal relationship of CD on RA. Abbreviation: MR, Mendelian randomization; CD, celiac disease; RA, rheumatoid arthritis; SNP, single nucleotide polymorphism.

supporting that genetically predicted CD was associated with an increased risk of RA.

Previous observational studies have investigated the association between RA and CD, but the relational literature based on the European population is sparse. Conclusions from these studies have been varied and, at times conflicting. For instance, a study conducted on Italian rheumatological patients concluded that RA patients had a higher risk for CD (Caio et al., 2018), which was consistent with our findings. However, other studies yielded conflicting results regarding the effect pattern (Francis et al., 2002; Moghtaderi et al., 2016). In the reverse relationship, our MR estimates contradict the available observational study, which suggested that the prevalence of autoimmune-related comorbidities (including RA) was more than three times higher among CD patients compared with a representative sample of the general Danish population (Grode et al., 2018). Furthermore, several other studies have not clarified a specific relationship between both diseases but rather explained a relationship of coexistence because of sharing a similar pathogenic mechanism and potential triggers, having a common genetic predisposition and a possible symptomatic overlap (Lerner and Matthias, 2015; Warjri et al., 2015; Therrien et al., 2020). Nevertheless, our MR study does not support a bidirectional causality between RA and CD. One explanation could be that the previously observed association of CD with RA is coincidental or thwarted by unknown confounders.

The causal effect of RA on CD in our study is of great significance for the diagnosis and treatment of CD. The NICE guidelines recommend testing high-risk adults with celiac serology (Downey et al., 2015). Immunoglobulin-A anti-tissue transglutaminase (IgA-TTG) testing is the recommended first-line approach for the diagnosis of CD in adults unless IgA-TTG is weakly positive, under which circumstance endomysial antibodies (EMA) concentration should also be tested (Lebwohl et al., 2018). Since we found that RA is a cause of CD, we recommend that patients with RA should be included as the high-risk population of CD and emphasize the significance of RA in the updated guidelines. Our research contributes to the existing body of knowledge about RA and CD, and the finding has substantial implications for public health, as it will anticipate the occurrence of CD in RA patients and give prevention and treatment measures for CD in RA patients. For example, surveillance examinations for RA patients should include not only regular rheumatological laboratory tests such as erythrocyte sedimentation rate (ESR) or concentrations of C-reactive protein (CRP) but also IgA-TTG testing and/or duodenal biopsy. In addition, the diet inflammatory potential has been demonstrated positively correlated with the risk of RA and increasing the probability of the risk of disease *via* superimposing effects with other risk factors (Xiang et al., 2022). So we also suggest RA patients adhere to a gluten-free diet (GFD), an anti-inflammatory diet, which is not only beneficial to prevent the development of CD but has also proven to reduce arthritic pain perception, control inflammation and improve the quality of life in RA patients (Guagnano et al., 2021).

There are several advantages of this research. First, the MR study design minimizes the residual confounding and reverse causality inherent in observational and epidemiological studies. Second, the genetic instruments explained 3% and 6.9% of the variation of RA and CD, with minimum *F* statistics of 210 and 205 respectively, consistent with the absence of weak instrument bias. Third, the MR analysis, IVW in particular, is precise enough to detect causal effects when all the IVs are valid, and produce consistent estimates using different MR techniques. Last, we provide evidence intensely supporting the causality of RA on CD from a genetic standpoint, the bidirectional analysis guaranteed the causality inference between RA and CD in both directions. Nevertheless, the limitations of the current study need to be considered. First, the RA and CD GWAS data were derived from patients of European ancestry, which may partially bias the outcomes. Applying the conclusions to populations of other ethnicities requires caution. Second, as the demographic data of all the GWAS participants are unavailable, the current study did not perform a gender-specific MR analysis although RA and CD are more prevalent in women than in men (Grode et al., 2018; Safiri et al., 2019). Third, there were likely overlapping involvers in the exposure and outcome research, but it is challenging to appraise the degree of sample overlap. Reassuringly, the strong IVs (*F* statistic much greater than 10) used in the study could minimize potential bias on sample overlap (Pierce and Burgess, 2013).

The results of this study demonstrated a causal association between genetically predicted RA on CD but did not indicate a causal effect of CD on RA. It is challenging to diagnose CD in patients with RA since their symptoms overlap in some ways. We need to keep in mind that patients with RA can have latent CD, in particular those with gastrointestinal symptoms. At the same time, we should not ignore symptoms of non-intestinal for CD and extra-articular manifestations for RA, like chronic fatigue, osteoporosis and anemia, which are important factors contributing to poor life quality for both RA and CD patients. After all, controlling disease activity, improving quality of life and enhancing subjective well-being is more important than curing the primary disease for patients with lifelong chronic diseases. In summary, the symptoms and indicators of CD need to be considered during diagnosing and managing any RA patients. Monitoring the intestinal mucosal events related to articular and extra-articular etiological pathways of RA may reduce the risk of CD in RA patients. GFD is a beneficial treatment and prevention measure that should be considered in RA and CD patients. Subsequent further studies or MR analysis based on updated and more extensive GWAS data are warranted to verify the mentioned results and elucidate the possible underlying mechanism.

Data availability statement

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

Author contributions

RW and QS conceived the idea and contributed to designing the research. LH, SX, RX, and XX drafted the manuscript and visualized the results. TL, YS, and LW assisted in performing the computations and supervised the method. All authors contributed to the manuscript revision and approved the submitted version.

Funding

This work was supported by University Students Science and Technology Innovation Activity Plan of Zhejiang Province (No. 2022R410B049).

Acknowledgments

We thank all the volunteers who participated in the original research of this MR study. We thank the investigators for sharing these data.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.976579/full#supplementary-material

References

Andersen, V., Pedersen, A. K., Möller, S., and Green, A. (2021). Chronic inflammatory diseases – diabetes mellitus, rheumatoid arthritis, coeliac disease, crohn's disease, and ulcerative colitis among the offspring of affected parents: A Danish population-based registry study. *Clin. Epidemiol.* 13, 13–20. doi:10.2147/CLEP.S286623

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

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

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., et al. (2020). Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res.* 4, 186. doi:10.12688/wellcomeopenres.15555.2

Caio, G., De Giorgio, R., Ursini, F., Fanaro, S., and Volta, U. (2018). Prevalence of celiac disease serological markers in a cohort of Italian rheumatological patients. *Gastroenterol. Hepatol. Bed Bench* 11, 244–249.

Choi, S. T., Kwon, S.-R., Jung, J.-Y., Kim, H.-A., Kim, S.-S., Kim, S. H., et al. (2018). Prevalence and fracture risk of osteoporosis in patients with rheumatoid arthritis: A multicenter comparative study of the frax and WHO criteria. *J. Clin. Med.* 7, 507. doi:10.3390/jcm7120507

Ciccia, F., Guggino, G., Rizzo, A., Alessandro, R., Luchetti, M. M., Milling, S., et al. (2017). Dysbiosis and zonulin upregulation alter gut epithelial and vascular barriers in patients with ankylosing spondylitis. *Ann. Rheum. Dis.* 76, 1123–1132. doi:10.1136/annrheumdis-2016-210000

Cutolo, M., Kitas, G. D., and van Riel, P. L. C. M. (2014). Burden of disease in treated rheumatoid arthritis patients: Going beyond the joint. *Semin. Arthritis Rheum.* 43, 479–488. doi:10.1016/j.semarthrit.2013.08.004

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

Downey, L., Houten, R., Murch, S., and Longson, D.Guideline Development Group (2015). Recognition, assessment, and management of coeliac disease: Summary of updated NICE guidance. *BMJ* 351, h4513. doi:10.1136/bmj.h4513

Drago, S., El Asmar, R., Di Pierro, M., Grazia Clemente, M., Tripathi, A., Sapone, A., et al. (2006). Gliadin, zonulin and gut permeability: Effects on celiac and non-celiac intestinal mucosa and intestinal cell lines. *Scand. J. Gastroenterol.* 41, 408–419. doi:10.1080/00365520500235334

Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302. doi:10.1038/ng.543

Ebrahim, S., and Davey Smith, G. (2008). Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology? *Hum. Genet.* 123, 15–33. doi:10.1007/s00439-007-0448-6

El Asmar, R., Panigrahi, P., Bamford, P., Berti, I., Not, T., Coppa, G. V., et al. (2002). Host-dependent zonulin secretion causes the impairment of the small intestine barrier function after bacterial exposure. *Gastroenterology* 123, 1607–1615. doi:10.1053/gast.2002.36578

Elhami, E., Zakeri, Z., Sadeghi, A., Rostami-Nejad, M., Volta, U., and Zali, M. R. (2018). Prevalence of celiac disease in Iranian patients with rheumatologic disorders. *Gastroenterol. Hepatol. Bed Bench* 11, 239–243.

Fasano, A. (2020). All disease begins in the (leaky) gut: Role of zonulin-mediated gut permeability in the pathogenesis of some chronic inflammatory diseases. *F1000Res.* 9. Faculty Rev-69. doi:10.12688/f1000research.20510.1

Fayyaz, B., Gunawan, F., and Rehman, H. J. (2019). 'Preclinical' rheumatoid arthritis in patients with celiac disease: A cross-sectional study. *J. Community Hosp. Intern. Med. Perspect.* 9, 86–91. doi:10.1080/20009666.2019.1593777

Francis, J., Carty, J. E., and Scott, B. B. (2002). The prevalence of coeliac disease in rheumatoid arthritis. *Eur. J. Gastroenterol. Hepatol.* 14, 1355–1356. doi:10.1097/00042737-200212000-00011

Ganji, R., Moghbeli, M., Sadeghi, R., Bayat, G., and Ganji, A. (2019). Prevalence of osteoporosis and osteopenia in men and premenopausal women with celiac disease: A systematic review. *Nutr. J.* 18, 9. doi:10.1186/s12937-019-0434-6

Ghozzi, M., Melayah, S., Adaily, N., and Ghedira, I. (2022). Frequency of serological markers of rheumatoid arthritis in adult patients with active celiac disease. *J. Clin. Lab. Anal.* 36, e24249. doi:10.1002/jcla.24249

Gnodi, E., Meneveri, R., and Barisani, D. (2022). Celiac disease: From genetics to epigenetics. *World J. Gastroenterol.* 28, 449–463. doi:10.3748/wjg.v28.i4.449

Grode, L., Bech, B. H., Jensen, T. M., Humaidan, P., Agerholm, I. E., Plana-Ripoll, O., et al. (2018). Prevalence, incidence, and autoimmune comorbidities of celiac disease: A nation-wide, population-based study in Denmark from 1977 to 2016. *Eur. J. Gastroenterol. Hepatol.* 30, 83–91. doi:10.1097/MEG.0000000000000992

Guagnano, M. T., D'Angelo, C., Caniglia, D., Di Giovanni, P., Celletti, E., Sabatini, E., et al. (2021). Improvement of inflammation and pain after three months' exclusion diet in rheumatoid arthritis patients. *Nutrients* 13, 3535. doi:10.3390/nu13103535

Hartwig, F. P., Smith, G. D., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998. doi:10.1093/ije/dyx102

Hsieh, P.-H., Wu, O., Geue, C., McIntosh, E., McInnes, I. B., and Siebert, S. (2020). Economic burden of rheumatoid arthritis: A systematic review of literature in biologic era. *Ann. Rheum. Dis.* 79, 771–777. doi:10.1136/annrheumdis-2019-216243

Koning, F., Thomas, R., Rossjohn, J., and Toes, R. E. (2015). Coeliac disease and rheumatoid arthritis: Similar mechanisms, different antigens. *Nat. Rev. Rheumatol.* 11, 450–461. doi:10.1038/nrrheum.2015.59

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

Lebwohl, B., Sanders, D. S., and Green, P. H. R. (2018). Coeliac disease. *Lancet* 391, 70–81. doi:10.1016/S0140-6736(17)31796-8

Lerner, A., and Matthias, T. (2015). Rheumatoid arthritis-celiac disease relationship: Joints get that gut feeling. *Autoimmun. Rev.* 14, 1038–1047. doi:10.1016/j.autrev.2015.07.007

Moghtaderi, M., Farjadian, S., Aflaki, E., Honar, N., Alyasin, S., and Babaei, M. (2016). Screening of patients with juvenile idiopathic arthritis and those with rheumatoid arthritis for celiac disease in southwestern Iran. *Turk. J. Gastroenterol.* 27, 521–524. doi:10.5152/tjg.2016.16354

Molberg, O., and Sollid, L. M. (2006). A gut feeling for joint inflammation - using coeliac disease to understand rheumatoid arthritis. *Trends Immunol.* 27, 188–194. doi:10.1016/j.it.2006.02.006

Mukamal, K. J., Stampfer, M. J., and Rimm, E. B. (2020). Genetic instrumental variable analysis: Time to call mendelian randomization what it is. The example of alcohol and cardiovascular disease. *Eur. J. Epidemiol.* 35, 93–97. doi:10.1007/s10654-019-00578-3

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

Pierce, B. L., Ahsan, H., and Vanderweele, T. J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int. J. Epidemiol.* 40, 740–752. doi:10.1093/ije/dyq151

Pierce, B. L., and Burgess, S. (2013). Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* 178, 1177–1184. doi:10.1093/aje/kwt084

Rej, A., and Sanders, D. S. (2021). An update on coeliac disease from the NHS england national centre for refractory coeliac disease. *Clin. Med.* 21, 127–130. doi:10.7861/clinmed.2021-0025

Safiri, S., Kolahi, A. A., Hoy, D., Smith, E., Bettampadi, D., Mansournia, M. A., et al. (2019). Global, regional and national burden of rheumatoid arthritis 1990-2017: A systematic analysis of the global burden of disease study 2017. *Ann. Rheum. Dis.* 78, 1463–1471. doi:10.1136/annrheumdis-2019-215920

Smolen, J. S., Aletaha, D., and McInnes, I. B. (2016). Rheumatoid arthritis. *Lancet* 388, 2023–2038. doi:10.1016/S0140-6736(16)30173-8

Teng, F., Klinger, C. N., Felix, K. M., Bradley, C. P., Wu, E., Tran, N. L., et al. (2016). Gut microbiota drive autoimmune arthritis by promoting differentiation and migration of peyer's patch T follicular helper cells. *Immunity* 44, 875–888. doi:10.1016/j.immuni.2016.03.013

Therrien, A., Kelly, C. P., and Silvester, J. A. (2020). Celiac disease: Extraintestinal manifestations and associated conditions. *J. Clin. Gastroenterol.* 54, 8–21. doi:10.1097/MCG.0000000000001267

Warjri, S. B., Ete, T., Beyong, T., Barman, B., Lynrah, K. G., Nobin, H., et al. (2015). Coeliac disease with rheumatoid arthritis: An unusual association. *Gastroenterol. Res.* 8, 167–168. doi:10.14740/gr641w

Xiang, S., Wang, Y., Qian, S., Li, J., Jin, Y., Ding, X., et al. (2022). The association between dietary inflammation index and the risk of rheumatoid arthritis in Americans. *Clin. Rheumatol.* 41, 2647–2658. doi:10.1007/s10067-022-06217-9

Yarmolinsky, J., Wade, K. H., Richmond, R. C., Langdon, R. J., Bull, C. J., Tilling, K. M., et al. (2018). Causal inference in cancer epidemiology: What is the role of mendelian randomization? *Cancer Epidemiol. Biomarkers Prev.* 27, 995–1010. doi:10.1158/1055-9965.EPI-17-1177

Zaiss, M. M., Joyce Wu, H.-J., Mauro, D., Schett, G., and Ciccia, F. (2021). The gut-joint axis in rheumatoid arthritis. *Nat. Rev. Rheumatol.* 17, 224–237. doi:10.1038/s41584-021-00585-3

Zhernakova, A., Stahl, E. A., Trynka, G., Raychaudhuri, S., Festen, E. A., Franke, L., et al. (2011). Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* 7, e1002004. doi:10.1371/journal.pgen.1002004

frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 10 November 2022
DOI 10.3389/fgene.2022.1034946

Probabilistic edge inference of gene networks with markov random field-based bayesian learning

Yu-Jyun Huang[1], Rajarshi Mukherjee[2] and Chuhsing Kate Hsiao[1,3]*

[1]Division of Biostatistics and Data Science, Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan, [2]Department of Biostatistics, Harvard University, Boston, MA, United States, [3]Bioinformatics and Biostatistics Core, Center of Genomic Medicine, National Taiwan University, Taipei, Taiwan

Check for updates

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Wenan Chen,
St. Jude Children's Research Hospital,
United States
Xuan Cao,
University of Cincinnati, United States

*CORRESPONDENCE
Chuhsing Kate Hsiao,
ckhsiao@ntu.edu.tw

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

RECEIVED 02 September 2022
ACCEPTED 24 October 2022
PUBLISHED 10 November 2022

CITATION
Huang Y-J, Mukherjee R and Hsiao CK (2022), Probabilistic edge inference of gene networks with markov random field-based bayesian learning.
Front. Genet. 13:1034946.
doi: 10.3389/fgene.2022.1034946

Current algorithms for gene regulatory network construction based on Gaussian graphical models focuses on the deterministic decision of whether an edge exists. Both the probabilistic inference of edge existence and the relative strength of edges are often overlooked, either because the computational algorithms cannot account for this uncertainty or because it is not straightforward in implementation. In this study, we combine the Bayesian Markov random field and the conditional autoregressive (CAR) model to tackle simultaneously these two tasks. The uncertainty of edge existence and the relative strength of edges can be measured and quantified based on a Bayesian model such as the CAR model and the spike-and-slab lasso prior. In addition, the strength of the edges can be utilized to prioritize the importance of the edges in a network graph. Simulations and a glioblastoma cancer study were carried out to assess the proposed model's performance and to compare it with existing methods when a binary decision is of interest. The proposed approach shows stable performance and may provide novel structures with biological insights.

KEYWORDS

Bayesian markov random field, edge prioritization, existence probability, gene regulatory network, network structure, probabilistic association

1 Introduction

The network analysis of multi-dimensional data for structural information learning has attracted much attention in the biomedical research community. Examples include gene regulatory networks, brain connectivity networks, and microbial networks (Zhang et al., 2019; Huang et al., 2020). An undirected graphical model, the Markov random field (MRF), is a common approach to describe the network structure of a group of genetic variables, because of its direct interpretation of edges with the conditional dependence between nodes. The Gaussian MRF, also known as the Gaussian graphical model (GGM), imposes a multivariate distribution for gene regulatory networks, assuming the

p-dimensional vector $\mathbf{X} = (X_1, X_2, \ldots X_p)^T \in \mathbb{R}^p$ follows a multivariate normal distribution $\mathbf{X} \sim \mathrm{MVN}(\mu, \Omega = \sum^{-1})$ with X_i denoting the gene expression value of the˜ i-th gene node. A zero-entry in the precision matrix Ω corresponds to conditional independence and no connecting line between nodes. In other words, if the off-diagonal (i, j)-th element ω_{ij} in Ω is zero, then the partial correlation $|_{ij} = cor(X_i, X_j|X_{-(i,j)})$ is zero; namely, the X_i and X_j are conditionally independent given the remaining variables, and there exists no edge between these paired nodes in the network. Therefore, under GGM, the problem of network construction becomes the inference of a sparse precision matrix or the selection of non-zero partial correlation.

Recent work on inferring network structure with GGM can be categorized into two groups. Methods in the first group focus on determining if an edge exists between nodes using the idea of "covariance selection". When p is large, these methods follow the principle of variable selection with a regularization procedure to complete the binary decision about whether ω_{ij} or $|_{ij}$ is zero. Various methods of this regularization approach have been developed that adopt different objective functions and/or L_1 penalty, including neighborhood selection with lasso (M&B) by Meinshausen and Buhlmann (2006), graphical lasso (Glasso) in Friedman et al. (2008), the space partial correlation estimation (SPACE) in Peng et al. (2009), and the constrained l_1 minimization for inverse matrix estimation (CLIME) in Cai et al. (2011). These penalized optimization methods can be applied straightforwardly, but they are not designed to infer the intensity of edges or to interpret the dependence between nodes, although this information may be influential in biological experiments (Ni et al., 2020). If the inference, such as the estimation of the non-zero partial correlation, is based on a given network, then the network structure needs to be fixed first with one of the methods mentioned above. Therefore, this estimation procedure relies heavily on the choice of the selected network structure, which may cause concern about subsequent inference if the validity of this structure is in question.

Methods in the second group, usually under the Bayesian framework, explicitly adopt the uncertainty in the network graph, through a prior on the precision matrix, such as the G-Wishart, spike-and-slab lasso (SSL), and a subset-specific prior (Wang and Pillai, 2013; Mohammadi and Wit, 2015; Gan et al., 2019; Williams 2021; JalaliKhare and Michailidis, 2022). To enhance computational efficiency, researchers have proposed various tools, such as the double Metropolis-Hasting algorithm and birth-death Markov chain Monte Carlo methods, and the Bayes EM to estimate the maximum *a posteriori* (MAP) to avoid complex computation. These analyses provide a posterior probability for each candidate graph and a posterior inclusion probability for each edge. The inclusion probability, in this case, can be a good indication of its existence, but the strength of the edge is not considered in the computation. One solution may be to average the estimates of precision matrices in an element-wise way and weigh by the posterior probability of the matrix and the corresponding candidate graph. For instance, the BDgraph in Mohammadi and Wit (2015) can be utilized to perform this analysis. The computational burden in these procedures is fairly heavy due to the large number of nodes and the even more significant number of candidate graphs.

To relieve the computational burden, Gan et al. (2019) proposed a novel EM algorithm, called BAGUS, that first estimates the maximum *a posteriori* (MAP) of the precision matrix and then approximates the probability of edges with the precision matrix fixed at the MAP to learn the graph structure. BAGUS outperformed existing methods in terms of computation time, accuracy in recovering graph structure, and prediction error of the precision matrix. However, the uncertainty of the network graph and the posterior distribution of the edges are not accounted for in the BAGUS algorithm.

The inference of the strength of the edges has not been the target of these aforementioned algorithms. This inference requires a fully Bayesian approach and can be complicated in computation. In a recent research, Williams (2021) discussed the importance and implication of this topic. In that study, the edge inference was carried out with a fully Bayesian approach and the posterior probability of the precision element is used to infer the dependence between nodes. The conjugate Wishart prior was adopted to save computation time. If the SSL prior with a latent variable indicating the randomness in the edge existence is considered, further computational complexity will be incurred.

This research adopts the Bayesian learning approach for its ability to incorporate *a priori* information and to offer probabilistic inference, and for its wide application in bioinformatic research, including the Bayesian scoring rule for metabolite molecules (Ludwig et al., 2018), peak calling with Hi-C data (Xu et al., 2016), and pathway prioritization with posterior probability (Lin et al., 2018). The rationale of this research is twofold. First, an informative metric to quantify the strength of an edge is needed, which can provide more information beyond its existence. This is crucial when decoding the interplay between nodes or prioritizing intervention in a gene regulatory network. Second, since most genes do not work alone, the strength or intensity of the relationship between any two nodes should account for the presence of other genes when learning the network structure of a given set of genetic nodes. In this study, we start with the Bayesian MRF combining the conditional autoregressive (CAR) model to estimate the strength of the edge and its existence probability. Under the Gaussian CAR model, the conditional mean $E(X_j|X_{(-j)})$ is expressed as $\sum_{k \neq j} \beta_{jk} X_k$ for $j = 1, 2, \ldots, p$, where $X_{(-j)} \triangleq \{X_k\colon k \neq j\}$ represents the set containing all variables except X_j. Following Besag (1974) and Besag and Kooperberg (1995), the coefficient β_{jk} is a function of elements in the precision matrix Ω, and is connected to the partial correlation $|_{jk}$ between X_j and X_k. That is, the β_{jk} can be used to characterize the strength of dependence between these two genes. In addition, the Spike-and-Slab Lasso (SSL) prior proposed by Ročková and George (2018) is adopted for β_{jk}. Then, the regularization procedure on these β_{jk}'s functions similarly to the "covariance selection" procedure in previous literature and provides a direct and intuitive interpretation of the intensity and relationship between nodes.

The rest of this article is organized as follows. The rationale and complete model of the Bayesian MRF and the implementation of prior knowledge are introduced in Section 2. In Section 3, extensive simulation studies are conducted to demonstrate the performance of the proposed model and comparison with other state-of-the-art methods. In Section 4, the proposed model is applied to a glioblastoma study with gene expression values from TCGA (Hutter and Zenklusen, 2018). Some biologically relevant findings will be highlighted. We then conclude with a discussion.

2 Methods

2.1 Learning network structure

To introduce the proposed Bayesian Markov Random field (BMRF) model, we first let the $n \times p$ matrix $\mathbf{X}$ represent the observed gene expression values of the p genes from the n subjects, where x_{ij} is the expression value of the j-th gene $(j = 1, 2, \ldots, p)$ from the i-th subject $(i = 1, 2, \ldots, n)$. Without loss of generality, the values across subjects per gene are standardized so that $E(X_j) = 0$ and $Var(X_j) = 1$. Under GGM, the $p-$ dimensional random vector $(X_1, X_2, \ldots, X_p)^T$ follows a multivariate normal distribution (MVN) with the following conditional distribution (Besag 1974),

$$X_j \big| X_{(-j)} \sim N\left(\sum_{k \neq j} \beta_{jk} X_k, \sigma_j^2\right), j = 1, 2, \ldots, p. \qquad (1)$$

Following Besag (1974) and Besag and Kooperberg (1995), the coefficients can be expressed as $\beta_{jk} = \frac{-\omega_{jk}}{\omega_{jj}}$ if $j \neq k$. This is related to the partial correlation $|_{jk}$ between X_j and X_k where $|_{jk} = \frac{-\omega_{jk}}{\sqrt{\omega_{jj}\omega_{kk}}}$. When the diagonal elements in Ω are equal, then $\beta_{jk} = \beta_{kj}$ and the underlying coefficients in the CAR model can be expressed as $\boldsymbol{\beta} = \{\beta_{jk}\colon 1 \leq j < k \leq p\}$ where $\| \boldsymbol{\beta} \| = p(p-1)/2$ is the number˜ of unknown parameters to˜ be estimated. Moreover, when $\beta_{jk} = 0$, the corresponding $|_{jk} = 0$, implying no edge between two gene nodes. These properties provide two advantages in supporting β_{jk} as promising candidates in inferring the network structure. First, the selection of non-zero elements of $\beta_{jk} \in \boldsymbol{\beta}$ is equivalent to the decision of the existence of the edge. Sec̃ond, the magnitude of these coefficients can quantify the relative intensity of the partial correlation between nodes. Their estimates can be derived based on the CAR model and thus the regression model. Such an approach would be easier than directly estimating the correlation coefficient matrix, especially when a direct estimate of the matrix is not straightforward due to the curse of dimensionality and the requirement of positive definiteness.

This CAR model is more general than those used in spatial statistics, where only neighboring "areas" are included in the mean structure. Here all genetic nodes are included first as a fully connected model. Then the procedures and computations below will decide which β_{jk} remain and how strong the evidence is. In addition, this conditional distribution is also similar to node-wise regression where constraints are imposed to ensure symmetry in the β_{jk}'s (Ha et al., 2021).

2.2 Spike-and-slab lasso prior: Probabilistic estimation of edge

For the inference of β_{jk}, we consider the Spike-and-Slab Lasso (SSL) prior (Rockova and George, 2018),

$$\pi\left(\beta_{jk} \middle| \gamma_{jk}\right) = \gamma_{jk} \times \psi_1\left(\beta_{jk}\right) + \left(1 - \gamma_{jk}\right) \times \psi_0\left(\beta_{jk}\right). \qquad (2)$$

where the slab distribution $\psi_1(\beta_{jk}) = \frac{\tau_1}{2}\exp(-\tau_1|\beta_{jk}|)$ and the spike $\psi_0(\beta_{jk}) = \frac{\tau_0}{2}\exp(-\tau_0|\beta_{jk}|)$ are both double exponential (Laplace) with a small τ_1 and large τ_0, respectively. The binary γ_{jk} takes the value of one if β_{jk} represents a large effect, and $\gamma_{jk} = 0$ if the effect is around zero. Therefore, the marginal posterior probability of $\gamma_{jk} = 1$ can represent the probability of the edge existence.

The SSL prior is considered a fundamental variable selection tool in the Bayesian framework for sparse models. This differs from the previously mentioned penalized optimization methods for variable selection, where the estimated effect size is biased. In addition, the SSL prior is flexible because it allows the shrinkage effects to vary among different edges. For instance, a substantial shrinkage penalty can be deployed for those edges with weak partial correlation, while for those with strong partial correlation, a non-shrinkage effect can be considered. Other studies have used the SSL prior in the matrix inference (Peterson et al., 2015; Deshpande et al., 2019; Gan et al., 2019). For instance, Gan et al. (2019) assumed this prior for the off-diagonal entries in the precision matrix, the ω_{jk} in our case, and Deshpande et al. (2019) adopted this prior for the regression parameter, the β_{jk} in our case. In Peterson et al. (2019), the SSL prior was incorporated to model the network similarity.

By adopting the SSL prior, we can select the influential edges and perform statistical inference with β_{jk}. The BMRF model specification is completed with a Bernoulli prior for γ_{jk} , $\gamma_{jk} \sim Ber(p_{jk})$, where p_{jk} follows a conjugate beta distribution. Specifically, in contrast to previous studies investigating if the edge exists, here we are interested in constructing the posterior distributions of β_{jk} and γ_{jk}, respectively, to model the strength of the edge and its existence probability.

2.3 Computation

Since the posterior distributions of γ_{jk} and β_{jk} are the bases of the probabilistic inference, one can obtain the posterior samples of γ_{jk} and β_{jk} with Markov chain Monte Carlo (MCMC) methods implemented in any standard Bayesian software. In the following simulation studies and applications, the R package *R2OpenBUGS* is used to carry out the computations.

When the number of gene nodes is large, the number of possible edges and parameters increases rapidly. Fortunately, most genetic networks/pathways are sparse. For instance, the sparsity of the signaling pathway networks in KEGG ranges between 5% and 10%. Liu et al. (2009), Zhao et al. (2012), and Mohammadi and Wit (2015) have adopted similar values in their simulation studies. Such *a priori* information can be utilized in a $p \times p$ adjacency matrix G^{*}, where elements $g_{jk} = 1$ if two genes X_j and X_k are known biologically to be associated and $g_{jk} = 0$ otherwise. By imposing the matrix of domain knowledge G^{*} on $\boldsymbol{\beta} = \{\beta_{jk}\colon 1 \leq j < k \leq p\}$, one can save computational cost from es̃timating the edges known to be non-existent. Similarly, another $p \times p$ adjacency matrix M^{*} can be introduced to contain elements $m_{jk} = 1$ if the corresponding interrelation is of interest to particular experts. This would force the inclusion of the edge in the network, yet the flexibility remains when later inference does not favor its existence. Inclusion of these two matrices and the distribution of p_{jk} can account for all the cases described here. For example, this matrix M^{*} can be derived first and the data-driven prior on γ_{jk} can be further established. The BMRF with this setup will be denoted as BMRF.P in later sections.

3 Numerical simulation experiments

For performance evaluation and comparison with existing methods, three types of network graph are considered in the simulation studies: the random network (M1), random scale-free network (M2), and fixed network structure (M3). In M1, edges are considered exchangeable, and all nodes in a network are equally important. The scale-free network in M2 is commonly adopted for genetic pathways, where the edges are not exchangeable because hub nodes may exist in the network. These two are designed to compare with the traditional approach of variable selection, where only the number of true edges successfully detected is of concern. While in M3, with a fixed and known structure, further comparison between the inclusion probability in previous Bayesian methods and the existence probability in current BMRF can be carried out, and the strength of edge is demonstrated. In other words, in M3, in addition to the number of true edges successfully detected, both the probability of existence and strength of edges will be emphasized.

3.1 Simulation settings

In the random network setting M1, the GGM is generated with the following steps, similar to the procedures in Fan et al. (2009), and Peng et al. (2009).

1) Set up the network sparsity S, $0 \leq S \leq 1$
2) Construct the true network E by randomly sampling the Bernoulli e_{ij} with probability S. If $e_{ij} = 1$, then there is an edge between the node i and j, and 0 otherwise.
3) Generate the precision matrix $\Omega = (\omega_{ij})$ according to E by

$$\omega_{ij} = \begin{cases} 1, \; i = j \\ 0, \; i \neq j, \; e_{ij} = 0 \\ U(W), \; i \neq j, \; e_{ij} = 1 \end{cases}$$

where $W = [-1, -0.05] \cup [0.05, 1]$ and $U(.)$ denotes the uniform distribution.

4) To assure the positive definiteness of Ω, each off-diagonal ω_{ij} in Ω is replaced by the original ω_{ij} divided by $1.5 \times \sum_{j=1, j \neq i}^{p} |\omega_{ij}|$.
5) Average the rescaled matrix calculated in (4) with its transpose matrix to ensure symmetry. The values of the nodes are generated from a multivariate normal distribution (MVN) with a zero mean vector and the precision matrix.

Note that different combinations of p and S have been considered, denoted as M1.1 for $p = 25$, $S = 0.05$, M1.2 for $p = 25$, $S = 0.10$, M1.3 for $p = 50$, $S = 0.05$, and M1.4 for $p = 50$, $S = 0.10$. The number of edges in each network is about $\begin{pmatrix} p \\ 2 \end{pmatrix} \times S$.

In the random scale-free network setting M2, the R package huge was used to generate the scale-free networks. Two settings (M2.1) $p = 25$ and (M2.2) $p = 50$ were considered. The average number of edges in the scale-free network is $p - 1$.

In M3, the fixed network structure setting, a scale-free network graph containing 50 nodes and 49 edges was selected, and the node values were generated with the *huge* package with the partial correlation in the network set at −0.216.

For all stimulations, the hyper-parameters were specified as $\tau_1 = 2$ and $\tau_0 = 20$, the sample size was $n = 250$, and the number of replications in each setting was 100. More detailed information, including the network sparsity and number of true edges, is summarized in the Supplementary Table S1. For the proposed BMRF, the corresponding edge is selected for the network if the posterior probability of $\gamma_{jk} = 1$ is greater than 0.5. This choice is used in simulation studies when comparing different regularized methods for variable selection.

3.2 Comparing methods and evaluation criteria

The proposed BMRF model was compared with M&B, Glasso, SPACE, and CLIME, as well as with the Bayesian approach BDgraph using the Bayesian model averaging procedure (denoted as BD_BMA), the Maximum a posterior probability procedure (BD_MAP), and BAGUS. M&B and Glasso were performed with the R package huge, and the tuning parameter used in these two methods was chosen through the rotation information criterion (ric). The SPACE approach was performed with the R package space with the tuning parameter set by default. The package flare was used for the estimator CLIME with tuning parameters obtained by 5-fold cross-validation. The R package BDgraph was used for BDgraph. BAGUS was performed with the R code provided in the online supplementary material in Gan et al. (2019).

Several criteria were used to compare performance, including the total number of true positives (TP), the sensitivity (SEN), the specificity (SPE), the false discovery rate (FDR), the Matthew correlation coefficient (MCC), and the F1-score (F1). These quantities are calculated based on TP and the total number of false negatives (FN), where TP is defined as the total number of true edges that were successfully identified, and FN as the total number of true edges that failed to be detected.

3.3 Implementations

When handling a large set of gene nodes with BMRF, we recommend two modeling strategies, one with a non-informative prior and the other with a data-driven prior. The former is denoted as BMRF.O, corresponding to the prior distribution $\gamma_{jk} \sim Ber(p_{jk})$ with p_{jk} from a beta distribution with mean 0.5. The latter, denoted as BMRF.P, models the network edges with $p_{jk} \sim Beta(\alpha^{\star}, \beta^{\star})$, an informative prior with a mean larger than 0.5 if $e_{ij} \in M^{\star} \cap G^{\star}$, or $p_{jk} \sim Beta(\alpha^{\dagger}, \beta^{\dagger})$, a non-informative prior with a mean around 0.5. As stated earlier, the matrices $M^{\star}$ and $G^{\star}$ can be elicited by experts, with domain knowledge, with a screening scheme based on sparsity or sample correlation, or with SPACE proposed in Peng et al. (2009), which outperforms other methods when dealing with a scale-free network structure. In the following analysis, the matrix $G^{\star}$ containing the edges corresponding to the largest 10% absolute sample correlations was determined first when the network sparsity was set at 0.05 (or the top 15% if set at 0.10). For the matrix $M^{\star}$, we incorporated the information from SPACE to accelerate the computational efficiency. The mean of the informative prior $p_{jk} \sim Beta(\alpha^{\star}, \beta^{\star})$ was set at 0.8.

3.4 Results

3.4.1 Existence or not: Random network (M1 and M2)

To compare performance, Table 1 lists the values of several evaluation criteria under settings M1.1, M1.3, and M2.2. A quick look shows that, except for BD_MAP, the other four Bayesian algorithms perform equivalently or slightly better than the rest. In most cases, BAGUS is the best in terms of F1-score and MCC, but is less satisfactory in the number of true positives (TP) and sensitivity (SEN). Other Bayesian algorithms achieve larger TP and sensitivity. Among the Bayesian methods, BD_BMA and BD_MAP tended to identify more edges, leading to larger TP and SEN but lower F1 and MCC. Consequently, these two often produce a larger FDR. BD_MAP was usually the worst in this regard due to the lack of consideration of model uncertainty. In M1.3 and M2.2, BAGUS, M&B and SPACE perform similarly well. Generally, the proposed BMRF.O and BMRF.P are comparable to the best performers. The performances under other settings are displayed in the Supplementary Table S2.

One metric among the evaluation criteria, the F1-score, is displayed in Figure 1. When the number of nodes p is as large as 50, most methods are still satisfactory if the graph is sparse, such as when the case sparsity = 0.05 in Figures 1C,F. The Bayesian approaches, both BMRF and BDgraph, tend to identify more edges when compared with the frequentist approach to variable selection, therefore leading to a higher F1-score. These results highlight the advantages of probabilistic inference on the conditional dependence in network analysis, in contrast to the detection of whether or not the edge exists.

3.4.2 Existence probability: Fixed network (M3)

In setting M3, a fixed network structure with two hub nodes was determined first, as shown in Figure 2A, and then the node values were generated from MVN. The numbers of edges connecting to the two hubs, Node-2 and Node-4, are 14 and 7, respectively. Various methods were then applied to infer the network structure. Across 100 replications, the average number of edges estimated by each method is listed in Table 2. Four methods, BMRF.P, BD_BMA, M&B, and SPACE, performed the best, with the first two being slightly better with a smaller standard error. When examining the F1-score in Figure 2B, BAGUS performed best.

For the probabilistic inference of edge existence, we first stratify the edges into two groups, truly *No Edge* and *Edge exists*, and display in Figure 2C the estimated edge existence probability or the inclusion probability derived from the four competing methods, BMRF.O, BMRF.P, BD_BMA, and BAGUS. As indicated in the figure, when there exists no edge (labeled *No Edge* on *X*-axis in the figure), BMRF.O and BMRF.P provide very low probabilities while BD_BMA and BAGUS show slightly larger probabilities. When the edge truly exists, labeled *Edge exists* on *X*-axis in the right group in

TABLE 1 Values of six evaluation criteria (F1, MCC, FDR, TP, SEN, and SPE) under simulation settings M1.1, M1.3, and M2.2. Each value is the average of 100 replications with standard error (SE) in parentheses.

M1.1	F1	MCC	FDR	TP	SEN	SPE
BMRF.O	0.89 (0.06)	0.89 (0.07)	0.12 (0.09)	13.6 (3.0)	0.91 (0.08)	0.99 (0.005)
BMRF.P	0.88 (0.06)	0.87 (0.06)	0.16 (0.09)	13.8 (3.1)	0.92 (0.07)	0.99 (0.005)
BD_BMA	0.87 (0.06)	0.86 (0.06)	0.19 (0.09)	14.1 (3.1)	0.94 (0.07)	0.99 (0.006)
BD_MAP	0.58 (0.09)	0.60 (0.08)	0.58 (0.10)	14.1 (3.2)	0.94 (0.07)	0.93 (0.020)
BAGUS	0.94 (0.05)	0.94 (0.05)	0.02 (0.04)	13.6 (2.9)	0.91 (0.08)	0.99 (0.002)
Glasso	0.83 (0.06)	0.82 (0.06)	0.25 (0.09)	14.1 (3.2)	0.94 (0.07)	0.98 (0.009)
CLIME	0.88 (0.08)	0.89 (0.08)	0.01 (0.02)	11.9 (2.5)	0.81 (0.13)	0.99 (0.001)
M&B	0.90 (0.06)	0.90 (0.06)	0.10 (0.09)	13.8 (3.0)	0.92 (0.07)	0.99 (0.005)
SPACE	0.89 (0.06)	0.88 (0.06)	0.14 (0.08)	13.8 (3.0)	0.92 (0.07)	0.99 (0.005)
M1.3	**F1**	**MCC**	**FDR**	**TP**	**SEN**	**SPE**
BMRF.O	0.78 (0.05)	0.78 (0.05)	0.13 (0.05)	44.2 (3.7)	0.72 (0.07)	0.99 (0.002)
BMRF.P	0.79 (0.05)	0.79 (0.05)	0.14 (0.04)	45.6 (4.0)	0.74 (0.07)	0.99 (0.002)
BD_BMA	0.76 (0.04)	0.75 (0.05)	0.25 (0.06)	47.7 (4.2)	0.77 (0.06)	0.99 (0.004)
BD_MAP	0.50 (0.03)	0.51 (0.04)	0.63 (0.03)	48.8 (4.3)	0.79 (0.06)	0.93 (0.007)
BAGUS	0.80 (0.05)	0.80 (0.05)	0.04 (0.03)	42.3 (3.8)	0.69 (0.07)	0.99 (0.001)
Glasso	0.78 (0.05)	0.78 (0.05)	0.10 (0.05)	43.0 (3.7)	0.70 (0.08)	0.99 (0.002)
CLIME	0.63 (0.11)	0.67 (0.09)	0.01 (0.02)	29.2 (6.7)	0.48 (0.11)	0.99 (0.001)
M&B	0.79 (0.05)	0.80 (0.05)	0.04 (0.03)	41.9 (3.4)	0.68 (0.08)	0.99 (0.001)
SPACE	0.80 (0.05)	0.80 (0.05)	0.09 (0.04)	43.9 (3.8)	0.71 (0.07)	0.99 (0.002)
M2.2	**F1**	**MCC**	**FDR**	**TP**	**SEN**	**SPE**
BMRF.O	0.78 (0.09)	0.77 (0.09)	0.24 (0.06)	39.2 (6.1)	0.80 (0.13)	0.99 (0.002)
BMRF.P	0.84 (0.04)	0.84 (0.04)	0.23 (0.04)	45.3 (2.8)	0.92 (0.06)	0.99 (0.002)
BD_BMA	0.83 (0.04)	0.83 (0.04)	0.27 (0.05)	46.0 (2.3)	0.94 (0.05)	0.99 (0.003)
BD_MAP	0.52 (0.03)	0.55 (0.03)	0.64 (0.02)	45.7 (2.3)	0.93 (0.05)	0.93 (0.006)
BAGUS	0.89 (0.05)	0.89 (0.05)	0.04 (0.03)	41.5 (3.6)	0.85 (0.07)	0.99 (0.01)
Glasso	0.83 (0.06)	0.82 (0.07)	0.19 (0.07)	41.7 (4.1)	0.85 (0.08)	0.99 (0.004)
CLIME	0.63 (0.11)	0.65 (0.09)	0.52 (0.13)	47.3 (2.1)	0.97 (0.04)	0.95 (0.025)
M&B	0.88 (0.05)	0.88 (0.05)	0.08 (0.05)	41.6 (4.2)	0.85 (0.09)	0.99 (0.002)
SPACE	0.86 (0.05)	0.86 (0.05)	0.16 (0.05)	43.7 (3.1)	0.89 (0.06)	0.99 (0.003)

the figure, the BD_BMA performs the best and is followed by BMRF.P. It needs to be clarified, however, that it may not be fair to compare the edge existence probability against the inclusion probability because of the different definitions. In BMRF, the existence probability of the edge is the posterior probability of $\gamma_{jk} = 1$; while in BD_BMA, the inclusion probability is the sum of all posterior probabilities of networks containing the edge. The inclusion probability in this sense can be viewed as the expected value of the existence probability if all possible network structures are accounted for. In BAGUS, the inclusion probability is estimated with a conditional probability, conditioning on the Bayes EM estimates of the other parameter values. In other words, the BAGUS estimate assumes a fixed network structure rather than estimating across all possible structures.

The association between the existence probability from BMRF.P and the inclusion probability from BM_BMA is further examined in Figure 2D. The blue circles represent true edges and the red circles indicate non-existent edges. These two are fairly consistent, except that BD_BMA seems to detect more non-existent edges (red circles) than BMRF.P. The values of the other criteria are summarized in the Supplementary Table S3.

3.4.3 Accuracy of probabilistic inference

An alternative way to evaluate the probabilistic inference of the edge existence is the Brier score (Brier, 1950), which can

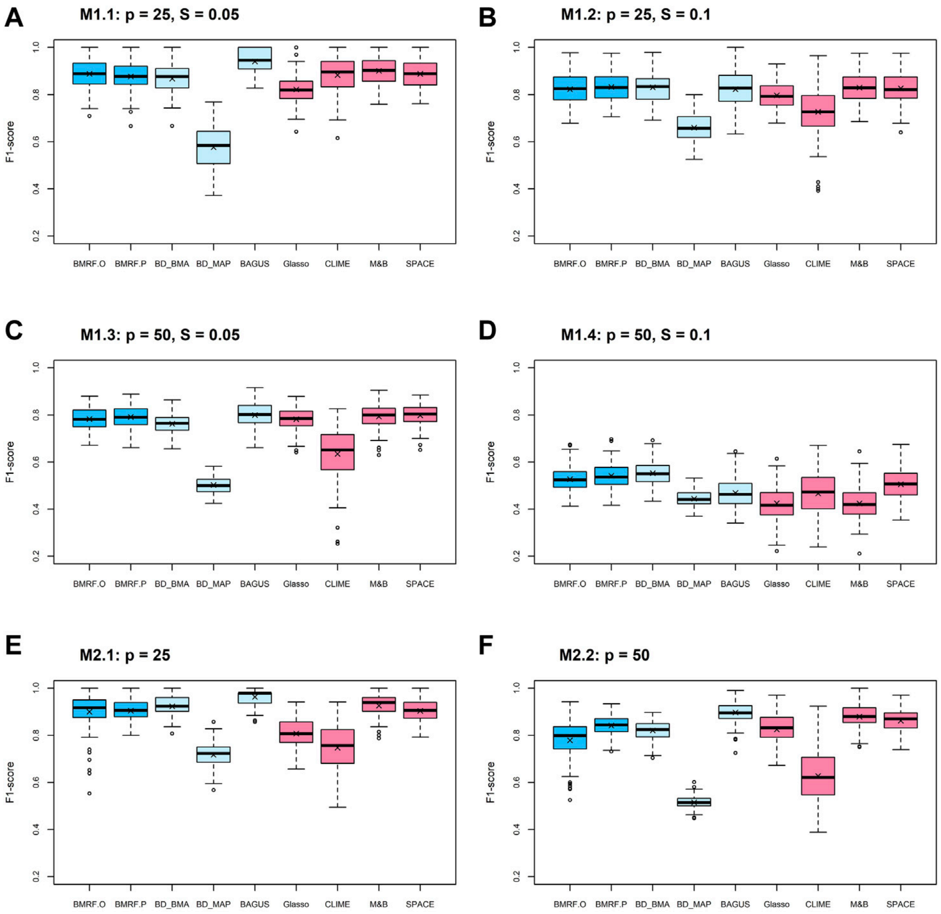


FIGURE 1
Boxplots of F1-scores from 100 replications under each method. Each subfigure corresponds to a setting with a combination of *p* and *S*. Note that the blue boxplots correspond to the five Bayesian algorithms and the pink ones correspond to the four penalized methods.

be calculated for each of the Bayesian estimates. The Brier score, ranging between 0 and 1, is a mean squared difference between the true class label (edge exists or not) and the estimated probability. Smaller values of the Brier score indicate better estimates. This score has become a common measure to assess the accuracy of the probabilistic estimates of binary outcomes, especially when comparing performance of machine learning algorithms (Dinga et al., 2019; Ovadia et al., 2019).

The boxplots of the Brier score for the four Bayesian estimates under different simulation settings are displayed in Figure 3. Every boxplot is composed of 100 Brier scores, each from a replication in the simulations. In all the subfigures, it can be observed that all four methods provide small Brier scores, mostly below 0.07, indicating good accuracy. In other words, they provide large probability estimates when the edge truly exists and small probability estimates when the edge does not exist. This pattern is consistent with that in Figure 2C under simulation setting M3. Note that the average Brier scores of the four Bayesian estimates under M3 are 0.01, 0.01, 0.02, and 0.03 for BMRF.O, BMRF.P, BD_BMA, and BAGUS, respectively. The second observation in the figure is that the probabilistic estimates of BAGUS are more variable and usually slightly larger than the rest. This could result from the utilization of MAP in the BAGUS probability estimate, where the estimate is a probability conditioning on MAP estimates of the other parameters and therefore incurs further estimation errors in the graph structure.

4 Applications in two glioblastoma studies

In this section, we consider two data types, array and sequencing gene expression values, collected from Glioblastoma (GBM) patients. GBM is a grade IV

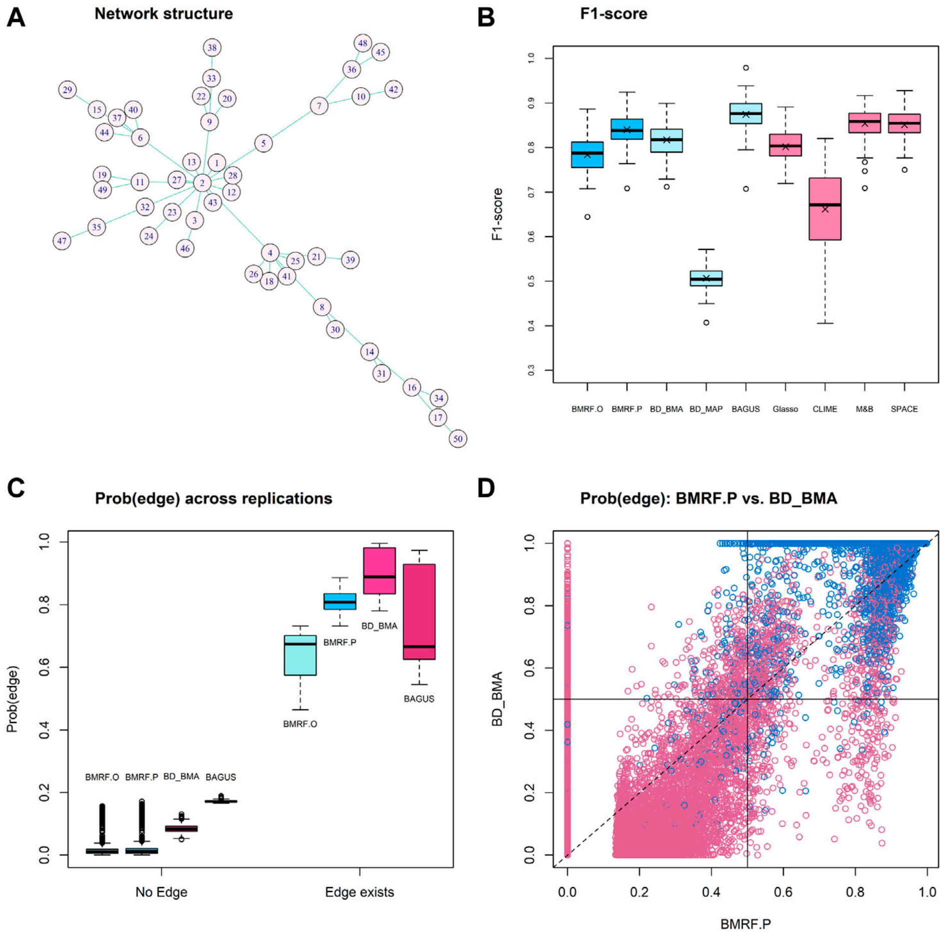


FIGURE 2
Results of competing methods under simulation setting M3. **(A)** The true network structure with 49 true edges; **(B)** Boxplots of the F1-score from 100 replications under each method; **(C)** Boxplots of average existence probability over 100 replications under each of the four Bayesian algorithms. The left group No Edge corresponds to the case when there is truly no edge, and the right group Edge exists corresponds to the case when the edge truly exists. Each boxplot in the right group is composed of 49 average probabilities; **(D)** The edge existence probability from BMRF.P *versus* the inclusion probability from BD_BMA for each edge across replications. Blue circles indicate true edges and red indicates no edge. The vertical and horizontal solid lines denote the cut-off values for BMRF.P and BD_BMA, respectively.

malignant brain tumor, usually in adults. After being diagnosed, patients have a median survival time of about 12–15 months and generally respond poorly to treatments (Stupp et al., 2005; The Cancer Genome Atlas Research Network, 2008). Although several molecular biomarkers have been identified, such as TP53 mutation and overexpression in EGFR (Bralten and French, 2011; Zhang et al., 2018), targeted therapy shows a limited effect (Shergalis et al., 2018; Banerjee et al., 2021). Recent interest has focused on the molecular mechanism of the Janus kinase/signal transducer and activator of transcription (JAK-STAT) signaling pathway (Jain et al., 2012; Ou et al., 2021).

Here we aim at constructing relationships within two networks, EGFR and JAK-STAT, based on RNA sequencing and array data, respectively. The BMRF model is applied to two pathways to examine the conditional dependence among gene nodes and detect influential molecular relationships to understand the underlying biological mechanism better. The expression values were downloaded from the University of California Santa Cruz (UCSC Xena) TCGA Hub and TCGA GDC data portal. The array gene expressions were generated from the Affymetrix HT Human Genome U133a microarray platform with mRNA values in the log two scale, and the sequencing data from Illumina HTSeq. The nodes in the JAK-STAT network were collected with the procedures in Chang et al. (2020). The EGFR network was determined based on the protein-protein interaction (PPI) network in STRING. The final array data consist of 27 gene expression values from

TABLE 2 The listed values are the average number of estimated edges connecting to each of the two hub nodes (Node-2 and Node-4) across 100 replications under M3. The number in parenthesis is the standard error. The true number of edges connecting to Node-2 is 14 and to Node-4 is 7.

	Node-2 (true = 14)	Node-4 (true = 7)
BMRF.O	8.5 (1.3)	5.7 (1.0)
BMRF.P	14.1 (0.8)	7.0 (0.9)
BD_BMA	14.4 (0.8)	7.2 (0.8)
BD_MAP	16.4 (1.5)	9.5 (1.6)
BAGUS	14.1 (0.4)	6.3 (0.7)
Glasso	14.7 (0.8)	7.6 (1.8)
CLIME	17.6 (2.3)	9.9 (1.9)
M&B	14.3 (0.6)	6.5 (1.2)
SPACE	14.7 (0.9)	6.8 (1.1)

253 primary tumor tissues, and the sequencing data contain 30 genes from 83 tissues. All are primary tumor tissues from male patients aged 40 and 75. The procedures (computing sample correlation, SPACE, and taking union) discussed earlier were carried out and resulted in 99 possible edges in the JAK-STAT network and 80 edges in the EGFR network, respectively, as the starting sets of edges for further analysis. More information about the selection procedures is in the Supplementary Sections S2, S3.

4.1 Edges in the JAK-STAT network with gene expression arrays

Based on the GBM array data, the BMRF.P identified 69 edges in the network with probabilities greater then 0.5, 15 of which were associated with a posterior existence probability greater than 0.9. Figure 4A plots the posterior probabilities of all 99 edges, from the largest to the smallest. Figure 4B shows the resulting gene regulatory network, where the 15 edges are represented with thick lines and the others with thin lines. The corresponding magnitudes of the 15 existence probabilities are displayed in Figure 4C, where the width denotes the

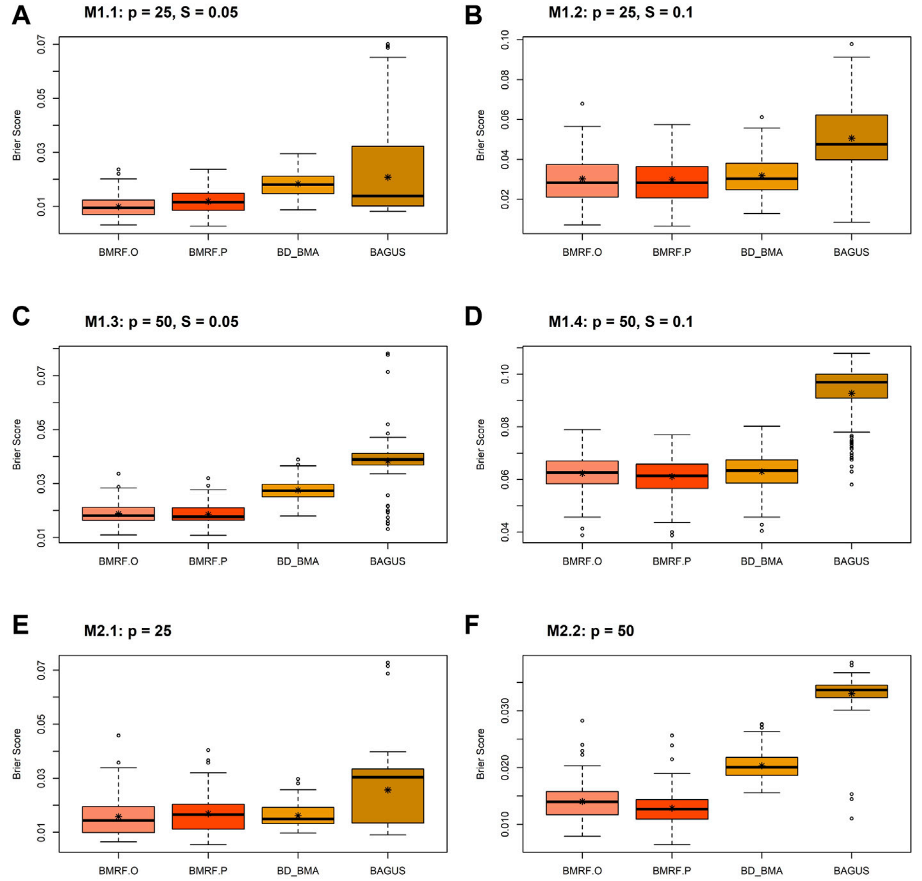


FIGURE 3
Boxplots of the Brier scores of the four Bayesian estimates, BMRF.O, BMRF.P, BD_BMA, and BAGUS, under six different simulation settings: **(A)** M1.1; **(B)** M1.2; **(C)** M1.3; **(D)** M1.4; **(E)** M2.1; **(F)** M2.2.

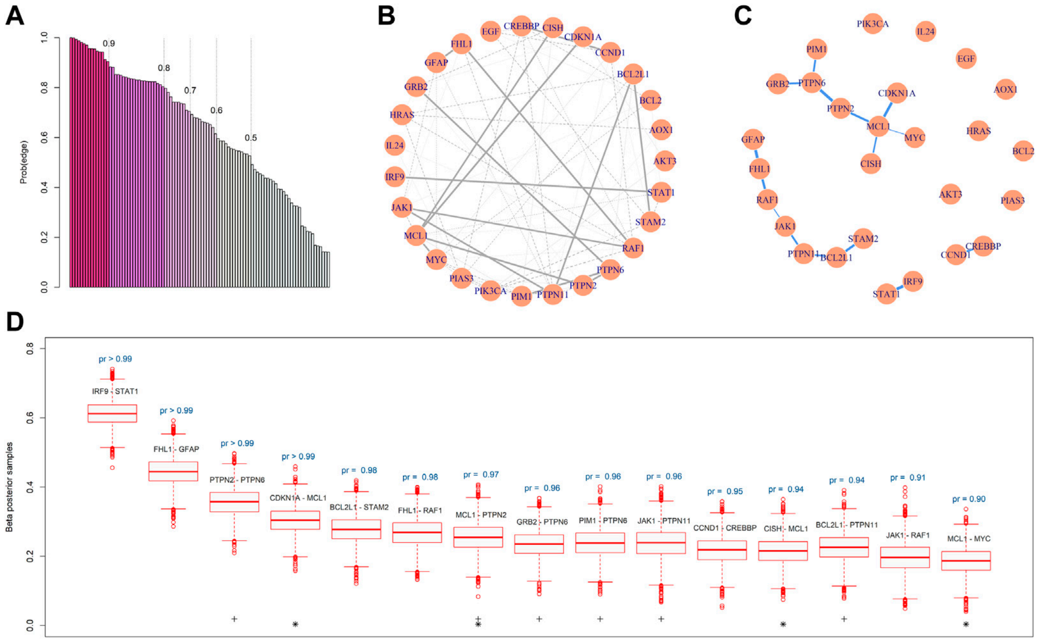


FIGURE 4
Gene regulatory network constructed by BMRF.P. **(A)** The ordered probabilities of the 99 edges are estimated by BMRF.P, and different colors correspond to different thresholds. The first 69 are the edges with a probability greater than 0.5; **(B)** The estimated genetic network. The 15 thick lines are edges with an estimated existence probability greater than 0.9; **(C)** The network structure containing only the 15 edges, where the width of the edge corresponds to the magnitude of the existence probability; **(D)** Boxplots of the posterior samples of the strength coefficient corresponding to each one of the 15 edges. The text above the boxplot represents the estimated existence probability. The '+' indicates edges involving genes in the PTPN family and '*' involves *MCL1*.

magnitude of the probability. The boxplots in Figure 4D show the posterior samples of the strength of each edge, all displaying positive conditional correlations between paired nodes. This is consistent with the pattern of co-expression, and the first two pairs seem to be strongly correlated with each other.

Note that the ordered existence probabilities in Figure 4A may be useful if prioritization is of interest. When comparing the top leading 15 edges with the lines in KEGG, we note that two edges (*JAK1-PTPN11* and *IRF9-STAT1*) are listed in KEGG. These two each have a probability greater than 0.95. The other thirteen edges with such a large probability were not listed in KEGG and may deserve further validation and investigation. For the connecting lines in KEGG, the BMRF posterior probabilities can be adopted to provide relative degrees of conditional dependence.

The proposed BMRF detected several influential biomarkers and biomarker pairs in the JAK-STAT network. First, the node *MCL1* is clearly crucial in this network since it appears in four edges (indicated with '*') among the 15 in Figure 4C. This hub node has been reported as one of the cell apoptosis inhibitors associated with the progression of GMB and participates in the signaling of the maintenance of neural stem cells (Fassl et al., 2012; Murphy et al., 2014). Second, in the constructed network by BMRF.P, the *PTPN2*, *PTPN6*, and *PTPN11* in the Protein-Tyrosine Phosphatase Non-Receptor (PTPN) family play critical roles. They appear in six edges (indicated with '+') among the 15 in Figure 4C. This is not surprising since the expression level of the immunotherapy target PTP2 has been shown to associate with the grade of glioma (Wang et al., 2018). Liu and others (Liu et al., 2011) have suggested *PTPN11* as a functional target for treating glioblastomas in human and animal studies, and Cerami et al. (2010) have identified *PTPN11* as associated with an oncogenic process in GBM patients. Members of the PTPN family induce dephosphorylation of *JAK*, thereby regulating JAK-STAT signaling (Xu and Qu, 2008; Jain et al., 2012; Hammarén et al., 2019). Third, the top-ranking pair shows the largest conditional dependence between *IRF9* and *STAT1*. This interaction was found to involve in type I interferon (IFN) signaling and anti-viral immune response (Au-Yeung et al., 2013). Fourth, BMRF.P identified the relationship between *MYC* and *MCL1*, where the transcription factor c-Myc of *MYC* was associated with the regulation the proliferation and survival of glioblastoma stem cells (Wang et al., 2008; Ha et al., 2015).

Other summary statistics regarding these 15 edges and all 69 edges are provided in the Supplementary Table S4; Supplementary Figure S3, respectively; and other interactions

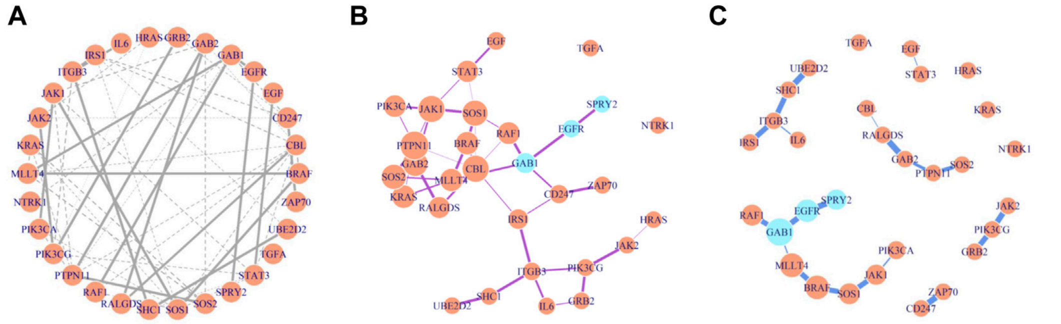


FIGURE 5
Gene regulatory network constructed by BMRF.P based on different thresholds. The width of edges is proportional to the existence probability and node size in **(B)** and **(C)** is proportional to number of immediate neighbors. **(A)** These 55 edges are of estimated existence probability greater than 0.5; **(B)** These 41 edges are of estimated existence probability greater than 0.7; **(C)** The 20 edges are of estimated existence probability greater than 0.9.

are summarized in the Supplementary Table S5. The findings of BMRF.P are compared with those of alternative procedures in the Supplementary Figures S4, S5. All edges identified by BMRF. P overlap with those identified by other procedures. Similar to the simulation studies, the edges identified by CLIME and BD_BMA overlap the least with the other procedures. This demonstrates again that the BMRF.P can provide more information than previous algorithms.

4.2 Edges in EGFR network with RNA-Seq

The BMRF model was next applied to the RNA sequencing gene expression of the 30 genes in the EGFR network. Figures 5A–C demonstrate the structure and relative strength of edges among these gene nodes, when different thresholds for the probability of existence are adopted. For instance, with the 0.5 threshold, 55 edges were identified, and with 0.90, 20 edges were detected. Three genes, *GAB1*, *EGFR*, and *SPRY2*, are colored differently to indicate that relatively *EGFR* depends more on the other two, if the conditional dependence inside this network is quantified and prioritized. Studies have shown that *GAB1* is involved in the cell proliferation and signaling process of positive feedback activation to *EGFR* (Kapoor and DM O'Rourke, 2010; Azuaje et al., 2015) and *SPRY2* knockdown is related to the negative prognosis and drug resistance of GBM (Walsh et al., 2015; Park et al., 2018; Day et al., 2020).

Another interesting observation is about the genes *GAB1* and *GAB2*. These two are crucial in the constructed network, appearing in five edges among 20 (Supplementary Figure S7). The probability of connection between these genes is strong (>0.9). The *GAB1* is connected to *EGFR* in the lower left in Figure 5C, and *GAB2* appears in the middle in Figure 5C. They apparently deserve more attention when studying the activity of this network.

In addition, note in Figure 5B where both *PTPN11* and *CBL* have six neighbors and are displayed with larger circles, indicating more connection with other gene nodes. When examining the edges with an existence probability greater than 0.9 in Figure 5C, these two genes interact with *GAB2*, *RALGDS*, and *SOS2* (in the middle of Figure 5C). These genes have been reported in the literature to associate with immune function and GBM. The findings here are not just reproducible results but also support that further investigation in the collective effect of these genes may be warranted. The hub nodes identified here and by other methods are consistent, as listed in Supplementary Table S7. More details can be found in Supplementary Section S3.

5 Discussion

In addition to the binary decision of edge existence, the proposed BMRF algorithm offers a probability measure of this existence, and is able to quantify the relative strength of edges, through the conditional autoregressive model and SSL prior. Its novelty lies in the Bayesian inference of the relative strength of the edges so that the conditional dependence can be prioritized. Simulation studies have demonstrated that, for the scale-free network, the performance of BMRF can be significantly improved when prior information is incorporated. Even when only the existence is of interest, the BMRF model can provide performance comparable with existing methods. In the two glioblastoma studies, the proposed algorithm highlights highly dependent subsets in the network that are worth for further investigation.

In contrast to other Bayesian network approaches, BMRF focuses on inference of the relative strength of the conditional dependence, while others are more interested in identifying non-zero elements in the precision matrix (Huang, 2022). The proposed method provides a complimentary tool when more interpretations of the relationship among genes is needed. That is, this BMRF can be executed with other Bayesian models, including ones that assign for the precision matrix a prior distribution composed of a product of all probability distributions of each element (Wang, 2012; Peterson et al., 2013; Gan et al., 2019), so that the post-processing computation can be saved. Another good choice is the BAGUS algorithm proposed by Gan et al. (2019). It provides a fast and accurate estimate of the graph structure, including the MAP estimate of the precision matrix with EM and the approximate inclusion probability of each edge. The implementation of the frequentist perspective may increase the scalability of BMRF. For example, these estimates may be utilized as baseline information to determine which edges to initially include for the inference of edge strength, or to tune the hyper-parameter values in the prior distributions of β_{jk} and γ_{jk}. Incorporation of such information may reduce the number of iterations required in the MCMC algorithm to save computational burden. The choice of the hyperparameter values τ_0 and τ_1 in the prior distribution does not change the basic outcome. The posterior distributions of β_{jk} corresponding to different hyperparameters are very similar, leading to the same conclusions based on the posterior distributions. Similarly, the order of the relative strength remains the same. In other words, the prioritization is not affected by the hyperparameter values. The magnitudes of the existence probability are linearly correlated, though the value may differ slightly. These observations are based on our limited experiments with the GBM application. Further studies may be warranted.

The computation time for the BMRF can be as long as 30 min per replication, especially under the current R package *R2OpenBUGS*. This is slow and can hinder the use of the proposed model. In contrast, the computation for the frequentist methods discussed here and the BAGUS is much faster. This is a reason why we did not consider a graph with more than 100 nodes in simulation studies. This limitation also restricts the use of the BMRF model to screen pairwise relationship among a large group of genes. Further research in tailoring a fast computation algorithm is worth investigating.

The proposed algorithm can be extended to integrative network analysis. With a graphical model comprised of biomarkers from different platforms, it is possible to reveal the underlying complex biological structure among various forms of molecules (Peng et al., 2010; Yin and Li, 2011; Ha et al., 2021). In this case, adjustments in the CAR model would be needed to account for the genetic variables at different levels. However, this approach would be computationally intensive when facing the enormous number of all parameters combined.

Another generalization of the BMRF is to relax the distributional assumption in the CAR model. The GGM for the gene network assumes the MVN as the joint distribution, and the conditional and marginal distribution are also Gaussian. This assumption may not be valid generally, particularly for gene expression data. Ho et al. (2022) performed a systematic study to investigate the multivariate normality of gene expression values. Several parametric and nonparametric multivariate tests were considered and applied on more than twenty sets of empirical data. It was concluded that the normality assumption is not guaranteed. Classical research has addressed non-Gaussian Markov random fields (Besag, 1974), but these studies are not designed for sparse neighborhood selection. One solution would be to combine the non-paranormal distribution in Liu et al. (2009) or the exponential family graphical model (Yang et al., 2015) with BMRF for further investigation.

When comparing the relative strength estimated by BMRF with the connecting lines in current pathway/network databases, two issues should be noted. First, databases like KEGG collect current knowledge of relationships, such as interactions and reactions, between molecules, and the resulting pathways/networks represent a collection of research findings from multiple studies involving various types of genetic markers. These studies are not necessarily comparable. In other words, although KEGG can be a good source to examine if the conditional dependence detected by BMRF has been identified before, one should bear in mind that the comparison may not be fair, since the data sets as well as the genetic biomarkers can be very different. Second, since the curation of pathways/networks is based on published literature, the definition of their connecting lines differs from the existence probability and the inclusion probability considered in this study. Therefore, a validation study of the findings here, especially for the two GBM studies, would need to be carefully designed. Disease status, tissue sample source and conditions, and genetic markers would all need to be incorporated for consideration.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: The glioblastoma dataset can be downloaded from TCGA hub and GDC hub in https://xenabrowser.net/datapages/. The R code for the implementation is available in https://github.com/YJGene0806/BMRF_Code.

Author contributions

Y-JH, RM, and CH contributed to the conceptualization of the study. RM and CH were major principal investigators in the funded projects. Y-JH contributed to statistical data analyses and machine learning modeling. Y-JH and CH prepared the original draft. All authors critically reviewed the draft and approved the final version.

Funding

This work was partially supported by the MOST 109-2314-B-002-152 and MOST 110-2314-B-002-078-MY3.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1034946/full#supplementary-material

References

Au-Yeung, N., Mandhana, R., and Horvath, C. M. (2013). Transcriptional regulation bySTAT1 and STAT2 in the interferon JAK-STAT pathway. *JAK-STAT* 2, e23931. doi:10.4161/jkst.23931

Azuaje, F., Tiemann, K., and Niclou, S. P. (2015). Therapeutic control and resistance of theEGFR driven signaling network in glioblastoma. *Cell Commun. Signal.* 13, 23. doi:10.1186/s12964-015-0098-6

Banerjee, K., Núñez, F. J., Haase, S., McClellan, B. L., Faisal, S. M., Carney, S. V., et al. (2021). Current approaches for glioma gene therapy and virotherapy. *Front. Mol. Neurosci.* 14, 621831. doi:10.3389/fnmol.2021.621831

Besag, J., and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* 82, 733–746. doi:10.1093/biomet/82.4.733

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B* 36, 192–225. doi:10.1111/j.2517-6161.1974.tb00999.x

Bralten, L. B. C., and French, P. J. (2011). Genetic alterations in glioma. *Cancers* 3, 1129–1140. doi:10.3390/cancers3011129

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* 78, 1–3. doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ 1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* 106, 594–607. doi:10.1198/jasa.2011.tm10155

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLOS ONE* 5, e8918. doi:10.1371/journal.pone.0008918

Chang, H.-C., Chu, C.-P., Lin, S.-J., and Hsiao, C. K. (2020). Network hub-node prioritization of gene regulation with intra-network association. *BMC Bioinforma.* 21, 101. doi:10.1186/s12859-020-3444-7

Day, E. K., Sosale, N. G., Xiao, A., Zhong, Q., Purow, B., and Lazzara, M. J. (2020). Glioblastoma cell resistance to EGFR and MET inhibition can be overcome via blockade of FGFR SPRY2 bypass signaling. *Cell Rep.* 30, 3383–3396.e7. doi:10.1016/j.celrep.2020.02.014

Deshpande, S. K., Ročková, V., and George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and slab lasso. *J. Comput. Graph. Stat.* 28, 921–931. doi:10.1080/10618600.2019.1593179

Dinga, R., Penninx, B. W., Veltman, D. J., Schmaal, L., and Marquand, A. F. (2019). Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. BioRxiv. Available at: https://www.biorxiv.org/content/10.1101/743138v1.full (Accessed August 22, 2019).743138

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* 3, 521–541. doi:10.1214/08-AOAS215SUPP

Fassl, A., Tagscherer, K. E., Richter, J., Berriel Diaz, M., Alcantara Llaguno, S. R., Campos, B., et al. (2012). Notch1 signaling promotes survival of glioblastoma cells via EGFR mediated induction of anti-apoptotic Mcl-1. *Oncogene* 31, 4698–4708. doi:10.1038/onc.2011.615

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441. doi:10.1093/biostatistics/kxm045

Gan, L., Narisetty, N. N., and Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *J. Am. Stat. Assoc.* 114, 1218–1231. doi:10.1080/01621459.2018.1482755

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015). Dingo: Differential network analysis in genomics. *Bioinformatics* 31, 3413–3420. doi:10.1093/bioinformatics/btv406

Ha, M. J., Stingo, F. C., and Baladandayuthapani, V. (2021). Bayesian structure learning in multilayered genomic networks. *J. Am. Stat. Assoc.* 116, 605–618. doi:10.1080/01621459.2020.1775611

Hammarén, H. M., Virtanen, A. T., Raivola, J., and Silvennoinen, O. (2019). The regulation of JAKs in cytokine signaling and its breakdown in disease. *Cytokine* 118, 48–63. doi:10.1016/j.cyto.2018.03.041

Ho, C.-H., Huang, Y.-J., Lai, Y.-J., Mukherjee, R., and Hsiao, C. K. (2022). The misuse of distributional assumptions in functional class scoring gene-set and pathway analysis. *G3* 12, jkab365. doi:10.1093/g3journal/jkab365

Huang, Y.-J. (2022). "Bayesian approaches to probabilistic genetic networks," (New Taipei, Taiwan: National Taiwan University). Doctoral Dissertation.

Huang, Y.-J., Lu, T.-P., and Hsiao, C. K. (2020). Application of graphical lasso in estimating network structure in gene set. *Ann. Transl. Med.* 8, 1556. doi:10.21037/atm-20-6490

Hutter, C., and Zenklusen, J. C. (2018). The cancer Genome Atlas: Creating lasting value beyond its data. *Cell* 173, 283–285. doi:10.1016/j.cell.2018.03.042

Jain, R., Dasgupta, A., Moiyadi, A., and Srivastava, S. (2012). Transcriptional analysis of JAK/STAT signaling in glioblastoma multiforme. *Curr. Pharmacogenomics Person. Med.* 10, 54–69. doi:10.2174/187569212800166648

Jalali, P., Khare, K., and Michailidis, G. (2022). A Bayesian subset specific approach to joint selection of multiple graphical models. *Stat. Sin.* doi:10.5705/ss.202021-0245

Kapoor, G. S., and O'Rourke, D. M. (2010). SIRPalpha1 receptors interfere with the EGFRvIII signalosome to inhibit glioblastoma cell transformation and migration. *Oncogene* 29, 4130–4144. doi:10.1038/onc.2010.164

Lin, S.-J., Lu, T.-P., Yu, Q.-Y., and Hsiao, C. K. (2018). Probabilistic prioritization of candidate pathway association with pathway score. *BMC Bioinforma.* 19, 391. doi:10.1186/s12859-018-2411-z

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* 10, 2295–2328.

Liu, K.-W., Feng, H., Bachoo, R., Kazlauskas, A., Smith, E. M., Symes, K., et al. (2011). SHP-2/PTPN11 mediates gliomagenesis driven by PDGFRA and INK4A/ARF aberrations in mice and humans. *J. Clin. Invest.* 121, 905–917. doi:10.1172/JCI43690

Ludwig, M., Dührkop, K., and Böcker, S. (2018). Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics* 34, i333–i340. doi:10.1093/bioinformatics/bty245

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* 34, 1436–1462. doi:10.1214/009053606000000281

Mohammadi, A., and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* 10, 109–138. doi:10.1214/14-BA889

Murphy, Á. C., Weyhenmeyer, B., Noonan, J., Kilbride, S. M., Schimansky, S., Loh, K. P., et al. (2014). Modulation of Mcl-1 sensitizes glioblastoma to TRAIL-induced apoptosis. *Apoptosis* 19, 629–642. doi:10.1007/s10495-013-0935-2

Ni, Y., Baladandayuthapani, V., Vannucci, M., and Stingo, F. C. (2021). Bayesian graphical Models for modern biological applications. *Stat. Methods Appt.* 31, 197–225. doi:10.1007/s10260021-00572-8

Ou, A., Ott, M., Fang, D., and Heimberger, A. B. (2021). The role and therapeutic targeting of JAK/STAT signaling in glioblastoma. *Cancers* 13, 437. doi:10.3390/cancers13030437

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., et al. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* 32.

Park, J.-W., Wollmann, G., Urbiola, C., Fogli, B., Florio, T., Geley, S., et al. (2018). Sprouty2 enhances the tumorigenic potential of glioblastoma cells. *Neuro. Oncol.* 20, 1044–1054. doi:10.1093/neuonc/noy028

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.* 104, 735–746. doi:10.1198/jasa.2009.0126

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., et al. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* 4, 53–77. doi:10.1214/09-AOAS271SUPP

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Am. Stat. Assoc.* 110, 159–174. doi:10.1080/01621459.2014.896806

Peterson, C., Vannucci, M., Karakas, C., Choi, W., Ma, L., and MaletićSavatić, M. (2013). Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. *Stat. Interface* 6, 547–558. doi:10.4310/SII.2013.v6.n4.a12

Ročková, V., and George, E. I. (2018). The Spike-and-Slab lasso. *J. Am. Stat. Assoc.* 113, 431–444. doi:10.1080/01621459.2016.1260469

Shergalis, A., Bankhead, A., Luesakul, U., Muangsin, N., and Neamati, N. (2018). Current challenges and opportunities in treating glioblastoma. *Pharmacol. Rev.* 70, 412–445. doi:10.1124/pr.117.014944

Stupp, R., Weller, M., Belanger, K., Bogdahn, U., Ludwin, S. K., Lacombe, D., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* 10, 987–996. doi:10.1056/NEJMoa043330

The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi:10.1038/nature07385

Walsh, A. M., Kapoor, G. S., Buonato, J. M., Mathew, L. K., Bi, Y., Davuluri, R. V., et al. (2015). Sprouty2 drives drug resistance and proliferation in glioblastoma. *Mol. Cancer Res.* 13, 1227–1237. doi:10.1158/1541-7786.MCR-14-0183-T

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* 7, 867–886. doi:10.1214/12-BA729

Wang, H., and Pillai, N. S. (2013). On a class of shrinkage priors for covariance matrix estimation. *J. Comput. Graph. Stat.* 22, 689–707. doi:10.1080/10618600.2013.785732

Wang, J., Wang, H., Li, Z., Wu, Q., Lathia, J. D., McLendon, R. E., et al. (2008). c-Myc is required for maintenance of glioma cancer stem cells. *PLOS ONE* 3, e3769. doi:10.1371/journal.pone.0003769

Wang, P., Cai, H., Zhang, C., Li, Y.-M., Liu, X., Wan, J., et al. (2018). Molecular and clinical characterization of PTPN2 expression from RNA-seq data of 996 brain gliomas. *J. Neuroinflammation* 15, 145. doi:10.1186/s12974-018-1187-4

Williams, D. R. (2021). Bayesian estimation for Gaussian graphical models: Structure learning, predictability, and network comparisons. *Multivar. Behav. Res.* 56, 336–352. doi:10.1080/00273171.2021.1894412

Xu, D., and Qu, C.-K. (2008). Protein tyrosine phosphatases in the JAK/STAT pathway. *Front. Biosci.* 13, 4925–4932. doi:10.2741/3051

Xu, Z., Zhang, G., Jin, F., Chen, M., Furey, T. S., Sullivan, P. F., et al. (2016). A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics* 32, 650–656. doi:10.1093/bioinformatics/btv650

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* 16, 3813–3847.

Yin, J., and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* 5, 2630–2650. doi:10.1214/11-AOAS494

Zhang, Y., Dube, C., Gibert, M., Cruickshanks, N., Wang, B., Coughlan, M., et al. (2018). The p53 pathway in glioblastoma. *Cancers* 10, 297. doi:10.3390/cancers10090297

Zhang, Z., Allen, G. I., Zhu, H., and Dunson, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage* 197, 330–343. doi:10.1016/j.neuroimage.2019.04.027

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* 13, 1059–1062.

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 December 2022
DOI 10.3389/fgene.2022.1015531

Biomarker-driven drug repurposing on biologically similar cancers with DNA-repair deficiencies

Seeya Awadhut Munj[1], Tasnimul Alam Taz[1], Suzan Arslanturk[1]* and Elisabeth I. Heath[2,3]

[1]Department of Computer Science, Wayne State University, Detroit, MI, United States, [2]Department of Oncology, Wayne State University, Detroit, MI, United States, [3]Molecular Therapeutics Program, Barbara Ann Karmanos Cancer Institute, Detroit, MI, United States

Similar molecular and genetic aberrations among diseases can lead to the discovery of jointly important treatment options across biologically similar diseases. Oncologists closely looked at several hormone-dependent cancers and identified remarkable pathological and molecular similarities in their DNA repair pathway abnormalities. Although deficiencies in Homologous Recombination (HR) pathway plays a significant role towards cancer progression, there could be other DNA-repair pathway deficiencies that requires careful investigation. In this paper, through a biomarker-driven drug repurposing model, we identified several potential drug candidates for breast and prostate cancer patients with DNA-repair deficiencies based on common specific biomarkers and irrespective of the organ the tumors originated from. Normalized discounted cumulative gain (NDCG) and sensitivity analysis were used to assess the performance of the drug repurposing model. Our results showed that Mitoxantrone and Genistein were among drugs with high therapeutic effects that significantly reverted the gene expression changes caused by the disease (FDR adjusted p-values for prostate cancer =1.225e-4 and 8.195e-8, respectively) for patients with deficiencies in their homologous recombination (HR) pathways. The proposed multi-cancer treatment framework, suitable for patients whose cancers had common specific biomarkers, has the potential to identify promising drug candidates by enriching the study population through the integration of multiple cancers and targeting patients who respond poorly to organ-specific treatments.

KEYWORDS

drug repurposing, DNA repair, personalized medicine, multi cancer treatment, mitoxantrone, homologous recombination

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Shyam Nyati,
Henry Ford Health System,
United States
Carlo Cesar Cortes,
National Cancer Institute, Mexico

*CORRESPONDENCE
Suzan Arslanturk,
✉ suzan.arslanturk@wayne.edu

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

RECEIVED 09 August 2022
ACCEPTED 15 November 2022
PUBLISHED 13 December 2022

CITATION
Munj SA, Taz TA, Arslanturk S and Heath EI (2022), Biomarker-driven drug repurposing on biologically similar cancers with DNA-repair deficiencies. *Front. Genet.* 13:1015531.
doi: 10.3389/fgene.2022.1015531

1 Introduction

Developing a new drug for a condition can take around 10–13 years and close to 2.8 billion dollars (DiMasi et al., 2016). Despite this, 90% of the drug candidates entering clinical trials fail (Sun et al., 2022). Human body is a complex system, with myriad interactions taking place simultaneously, interdependent on each other. The same pathway or mechanism involving certain genes, may be responsible for different diseases. A drug developed for a particular condition, therefore, could be a potential candidate for another condition. Drug repurposing can drastically reduce the time and cost of developing new drugs by searching for FDA-approved drugs, drugs under trial, or other chemicals that have a therapeutic effect on conditions outside the scope of the original medical indication (Pushpakom et al., 2019). Drug repurposing minimizes the chances of failure in clinical trials and reduces time for approval.

Similar molecular and genetic aberrations among diseases can lead to the discovery of jointly important treatment options across biologically similar diseases. Oncologists have closely looked at prostate, ovarian and breast cancers and identified that the tumors arising from these cancers are typically hormone-dependent and have remarkable underlying pathological and molecular similarities in their DNA repair pathway abnormalities (Risbridger et al., 2010). Analyzing patient data from biologically similar cancers together provides insights into their similarities as well as knowledge about individual cancers, which may not have been possible by analyzing individual cancer data separately. Zhou et al. (2021) identified jointly important biomarkers across breast, prostate and ovarian cancers by utilizing patient data from the three cancers using a cross-cancer learning approach. This reiterates that the same pathway or a gene is responsible for multiple diseases. These biological similarities have led to remarkably similar treatment options. For instance, combining the androgen deprivation therapy (ADT) with PARP inhibitors (i.e. drugs already used in breast cancer treatment) showed to be an effective approach in reducing the progression and recurrence of prostate cancer. Several single agent activity PARP inhibitors (PARPi) were recently approved for treating certain ovarian and breast cancers (Asim et al., 2017). The US Food and Drug Administration (FDA) approved the first multi-cancer treatment (Keytruda®), for patients whose cancers had a common specific biomarker. FDA, for the first time, approved a drug based on a common biomarker, instead of the organ the tumor had originated. Despite this, majority of studies still consider each cancer disease in isolation from the rest and identify the treatment options that are cancer-type specific. Hence, the critical need is to discover multi-cancer treatment options through the exploitation of cancers with similar molecular and genetic aberrations.

Mutations in several genes within the homologous recombination (HR) pathway occur in around 20%–25% of advanced prostate cancers (Marshall et al., 2019). There is accumulating evidence that depicts a considerable proportion of individuals with metastatic breast cancer are HR deficient with mutations in *BRCA1/BRCA2* genes (den Brok et al., 2017). Base excision repair (BER) pathway genes limit the ability of DNA repair in prostate cancer (PCa, henceforth) patients, which leads to an increased risk of PCa. (Mittal et al., 2012). Further, *APEX1*, which is a BER gene, has shown a compelling effect indicating an increased risk of breast cancer through a gene-gene interactivity analysis (Kim et al., 2013). In an effort to understand the effect of mismatch repair (MMR) genes in the progression of PCa, gene expression-based analysis were conducted within the cancer cell lines and in tumor specimens, which indicated a loss of *MSH2* and *MLH1* genes in different cell lines (Chen et al., 2001). The deficiency of MMR genes was observed across most of the subtypes of breast cancers with high-grade tumor-infiltrating lymphocyte counts (Cheng et al., 2020). All these findings confirmed that there were significant commonalities across breast and prostate cancers in their DNA repair pathway abnormalities that could lead to common and jointly important treatment options.

Drug repurposing strategies can be classified into drug-based and disease-based, depending on the substantial availability of data and the intent of the research (Jarada et al., 2020) (Dudley et al., 2011). Several computational approaches proposed in recent years have used both disease and drug data (Peyvandipour et al., 2018) (Sirota et al., 2011) (Chiang and Butte, 2009) (Gottlieb et al., 2011). In a systems biology approach proposed by Peyvandipour et al. (2018) a drug-disease network (DDN) was constructed by considering drug targets, disease-related genes and all signalling pathways that were then integrated with disease gene expression signatures and drug-exposure gene expression signatures to discover novel therapeutic roles for established drugs. Nafiseh et al. used a machine learning approach to find anti-similarities between drugs and disease (Saberian et al., 2019). In their approach, they used drug exposure gene expression data, disease gene expression data and the associations between FDA-approved drugs and diseases. They used a distance metric learning (DML) algorithm where disease and the associated FDA-approved drugs had smaller distances compared to drugs not associated with disease. Luo et al. (2016) proposed a novel approach that computed the similarity between drugs and diseases. In particular, they constructed a heterogeneous network consisting of drug and disease similarity networks and drug–disease interactions and then used a Bi-Random walk (BiRW) algorithm to rank the drugs (Xie et al., 2012). Hu and Agarwal. (2009) generated a disease-drug network based on extensive drug and disease gene expression profiles which was used for identifying new indications for drugs and side effects of drugs.

In this paper, we used several state-of-the-art drug repurposing approaches to determine potential drug

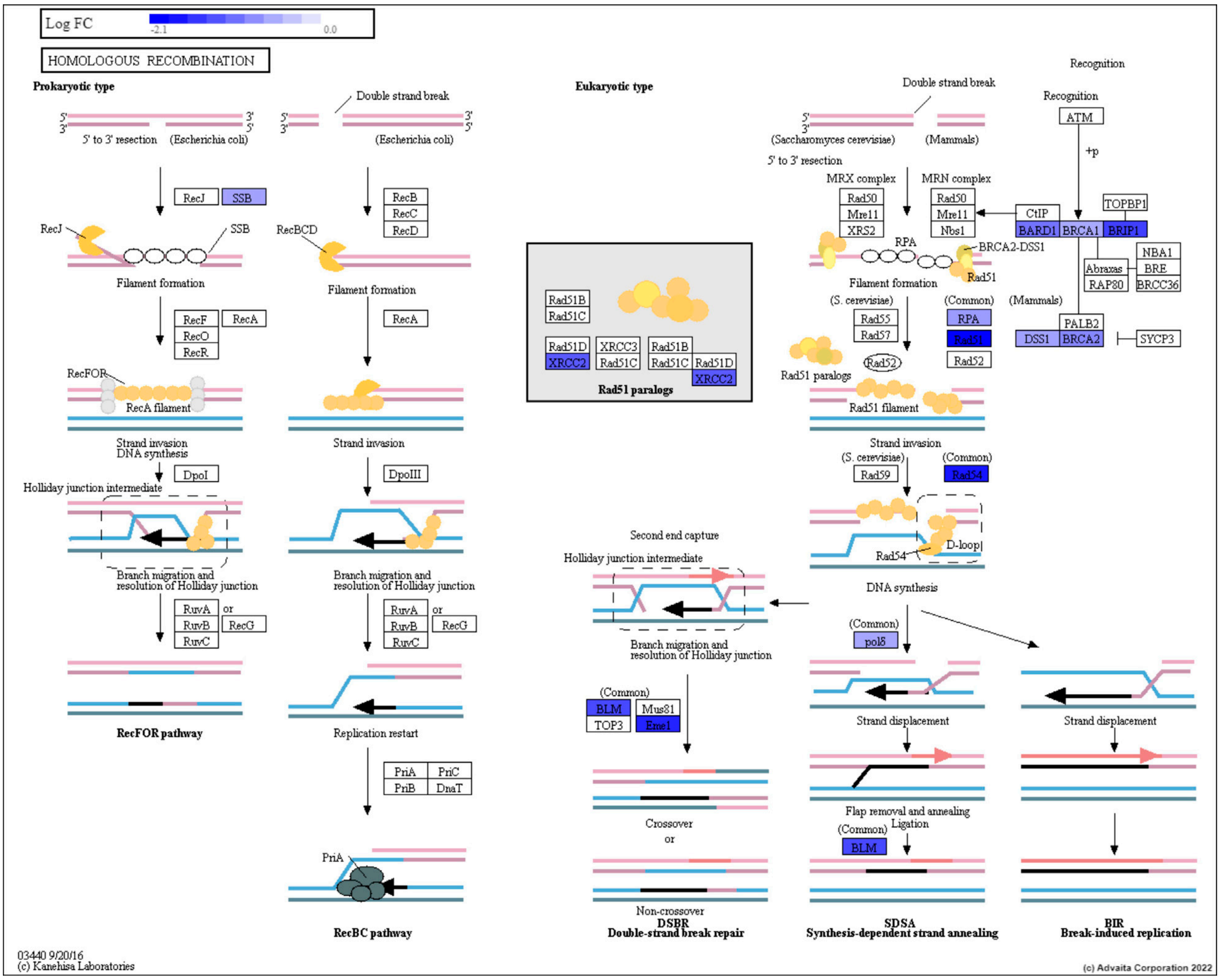


FIGURE 1
The Homologous Recombination pathway. The genes are represented in the rectangular boxes, with the shades of blue representing down-regulated genes for prostate cancer patients.

candidates for patients with breast or prostate cancers with common specific biomarkers. More specifically, we identified drugs with potential therapeutic effects on patients with DNA repair deficiencies.

Our contribution in this study is three-fold: 1) We initially developed a data-driven approach able to enrich the study population by integrating data from biologically similar cancers and using patient subpopulations with different types of DNA repair deficiencies which will enable personalized treatment strategies. We then used an existing approach referred to as drug-disease similarity to come up with novel treatments on the integrated data by identifying drugs that may have a therapeutic effect on patients irrespective of their cancer type. 2) We revisited our previously published deep cross cancer learning approach to identify jointly important biomarkers among breast, prostate and ovarian cancers. These biomarkers were used to identify common treatment options among those cancers through network interactions-based drug repositioning. 3) We presented the associations between the proposed drug target genes and biological functions (e.g., cell cycle) and investigated the drug target genes within the HR pathway and their interactions with the proposed drugs.

2 Materials and methods

2.1 Data preparation

The variant data and the disease gene expression data for breast and prostate cancers were obtained from The Cancer

TABLE 1 The number of breast and prostate cancer patients with deficiencies in their DNA repair pathways. Note that, different types of DNA-repair deficiencies has formed several subpopulations, that were analyzed separately.

DNA repair pathway	Number of patients	
	Breast cancer	Prostate cancer
Homologous Recombination (HR)	36	14
Base Excision Repair (BER)	23	7
Mismatch Repair (MMR)	73	31
Nucleotide Excision Repair (NER)	55	23
Non-Homologous End Joining (NHEJ)	23	6
Total	210	147

Genome Atlas (TCGA). The number of samples for breast and prostate tumors were 1,091 and 495, respectively with 120 and 53 samples with adjacent normal tissues. All expression datasets were log2 transformed. We obtained the signalling pathways from Kyoto Encyclopedia of Genes Genomics (KEGG) (Kanehisa et al., 2016). The signalling pathways are represented in the form of a directed graph, where each node represents the genes (or proteins) and the associations including activation, inhibition, etc. between the genes were represented by the edges. The large scale drug-exposure gene expression data were obtained from the Connectivity Map and the Library of Integrated Network-Based Cellular Signatures (LINCS) (Subramanian et al., 2017).

We initially identified all genes within each DNA repair pathway separately using the KEGG database. The DNA repair pathways used were: homologous recombination (HR), base excision repair (BER), mismatch repair (MMR), nucleotide excision repair (NER) and non-homologous end joining pathway (NHEJ). As an example, the set of genes (or proteins) that exist within the HR pathway can be seen in Figure 1. Using the variant data collected from TCGA, a subset of breast and prostate cancer patients with mutations in any of their DNA repair genes were identified and grouped according to their type of DNA repair deficiency. This resulted in multiple cohorts of homogeneous subpopulations with common biomarkers. Table 1 shows the distribution of the breast and prostate cancer patients within each cohort. Note that, the same patient may fall into multiple cohorts.

Next, we identified the differentially expressed genes (DEGs) through a moderated *t*-test by comparing the tumor samples with their adjacent normal tissues on each cohort separately. The resulting *p*-values were FDR adjusted to correct for multiple comparisons. Including ovarian cancer samples would have been optimal as ovarian cancer is known to also have biological similarities with breast and prostate cancers. However, due to not having access to TCGA ovarian cancer gene expression data of *adjacent normal tissue*, we were unable to run the differential expression analysis on ovarian cancer samples in this study. An alternative approach we considered was to run experiments on ovarian cancer data collected from different data sources, however this requires extensive preprocessing due to different representation, distribution, scale, and density of data.

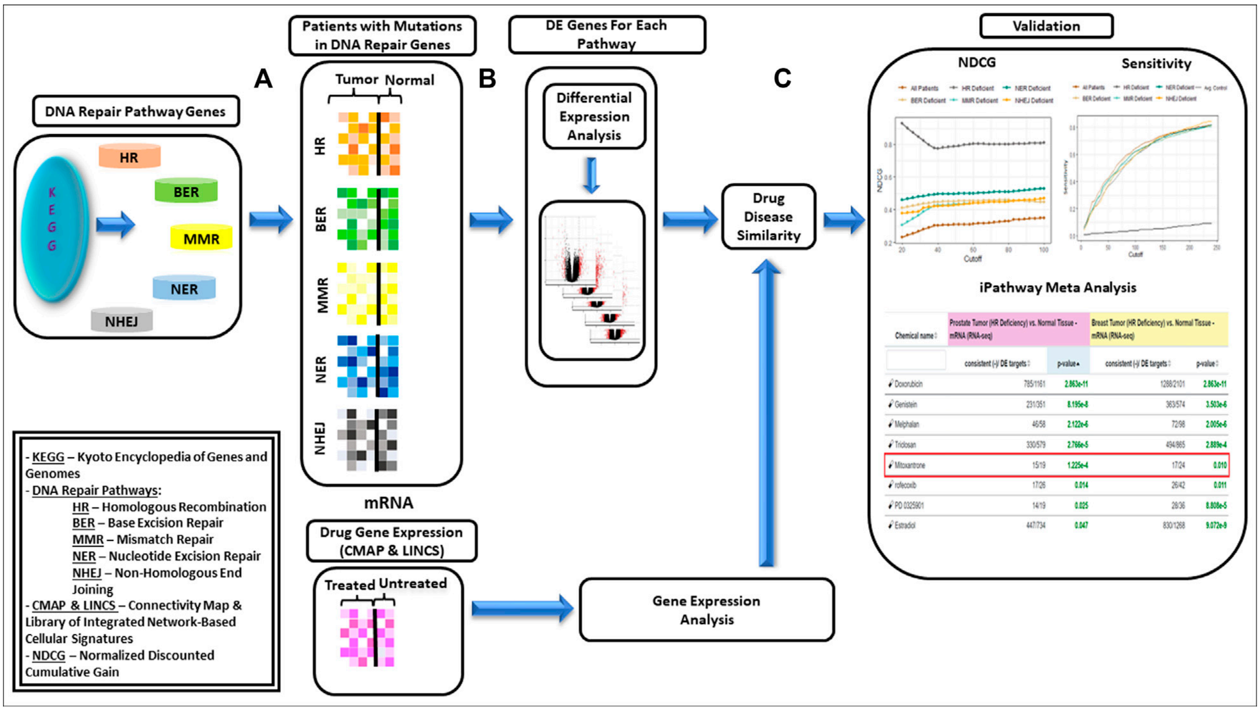
Our previously published deep cross cancer learning approach discussed in Section 3.3 identified jointly important biomarkers among breast, prostate, and ovarian cancers (Zhou et al., 2021). We were then able to identify drug candidates common among the three cancers using the proposed biomarkers. As this was a multi-label classification based neural network, we were able to conduct the analysis without the presence of ovarian normal tissue.

The methodology used for data preparation described above has been shown in Figures 2A,B. Prediction of drugs using drug-disease similarity and validation shown in Figure 2C has been described in subsequent sections.

2.2 The prediction of drugs using drug-disease similarities

Sirota et al. (2011) proposed a systematic computational drug repurposing approach to predict novel therapeutic indications by understanding drug and disease relationships. The association between every pairing of drug and disease is represented by a similarity score ranging from +1 to −1, with +1 indicating perfect correlation and −1 indicating an opposite effect. The largest negative score representing a reverse set of changes with exposure to a drug, indicates that the drug may have a therapeutic effect on the disease.

Here, we used the preprocessed expression data as discussed in Section 2.1 for breast and prostate cancer and the drug expression signatures from CMap to calculate the similarity scores. We only considered those drugs with FDR-adjusted *p*-values less than 0.05. This shortened list was then arranged in the ascending order based on the enrichment scores. The largest negative score implied the best drug candidates with highest therapeutic effects.



- KEGG – Kyoto Encyclopedia of Genes and Genomes
- DNA Repair Pathways:
  - HR – Homologous Recombination
  - BER – Base Excision Repair
  - MMR – Mismatch Repair
  - NER – Nucleotide Excision Repair
  - NHEJ – Non-Homologous End Joining
- CMAP & LINCS – Connectivity Map & Library of Integrated Network-Based Cellular Signatures
- NDCG – Normalized Discounted Cumulative Gain

Chemical name	Prostate Tumor (HR Deficiency) vs. Normal Tissue - mRNA (RNA-seq)		Breast Tumor (HR Deficiency) vs. Normal Tissue - mRNA (RNA-seq)	
	consistent (-)/ DE targets	p-value	consistent (-)/ DE targets	p-value
Doxorubicin	785/1161	2.863e-11	1288/2101	2.863e-11
Genistein	231/351	8.195e-8	363/574	3.503e-6
Melphalan	46/58	2.122e-6	72/98	2.005e-6
Triclosan	330/579	2.766e-5	494/865	2.889e-4
Mitoxantrone	15/19	1.225e-4	17/24	0.010
rofecoxib	17/26	0.014	26/42	0.011
PD 0325901	14/19	0.025	28/36	8.808e-5
Estradiol	447/734	0.047	830/1268	9.072e-9

FIGURE 2
Framework proposed for data-driven drug repurposing for biologically similar cancers—**(A)**: Genes within each of the DNA repair pathways, i.e., HR (Homologous Recombination), BER (Base Excision Repair), MMR (Mismatch Repair, NER(Nucleotide Excision Repair) and NHEJ (Non-Homologous End Joining) were identified using KEGG database. Subset of breast and prostate cancer patients with mutations in DNA repair genes were identified and grouped based on DNA repair deficiency. **(B)**: Differentially expressed genes (DEGs) were identified on each cohort separately. **(C)**: Drugs for each cohort were identified using Drug-Disease Similarity. Framework was validated using NDCG (Normalized Discounted Cumulative Gain) and sensitivity scores; and network interaction analysis was used for validating the utility of the drugs.

In an effort to evaluate the results obtained through the drug-disease similarity model, we performed sensitivity-based validation only (SV) and calculated the normalized discounted cumulative gain (NDCG). The best strategy for analytic validation of drug repurposing is through sensitivity based validation techniques. Sensitivity and specificity based validation, although ideal, is not practical to assess the model performance due to the lack of access to true negatives (TNs) as discussed by Adam et al. (Brown and Patel, 2018). The discounted cumulative gain was constructed under the assumption that top rank drugs were more relevant and more likely to be of interest (Schuler et al., 2022). The NDGC score was calculated as follows:

$$DCG = \sum_{i=1}^{p} \frac{2^{\mathrm{rel}_i - 1}}{\log_2 (i+1)} \tag{1}$$

$$IDCG = \sum_{i=1}^{|REL_p|} \frac{2^{\mathrm{rel}_i - 1}}{\log_2 (i+1)} \tag{2}$$

$$NDGC = DCG/IDCG \tag{3}$$

where i is the rank of the drug of interest, up to rank p, and rel_i denotes the relevance of the drug to the indication, 0 indicating non-relevance and 1 indicating relevance, REL_p is the list of associated drugs in the set up to a cutoff position of p, and $|REL_p|$ is the cardinality of the list.

2.3 The validation of proposed drugs using network interactions

Here, we used a drug repurposing analysis module to identify FDA-approved drugs that could be used to revert a given pattern of gene expression changes caused by a disease. The prediction of upstream Chemicals, Drugs, Toxicants (CDTs) is based on two types of information: 1) the enrichment of differentially expressed genes from the experiment and 2) a network of interactions from the Advaita Knowledge Base (AKB v2006). The network is a directed graph in which the source node represents either a chemical substance or compound, a drug, or a toxicant. The edges represent known effects that these CDTs have on various genes. A signed edge in this graph consists of a source CDT, a target gene, and a sign to indicate the type of effect: activation (+) or inhibition (−). To generate the network, the analysis selects only those edges observed in the literature with at least a medium confidence. The analysis considers two

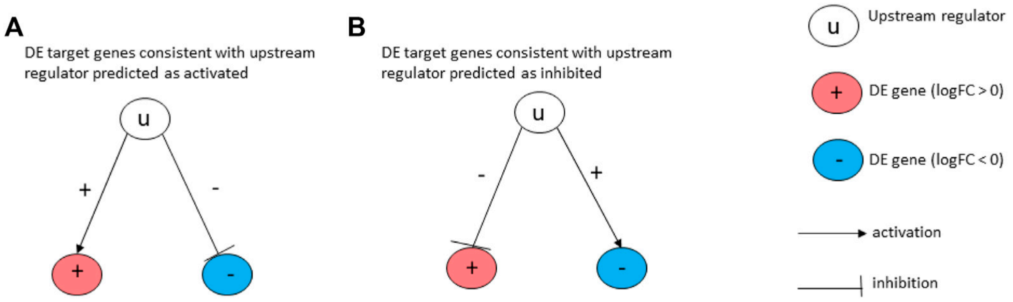

A
DE target genes consistent with upstream regulator predicted as activated
u
+
-
+
-
B
DE target genes consistent with upstream regulator predicted as inhibited
u
-
+
+
-
u
Upstream regulator
+
DE gene (logFC > 0)
-
DE gene (logFC < 0)
activation
inhibition

FIGURE 3
Target genes consistent with the hypothesis considered: In **(A)**, the signs of the DE genes shown in red (+) and blue (–) match the signs of their respective incoming edges, suggesting that the upstream regulator u is activated. In **(B)**, the signs of the DE genes shown in red (+) and blue (–) are opposite to the signs of their edges, suggesting that the upstream regulator u is inhibited.

hypotheses: HA: The upstream regulator is activated in the condition studied. HI: The upstream regulator is inhibited in the condition studied. The set of genes from National Center for Biotechnology Information (NCBI) Gene database is divided into many subsets by the analysis based on the measurements from the experiment and the definitions shown in Figure 3. The (+) sign in the figure indicates up-regulated genes while (−) sign indicates down-regulated genes. If a gene has at least one incoming edge, then it is considered as a target gene in the network. The gene g is consistent with hypothesis HA if there is an incoming edge e and if sign(g) = sign(e). This implies that when upstream regulator is activated, the signal is an activation and gene is up-regulated or signal is an inhibition, and the gene is down-regulated. (see Figure 3A). The gene g is consistent with hypothesis HI if there is an incoming edge e and if sign(g) does not match sign(e). This implies that when upstream regulator is inhibited the signal is inhibition and gene is up-regulated or signal is activation and gene is down-regulated. (see Figure 3B).

Herein, we focused on drugs that could reverse the changes induced by the disease. For this purpose, we hypothesized that the disease is considered as a state in which the changes are associated with the absence of a drug. Given the interactions between a specific drug A and its downstream DE genes, the Z-score was computed as follows:

$$z(A) = \frac{\sum_{e,g} w(g).s(e).s(g)}{\sqrt{\sum |w(g)|^2}} \tag{4}$$

where $s(e)$ represents the type of the edge (−1 for inhibition and +1 for activation), $s(g)$ is the sign of expression change of the gene (−1 for down-regulated and +1 for up-regulated), and $w(g)$ the confidence score of the edge g. The Z-score p-value for each drug was then calculated by mapping the z-score on a p-value using the normal distribution. (Draghici et al., 2020).

Note that, the drugs identified through drug-disease similarities as discussed in Section 2.2, though powerful, do not consider the network of interactions between drugs and their associated downstream genes. On the other hand, the network interactions as discussed in this section may still not be able to detect all significant drugs as only direct interactions between drug and disease is considered, rather than investigating indirect interactions due to co-expressions of genes. Hence in order to identify drugs with high therapeutic effects, we relied on the intersecting drugs among multiple approaches.

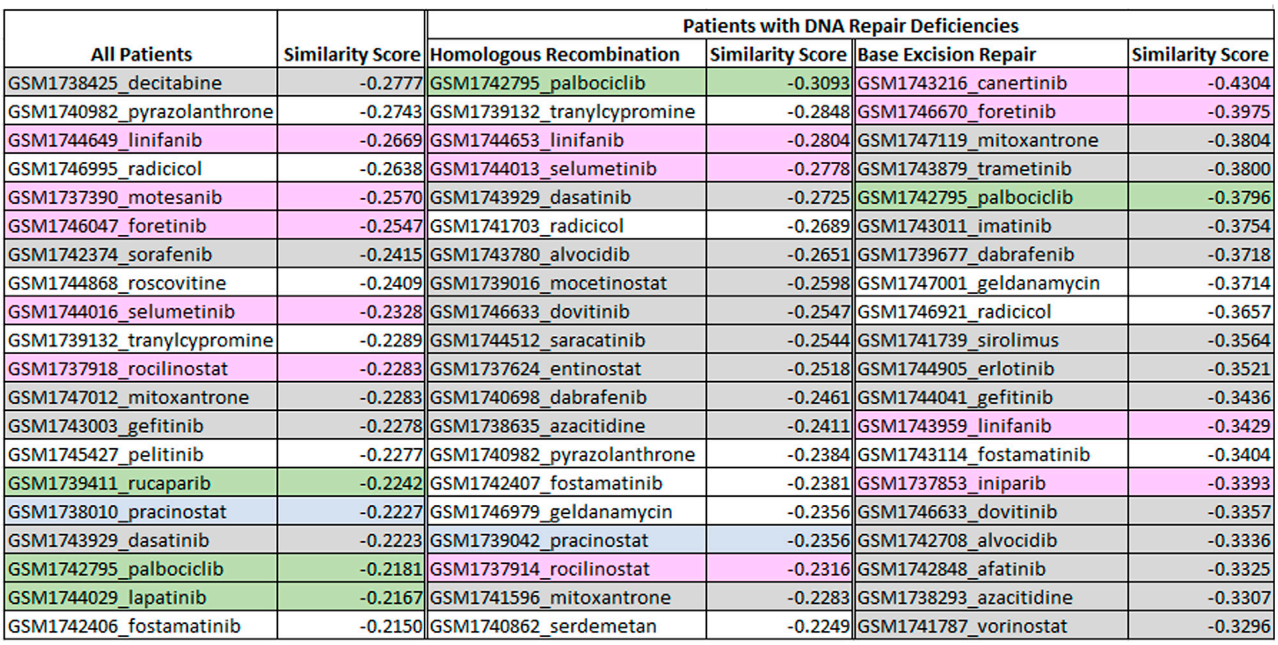
3 Results

3.1 Drug-disease similarity results

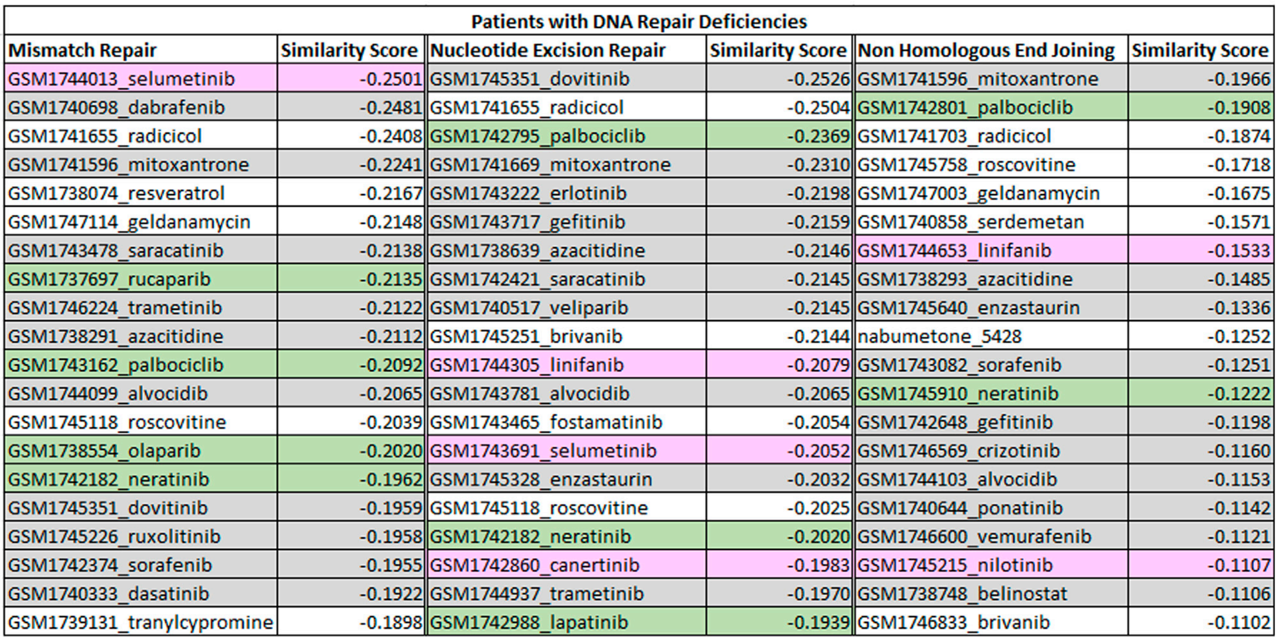
The results obtained through the drug-disease similarity analysis are shown in Table 2. Initially, all breast and prostate cancer patients were included in the analysis which resulted in a list of drugs presented in the first column of the table (see column: All Patients). In essence, a good repurposing approach on a truly homogeneous data should place the already FDA-approved drugs (i.e., the gold standard) at the very top of the list for that particular disease. Note that, since we focussed on multiple biologically similar diseases, we expected to see drugs approved for either or both of the conditions at the very top of the list.

Results showed that six investigational drugs (two of which are under investigation for breast and prostate cancers, and four of which are under investigation for breast cancer only) and no FDA-approved drugs appeared within the top 10 ranked drugs. Cancer being a heterogeneous disease with large genetic diversity even between tumors of the same cancer types, it is common for the patients to have significant differences between their molecular profiles (Arslanturk et al., 2020). Our results clearly showed that the data needed to be further refined to identify more homogeneous subpopulations for more optimal and targeted treatment decisions. Hence, as the next step, we

TABLE 2 The list of top ranked drugs identified through the drug-disease score analysis for subsets of patients with different types of DNA repair deficiencies. The cells highlighted in green, grey, blue and pink are the FDA-approved drugs, investigational drugs for breast and prostate cancers, investigational drugs for prostate cancer, and investigational drugs for breast cancer, respectively along with their respective similarity scores that was calculated. Results demonstrated that although there are certain drugs that are common across subpopulations, the top ranked drugs differed between different DNA-repair pathways. Hence, the identification of biomarkers associated with a specific subpopulation can change the course of treatment and enable personalized treatment strategies among individuals.



		Patients with DNA Repair Deficiencies			
All Patients	Similarity Score	Homologous Recombination	Similarity Score	Base Excision Repair	Similarity Score
GSM1738425_decitabine	-0.2777	GSM1742795_palbociclib	-0.3093	GSM1743216_canertinib	-0.4304
GSM1740982_pyrazolanthrone	-0.2743	GSM1739132_tranylcypromine	-0.2848	GSM1746670_foretinib	-0.3975
GSM1744649_linifanib	-0.2669	GSM1744653_linifanib	-0.2804	GSM1747119_mitoxantrone	-0.3804
GSM1746995_radicicol	-0.2638	GSM1744013_selumetinib	-0.2778	GSM1743879_trametinib	-0.3800
GSM1737390_motesanib	-0.2570	GSM1743929_dasatinib	-0.2725	GSM1742795_palbociclib	-0.3796
GSM1746047_foretinib	-0.2547	GSM1741703_radicicol	-0.2689	GSM1743011_imatinib	-0.3754
GSM1742374_sorafenib	-0.2415	GSM1743780_alvocidib	-0.2651	GSM1739677_dabrafenib	-0.3718
GSM1744868_roscovitine	-0.2409	GSM1739016_mocetinostat	-0.2598	GSM1747001_geldanamycin	-0.3714
GSM1744016_selumetinib	-0.2328	GSM1746633_dovitinib	-0.2547	GSM1746921_radicicol	-0.3657
GSM1739132_tranylcypromine	-0.2289	GSM1744512_saracatinib	-0.2544	GSM1741739_sirolimus	-0.3564
GSM1737918_rocilinostat	-0.2283	GSM1737624_entinostat	-0.2518	GSM1744905_erlotinib	-0.3521
GSM1747012_mitoxantrone	-0.2283	GSM1740698_dabrafenib	-0.2461	GSM1744041_gefitinib	-0.3436
GSM1743003_gefitinib	-0.2278	GSM1738635_azacitidine	-0.2411	GSM1743959_linifanib	-0.3429
GSM1745427_pelitinib	-0.2277	GSM1740982_pyrazolanthrone	-0.2384	GSM1743114_fostamatinib	-0.3404
GSM1739411_rucaparib	-0.2242	GSM1742407_fostamatinib	-0.2381	GSM1737853_iniparib	-0.3393
GSM1738010_pracinostat	-0.2227	GSM1746979_geldanamycin	-0.2356	GSM1746633_dovitinib	-0.3357
GSM1743929_dasatinib	-0.2223	GSM1739042_pracinostat	-0.2356	GSM1742708_alvocidib	-0.3336
GSM1742795_palbociclib	-0.2181	GSM1737914_rocilinostat	-0.2316	GSM1742848_afatinib	-0.3325
GSM1744029_lapatinib	-0.2167	GSM1741596_mitoxantrone	-0.2283	GSM1738293_azacitidine	-0.3307
GSM1742406_fostamatinib	-0.2150	GSM1740862_serdemetan	-0.2249	GSM1741787_vorinostat	-0.3296



Patients with DNA Repair Deficiencies					
Mismatch Repair	Similarity Score	Nucleotide Excision Repair	Similarity Score	Non Homologous End Joining	Similarity Score
GSM1744013_selumetinib	-0.2501	GSM1745351_dovitinib	-0.2526	GSM1741596_mitoxantrone	-0.1966
GSM1740698_dabrafenib	-0.2481	GSM1741655_radicicol	-0.2504	GSM1742801_palbociclib	-0.1908
GSM1741655_radicicol	-0.2408	GSM1742795_palbociclib	-0.2369	GSM1741703_radicicol	-0.1874
GSM1741596_mitoxantrone	-0.2241	GSM1741669_mitoxantrone	-0.2310	GSM1745758_roscovitine	-0.1718
GSM1738074_resveratrol	-0.2167	GSM1743222_erlotinib	-0.2198	GSM1747003_geldanamycin	-0.1675
GSM1747114_geldanamycin	-0.2148	GSM1743717_gefitinib	-0.2159	GSM1740858_serdemetan	-0.1571
GSM1743478_saracatinib	-0.2138	GSM1738639_azacitidine	-0.2146	GSM1744653_linifanib	-0.1533
GSM1737697_rucaparib	-0.2135	GSM1742421_saracatinib	-0.2145	GSM1738293_azacitidine	-0.1485
GSM1746224_trametinib	-0.2122	GSM1740517_veliparib	-0.2145	GSM1745640_enzastaurin	-0.1336
GSM1738291_azacitidine	-0.2112	GSM1745251_brivanib	-0.2144	nabumetone_5428	-0.1252
GSM1743162_palbociclib	-0.2092	GSM1744305_linifanib	-0.2079	GSM1743082_sorafenib	-0.1251
GSM1744099_alvocidib	-0.2065	GSM1743781_alvocidib	-0.2065	GSM1745910_neratinib	-0.1222
GSM1745118_roscovitine	-0.2039	GSM1743465_fostamatinib	-0.2054	GSM1742648_gefitinib	-0.1198
GSM1738554_olaparib	-0.2020	GSM1743691_selumetinib	-0.2052	GSM1746569_crizotinib	-0.1160
GSM1742182_neratinib	-0.1962	GSM1745328_enzastaurin	-0.2032	GSM1744103_alvocidib	-0.1153
GSM1745351_dovitinib	-0.1959	GSM1745118_roscovitine	-0.2025	GSM1740644_ponatinib	-0.1142
GSM1745226_ruxolitinib	-0.1958	GSM1742182_neratinib	-0.2020	GSM1746600_vemurafenib	-0.1121
GSM1742374_sorafenib	-0.1955	GSM1742860_canertinib	-0.1983	GSM1745215_nilotinib	-0.1107
GSM1740333_dasatinib	-0.1922	GSM1744937_trametinib	-0.1970	GSM1738748_belinostat	-0.1106
GSM1739131_tranylcypromine	-0.1898	GSM1742988_lapatinib	-0.1939	GSM1746833_brivanib	-0.1102

investigated potential treatment options based on common biomarkers, specifically for patients with aberrations in genes within different DNA repair mechanisms. Results showed Palbociclib, an endocrine-based chemotherapeutic agent approved for treating HER2-negative and HR-positive advanced or metastatic breast cancers (McCain, 2015) (Walker et al., 2016) (Beaver et al., 2015), appeared at the top of the list for patients with HR-deficiencies. Results further suggested that tranylcypromine, a monoamine oxidase inhibitor, mainly approved for the treatment of major depressive episodes without melancholia (Ricken et al., 2017), showed promise as a multi-cancer treatment, specifically for

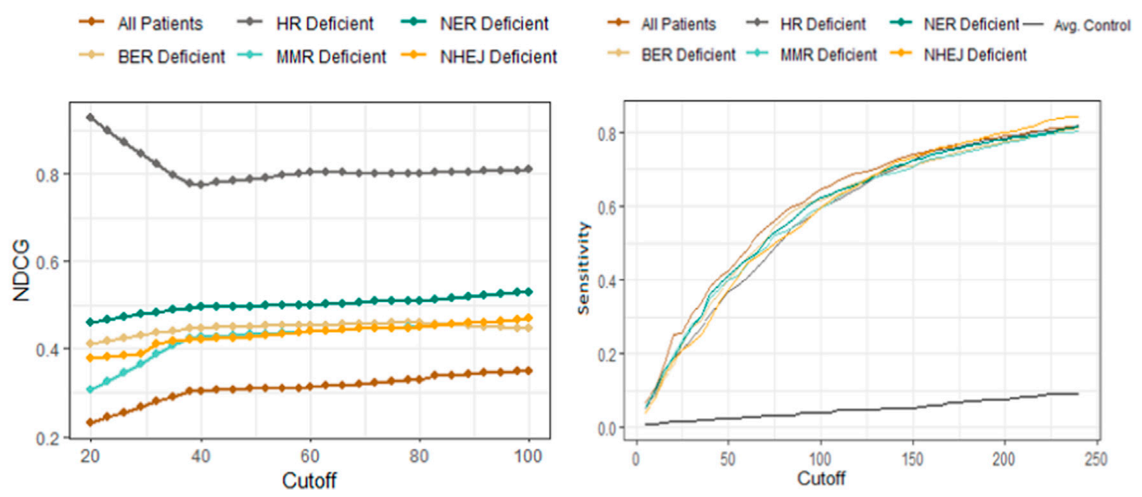


FIGURE 4
Performance comparison of the drug-disease similarity model on DNA-repair deficient patient subpopulations using NDCG (left) and sensitivity analysis (right). The NDCG/sensitivity values (vertical axes) of all drug–indication associations using different DNA repair deficient subpopulations are shown according to different cutoff values (horizontal axis). The NDCG results clearly demonstrate that the HR-deficient subpopulations result in drugs that are clinically more relevant with more FDA-approved/investigational drugs compared with other DNA-repair pathway deficiencies. The plot has further shown that identifying homogeneous subpopulations through common biomarkers result in better performances when compared to all patients combined. The sensitivity values demonstrate that the list of breast/prostate cancer drugs retrieved for all cutoff levels are clinically relevant and indicates an overall better performance relative to random controls (shown as the black curve).

breast and prostate cancers. The top ranked drugs further consisted of several chemotherapy drugs including linifanib, selumetinib and dasatinib. The top ranked drugs for all other DNA repair deficient patients are listed in Table 2. A detailed description of all the top ranked drugs for each pathway along with their clinical relevance is reported in the discussion section of the paper.

The sensitivity and NDCG scores of the proposed drugs are shown in Figure 4. The sensitivity values of all drug-disease associations for different subsets of patients based on their types of DNA repair deficiencies were compared with several random control runs. The sensitivity values were reported for different rank/cutoff levels. The SV results as shown in Figure 4B demonstrates that the list of drugs retrieved for all cutoff levels for breast and prostate cancer patients were clinically relevant and indicated an overall better performance relative to random controls. The NDCG scores as shown in Figure 4A show that the identification of homogeneous sub populations with common biomarkers resulted in drugs that were clinically more relevant with more FDA-approved/investigational drugs appearing at the very top of the list when compared with all patients combined. Results further showed that drugs proposed for patients with aberrations in their HR pathway outperformed all other pathways. This is mainly due to hormone driven cancers' significant molecular similarities within HR pathways (Toh and Ngeow, 2021) (Watkins et al., 2014). Less is known about the similarities between those cancers in other DNA-repair pathways.

3.2 Drugs proposed through network interactions

The drugs proposed through network interactions using iPathwayGuide (Advaita) are listed in Table 3. Note that, this table includes only the drugs that have a significant therapeutic effect ($p < 0.05$) on both breast and prostate cancers. The number of DE genes that would be reverted by each drug is listed. For instance, the 15/19 notation next to mitoxantrone demonstrates that there were 19 downstream genes that mitoxantrone is interacting with that were DE for prostate cancer (vs. adjacent normal tissue), 15 of which were consistent with our hypothesis as described in Section 2.3.

The SV and NDCG are metrics used to evaluate the drug repurposing models' ability to identify clinically relevant treatment options. In order to validate the utility of the drugs proposed, we investigated the mechanisms through which the drugs act on genes measured to be DE for the disease studied. Figure 5 generated using network interactions shows the mechanisms of mitoxantrone on the DE genes for prostate cancer. Mitoxantrone was able to activate the down-regulated genes and inhibit the up-regulated genes 15 out of 19 times ($p < 1.225e-4$) as described in Section 2.3 In an effort to confirm the changes in the downstream genes, we have reported the fold-changes of those genes using cell lines treated with Mitoxantrone as shown in Figure 5B. The upregulated genes are highlighted in red, and the downregulated genes are highlighted in blue.

TABLE 3 The top eight drugs proposed for repurposing using the network interactions approach. The table shows the *p*-values (sorted based on the prostate tumor vs. adjacent normal tissue experiment), as well as the number of DE genes that would be reverted by each drug (i.e., the number of genes consistent with the hypothesis) for patients with HR–deficiencies. Doxorubicin slows or stops the growth of cancer cells, and is used to treat certain neoplastic conditions such as acute lymphoblastic leukemia, soft tissue and bone sarcomas, breast carcinoma and ovarian carcinoma. Genistein is currently under clinical trials for the treatment of prostate cancer. Melphalan and Estradiol are also among drugs used to treat certain cancers. Mitoxantrone is highlighted as a promising drug candidate as it appears to be a top drug using both network interactions and drug-disease similarity scores.

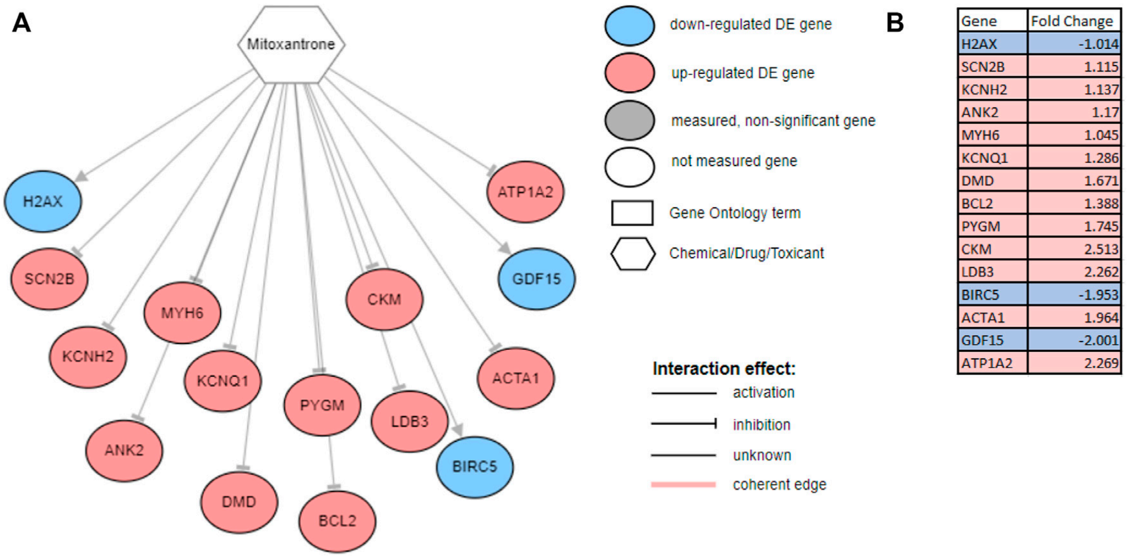
Chemical name	Prostate tumor (HR deficiency) vs. Normal tissue - mRNA (RNA-seq)		Breast tumor (HR deficiency) vs. Normal tissue - mRNA (RNA-seq)	
	Consistent (-)/DE targets	*p*-value	Consistent (-)/DE targets	*p*-value
Doxorubicin	785/1161	2.863e-11	1288/2101	2.863e-11
Genistein	231/351	8.195e-8	363/574	3.503e-6
Melphalan	46/58	2.122e-6	72/98	2.005e-6
Triclosan	330/579	2.766e-5	494/865	2.889e-4
Mitoxantrone	15/19	1.225e-4	17/24	0.01
rofecoxib	17/26	0.014	26/42	0.011
PD 0325901	14/19	0.025	28/36	8.808e-5
Estradiol	447/734	0.047	830/1268	9.072e-9

3.3 Drugs proposed using novel biomarkers discovered using cross cancer learning approach

We utilized our previously published approach that discovered jointly important novel biomarkers across breast, prostate and ovarian cancers through a data-driven, deep learning approach referred to as cross-cancer learning (Zhou et al., 2021). This approach exploited patient data from multiple cancers to discover prostate cancer biomarkers and jointly important biomarkers across breast, prostate and ovarian cancers by leveraging pathological and molecular similarities in their DNA repair pathways. Different cancers share common genomic instabilities. Exploring cancers having similarities can help discover previously unknown biomarkers and pathways. In addition, this helps in alleviating the problem of limited patient samples availability and underestimation of various genes previously not known to be involved. This cross cancer learning framework utilized a multi-label classification autoencoder (MLC-AE) that used lower dimensional latent representation of the mRNA gene expression profiles to predict the tissue type (breast, prostate, ovarian) and the disease state (solid tumor vs. adjacent normal tissue) as separate output layers. To explain and interpret the MLC-AE model, SHapley Additive exPlanations (SHAP) was used. This method uses SHAP values to extract feature importance across three cancers. SHAP method used each feature to calculate the change in performance in the presence and absence of each feature. The features whose absence lead to reduction in the performance were given the highest score. The cross cancer framework has been shown in Figure 6. Figure 7 A shows the most significant genes based on their contribution towards prediction using breast, prostate, and ovarian tissues. The biomarkers discovered using this approach were further used to find disrupted pathways using the impact analysis. The drugs identified using cross cancer genes are listed in Figure 7B and are discussed in detail in the Discussion section.

In order to validate our results further, additional experiments were conducted using the cell lines obtained from CMap. Table 4 shows the fold changes that were calculated using the cell lines treated with the drugs shown on each column. Specifically, the drugs investigated were Genistein, Mitoxantrone, Palbociclib, Tranylcypromine, Linifanib and Selumetinib. Threshold parameters used for the analysis were an absolute fold-change greater than 0.6 and false discovery rate (FDR) adjusted *p*-value less than 0.05. All genes presented in the table are differentially expressed, with the genes associated with DNA repair pathways being color-coded. Specifically, the red, yellow, green, blue, and gray colors represent significant changes in the genes associated with BER, HR, MMR, NER and a combination of multiple DNA repair pathways, respectively.

Note that there are no differentially expressed genes involved in the DNA repair process for Genistein. However, expression changes obtained from CMap includes an arbitrary selection of patients and is not filtered based on homogeneous subpopulations identified through specific DNA repair deficiencies. Instead, our proposed drug candidates have been derived by filtering a list of patients with specific types of DNA repair deficiencies, and therefore, is a preprocessed dataset with a more homogenous population than the CMap patient set.



Gene	Fold Change
H2AX	-1.014
SCN2B	1.115
KCNH2	1.137
ANK2	1.17
MYH6	1.045
KCNQ1	1.286
DMD	1.671
BCL2	1.388
PYGM	1.745
CKM	2.513
LDB3	2.262
BIRC5	-1.953
ACTA1	1.964
GDF15	-2.001
ATP1A2	2.269

FIGURE 5
(A) The mechanism through which Mitoxantrone act on the genes measured to be DE for prostate cancer. Note that, out of the 19 downstream DE genes that Mitoxantrone is interacting with, 15 were consistent with the hypothesis, i.e., the drug was able to revert the expression changes caused by disease 15 out of 19 times. All 15 genes were shown on the figure with three down-regulated genes (blue circles) being activated, and 12 up-regulated genes (red circles) being inhibited with the exposure of the drug. **(B)** Fold changes reported for cell lines treated with Mitoxantrone. The upregulated genes are highlighted in red, and the downregulated genes are highlighted in blue.

Although this could explain the lack of gene changes in DNA repair pathways when Genistein is administered, additional analyses would be required to further confirm the therapeutic effect of this drug.

In order to understand the effect of our proposed drugs on the nodes within the HR pathway, we explored the drug-gene interactions. The results are shown in Figure 8. The differentially expressed genes highlighted in this figure are based on patients with HR deficient breast cancer vs. adjacent normal tissue. This figure clearly shows that several HR genes are indeed drug targets and our proposed drugs are indeed interacting with such genes.

In summary, our results showed several promising drug candidates including Mitoxantrone, Palbociclib and Genistein for multi-cancer treatment as supported by multiple approaches. Mitoxantrone appeared to be a top drug using drug-disease similarity scores and network interactions approaches, and Genistein appeared to be a top drug using cross-cancer biomarkers and network interactions.

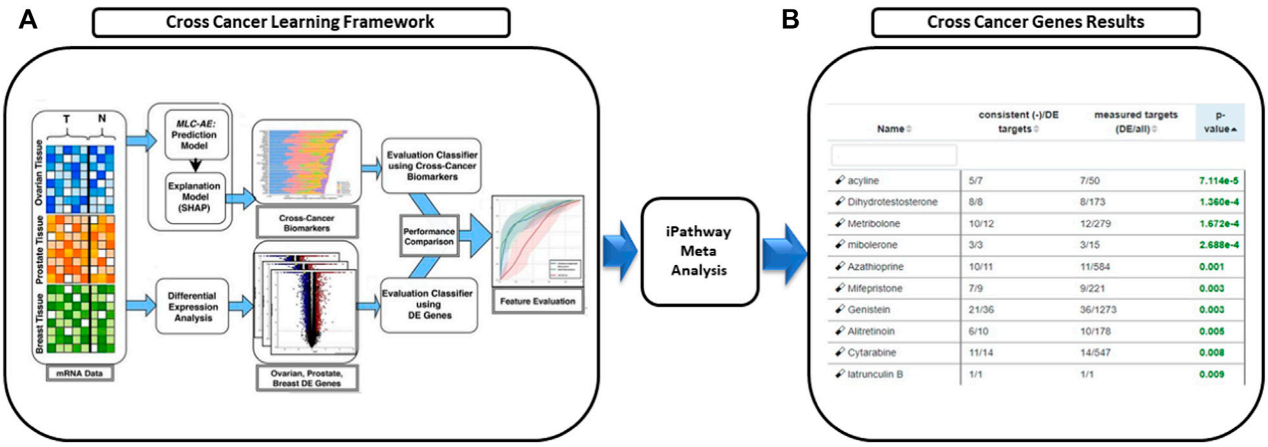
Experiments conducted by Tang et al. (2018) and Siddiqui et al. (2021) suggest that genistein and mitoxantrone in combination with other drugs can influence the cell cycle of the cancer cells. In order to understand the effects of these drugs on cell-cycle, we ran an experiment on iPathwayGuide, to understand the associations between the downstream genes of Genistein and Mitoxantrone and their associations with biological processes including the cell cycle. Results are presented in Figure 9.

4 Discussion

DNA damage is not uncommon and results in tens of thousands of damages everyday (Jackson and Bartek, 2009; O'Connor, 2015). This genomic instability is the key feature of carcinogenesis. DNA damage response (DDR) collectively refers to all the mechanisms that are responsible for the DNA damage repair. O'Connor. (2015) discussed targeted therapies based on DNA damage response of patients to tailor targeted therapy. They further mentioned various drugs under clinical trials for different types of cancers targeting DNA repair pathways.

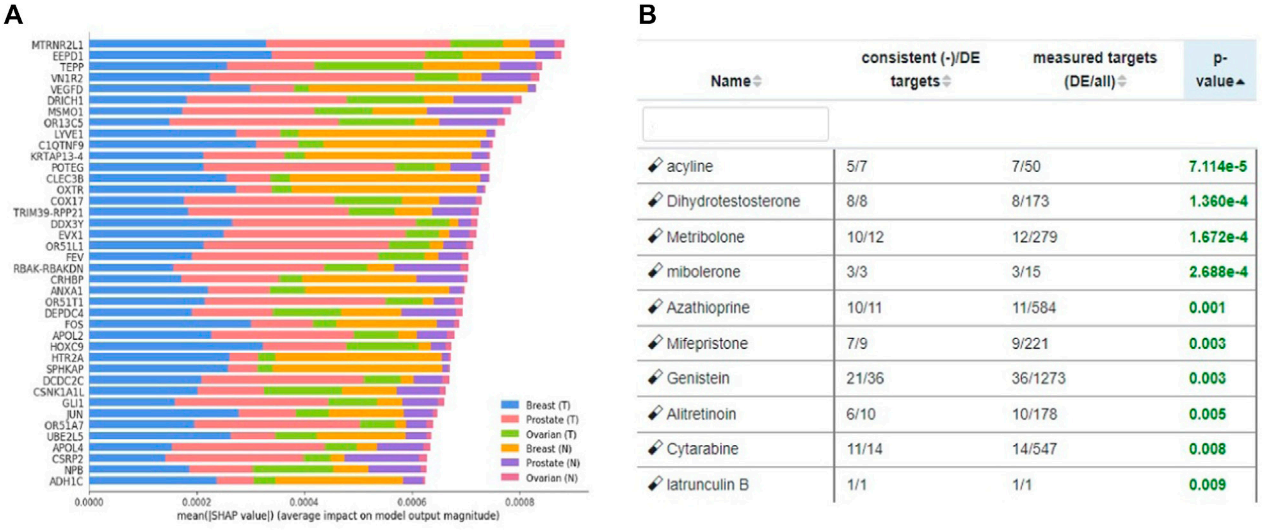
Homologous Recombination is responsible for the repair of DNA double stranded breaks (DSBs) during G2/M phase (Saleh-Gohari and Helleday, 2004). Li and Heyer. (2008); Al-Mugotir et al. (2021) showed that doxorubicin, and quinacrine, along with mitoxantrone were effective in HR deficient cells by recruiting *RAD52* to repair sites of DNA damage.

Table 2 shows the drugs that were identified using drug-disease score analysis for the subset of patients who had deficiencies in their DNA repair pathways for prostate cancer and breast cancer. In the list of drugs identified for HR pathway, palbociclib came as significant. Palbociclib is approved for HER2-negative and HR-positive advanced or metastatic breast cancer. It is known that *BRCA1* and *BRCA2* mutations are involved in the HR deficiency. Hence, this could be a promising drug for the prostate cancer patients who at


A
Cross Cancer Learning Framework
T
N
Ovarian Tissue
Prostate Tissue
Breast Tissue
mRNA Data
MLC-AE: Prediction Model
Explanation Model (SHAP)
Cross-Cancer Biomarkers
Differential Expression Analysis
Ovarian, Prostate, Breast DE Genes
Evaluation Classifier using Cross-Cancer Biomarkers
Performance Comparison
Evaluation Classifier using DE Genes
Feature Evaluation
iPathway Meta Analysis
B
Cross Cancer Genes Results

Name	consistent (-)/DE targets	measured targets (DE/all)	p-value
acyline	5/7	7/50	7.114e-5
Dihydrotestosterone	8/8	8/173	1.360e-4
Metribolone	10/12	12/279	1.672e-4
mibolerone	3/3	3/15	2.688e-4
Azathioprine	10/11	11/584	0.001
Mifepristone	7/9	9/221	0.003
Genistein	21/36	36/1273	0.003
Alitretinoin	6/10	10/178	0.005
Cytarabine	11/14	14/547	0.008
latrunculin B	1/1	1/1	0.009

FIGURE 6
Drugs proposed using cross cancer genes. **(A)** Breast, prostate and ovarian cancer expression data was used to predict the tissue type and the disease type using multi-label classification—auto encoder (MLC-AE). SHAP Explanation model was used to identify the contribution of each gene towards the prediction using SHAP values that rank the genes. **(B)** Network interaction analysis was used to perform meta analysis and predict novel drugs.



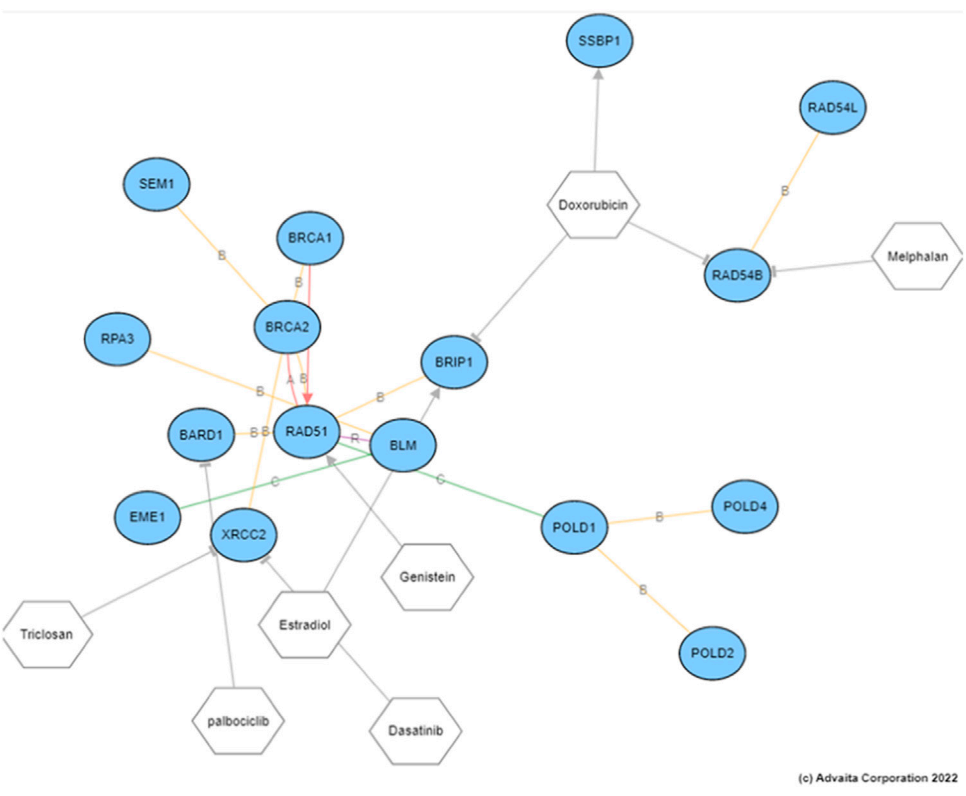
Name	consistent (-)/DE targets	measured targets (DE/all)	p-value
acyline	5/7	7/50	7.114e-5
Dihydrotestosterone	8/8	8/173	1.360e-4
Metribolone	10/12	12/279	1.672e-4
mibolerone	3/3	3/15	2.688e-4
Azathioprine	10/11	11/584	0.001
Mifepristone	7/9	9/221	0.003
Genistein	21/36	36/1273	0.003
Alitretinoin	6/10	10/178	0.005
Cytarabine	11/14	14/547	0.008
latrunculin B	1/1	1/1	0.009

FIGURE 7
(A) Significant genes identified using SHAP based on contribution scores for all three tissues (breast, prostate and ovary). (T) Denotes solid tumor and (N) denotes solid normal tissue. Figure utilized from Zhou et al. (2021) **(B)** Top eight drugs proposed for repurposing using cross cancer genes.

present do not respond to the current treatment. The network interactions approach shown in Table 3 came up with interesting set of drugs. Studies have shown that Genistein affects cell cycle during G2/M phase (Zhang et al., 2013). Genistein inhibits protein-tyrosine kinase and topoisomerase-II (DNA topoisomerases, type II) and is under investigation as an anti-cancer agent. *In vivo* experiments carried out by Tang et al. (2018). have showed that Genistein when combined with AG1024 (a tyrosine kinase inhibitor) led to a decrease in tumor size in prostate cancer patients. Genistein suppressed the homologous recombination (HR) and the non-homologous end joining (NHEJ) pathways by inhibiting the expression of *Rad51* and *Ku70* (Tang et al., 2018). Genistein, an isoflavone found in soy products and an integral part of the Asian diet, was found to be effective against various cancers and responsible for lowering the prostate and breast cancer rates in

TABLE 4 This table shows the expression changes of genes when the drugs that were found to be significant in our analysis were administered. Red, yellow, green, blue, and gray colors represent significant changes in the genes associated with BER, HR, MMR, NER, and a combination of multiple DNA repair pathways, respectively.

MITOXANTRONE			GENISTEIN			PALBOCICLIB			TRANYLCYPROMINE			LINIFANIB			SELUMETINIB	
Gene	Fold Change		Gene	Fold Change		Gene	Fold Change		Gene	Fold Change		Gene	Fold Change		Gene	Fold Change
POLE2	-16.0984		TFPI2	-0.601714122		PCNA	-10		OXA1L	-10		BIRC5	-10		PARP1	-10
RFC2	-12.9617		TACC3	0.666655568		POLE2	-7.882		CCNA1	-10		CD44	-10		CD58	-10
BRCA1	-9.5644		KRT12	0.648860924		PARP1	-5.49		DNM1L	-10		CDK1	-10		IER3	-10
RFC5	-6.14226667		MRPS14	-1.007313661		DDB2	-4.78		PHKA1	-9.618		DLD	-10		DNM1L	-10
TOPBP1	-4.7233		DET1	-0.640326879		BRCA1	-4.616		TJP1	-6.588		MCM3	-10		TXNDC9	-10
PARP1	-2.8768		CRYBB1	-0.614814657		RFC2	-4.372		MSH6	-2.472		LIG1	-2.806		CCNH	0.732
MSH6	-2.0806		OSR2	-0.827627071		LIG1	-3.853		RFC5	-0.747		TOPBP1	1.414		PARP2	1.32
RAD51C	-1.7007		NCAPD3	0.632002377		PARP2	-3.061		RFC2	-0.633		PARP2	1.51		CDK7	1.33
CCNH	-1.3119		MBIP	-0.619790329		RFC5	-2.443		CDK7	0.719		RFC5	2.55		RFC5	1.506
CDK7	1.4994		ASCC2	-0.685010907		MSH6	-1.715		POLD4	0.801		DUSP22	8.1212		ERCC6	2.187
ERCC6	2.5378		DUSP1	0.618809242		TOPBP1	-1.405		RAD51C	1.05		PRR15L	8.34		CCL2	10
POLD4	4.5108		TMEM186	-0.633685938		RAD51C	-1.186		BRCA1	1.68		POLD4	8.877		SUV39H1	10
DDB2	5.9346		LIPE	-0.995489484		CCNH	-0.761		P4HA2	4.113		FBXL12	9.742		TLR4	10
ADAT1	13.0098		GABARAPL3	-0.625627357		ERCC6	0.732		CLPX	4.227		CIAO3	10		MAP7	10
DDX42	13.9626		MAP7D1	-0.715569418		CDK7	1.634		PRR15L	4.285		C2CD2	10		CLPX	10
MAPKAPK2	15.32706667		SUPT16H	-0.628213568		POLD4	8.877		NUDT9	4.683		E2F2	10		PRR7	10



(c) Advaita Corporation 2022

FIGURE 8
The drug target genes within the HR pathway and their interactions with the proposed drugs. The differential expression analysis here were conducted on patients with breast cancer with HR deficiencies vs. adjacent normal tissue.

Asian countries. It inhibited the cell cycle proliferation and induces apoptosis. (Banerjee et al., 2008). Khan et al. (2021) described the emerging role of natural products in cancer treatment. Among them, soy isoflavones, were reported to target *BRCA* histones for repair. Through their *in vivo* experiments, Fan et al. (2006) found that genistein along with indoole-3-carbinol targeted both *BRCA1* and *BRCA2* genes in breast and prostate cancer cells. This research is useful in

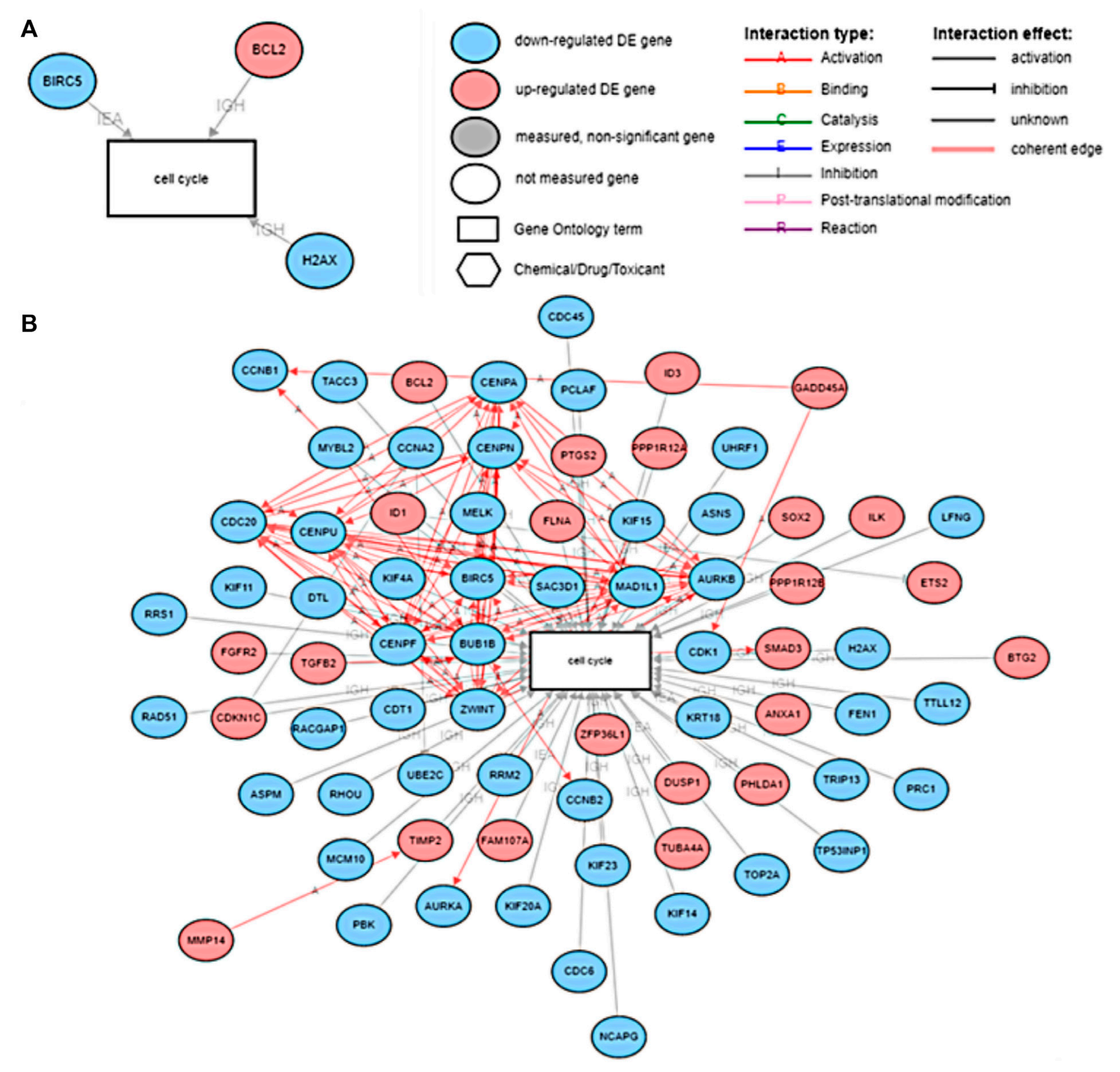


FIGURE 9
(A) The DE genes of prostate cancer associated with the cell cycle downstream of Mitoxantrone and **(B)** The DE genes of prostate cancer associated with the cell cycle downstream of Genistein.

suggesting that natural products can be potential therapeutics for cancer treatment.

Al-Mugotir et al. (2021) listed mitoxantrone as a potential drug for clinical use targeting Topoisomerase II. Siddiqui et al. (2021) showed that mitoxantrone along with imatinib could be used to suppress apoptosis. Their research specifically targeted treatment-resistant HR-proficient cancers. *RAD52*, a protein involved in the HR pathway, was found to be differentially expressed in BRCA-deficient cells. The changes in the expression of the gene *RAD52* is associated with HR activity and hence can affect the way cancer can be treated (Nogueira et al., 2019), (Lok and Powell, 2012). Al-Mugotir et al. (2021) reported that *RAD52* could be a potential target for the HR deficient cancers and further showed the effectiveness of mitoxantrone on such cancers. These findings further strengthen our proposed results of mitoxantrone as a potential candidate for patients with mutations in their HR repair pathways.

In a study conducted on COX-2 inhibitors and breast cancer patients between 1998–2004, it was shown that rofecoxib had the highest percentage (71%, $p < 0.01$) of breast cancer reduction as compared to other drugs including ibuprofen (63%) and 325 mg aspirin (49%). (Harris et al., 2014).

Estradiol is already in use for breast and prostate cancers for palliation therapy.

Figure 7B shows the drugs that were listed as siginificant for the novel biomarkers discovered using cross cancer learning approach by Zhou et al. (2021) Acyline showed as significant drug in our table. In a study conducted by Sofikerim et al. (2007) to find the hormonal predictors of the prostate cancer, follicle-stimulating hormone (FSH) was found to be significantly higher in patients with prostate cancer. Crawford et al. (2017) discussed about evidences of high levels of FSH in the advanced and metastatic prostate cancer. Christenson and Antonarakis. (2018) discussed the use of gonadotropin-releasing hormone (GnRH) agonists to inhibit FSH levels as an initial step once prostate cancer turns metastatic. In the first experiment conducted on humans, Herbst et al., (2002) found that acyline, a novel GnRH antagonist was found to suppress FSH levels. They discussed the use of acyline as a probable prostate cancer drug. O'Toole et al. (2007) discussed the potential use of acyline for breast cancer and prostate cancer. Limonta et al. (2012) discussed GnRH agonists decreasing the tumor growth and proliferation in prostate, ovarian and breast cancers. Genistein, which came up as significant for HR-deficient patients earlier, was listed as significant for cross cancer genes as well and has been discussed earlier.

Currently, there is a strong evidence that the biologically similar cancers have the same underlying genetic aberrations (Risbridger et al., 2010). Hence, providing jointly important treatments could drastically reduce the time invested in development of novel drugs as well as repurposing drugs for diseases separately. Our study exploited the prostate cancer and breast cancer patients with deficiencies in their DNA-repair pathways. There is not clear understanding of DNA repair pathways (excluding HR pathway) involved in the breast and prostate cancer, and hence may require further study. There is a strong evidence that a subset of prostate and breast cancer patients have deficiencies in their HR pathways. The drugs proposed using our approach for this pool of patients have strong evidence from literature and show strong promise.

5 Conclusion

DNA repair pathways are responsible for maintaining the genome stability by performing various mechanisms to reverse the damage caused. Failure to do so may result in various diseases, including cancer. Most malignancies arise from mutations caused by damage to the DNA that was not repaired. While some patients respond to treatments, a subset of patients do not respond to the standard treatments. This clearly concludes that there is heterogeneity within the same type of cancer that needs to be further refined.

In this paper, we identified commonalities and differences among multiple cancers by leveraging the abnormalities within the DNA repair pathways to identify potential drugs through repurposing. Often, a specific drug repurposing approach may not always provide optimal results due to its limitations. Hence, we employed multiple approaches and provided treatment options that were intersecting between the approaches.

Our multi cancer treatment model 1) integrated subsets of patients with common biomarkers in their DNA repair pathways and 2) provided promising drug candidates for patients with different DNA repair deficiencies. The results of the proposed framework can be further utilized as a personalized medicine option for patients who do not respond to regular and organ specific treatment options.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://cancergenome.nih.gov.

Author contributions

SM, EH, and SA. conceived and designed the project. SM performed the experiments. SM, TT, and SA analyzed the data and the results. SM and SA wrote the paper. All authors read and approved the final manuscript.

Funding

This research was funded by the National Science Foundation (NSF: 1948338), the Department of Defense (DoD: W81XWH-21-1-0570), and the National Institutes of Health (NIH: 2P50CA186786-06).

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.

References

Al-Mugotir, M., Lovelace, J. J., George, J., Bessho, M., Pal, D., Struble, L., et al. (2021). Selective killing of homologous recombination-deficient cancer cell lines by inhibitors of the rpa: Rad52 protein-protein interaction. *PloS one* 16, e0248941. doi:10.1371/journal.pone.0248941

Arslanturk, S., Draghici, S., and Nguyen, T. (2020). Integrated cancer subtyping using heterogeneous genome-scale molecular datasets. *Pac. Symp. Biocomput.* 25, 551–562.

Asim, M., Tarish, F., Zecchini, H. I., Sanjiv, K., Gelali, E., Massie, C. E., et al. (2017). Synthetic lethality between androgen receptor signalling and the parp pathway in prostate cancer. *Nat. Commun.* 8, 374. doi:10.1038/s41467-017-00393-y

Banerjee, S., Li, Y., Wang, Z., and Sarkar, F. H. (2008). Multi-targeted therapy of cancer by genistein. *Cancer Lett.* 269, 226–242. doi:10.1016/j.canlet.2008.03.052

Beaver, J. A., Amiri-Kordestani, L., Charlab, R., Chen, W., Palmby, T., Tilley, A., et al. (2015). Fda approval: Palbociclib for the treatment of postmenopausal patients with estrogen receptor–positive, her2-negative metastatic breast cancer. *Clin. Cancer Res.* 21, 4760–4766. doi:10.1158/1078-0432.CCR-15-1185

Brown, A. S., and Patel, C. J. (2018). A review of validation strategies for computational drug repositioning. *Brief. Bioinform.* 19, 174–177. doi:10.1093/bib/bbw110

Chen, Y., Wang, J., Fraig, M. M., Metcalf, J., Turner, W. R., Bissada, N. K., et al. (2001). Defects of dna mismatch repair in human prostate cancer. *Cancer Res.* 61, 4112–4121.

Cheng, A. S., Leung, S. C., Gao, D., Burugu, S., Anurag, M., Ellis, M. J., et al. (2020). Mismatch repair protein loss in breast cancer: Clinicopathological associations in a large British columbia cohort. *Breast Cancer Res. Treat.* 179, 3–10. doi:10.1007/s10549-019-05438-y

Chiang, A. P., and Butte, A. J. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* 86, 507–510. doi:10.1038/clpt.2009.103

Christenson, E. S., and Antonarakis, E. S. (2018). Parp inhibitors for homologous recombination-deficient prostate cancer. *Expert Opin. Emerg. Drugs* 23, 123–133. doi:10.1080/14728214.2018.1459563

Crawford, E. D., Schally, A. V., Pinthus, J. H., Block, N. L., Rick, F. G., Garnick, M. B., et al. (2017). “The potential role of follicle-stimulating hormone in the cardiovascular, metabolic, skeletal, and cognitive effects associated with androgen deprivation therapy,” in *Urologic Oncology: Seminars and original investigations* (Amsterdam, Netherlands: Elsevier), Vol. 35, 183–191.

den Brok, W. D., Schrader, K. A., Sun, S., Tinker, A. V., Zhao, E. Y., Aparicio, S., et al. (2017). Homologous recombination deficiency in breast cancer: A clinical review. *JCO Precis. Oncol.* 1, 1–13. doi:10.1200/PO.16.00031

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of r&d costs. *J. Health Econ.* 47, 20–33. doi:10.1016/j.jhealeco.2016.01.012

Draghici, S., Nguyen, T.-M., Sonna, L. A., Ziraldo, C., Vanciu, R., Fadel, R., et al. (2020). *Covid-19: Disease pathways and gene expression changes predict methylprednisolone can improve outcome in severe cases*. medRxiv. doi:10.1101/2020.05.06.20076687

Dudley, J. T., Deshpande, T., and Butte, A. J. (2011). Exploiting drug–disease relationships for computational drug repositioning. *Brief. Bioinform.* 12, 303–311. doi:10.1093/bib/bbr013

Fan, S., Meng, Q., Auborn, K., Carter, T., and Rosen, E. (2006). Brca1 and brca2 as molecular targets for phytochemicals indole-3-carbinol and genistein in breast and prostate cancer cells. *Br. J. Cancer* 94, 407–426. doi:10.1038/sj.bjc.6602935

Gottlieb, A., Stein, G. Y., Ruppin, E., and Sharan, R. (2011). Predict: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496. doi:10.1038/msb.2011.26

Harris, R. E., Casto, B. C., and Harris, Z. M. (2014). Cyclooxygenase-2 and the inflammogenesis of breast cancer. *World J. Clin. Oncol.* 5, 677–692. doi:10.5306/wjco.v5.i4.677

Herbst, K. L., Anawalt, B. D., Amory, J. K., and Bremner, W. J. (2002). Acyline: The first study in humans of a potent, new gonadotropin-releasing hormone antagonist. *J. Clin. Endocrinol. Metab.* 87, 3215–3220. doi:10.1210/jcem.87.7.8675

Hu, G., and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PloS one* 4, e6536. doi:10.1371/journal.pone.0006536

Jackson, S. P., and Bartek, J. (2009). The dna-damage response in human biology and disease. *Nature* 461, 1071–1078. doi:10.1038/nature08467

Jarada, T. N., Rokne, J. G., and Alhajj, R. (2020). A review of computational drug repositioning: Strategies, approaches, opportunities, challenges, and directions. *J. Cheminform.* 12, 46–23. doi:10.1186/s13321-020-00450-7

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070

Khan, H., Labanca, F., Ullah, H., Hussain, Y., Tzvetkov, N. T., Akkol, E. K., et al. (2021). Advances and challenges in cancer treatment and nutraceutical prevention: The possible role of dietary phenols in brca regulation. *Phytochem. Rev.* 2021, 385–400. doi:10.1007/s11101-021-09771-3

Kim, K.-Y., Han, W., Noh, D. Y., Kang, D., and Kwack, K. (2013). Impact of genetic polymorphisms in base excision repair genes on the risk of breast cancer in a Korean population. *Gene* 532, 192–196. doi:10.1016/j.gene.2013.09.069

Li, X., and Heyer, W.-D. (2008). Homologous recombination in dna repair and dna damage tolerance. *Cell. Res.* 18, 99–113. doi:10.1038/cr.2008.1

Limonta, P., Marelli, M. M., Mai, S., Motta, M., Martini, L., and Moretti, R. M. (2012). Gnrh receptors in cancer: From cell biology to novel targeted therapeutic strategies. *Endocr. Rev.* 33, 784–811. doi:10.1210/er.2012-1014

Lok, B. H., and Powell, S. N. (2012). Molecular pathways: Understanding the role of Rad52 in homologous recombination for therapeutic advancement. *Clin. Cancer Res.* 18, 6400–6406. doi:10.1158/1078-0432.CCR-11-3150

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.-X., et al. (2016). Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* 32, 2664–2671. doi:10.1093/bioinformatics/btw228

Marshall, C. H., Fu, W., Wang, H., Baras, A. S., Lotan, T. L., and Antonarakis, E. S. (2019). Prevalence of dna repair gene mutations in localized prostate cancer according to clinical and pathologic features: Association of gleason score and tumor stage. *Prostate Cancer Prostatic Dis.* 22, 59–65. doi:10.1038/s41391-018-0086-1

McCain, J. (2015). First-in-class cdk4/6 inhibitor palbociclib could usher in a new wave of combination therapies for hr+, her2- breast cancer. *P T.* 40, 511–520.

Mittal, R. D., Mandal, R. K., and Gangwar, R. (2012). Base excision repair pathway genes polymorphism in prostate and bladder cancer risk in north indian population. *Mech. Ageing Dev.* 133, 127–132. doi:10.1016/j.mad.2011.10.002

Nogueira, A., Fernandes, M., Catarino, R., and Medeiros, R. (2019). Rad52 functions in homologous recombination and its importance on genomic integrity maintenance and cancer therapy. *Cancers* 11, 1622. doi:10.3390/cancers11111622

O’Connor, M. J. (2015). Targeting the dna damage response in cancer. *Mol. Cell.* 60, 547–560. doi:10.1016/j.molcel.2015.10.040

O’Toole, E., Amory, J., Bremner, W., Page, S., Adamczyk, B., Lee, A., et al. (2007). Mer-104 tablets: A dose-ranging study of an oral formulation of a gonadotropin-releasing hormone antagonist, acyline. *Mol. Cancer Ther.* 6, B83.

Peyvandipour, A., Saberian, N., Shafi, A., Donato, M., and Draghici, S. (2018). A novel computational approach for drug repurposing using systems biology. *Bioinformatics* 34, 2817–2825. doi:10.1093/bioinformatics/bty133

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* 18, 41–58. doi:10.1038/nrd.2018.168

Ricken, R., Ulrich, S., Schlattmann, P., and Adli, M. (2017). Tranylcypromine in mind (part ii): Review of clinical pharmacology and meta-analysis of controlled studies in depression. *Eur. Neuropsychopharmacol.* 27, 714–731. doi:10.1016/j.euroneuro.2017.04.003

Risbridger, G. P., Davis, I. D., Birrell, S. N., and Tilley, W. D. (2010). Breast and prostate cancer: More similar than different. *Nat. Rev. Cancer* 10, 205–212. doi:10.1038/nrc2795

Saberian, N., Peyvandipour, A., Donato, M., Ansari, S., and Draghici, S. (2019). A new computational drug repurposing method using established disease–drug pair knowledge. *Bioinformatics* 35, 3672–3678. doi:10.1093/bioinformatics/btz156

Saleh-Gohari, N., and Helleday, T. (2004). Conservative homologous recombination preferentially repairs dna double-strand breaks in the s phase of the cell cycle in human cells. *Nucleic Acids Res.* 32, 3683–3688. doi:10.1093/nar/gkh703

Schuler, J., Falls, Z., Mangione, W., Hudson, M. L., Bruggemann, L., and Samudrala, R. (2022). Evaluating the performance of drug-repurposing technologies. *Drug Discov. Today* 27, 49–64. doi:10.1016/j.drudis.2021.08.002

Siddiqui, A., Tumiati, M., Joko, A., Sandholm, J., Roering, P., Aakko, S., et al. (2021). Targeting dna homologous repair proficiency with concomitant topoisomerase ii and c-abl inhibition. *Front. Oncol.* 11, 733700. doi:10.3389/fonc.2021.733700

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77. doi:10.1126/scitranslmed.3001318

Sofikerim, M., Eskicorapcı, S., Oruç, Ö., and Oezen, H. (2007). Hormonal predictors of prostate cancer. *Urol. Int.* 79, 13–18. doi:10.1159/000102906

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1, 000, 000 profiles. *Cell.* 171, 1437–1452. doi:10.1016/j.cell.2017.10.049

Sun, D., Gao, W., Hu, H., and Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* 12, 3049–3062. doi:10.1016/j.apsb.2022.02.002

Tang, Q., Ma, J., Sun, J., Yang, L., Yang, F., Zhang, W., et al. (2018). Genistein and ag1024 synergistically increase the radiosensitivity of prostate cancer cells. *Oncol. Rep.* 40, 579–588. doi:10.3892/or.2018.6468

Toh, M., and Ngeow, J. (2021). Homologous recombination deficiency: Cancer predispositions and treatment implications. *Oncologist* 26, e1526–e1537. doi:10.1002/onco.13829

Walker, A. J., Wedam, S., Amiri-Kordestani, L., Bloomquist, E., Tang, S., Sridhara, R., et al. (2016). Fda approval of palbociclib in combination with fulvestrant for the treatment of hormone receptor–positive, her2-negative metastatic breast cancer. *Clin. Cancer Res.* 22, 4968–4972. doi:10.1158/1078-0432.CCR-16-0493

Watkins, J. A., Irshad, S., Grigoriadis, A., and Tutt, A. N. (2014). Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. *Breast Cancer Res.* 16, 211–11. doi:10.1186/bcr3670

Xie, M., Hwang, T., and Kuang, R. (2012). "Prioritizing disease genes by bi-random walk," in *Pacific-asia conference on knowledge discovery and data mining* (Berlin, Germany: Springer), 292–303.

Zhang, Z., Wang, C.-Z., Du, G.-J., Qi, L.-W., Calway, T., He, T.-C., et al. (2013). Genistein induces g2/m cell cycle arrest and apoptosis via atm/p53-dependent pathway in human colon cancer cells. *Int. J. Oncol.* 43, 289–296. doi:10.3892/ijo.2013.1946

Zhou, K., Arslanturk, S., Craig, D. B., Heath, E., and Draghici, S. (2021). Discovery of primary prostate cancer biomarkers using cross cancer learning. *Sci. Rep.* 11, 10433. doi:10.1038/s41598-021-89789-x

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1065568

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU), United States

REVIEWED BY
Tianyuan Lu,
McGill University, Canada
Zhaoxia Yu,
University of California, Irvine,
United States

*CORRESPONDENCE
Brad McNeney,
✉ mcneney@sfu.ca

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

RECEIVED 09 October 2022
ACCEPTED 29 November 2022
PUBLISHED 04 January 2023

CITATION
Ratnasekera P, Graham J and McNeney B (2023), Inference of gene-environment interaction from heterogeneous case-parent trios. *Front. Genet.* 13:1065568. doi: 10.3389/fgene.2022.1065568

Inference of gene-environment interaction from heterogeneous case-parent trios

Pulindu Ratnasekera, Jinko Graham and Brad McNeney*

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

Introduction: In genetic epidemiology, log-linear models of population risk may be used to study the effect of genotypes and exposures on the relative risk of a disease. Such models may also include gene-environment interaction terms that allow the genotypes to modify the effect of the exposure, or equivalently, the exposure to modify the effect of genotypes on the relative risk. When a measured test locus is in linkage disequilibrium with an unmeasured causal locus, exposure-related genetic structure in the population can lead to spurious gene-environment interaction; that is, to apparent gene-environment interaction at the test locus in the absence of true gene-environment interaction at the causal locus. Exposure-related genetic structure occurs when the distributions of exposures and of haplotypes at the test and causal locus both differ across population strata. A case-parent trio design can protect inference of genetic main effects from confounding bias due to genetic structure in the population. Unfortunately, when the genetic structure is exposure-related, the protection against confounding bias for the genetic main effect does not extend to the gene-environment interaction term.

Methods: We show that current methods to reduce the bias in estimated gene-environment interactions from case-parent trio data can only account for simple population structure involving two strata. To fill this gap, we propose to directly accommodate multiple population strata by adjusting for genetic principal components (PCs).

Results and Discussion: Through simulations, we show that our PC adjustment maintains the nominal type-1 error rate and has nearly identical power to detect gene-environment interaction as an oracle approach based directly on population strata. We also apply the PC-adjustment approach to data from a study of genetic modifiers of cleft palate comprised primarily of case-parent trios of European and East Asian ancestry. Consistent with earlier analyses, our results suggest that the gene-environment interaction signal in these data is due to the self-reported European trios.

KEYWORDS

gene-environment interaction, case-parent trios, population structure, genome-wide association study, cleft palate, principal components

1 Introduction

We start by considering a log-linear model of population disease risk that includes main effects for genotypes G, environmental exposures E, and a gene-environment interaction term $G \times E$. The $G \times E$ term allows genotypes to modify the effect of the exposure or, equivalently, the exposure to modify the effect of genotypes on the relative risk of developing the disease. Including a $G \times E$ term can improve model accuracy and provide a more detailed picture of disease etiology compared to models with just G and E main effects (Hunter, 2005). $G \times E$ is also useful for identifying environmental exposures with greater disease-association in individuals who carry particular alleles at susceptibility loci (Thomas, 2010). For example, dietary fat intake is more highly associated with obesity in carriers than in non-carriers of the Pro12Ala allele in the PPAR-γ gene (Garaulet et al., 2011).

We suppose throughout that G is an unmeasured causal locus in linkage disequilibrium with a measured non-causal test locus G', and that the distribution of GG' haplotypes differs across population strata (i.e. genetic structure). Stratum-specific differences in the GG' haplotype frequencies can lead to differences in G' risk across the population strata where none exist for G (Zaykin and Shibata, 2008). Exposure-related genetic structure occurs when the distribution of E also differs across the population strata (Weinberg et al., 2011). Without some adjustment for the population strata, E will tag the stratum-specific differences in G' risk (Figure 1), suggesting that E modifies G' risk, even in the absence of $G \times E$ (Shi et al., 2011; Weinberg et al., 2011); we refer to this as spurious $G' \times E$.

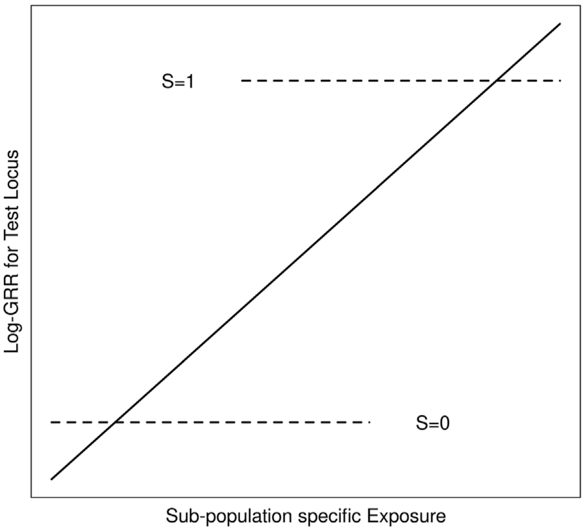


FIGURE 1
Schematic of log-GRRs for a non-causal test locus *versus* exposure in a structured population with two strata, S = 0 and S = 1. Dashed lines represent log-GRRs within each stratum. Horizontal positioning of these dashed lines indicates the support of the respective E distributions. High values of E are associated with S = 1, in which one of the alleles at the test locus is associated with increased disease risk. Low values of E are associated with S = 0 in which this same allele at the test locus is associated with low disease risk. Ignoring S yields the linear log-GRR curve indicated by the solid line, which erroneously suggests that E modifies the disease risk at the test locus.

A case-parent trio design can protect inference of genetic main effects from confounding bias due to genetic structure in the population (Weinberg, 1999). In this design, investigators collect information on G' and E in children affected with a disease of interest as well as the genotypes, G'_P, of their parents. To increase sample size, investigators may pool trios from multiple ancestral groups into one study; e.g., the GENEVA Oral Cleft Study (GENEVA, 2010) combined case-parent trios from recruitment sites in the United States, Europe and East Asia. Assuming G' and E are independent within families, a log-linear model of disease risk leads to a conditional likelihood for the G' and $G' \times E$ effects, based on the child's genotype given their exposure, affection status and parental genotypes (Shin et al., 2012). Unfortunately, when the genetic structure is exposure related, the protection against confounding bias for the genetic main effect does not extend to the gene-environment interaction term (Shi et al., 2011; Weinberg et al., 2011). Thus, spurious $G' \times E$ may be inferred from heterogeneous case-parent trio data in the absence of true $G \times E$.

Methods to mitigate this bias may be classified as design- or data-based. For a binary environmental exposure, the *design*-based tetrad approach of (Shi et al., 2011) augments the case-parent trio by adding the exposure of an unaffected sibling. These authors control the bias by including the sibship-averaged exposure in the log-linear model. They show that all information about the interaction in the tetrad design comes from the siblings, not the parents (Weinberg et al., 2011). Accordingly, they propose a sibling-augmented case-only design and analysis. By contrast, (Shin et al., 2012) takes a *data*-based approach, replacing the sibship-averaged exposure of (Shi et al., 2011) with the *predicted* exposure given ancestry. Predictions are obtained from a regression of exposure on principal components (PCs) computed from genetic markers that are unlinked to the test locus. This data-based approach may be applied to arbitrary exposures, including continuous exposures, and does not require siblings. However, its properties have not been evaluated in the case of more than two population strata.

We use the GENEVA Oral Cleft Study to motivate a new approach to unbiased inference of $G' \times E$ in case-parent trios. The analysis of (Beaty et al., 2011) found multiple single nucleotide polymorphisms (SNPs) that appeared to modify the effect of maternal smoking, maternal alcohol consumption or maternal multivitamin supplementation on the risk of cleft palate (CP). The self-reported ancestry of the study sample is primarily European or East Asian, and all three exposures are more common in self-reported Europeans than in self-reported East

TABLE 1 GG' **haplotype frequencies in four population strata.**

	Stratum			
GG'	$S = 0$	$S = 1$	$S = 2$	$S = 3$
R1	0.0	0.5	0.375	0.125
R0	0.5	0.0	0.125	0.375
N1	0.5	0.0	0.125	0.375
N0	0.0	0.5	0.375	0.125

Asians (Beaty et al., 2011, Table 2). If the frequencies of haplotypes spanning causal SNPs also vary by ancestral groups, exposure-related genetic structure may lead to spurious gene-environment interaction. (Ratnasekera and McNeney, 2021). focused on the self-reported Europeans and East Asians in the GENEVA Oral Cleft Study data. Applying the approach of (Shin et al., 2012), they confirmed the gene-environment interaction found by (Beaty et al., 2011), and concluded that the evidence for gene-environment interaction is predominantly from the data of self-reported Europeans. These authors also considered whether exposure-related genetic structure *within* self-reported Europeans could explain the apparent $G' \times E$. Their results were inconclusive, however, possibly owing to the methodology's limitation to just two ancestry groups. In modern datasets, the possibility of both inter- and intra-continental genetic structure necessitates methods that can more flexibly accommodate multiple ancestries. In this work we propose such an approach which relies on direct use of the genetic PCs to adjust for population structure.

The manuscript is structured as follows. In Section 2 we develop our direct PC-adjustment method and compare it to the indirect PC-based approach of (Shin et al., 2012). In Section 3 we present simulations to evaluate the statistical properties of both approaches. In Section 4 we re-analyze the GENEVA data. Section 5 includes a discussion and areas for future work.

2 Models and methods

2.1 Overview

We start with a log-linear model of disease risk parametrized in terms of genotype relative risks (GRRs) at a causal locus G. Under this model, G × E is equivalent to GRRs that depend on the exposure E. We then derive the GRRs at a non-causal test locus G' in linkage disequilibrium with G and show that, in the absence of G × E, the G'-GRRs can depend on E when there is dependence between E and GG' haplotypes in the population. Such dependence can lead to spurious inference of $G' \times E$ in the absence of G × E. However, valid inference is obtained if we adjust the risk model for any variable X for which E and GG' haplotypes are conditionally independent given X (Shin et al., 2012). We review the rationale for the adjustment used by (Shin et al., 2012) in this context, and propose an alternative adjustment based on inferred population structure. In particular, we use the method of (Gavish and Donoho, 2014) to select a parsimonious set of PCs with which to adjust the risk model. A key question is whether the PC-selection method yields a set of PCs that provide enough adjustment to maintain type 1 error in the absence of G × E, but not so much that we compromise power in the presence of G × E. The Models and Methods section concludes with a discussion of the simulation methods used to answer this question.

2.2 Risk model and likelihood

Let $G = 0$, 1 or 2 denote the number of copies of the variant allele at the causal locus and E denote the exposure variable. The disease-risk model of (Shin et al., 2012) can be obtained from a log-linear model of the GRRs

$$\begin{aligned}\log GRR_g(e) &= \log \frac{P(D=1|G=g,E=e)}{P(D=1|G=g-1,E=e)}\\ &= \beta_g + f_g(e) \quad \text{for } g = 1,2, \end{aligned} \tag{1}$$

and the log-disease risk for carriers of the baseline genotype $G = 0$

$$\log P(D=1|G=0,E=e) \equiv \eta(e).$$

The parameters β_g and $f_g(\cdot)$ are, respectively, genotype-specific main effects and functions that allow for $G \times E$ interaction. We can also write disease risk in terms of the baseline risk $\eta(e)$ and the GRRs as follows. First define $GRR_0(e) \equiv 1$. Next, note that

$$\frac{P(D=1|G=1,E=e)}{P(D=1|G=0,E=e)} = GRR_1(e) = GRR_1(e)GRR_0(e)$$

and

$$\begin{aligned}\frac{P(D=1|G=2,E=e)}{P(D=1|G=0,E=e)} &= \frac{P(D=1|G=2,E=e)}{P(D=1|G=1,E=e)}\frac{P(D=1|G=1,E=e)}{P(D=1|G=0,E=e)}\\ &= GRR_2(e)GRR_1(e)GRR_0(e).\end{aligned}$$

it follows that

$$P(D=1|G=g,E=e) = \eta(e)\prod_{i=0}^{g} GRR_i(e) \quad \text{for } g = 0, 1or2. \tag{2}$$

A likelihood for estimation of the GRR parameters β_g and $f_g(\cdot)$, $g = 1, 2$, from case-parent trio data can be derived under the assumption that G and E are conditionally independent given parental genotypes G_p. As shown in Supplementary Appendix S1, the likelihood is based on the conditional probability of the child's genotype given their exposure and parental genotypes. The function $\eta(\cdot)$ that parametrizes the environmental main effect drops out of the likelihood and cannot be estimated from case-parent trio data.

2.3 GRRs at a non-causal test locus

Let G' denote genotypes at a non-causal test locus in linkage disequilibrium with the causal locus G. We assume D and G' are conditionally independent given G and E, so that

$$P(D = 1|G = g, G' = g', E = e) = P(D = 1|G = g, E = e).$$

Therefore, the risk of disease given G' and E can be written as

$$\begin{aligned} &P(D = 1|G' = g', E = e) \\ &= \sum_{g=0}^{2} P(D = 1|G = g, E = e) P(G = g|G' = g', E = e). \end{aligned} \quad (3)$$

Eq. 3 is a latent-class model (Xu, 2017) with the unobserved causal locus G as the latent class having probabilities $P(G = g|G' = g', E = e)$. Eqs 2, 3 enable the log-GRRs at G' to be written in terms of the latent-class probabilities and the GRRs at G as follows:

$$\begin{aligned} \log GRR_{g'}(e) &\equiv \log \frac{P(D = 1|G' = g', E = e)}{P(D = 1|G' = g' - 1, E = e)} \\ &= \log \frac{\sum_{g=0}^{2} P(D = 1|G = g, E = e) P(G = g|G' = g', E = e)}{\sum_{g=0}^{2} P(D = 1|G = g, E = e) P(G = g|G' = g' - 1, E = e)} \\ &= \log \frac{\sum_{g=0}^{2} \left(\prod_{i=0}^{g} GRR_i(e)\right) P(G = g|G' = g', E = e)}{\sum_{g=0}^{2} \left(\prod_{i=0}^{g} GRR_i(e)\right) P(G = g|G' = g' - 1, E = e)}. \end{aligned} \quad (4)$$

Without $G \times E$, GRRs at G do not depend on E. Importantly, though, the log-GRRs at G' *can* depend on E through the latent-class probabilities $P(G = g|G' = g', E = e)$. In fact, as shown in Supplementary Appendix S2, these latent-class probabilities will depend on E whenever GG' haplotypes and E are associated, as happens when the population has exposure-related genetic structure. Since $G' \times E$ is equivalent to $GRR_{g'}$ varying with E, Eq. 4 gives insight into how exposure-related genetic structure creates spurious $G' \times E$.

2.4 Augmented risk model

The development so far has considered a disease-risk model that depends only on E and a causal locus G. We now consider an augmented disease-risk model that depends on E, G and a third variable X:

$$\begin{aligned} \log GRR_g(e, x) &\equiv \log \frac{P(D = 1|G = g, E = e, X = x)}{P(D = 1|G = g - 1, E = e, X = x)} \\ &= \beta_g + f_g(e, x) \quad \text{for } g = 1, 2, \end{aligned}$$

where β_g and $f_g(\cdot, x)$ are, respectively, genotype-specific main effects and functions that allow for $G \times E \times X$ interaction. Defining

$$GRR_0(e, x) \equiv 1,$$

an analogous development to Section 2.3 leads to the following X-adjusted log-GRRs at G':

$$\begin{aligned} \log GRR_{g'}(e, x) &\equiv \log \frac{P(D = 1|G' = g', E = e, X = x)}{P(D = 1|G' = g' - 1, E = e, X = x)} \\ &= \log \frac{\sum_{g=0}^{2} \left(\prod_{i=0}^{g} GRR_i(e, x)\right) P(G = g|G' = g', E = e, X = x)}{\sum_{g=0}^{2} \left(\prod_{i=0}^{g} GRR_i(e, x)\right) P(G = g|G' = g' - 1, E = e, X = x)}. \end{aligned} \quad (5)$$

In the next section we discuss choices for X that eliminate E from the latent-class probabilities for G, and hence eliminate spurious $G' \times E$ arising from exposure-related genetic structure.

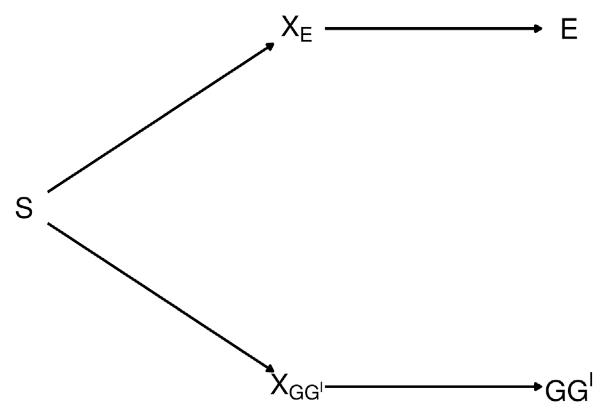


FIGURE 2
Diagram depicting exposure-related genetic structure. The latent population strata S induce dependence beetween E and GG'. Latent factors X_E and $X_{GG'}$ encode different distributions of E and GG', respectively. E and GG' are conditionally independent given any of the three variables that lie on the path between them.

2.5 Removing dependence of the latent-class probabilities on *E*

The diagram in Figure 2 depicts the dependence between GG' haplotypes and E from exposure-related genetic structure in the population. In the figure, S is a categorical variable that indicates population strata. The categorical variable X_E is a "coarsening" of S such that different levels of X_E correspond to different E distributions, and, similarly, $X_{GG'}$ is a coarsening of S such that different levels of $X_{GG'}$ correspond to different GG' haplotype distributions.

The path connecting E and GG' in Figure 2 is said to be *blocked* by each of the variables X_E, S and $X_{GG'}$ [(Pearl, 2009), Definition 1]. Therefore, E and GG' are conditionally independent given any of the blocking variables X_E, S or $X_{GG'}$ (Pearl, 1998). As shown in Supplementary Appendix S2, a consequence is that conditioning on any of these variables removes the dependence of the latent-class probabilities on E.

That is, letting X denote any of X_E, S or $X_{GG'}$, $P(G = g|G' = g', E = e, X = x) = P(G = g|G' = g', X = x)$. Consequently, from Eq. 5,

$$\log GRR_{g'}(e,x) \equiv \log \frac{P(D=1|G'=g',E=e,X=x)}{P(D=1|G'=g'-1,E=e,X=x)}$$
$$= \log \frac{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e,x)\right)P(G=g|G'=g',X=x)}{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e,x)\right)P(G=g|G'=g'-1,X=x)}. \quad (6)$$

GRRs at G' will thus depend on E if and only if GRRs at G do.

2.6 Linear model for the log GRRs

From Eq. 6 we see that, for fixed g' and x, log $GRR_{g'}(e, x)$ varies with e if and only if the $GRR_g(e, x)$ do. We can therefore test for $G \times E$ by fitting a model for log $GRR_{g'}(e, x)$ that allows separate curves in e for each combination of g' and x (Shin et al., 2014). We take these curves to be straight lines, and test whether any of them have non-zero slope. For a fixed value x of the adjustment variable X and a fixed value e of the environmental exposure E, the log-GRR is:

$$\log GRR_{g'}(e,x) = \beta_{g'} + \beta_{g'X}x + \beta_{g'E} \times e + \beta_{g'EX}x \times e;$$
$$g' = 1, 2. \quad (7)$$

The generalization of the above model to a vector X is to replace $\beta_{g'X}x$ with $\beta_{g'X}^T x$ and $\beta_{g'EX}$ with $\beta_{g'EX}^T x$ for coefficient vectors $\beta_{g'X}$ and $\beta_{g'EX}$. The intercepts of the log-GRR curves, $\beta_{g'}+\beta_{g'X}x$, are the genetic main effects in stratum x (i.e. when $e = 0$). The slopes, $\beta_{g'E}+\beta_{g'EX}x$, are the $G' \times E$ interaction terms in stratum x. We use a likelihood-ratio test of the null hypothesis that $\beta_{g'E} = \beta_{g'EX} = 0$ for $g' = 1, 2$, versus the alternative hypothesis that at least one of these slope parameters is non-zero to detect $G \times E$. We emphasize that the simplified log-GRR curves in e characterize $G \times E$ rather than environmental main effects, which are not estimable from case-parent trio data. Genetic main effects *are* estimable however and flexibly parametrized by the intercept terms of the log-GRR curves. The flexibility in the intercept terms avoids mis-specification of the genetic main effects which can lead to biased inference of interaction effects (Yu et al., 2015).

2.7 Choice of X

Following (Shi et al., 2011), (Shin et al., 2012) set X to be the categorical variable X_E that distinguishes E distributions among the genetic strata of the population. Since X_E is unobserved, (Shin et al., 2012), consider the expectation of E given genetic markers (EEGM) as a surrogate $\hat{X}_E$. The idea behind their EEGM approach is to distinguish exposure distributions by their mean, which may vary across genetic strata, S. Though S is not known, it is reflected in principal components (PCs), $\hat{S}$, computed from a set of genetic markers that are unlinked to G'. The expectation of E given $\hat{S}$ can be estimated by linear regression of E on $\hat{S}$ when E is continuous, or by logistic regression when E is binary. For EEGM adjustment, the expected exposure within genetic strata is estimated by $\hat{X}_E = \widehat{E(E|\hat{S})}$. (Shin et al., 2012). showed that EEGM adjustment works well where there are two population strata, but our simulation results (Section 3) indicate that it works poorly for more than two strata. We therefore propose to adjust for population strata directly; i.e., to take $X = S$. In particular, if the population has K+1 genetic strata, indexed 0, . . . , K, we let S denote a vector of K dummy variables that distinguish these strata such that the kth element $S_k = 1$ for trios in stratum $k > 0$ and 0 otherwise, for $k = 1, \ldots, K$.

2.8 Inferred population strata

The population stratum variable S reflects genetic ancestry and is not generally known. Since adjustment for self-reported ancestry can lead to bias (Wang et al., 2010) we use marker-based PCs, $\hat{S}$. An advantage of PC-adjustment is that it does not enforce discrete strata, and individuals whose PC values lie between those of clusters on the PC plot (e.g. admixed individuals) will have intermediate values of the slope and intercept of their log-GRR curve.

Standard PC adjustment in genetic association analyses relies on a relatively large set of PCs. For K PCs the degrees of freedom of the test for $G' \times E$ is equal to 2(K+1). Thus, using more PCs than are necessary reduces the power of the test for $G' \times E$. We seek methods to select a parsimonious set of PCs that provides enough adjustment to control type 1 error rate, without sacrificing power. We consider three PC-selection methods. The first (Zhu and Ghodsi, 2006) is an automated version of the graphical approach of looking for an "elbow" in the scree plot of variance explained by the PCs as a function of their number. The second (Gavish and Donoho, 2014) is an estimator of the rank of a matrix under a model in which the data matrix is a noisy version of a low-rank matrix. The third (Patterson et al., 2006) is to select PCs corresponding to eigenvalues that exceed a significance threshold determined from the distribution of the largest eigenvalue of an unstructured random matrix.

2.9 Simulation methods

2.9.1 Simulating G, G' and E on case-parent trios

To study the statistical properties of our proposed approach and compare it to the method of (Shin et al., 2012), we generated 5,000 data sets of 3,000 informative case-parent trios. Trios were sampled from one of four population strata labelled $S = 0, 1, 2$ or 3. We assumed random mating within and no mixing between strata. We performed some simulations using equal-sized strata and others using unequal-sized strata. In the case of unequal

TABLE 2 Estimated type 1 error rates (top entry) and corresponding 95% confidence intervals (bottom entry) when data are simulated from 2, 3 or 4 strata with equal (top three rows) or unequal (bottom three rows) stratum sizes.

Equal stratum sizes			
Number of strata			
Adjustment	2	3	4
S	0.0556	0.0524	0.0498
	(0.049, 0.062)	(0.046, 0.0586)	(0.044, 0.056)
EEGM	0.0538	1.0000	1.0000
	(0.048, 0.060)	NA	NA
PC	0.0546	0.0534	0.0496
	(0.048, 0.061)	(0.047, 0.060)	(0.044, 0.056)
Unequal stratum sizes			
	2	3	4
S	0.0524	0.0482	0.0536
	(0.046, 0.058)	(0.042 0.054)	(0.047,0.059)
EEGM	0.0536	1.0000	1.0000
	(0.047, 0.060)	NA	NA
PC	0.0540	0.0508	0.0527
	(0.048, 0.060)	(0.045, 0.057)	(0.046, 0.059)

stratum sizes, the split was 60%, 40% for two strata; 50%, 30% and 20% for three strata; and 40%, 30%, 20% and 10% for four strata.

For a given stratum, informative trios were simulated following the methods proposed by (Shin et al., 2013; Shin et al., 2014). Briefly, GG' haplotypes are first simulated on parents in a random-mating population according to the stratum-specific GG' haplotype distributions in Table 1. Child haplotypes are then simulated following Mendel's laws and assuming no recombination between G and G'. The child's exposure E is also simulated according to the stratum-specific distributions described below. Finally, the child's disease status is simulated based on the disease-risk model (1). Trios with an affected child and at least one heterozygous parent at the test locus are retained. The data recorded on each trio are G_p', G', and E, where G_p' is the pair of parental genotypes at the test locus.

Spurious $G' \times E$ is induced by specifying different distributions of E and GG' haplotypes in the four strata of Table 1. The GG' distributions for strata $S = 0$ and $S = 1$ are as in (Shin et al., 2012). Alleles at G are denoted R (risk) and N (non-risk), while alleles at G' are denoted 1 and 0. We summarize the haplotype distributions by the implied allelic correlations between the index alleles R and 1. Under the GG' haplotype frequencies given in Table 1, these correlations are $r_0 = -1$ in stratum $S = 0$, $r_1 = 1$ in stratum $S = 1$, $r_2 = 0.5$ in stratum $S = 2$ and $r_3 = -0.5$ in stratum $S = 3$.

The stratum-specific distributions of E are chosen to be normal with common variance $\sigma^2 = 0.36$, and means $\mu_0 = -0.8$, $\mu_1 = 0.8$, $\mu_2 = 2.4$ and $\mu_3 = 4.0$ in strata 0, 1, 2 and 3, respectively. The E distributions for strata $S = 0$ and $S = 1$ are as in (Shin et al., 2012).

The disease-risk model is specified as follows. The genetic main effect is $\beta_g = \log(3)/2$ for $g = 1, 2$, corresponding to a $\sqrt{3}$-fold increase in relative risk for each copy of the risk allele (R) in the absence of $G \times E$. To evaluate the type 1 error rate of the $G \times E$ test we set $f_g(e) = 0$ in our simulations. To investigate power we choose a linear interaction model for the $G \times E$ term, setting $f_g(e) = \beta_{gE} e$ with $\beta_{gE} = -0.10, -0.15, -0.20$ or -0.25.

2.9.2 Simulating markers for PC adjustment

A standard method of PC adjustment is to calculate PCs from a genomic region that is unlinked to the test locus. It is recommended that markers in this region be thinned, or LD pruned, to have pairwise correlations of $r^2 \le 0.1$ (Grinde, 2019). We simulated such panels of markers based on data from the 1,000 genomes project (Clarke et al., 2016) using two East Asian (Chinese Dai in Xishuangbanna, China [CDX] and Han Chinese in Bejing China [CHB]) and two European (Iberian population in Spain [IBS] and Finnish in Finland [FIN]) populations. From the initial download of the genome-wide data, we retained 6,929,035 diallelic, autosomal markers with minor allele frequency (MAF) 0.05 or greater in all four of the population groups.

Our initial approach to simulating markers for a given population stratum was to fit a hidden Markov model (HMM) to the haplotypes in that stratum, chromosome by chromosome, using fastPHASE (Scheet and Stephens, 2006), and use this fitted model to simulate individual multilocus genotypes using SNPknock (Sesia et al., 2019). The simulated data are then LD pruned and principal components are computed from the thinned panel of markers. However, the computation involved in this approach proved to be prohibitive. For example, fitting the HMMs took up to 5 h per chromosome. We therefore considered two computationally cheaper alternatives. In the first alternative, we started from an LD-pruned set of markers in the original data and fit HMMs to this set. In the second alternative, we used the same panel of pruned markers, but simulated genotypes *independently* based on the MAFs in the population strata. In what follows we refer to the first and second alternatives as *LD-based* and *independent* marker simulation, respectively.

Independent markers could contain more information about the population strata than markers in LD. As a result, PC adjustment with independent markers might control type 1 error more effectively than adjustment with markers in LD. To assess this possibility, we completed 100 preliminary simulation replicates using LD-based marker simulation and 5,000 replicates using independent marker simulation. We simulated trios from four population strata under the null hypothesis of no $G \times E$, used the PC selection method of (Gavish and Donoho, 2014) to adjust the risk model and estimated the resulting type 1 error rates. Estimated type 1 error rates and their 95% confidence intervals under the LD-based and independent simulation methods were 0.04 (0.002, 0.078) and 0.0496 (0.044, 0.056), respectively, and consistent with similar type 1 error rates for the two approaches. We therefore used the faster simulation of independent markers for the simulation study.

In Section 3.2, Section 3.3 we present type I error and power results for two, three or four population strata. For two strata ($S = 0$ and $S = 1$), marker simulations were based on the CHB and IBS population groups. For three strata ($S = 0$, $S = 1$ and $S = 2$), simulations were based on the CHB, IBS and CDX population groups.

3 Results

3.1 Selection of principal components

All PC selection methods performed well when the sizes of the population strata were equal (results not shown), but not when the sizes were unequal. We illustrate with simulation results involving datasets of 3,000 trios sampled from four unequal-sized strata. For $K+1 = 4$ populations we require $K = 3$ PCs. In 5,000 simulation replicates, the method of (Gavish and Donoho, 2014) always selected three, the method of (Zhu and Ghodsi, 2006) always selected one, and the method of (Patterson et al., 2006) selected three PCs 4,942 times and four PCs 58 times. Other simulation results with unequal-sized strata (not shown) yielded similar results. Therefore, in what follows we use the method of (Gavish and Donoho, 2014) to select PCs.

3.2 Type I error rate

We compared the type I error rates of the test for $G' \times E$ using (i) adjustment with the true stratum membership S, (ii) the EEGM adjustment of (Shin et al., 2012), and (iii) PC adjustment. Results for simulated datasets with equal or unequal stratum sizes are shown in Table 2. For both equal and unequal stratum sizes, adjustment by S or direct PCs maintains the nominal 5% error rate regardless of the number of strata. By contrast, EEGM adjustment leads to an inflated type I error rate when there are more than two strata. In light of the inflated size of the test, we do not consider EEGM adjustment in the following section on power.

3.3 Power

Table 3 provides a comparison of estimated power when data are simulated from two, three or four strata. Results are shown for simulations using both equal and unequal stratum sizes and for different values of the $G \times E$ effect. From these results we see that power increases with effect size, decreases with number of strata and tends to be slightly larger for unequal strata than equal strata. Importantly, the estimated power under PC adjustment is always within simulation error of that under adjustment for true stratum membership.

4 The GENEVA Oral Cleft study

4.1 Data and objectives

The GENEVA Oral Cleft study (GENEVA, 2010) is comprised of 550 case-parent trios from 13 different sites across the United States, Europe, Southeast and East Asia. Data were obtained through dbGAP at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1 with accession number phs000094.v1.p1. Of the 550 trios, only 462 were available for analysis. Summaries of the trios by ancestry and gender of the affected child are shown in Table 4. From this table we see the ancestry of the sample is predominantly European (46%) and East Asian (51%).

The objective of the GENEVA study is to discover genetic contributions to orofacial clefts, the most common type of craniofacial birth defect in humans, and to assess whether these genes modify the effect of exposures known to be associated with cleft palate. Maternal exposure to multivitamins, alcohol and smoking were assessed through maternal interviews focused on the peri-conceptual period (3 months prior to conception through the first trimester), which includes the first 8–9 weeks of gestation when palatal development is completed. Exposure status is summarized in Table 5. From this table we see that the three dichotomous exposures are all more common in Europeans. In contrast to the continuous exposures of the simulation study, the exposures we consider in the GENEVA study are all dichotomous.

4.2 GENEVA data analysis

4.2.1 PC selection

LD pruning of the genome-wide panel of SNPs at an r^2 threshold of 0.1 yielded 63,694 markers. In a principal component analysis of these markers, the first PC explains 6.3% of the total variance and all others explain less than 0.4%. Not surprisingly, the method of (Gavish and Donoho, 2014) selects one PC. A plot of the projections of the data onto the first two PCs is shown in Figure 3, with points colored by self-reported ancestry. Each PC has been shifted by subtracting the minimum value and scaled by the range so that the values are between zero and one. The first PC distinguishes those with self-reported East Asian ancestry from those with self-

TABLE 3 Estimated power (top entry) and corresponding 95% confidence intervals (bottom entry) of different adjustment schemes for different $G \times E$ interaction effects β_{gE}, number of strata and stratum-size distributions.

Equal stratum sizes					
		β_{gE}			
Num. Strata	Adjustment	−0.10	−0.15	−0.20	−0.25
2	S	0.2602	0.5660	0.8420	0.9558
		(0.248, 0.272)	(0.552, 0.580)	(0.832, 0.852)	(0.950, 0.961)
	PC	0.2580	0.5660	0.8404	0.9564
		(0.246, 0.270)	(0.552, 0.580)	(0.830, 0.850)	(0.951, 0.962)
3	S	0.1742	0.3844	0.6498	0.8288
		(0.164, 0.185)	(0.371, 0.398)	(0.636, 0.663)	(0.818, 0.839)
	PC	0.1788	0.3920	0.6616	0.8316
		(0.168, 0.189)	(0.378, 0.406)	(0.648, 0.675)	(0.821, 0.842)
4	S	0.1306	0.2766	0.5010	0.6970
		(0.121, 0.140)	(0.264, 0.289)	(0.487, 0.515)	(0.684, 0.710)
	PC	0.1396	0.2936	0.5088	0.6918
		(0.130, 0.149)	(0.281, 0.306)	(0.495, 0.523)	(0.679, 0.704)
Unequal stratum sizes					
		β_{gE}			
		−0.10	−0.15	−0.20	−0.25
2	S	0.2636	0.5724	0.8328	0.9518
		(0.251, 0.276)	(0.559, 0.586)	(0.822, 0.843)	(0.946, 0.958)
	PC	0.2648	0.5722	0.8322	0.9514
		(0.252, 0.277)	(0.558, 0.586)	(0.822, 0.842)	(0.945, 0.957)
3	S	0.1950	0.4322	0.7082	0.8640
		(0.184, 0.206)	(0.418, 0.446)	(0.696, 0.721)	(0.854, 0.874)
	PC	0.1936	0.4334	0.7054	0.8632
		(0.183, 0.204)	(0.420, 0.447)	(0.693, 0.718)	(0.854, 0.873)
4	S	0.1614	0.3470	0.6028	0.7894
		(0.151, 0.172)	(0.334, 0.360)	(0.589, 0.616)	(0.778, 0.801)
	PC	0.1598	0.3380	0.5872	0.7820
		(0.150, 0.170)	(0.325, 0.351)	(0.574, 0.601)	(0.770, 0.794)

reported European ancestry; hence, a value near zero corresponds to a hypothetical East Asian and a value near one corresponds to a hypothetical European. The second PC separates the single self-reported African child from all others.

4.2.2 Inference of $G \times E$

The conditional-likelihood methods outlined in Supplementary Appendix S1 were applied to the data. We focused on inference of $G \times E$ between maternal alcohol consumption and the six SNPs in the *MLLT3* gene that had

TABLE 4 Gender of 462 affected children by self-reported ancestry.

Ancestry	Males	Females	Total	%
European	103	111	214	46%
Asian	93	141	234	51%
Other/Afr	3	11	14	3%
Total	199	263	462	100%

TABLE 5 Exposure rates for maternal alcohol consumption, maternal smoking and maternal vitamin supplementation by self-reported ancestry in affected trios.

	Percent exposed to Maternal			
Ancestry	Alcohol	Smoking	Vitamin Supp	Affected children
European	41%	28%	57%	214
East Asian	4%	3%	21%	234
Other/Afr	14%	7%	71%	14
Total	21%	14%	39%	462

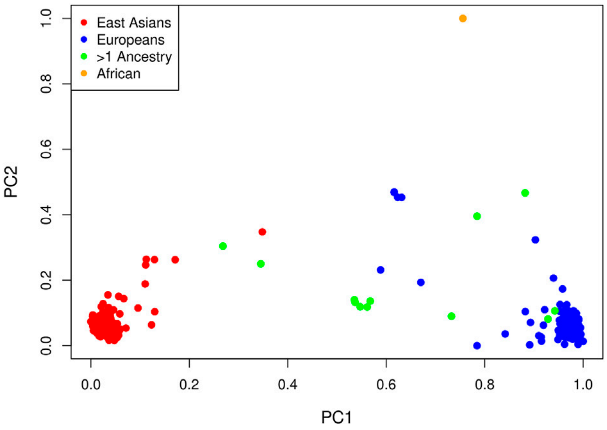


FIGURE 3
Projections of each affected child onto the first two PCs by self-reported ancestry: red = East Asian (234 trios), blue = European (214 trios), orange = African (one trio) and green = multiple ancestry/other (13 trios). Each PC has been shifted and scaled so that a PC1 value near zero corresponds to a hypothetical East Asian and a PC1 value near one corresponds to a hypothetical European.

significant $G \times E$ at the 5% level in the analysis of (Beaty et al., 2011). Displays of the LD between these SNPs and others nearby (Shin et al., 2006) are shown in Supplementary Figure S1, Supplementary Appendix S3, for self-reported European subjects and self-reported East Asian subjects. Table 6 shows the results of fitting three different log-linear models of $G' \times E$. Following (Beaty et al., 2011), each is based on an additive genetic model that specifies equal log-GRRs for genotypes $g' = 1$ or 2. Results based on fitting a more general co-dominant model (1) were similar (results not shown). The first model, as in (Beaty et al., 2011), makes no adjustment for exposure-related genetic structure in the population, the second uses EEGM adjustment and the third uses PC adjustment. From the table we see that, for each test SNP, p-values for the tests of $G' \times E$ are smallest when we make no adjustment. Comparing the EEGM and PC adjustment approaches we find that p-values from PC adjustment are similar to, but tend to be slightly smaller than, those from the EEGM adjustment. Of the six test SNPs show in the table, four retain significance at the 5% level after adjustment for exposure-related genetic structure.

The estimates shown in Table 6 are of the multiplicative factors by which maternal alcohol consumption modifies the GRRs at the six test SNPs. For a binary exposure such as maternal alcohol consumption, these modifying effects can be obtained by exponentiating the interaction term in the log-GRR model. With no adjustment for genetic structure there is a single interaction term and hence a single estimated modifying effect for all trios. For example, maternal alcohol consumption is estimated to increase the GRR at SNP rs4621895 by a factor of about 2.1 for all trios. By contrast, with EEGM or PC adjustment the interaction term depends on the value of the adjustment variable and we have reported estimates for hypothetical East Asian and European subjects in our sample. For example, maternal alcohol consumption is estimated to decrease the GRR at SNP rs4621895 by a factor of about 0.73 for East Asian trios and to increase the same GRR by a factor of about 2.4 for European trios. For these data, the adjustment variables used in the EEGM- and PC-adjustment approaches are highly correlated (Pearson correlation 0.996), and so the estimates for the two approaches are very similar. These estimates are also similar to those obtained from an analysis using self-reported ancestry (results not shown). The 95% confidence intervals for hypothetical East Asians cover one for each SNP but do not cover one for hypothetical Europeans, with the exception of SNP rs2780841. These results suggest that any $G \times E$ signal is from trios of European ancestry, where maternal alcohol consumption is more common.

5 Discussion

We consider a log-linear model of GRRs at a causal locus G. Under this model, $G \times E$ is equivalent to GRRs that vary with the exposure E. We show that exposure-related genetic structure in the population can lead to spurious $G' \times E$ at a non-causal test locus G' in LD with G. However, valid inference of $G' \times E$ can be obtained by augmenting the GRR model with a blocking variable X, such that GG' haplotypes and E are conditionally independent given X. We discuss the choice of X for inference of

TABLE 6 Estimated modifying effects of maternal alcohol consumption on GRRs, 95% confidence intervals and *p*-values from the analysis of the GENEVA data, at six SNPs in the MLLT3 gene (Chr 9) showing significant interaction with maternal alcohol consumption in (Beaty et al., 2011). Estimates, confidence intervals and tests are based on fitting an additive genetic model and use (i) no adjustment, (ii) EEGM adjustment or (iii) PC adjustment to control for exposure-related genetic structure in the population. The unadjusted analysis considers all trios without regard to genetic structure. The EEGM- and PC-adjusted analyses allow for genetic structure and we have reported estimates for hypothetical East Asian and European subjects.

	All			East Asian		European		
SNP	Adj	Est	95% CI	Est	95% CI	Est	95% CI	*p*-value
rs4621895	None	2.08	(1.36, 3.18)	–	–	–	–	0.0006
	EEGM	–	–	0.762	(0.214, 2.72)	2.44	(1.42, 4.20)	0.0047
	PC	–	–	0.701	(0.181, 2.72)	2.40	(1.42, 4.04)	0.0037
rs4977433	None	2.15	(1.40, 3.30)	–	–	–	–	0.0003
	EEGM	–	–	0.916	(0.244, 3.44)	2.47	(1.44, 4.25)	0.0036
	PC	–	–	0.854	(0.208, 3.45)	2.44	(1.45, 4.11)	0.0028
rs6475464	None	1.75	(1.13, 2.69)	–	–	–	–	0.0104
	EEGM	–	–	0.909	(0.271, 3.05)	2.25	(1.29, 3.95)	0.0158
	PC	–	–	0.840	(0.234, 3.02)	2.22	(1.29, 3.81)	0.0139
rs668703	None	2.02	(1.33, 3.07)	–	–	–	–	0.0008
	EEGM	–	–	0.588	(0.177, 1.95)	2.50	(1.45, 4.29)	0.0032
	PC	–	–	0.531	(0.148, 1.91)	2.43	(1.44, 4.09)	0.0025
rs623828	None	1.55	(1.00, 2.39)	–	–	–	–	0.0481
	EEGM	–	–	0.772	(0.239, 2.50)	1.77	(1.01, 3.11)	0.1368
	PC	–	–	0.757	(0.220, 2.60)	1.73	(1.00, 2.98)	0.1384
rs2780841	None	1.55	(1.01, 2.36)	–	–	–	–	0.0417
	EEGM	–	–	0.653	(0.217, 1.96)	1.71	(0.960, 3.04)	0.1613
	PC	–	–	0.620	(0.195, 1.97)	1.68	(0.965, 2.93)	0.1471

$G' \times E$ when data are collected from a study of case-parent trios. The population strata S would be an ideal choice for X but may not be known definitively. We propose to use principal components (PCs) instead. In particular, we calculate PCs from a genomic region unlinked to the test locus and select a parsimonious subset using the method of (Gavish and Donoho, 2014). We then specify a linear model for the log-GRRs whose intercept and slope depend on PC values. Slopes that vary with PC values allow the modifying effect of the exposure to vary with population strata, which can be important for maintaining power [20, Section 3.3]. Through simulations, we show that our PC adjustment maintains the nominal type-1 error rate and has nearly identical power to detect $G \times E$ as an oracle approach based directly on S. We illustrate our approach by applying it to an analysis of real data from case-parent trios in the GENEVA Oral Cleft Study. In our analysis of the GENEVA data, we focussed on SNPs and exposures identified by (Beaty et al., 2011). In a discussion of their results, these authors noted that the SNPs they identified are not in known cleft-palate susceptibility genes and are either intronic or are upstream/downstream of coding regions. This lack of compelling biological plausibility, coupled with the striking differences in exposure distributions between the self-reported European and East Asian strata, motivated our $G \times E$ analysis that adjusts for population structure. However, our results (Table 6) and those of (Ratnasekera and McNeney, 2021) do not contradict the hypothesis of $G \times E$, but rather suggest that any $G \times E$ signal is due to the self-reported European trios. Further data collection aimed at self-reported European trios may provide stronger conclusions regarding the presence of $G \times E$.

To reduce bias from exposure-related genetic structure, direct PC adjustment has advantages over the EEGM approach and design-based strategies such as the tetrad approach of (Shi et al., 2011) and the sibling-augmented case-only approach of (Weinberg et al., 2011). Unlike the EEGM approach, PC adjustment controls the type 1 error when there are more than two population strata. Unlike the design-based strategies, PC adjustment does not require siblings nor assume binary exposures.

Development of alternative approaches based on propensity scores is an area for future work. The EEGM approach is attractive in that it reduces the genetic principal components to a single score, $E(E|\hat{S})$. For binary exposures, such as those in the GENEVA study, the EEGM is a propensity score (Rosenbaum and Rubin, 1983). For continuous exposures, such as those in the simulation study, the analog to the EEGM is a continuous-treatment propensity score (Brown et al., 2021). With continuous exposures, we could predict E given the genetic markers and *then* convert the predictions to a Normal density score that takes low values for predictions far from their observed value. These density scores could be used either as predictors (Hirano and Imbens, 2004) or weights (Robins et al., 2000) in subsequent analyses. It would be interesting to explore the use of propensity-score methods in inference of $G' \times E$ from case-parent trios with continuous exposures, particularly when there are more than two population strata.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1

Author contributions

PR developed the statistical methods, performed the simulations and data analyses, and wrote the initial draft of the manuscript. BM and JG conceptualized the study and revised the manuscript. All authors proofread and approved the final version of the manuscript.

Funding

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), grants RGPIN/04296-2018 (JG) and RGPIN/05595-2019 (BM). This research was enabled in part by support provided by the BC DRI Group and the Digital Research Alliance of Canada (http://alliancecan.ca). Funding support for the study entitled “International Consortium to Identify Genes and Interactions Controlling Oral Clefts” was provided by several previous grants from the National Institute of Dental and Craniofacial Research (NIDCR). Data and samples were drawn from several studies awarded to members of this consortium. Funding to support original data collection, previous genotyping and analysis came from several sources to individual investigators. Funding for individual investigators include: R21-DE-013707 and R01-DE-014581 (Beaty); R37-DE-08559 and P50-DE-016215 (Murray, Marazita) and the Iowa Comprehensive Program to Investigate Craniofacial and Dental Anomalies (Murray); R01-DE-09886 (Marazita), R01-DE-012472 (Marazita), R01-DE-014677 (Marazita), R01-DE-016148 (Marazita), R21-DE-016930 (Marazita); R01-DE-013939 (Scott). Parts of this research were supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Wilcox, Lie). Additional recruitment was supported by the Smile Train Foundation for recruitment in China (Jabs, Beaty, Shi) and a grant from the Korean government (Jee). The genome-wide association study, also known as the Cleft Consortium, is part of the Gene Environment Association Studies (GENEVA) program of the trans-NIH Genes, Environment and Health Initiative (GEI) supported by U01-DE-018993. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University, contract number HHSN268200782096C. Funds for genotyping were provided by the NIDCR through CIDR’s NIH contract. Assistance with genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01-HG-004446) and by the National Center for Biotechnology Information (NCBI).

Acknowledgments

We sincerely thank all of the patients and families at each recruitment site for participating in this study, and we gratefully acknowledge the invaluable assistance of clinical, field and laboratory staff who contributed to this effort over the years.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1065568/full#supplementary-material

References

Beaty, T. H., Ruczinski, I., Murray, J. C., Marazita, M. L., Munger, R. G., Hetmanski, J. B., et al. (2011). Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genet. Epidemiol.* 35, 469–478. doi:10.1002/gepi.20595

Brown, D. W., Greene, T. J., Swartz, M. D., Wilkinson, A. V., and DeSantis, S. M. (2021). Propensity score stratification methods for continuous treatments. *Stat. Med.* 40, 1189–1203. doi:10.1002/sim.8835

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., et al. (2016). The international genome sample resource (igsr): A worldwide collection of genome variation incorporating the 1000 genomes project data. *Nucleic Acids Res.* 45, D854–D859. doi:10.1093/nar/gkw829

Garaulet, M., Smith, C. E., Hernández-González, T., Lee, Y. C., and Ordovás, J. M. (2011). PPARγ Pro12Ala interacts with fat intake for obesity and weight loss in a behavioural treatment based on the Mediterranean diet. *Mol. Nutr. Food Res.* 55, 1771–1779. doi:10.1002/mnfr.201100437

Gavish, M., and Donoho, D. L. (2014). The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory* 60, 5040–5053. doi:10.1109/TIT.2014.2323359

GENEVA (2010). *GENEVA oral clefts project imputation report - HapMap III reference panel [pdf file]*. [Dataset]. Available at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1.

Grinde, K. (2019). *Statistical inference in admixed populations*. Ph.D. thesis. University of Washington.

Hirano, K., and Imbens, G. W. (2004). *The propensity score with continuous treatments*. John Wiley & Sons, 73–84. chap. 7. doi:10.1002/0470090456.ch7

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nat. Rev. Genet.* 6, 287–298. doi:10.1038/nrg1578

Patterson, N., Prince, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 12, e190–e2093. doi:10.1371/journal.pgen.0020190

Pearl, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* 3, 96–146. doi:10.1214/09-SS057

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociol. Methods & Res.* 27, 226–284. doi:10.1177/0049124198027002004

Ratnasekera, P., and McNeney, B. (2021). Re-Analysis of a genome-wide gene-by-environment interaction study of case parent trios, adjusted for population stratification. *Front. Genet.* 11, 600232. doi:10.3389/fgene.2020.600232

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 11, 550–560. doi:10.1097/00001648-200009000-00011

Rosenbaum, P. R., and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects," in *Matched sampling for causal effects* (Cambridge University Press), 170–184. doi:10.1017/cbo9780511810725.016

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi:10.1086/502802

Sesia, M., Sabatti, C., and Candès, E. (2019). Gene hunting with hidden markov model knockoffs. *Biometrika* 106, 1–18. doi:10.1093/biomet/asy033

Shi, M., Umbach, D. M., and Weinberg, C. R. (2011). Family-based gene-by-environment interaction studies: Revelations and remedies. *Epidemiology* 22, 400–407. doi:10.1097/ede.0b013e318212fec6

Shin, J.-H., Blay, S., McNeney, B., and Graham, J. (2006). Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* 16. Code Snippet 3. doi:10.18637/jss.v016.c03

Shin, J.-H., Infante-Rivard, C., Graham, J., and McNeney, B. (2012). Adjusting for spurious gene-by-environment interaction using case-parent triads. *Stat. Appl. Genet. Mol. Biol.* 11, 1714. doi:10.2202/1544-6115.1714

Shin, J.-H., McNeney, B., and Graham, J. (2013). trioGxE: A data smoothing approach to explore and test gene-environment interaction in case-parent trio data. R package version 0.1-1

Shin, J. H., Infante-Rivard, C., McNeney, B., and Graham, J. (2014). A data-smoothing approach to explore and test gene-environment interaction in case-parent trios. *Stat. Appl. Genet. Mol. Biol.* 13, 159–171. doi:10.1515/sagmb-2013-0023

Thomas, D. (2010). Gene–environment-wide association studies: Emerging approaches. *Nat. Rev. Genet.* 11, 259–272. doi:10.1038/nrg2764

Wang, H., Haiman, C. A., Kolonel, L. N., Henderson, B. E., Wilkens, L. R., Le Marchand, L., et al. (2010). Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Hum. Genet.* 128, 165–177. doi:10.1007/s00439-010-0841-4

Weinberg, C. R. (1999). Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.* 65, 229–235. doi:10.1086/302466

Weinberg, C. R., Shi, M., and Umbach, D. M. (2011). A sibling-augmented case-only approach for assessing multiplicative gene-environment interactions. *Am. J. Epidemiol.* 174, 1183–1189. doi:10.1093/aje/kwr231

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Stat.* 45, 675–707. doi:10.1214/16-aos1464

Yu, Z., Demetriou, M., and Gillen, D. L. (2015). Genome-wide analysis of gene-gene and gene-environment interactions using closed-form wald tests. *Genet. Epidemiol.* 39, 446–455. doi:10.1002/gepi.21907

Zaykin, D. V., and Shibata, K. (2008). Genetic flip-flop without an accompanying change in linkage disequilibrium. *Am. J. Hum. Genet.* 82, 794–796. doi:10.1016/j.ajhg.2008.02.001

Zhu, M., and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Statistics Data Analysis* 51, 918–930. doi:10.1016/j.csda.2005.09.010

TYPE Methods
PUBLISHED 10 January 2023
DOI 10.3389/fgene.2022.1000290

Simultaneous detection of *G6PD* mutations using SNPscan in a multiethnic minority area of Southwestern China

Huagui Wei[1†], Chunfang Wang[1†], Weiyi Huang[1], Liqiao He[1], Yaqun Liu[2], Huiying Huang[2], Wencheng Chen[1], Yuzhong Zheng[2], Guidan Xu[1], Liyun Lin[2], Wujun Wei[1], Weizhong Chen[3], Liying Chen[1], Junli Wang [4*] and Min Lin [1,2,4*]

[1]Center for Clinical Laboratory Diagnosis and Research, The Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, China, [2]School of Biotechnology and Food Engineering, Hanshan Normal University, Chaozhou, China, [3]Department of Medical Laboratory, Chaozhou People's Hospital Affiliated to Shantou University Medical College, Chaozhou, China, [4]School of Laboratory Medicine, Youjiang Medical University for Nationalities, Baise, China

OPEN ACCESS

EDITED BY
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

REVIEWED BY
Germana Bancone,
Centre for Tropical Medicine and Global Health, University of Oxford,
United Kingdom
Richard Oscar Francis,
Columbia University, United States
Rajendranath Ramasawmy,
Universidade Nilton Lins, Brazil

*CORRESPONDENCE
Junli Wang,
✉ baisewangjunli@163.com
Min Lin,
✉ konfutea@hotmail.com

†These authors have contributed equally to this work

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

RECEIVED 22 July 2022
ACCEPTED 21 December 2022
PUBLISHED 10 January 2023

CITATION
Wei H, Wang C, Huang W, He L, Liu Y, Huang H, Chen W, Zheng Y, Xu G, Lin L, Wei W, Chen W, Chen L, Wang J and Lin M (2023), Simultaneous detection of *G6PD* mutations using SNPscan in a multiethnic minority area of Southwestern China.
Front. Genet. 13:1000290.
doi: 10.3389/fgene.2022.1000290

Objectives: Baise, a multiethnic inhabited area of southwestern China, is a historical malaria-endemic area with a high prevalence of *G6PD* deficiency. However, few studies of *G6PD* deficiency have been conducted in this region. Therefore, we performed a genetic analysis of *G6PD* deficiency in the Baise population from January 2020 to June 2021.

Methods: A SNPscan assay was developed to simultaneously detect 33 common Chinese *G6PD* mutations. 30 *G6PD*-deficient samples were used for the method's validation. Then, a total of 709 suspected *G6PD*-deficient samples collated from the Baise population were evaluated for *G6PD* status, type of mutation and effect of mutations.

Results: The SNPscan test had a sensitivity of 100% [95% confidence interval (CI): 94.87%–100%] and a specificity of 100% (95% CI: 87.66%–100%) for identifying *G6PD* mutations. A total of fifteen mutations were identified from 76.72% (544/709) of the samples. The most common mutation was discovered to be *G6PD* Kaiping (24.12%), followed by *G6PD* Canton (17.91%), and *G6PD* Gaohe (11.28%). We compared the *G6PD* mutation spectrum among Zhuang, Han and other Southeast Asian populations, and the Zhuang population's mutation distribution was quite similar to that in the Han population.

Conclusion: This study provided a detailed *G6PD* mutation spectrum in Baise of southwestern China and will be valuable for the diagnosis and research of *G6PD* deficiency in this area. Furthermore, the SNPscan assay could be used to quickly diagnose these *G6PD* mutations accurately.

KEYWORDS

G6PD deficiency, mutation spectrum, southwestern China, SNPscan assay, *G6PD* genotype

1 Introduction

Glucose-6-phosphate dehydrogenase (*G6PD*) deficiency is one of the most common enzymatic disorders of red blood cells, with a particularly high prevalence in tropical and subtropical regions, including southern China (Howell, 2006). According to the degree and extent of the enzyme deficiency, the World Health Organization (WHO) divided *G6PD* variants into four classifications in homozygous and hemizygous individuals (WHO, 2022). G6PD insufficiency manifests clinically as a range of conditions, ranging from severe enzyme deficiency to enhanced enzyme activity (Filosa et al., 1996). The most frequent clinical symptoms in patients are acute hemolysis, newborn hyperbilirubinemia, and chronic hemolysis, which are brought on by external factors including eating fava beans, taking specific medications, contracting an infection, or having a metabolic disorder (Jiang et al., 2006).

The *G6PD* gene (OMIM ID: 305900) spans 18 kb on the X chromosome (Xq28), contains an open reading frame of 1,545 bp, and encodes 515 amino acids (Tian et al., 2013; Wisnumurti et al., 2019). To date, approximately 217 mutations have been described worldwide (Gómez-Manzo et al., 2016). The *G6PD* mutation spectrum varies between different regions and ethnicities. The frequency distribution of these mutations closely correlates with populations that were exposed historically to endemic malaria (Dombrowski et al., 2017). Baise is a multiethnic inhabited area of southwestern China. The minority population accounts for 85% of the total population. It has a monsoon-influenced, humid subtropical climate and is a historical malaria-endemic area (Ji-Guang et al., 2017; Liang et al., 2020; Zheng et al., 2020). The spectrum of *G6PD* mutations, however, is poorly understood.

Currently, several analytical methods have been validated and developed to detect *G6PD* mutations, such as direct sequencing (Maloukh et al., 2021), reverse dot blot (RDB) assays (Chen et al., 2012; Duan et al., 2017; Zhang et al., 2016), high-resolution melting analysis (HRMA) (Boonyuen et al., 2021; Yang et al., 2015) and PCR-restriction fragment length polymorphism (PCR-RFLP) (Kumar et al., 2020). Although the aforementioned methods are powerful and exact, they are expensive, time-consuming and have low throughput (Zhang et al., 2016). The accuracy, sensitivity, and specificity of the SNPscan technology have been shown in numerous investigations. It is also high-throughput and cost-effective (Duan et al., 2017). Because of this, SNPscan is regarded as an acceptable method for the genetic diagnosis of G6PD deficiency.

In the present study, we established a SNPscan assay to identify 33 *G6PD* mutations. Combining the SNPscan assay with DNA sequence analysis for genotype detection and phenotypic screening, we studied the spectrum of *G6PD* mutations in Baise. Our research is essential for creating a community-based carrier screening and prevention program in the area.

2 Materials and methods

2.1 Subjects

A total of 709 suspected G6PD-deficient samples were enrolled from the Baise region of Guangxi Zhuang Autonomous Region between January 2020 and June 2021. These subjects included 346 males and 363 females, between the ages of 1 day old and ninety. Information on ethnic groups was collected. The Affiliated Hospital of Youjiang Medical University for Nationalities' Ethics Committee accepted the study. Informed written consent was obtained from all adult participants or the guardians of pediatric participants. Ethylenediaminetetraacetic acid (EDTA) tubes were used to collect blood samples, which were then brought to the lab and kept in storage at 4°C.

2.2 Quantitative G6PD enzyme activity

The G6PD enzyme activity was measured by a commercial G6PD Detection kit (Korfang Biotechnology Co., Guangzhou, Guangdong, China) according to the rate method (Zhong et al., 2018), which was approved by the China Food and Drug Administration (CFDA) (reg. no. CFDA (P) 20193400771). According to the National Inspection Operational Regulations, 1 mL solution (Korfang Biotechnology Co., Guangzhou, Guangdong, China) was added to a small cup, and then 20 μL of erythrocyte was accurately absorbed into the solution without the plasma layer. The activity of G6PD was detected by the rate method on Hitachi 7170A automatic biochemical analyzer (HITACHI, Japan), and the concentration of hemoglobin in hemolysis was detected by the HiCN method. This method can detect NADPH production in fixed time, which reflect G6PD activity in red blood cells. In each test run, the accuracy of the test findings was checked by calibration and the use of controls offered by KOFA Medical. The reference range of adults with values below 1.30 KU/L (1.30–3.60) and infants with values below 1.70 KU/L (1.70–4.00).

2.3 Genomic DNA extraction

According to the manufacturer's recommendations, genomic DNA was extracted from all samples using a QIAamp DNA Blood Mini kit (Qiagen, Hilden, Germany). The DNA concentration was measured using a Thermo Scientific Nanodrop™ 2000 spectrophotometer and subsequently adjusted to 50 ng/L.

2.4 SNPscan assay for *G6PD* mutations

A multiplex SNPscan assays were designed to detect 33 *G6PD* mutations reported in Chinese population (Wang et al., 2021) as follow: *G6PD* Gaohe (c.95A>G), *G6PD* Songklanagarind (c.196T>A), *G6PD* Asahi (c.202G>A), *G6PD* Chinese-4 (c.392G>T), *G6PD* Valladolid (c.406C>T), *G6PD* Liuzhou (c.442G>A), *G6PD* Shenzhe (c.473G>A), *G6PD* Mahidol (c.487G>A), *G6PD* Taipei (c.493A>G), *G6PD* Nankang (c.517T>C), *G6PD* Miaoli (c.519C>T/G), *G6PD* Mediterranean (c.563C>T), *G6PD* Shunde (c.592C>T), *G6PD* Nanning (c.703C>T), *G6PD* Haikou (c.835A>G/T), *G6PD* Viangchan (c.871G>A), *G6PD* Fushan (c.1004C>A/T), *G6PD* Chinese-5 (c.1024C>T), *G6PD* Beverly Hills (c.1160G>A), *G6PD* Santiago de Cuba (c.1339G>A), *G6PD* Jiangxi (c.1340G>T), *G6PD* Union (c.1360C>T), *G6PD* Canton (c.1376G>T), *G6PD* Yannan (c.1381G>A), *G6PD* Kamiube (c.1387C>T), *G6PD* Kaiping (c.1388G>A), *G6PD* Laibin (c.1414A>C), and four unnamed mutations (c.274C>T, c.371A>G, c.691G>C and c.1225C>T) and

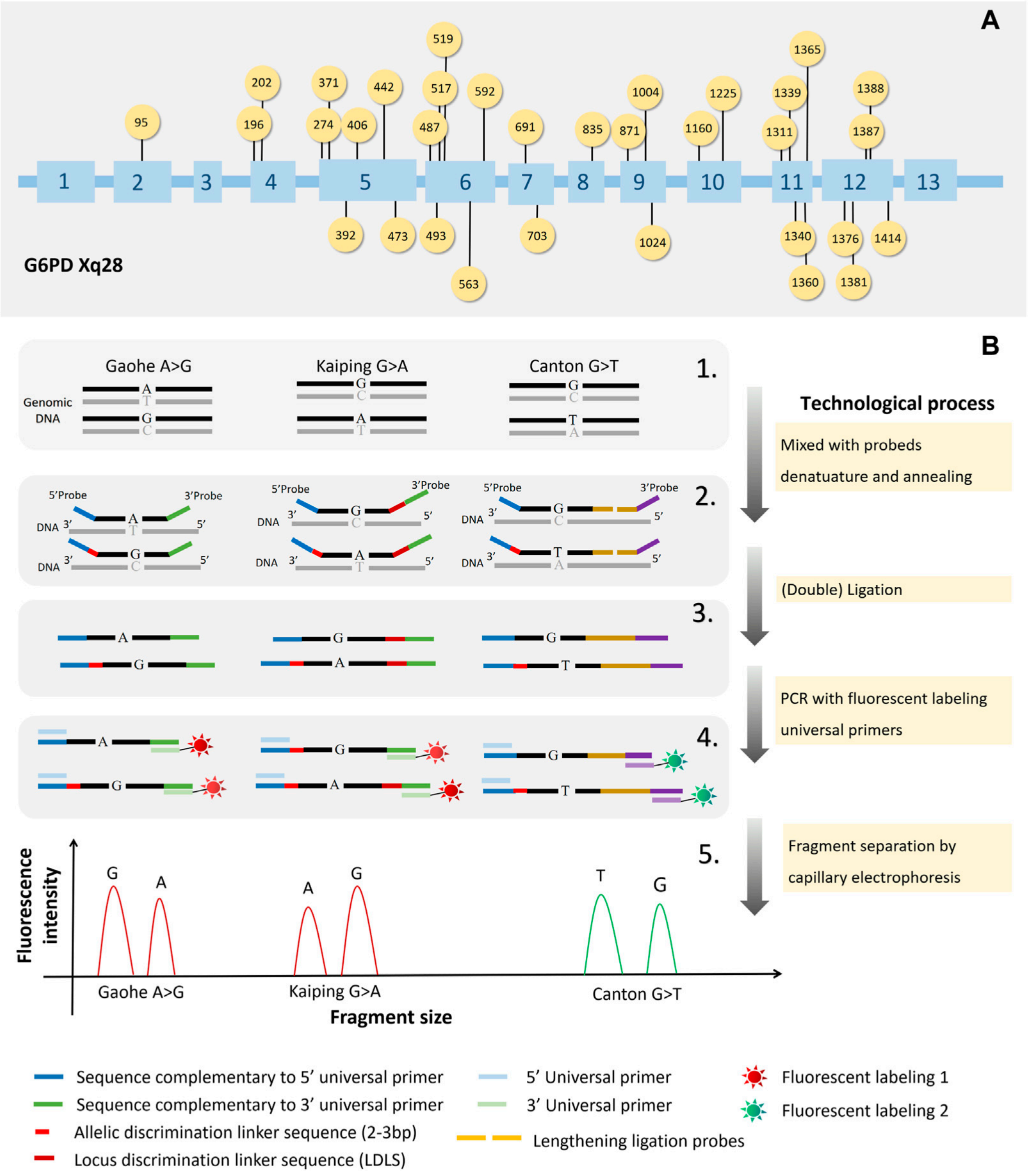


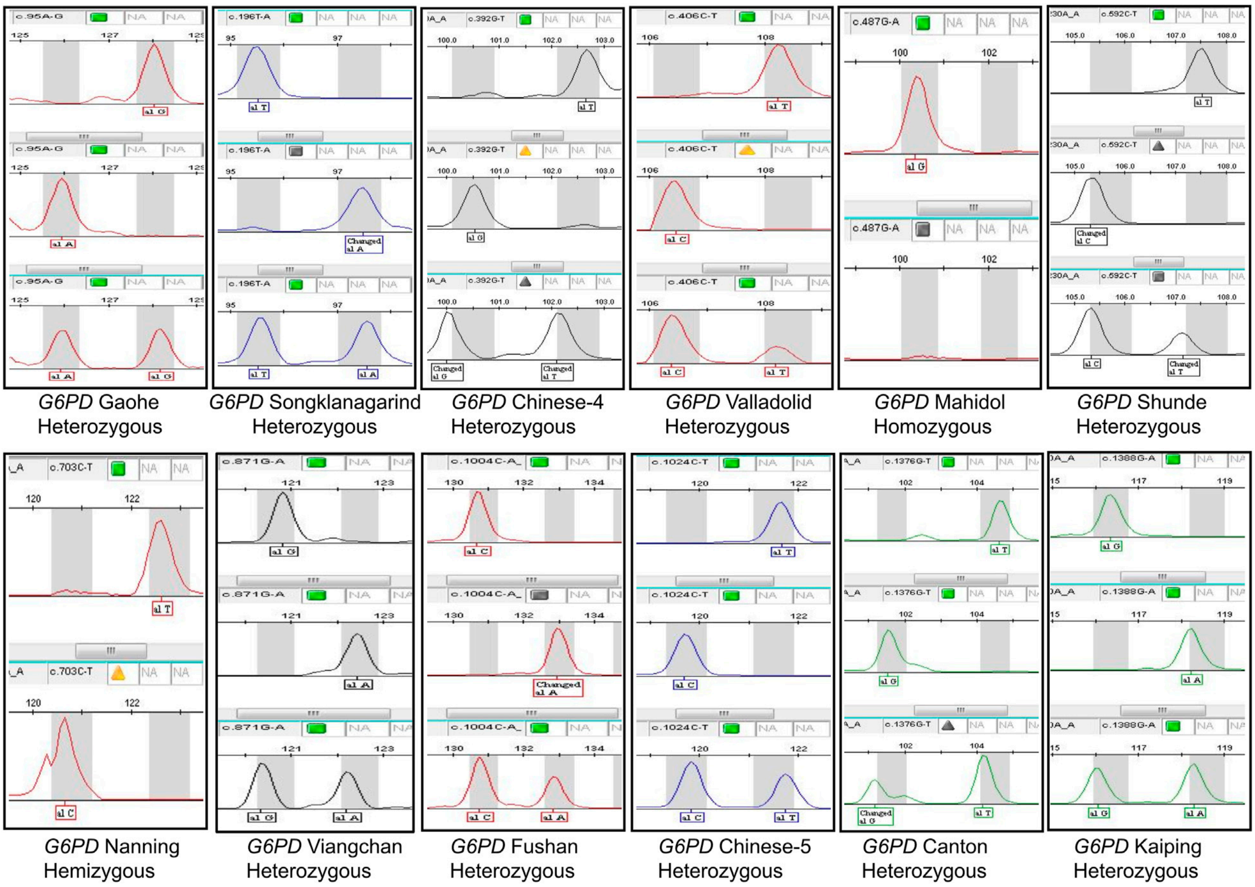
FIGURE 1
The workings of SNPscan technology and the locations of 33 *G6PD* gene mutations. The locations of the 33 mutations in the *G6PD* gene are shown in **(A)**. The principles of SNPscan technology are shown in **(B)**.

two Silent mutation (c.1311C>T, c.1365-13T>C). The 33 *G6PD* mutation sites in the *G6PD* gene are shown in Figure 1A.

As shown in Figure 1B, previously mentioned, the double ligation and multiplex fluorescence PCR serves as the foundation for the SNPscan test. (Wei et al., 2013). The primers and probes are listed in Supplementary Table S1. For each SNPscan assay, 12 μL of ligation mixture was first prepared to contain 2 μL of 10 × ligase buffer, 1 μL of 1 × probe mix, .5 μL of ligase, 7 μL of ddH_2O and 1 μL of 30–250 ng of DNA sample. The ligation reaction was performed on an ABI 2720 thermal cycler with the following cycling program: 98°C for 2 min; 5 cycles of 95°C for 1 min, 58°C for 3 h; 94°C for 2 min, hold at 72°C. Fluorescence in multiplex After that, PCRs were run on each ligation product. Every PCR mixture was made in 20 μL containing 2× PCR Buffer, 1 μL of primer mix, 8 μL of ddH_2O, and 1 μL of ligation product. The PCR program was as follows: 95°C for 2 min; 9 cycles of 94°C for 20 s, 62°C–.5°C/cycle for 40 s, and 72°C for 1.5 min; 26 cycles of 94°C for 20 s, 58°C for 40 s, and 72°C for 1.5 min; 60°C for 1 h; and hold at 4°C. Using a capillary electrophoresis system and an ABI 3730XL sequencer, PCR products were separated and identified. Raw data were analysed with GeneMapper 4.1 software (Applied Biosystems, United States), and the genotypes of each locus were determined.

2.5 DNA sequencing

In order to confirm the SNPscan assay results, PCR amplification and DNA sequencing of the entire *G6PD* coding region was performed as described in our earlier research (Pan



G6PD Gaohe Heterozygous

G6PD Songklanagarind Heterozygous

G6PD Chinese-4 Heterozygous

G6PD Valladolid Heterozygous

G6PD Mahidol Homozygous

G6PD Shunde Heterozygous

G6PD Nanning Hemizygous

G6PD Viangchan Heterozygous

G6PD Fushan Heterozygous

G6PD Chinese-5 Heterozygous

G6PD Canton Heterozygous

G6PD Kaiping Heterozygous

FIGURE 2
The results of *G6PD* positive mutation detected by SNPscan technology.

et al., 2013; Zheng et al., 2020). Purification and sequencing of PCR products were done by Shanghai Vebery Biotechnology (Shanghai, China). All primers are in Supplementary Table S4 (Pan et al., 2013).

2.6 Bioinformatics analysis of *G6PD* mutations

The bioinformatics software used in this work was used to analyze each *G6PD* mutation identified. Moreover, the pI of *G6PD* variants (i.e., monomers) was determined using Kozlowski's protein isoelectric point (IP) calculator (http://isoelectric.org/). Utilizing ConSurf (http://bental.tau.ac.il/new_ConSurfDB/), we looked at the evolutionary conservation of mutant amino acid residues. The pathogenicity of these potential variants was assessed by PolyPhen-2 (Polymorphism Phenotyping v2) (http://genetics.bwh.harvard.edu/pph2/) and Sorting Intolerant from Tolerant (SIFT) web server (http://sift.jcvi.org) prediction models.

2.7 Statistical analysis

The data are collated in Excel. All data were statistical using SPSS 22.0. Descriptive statistics were used to estimate the accuracy.

3 Results

3.1 Development and validation of the SNPscan assay

A SNPscan assay was developed to detect 33 *G6PD* mutations reported in Chinese individuals. As shown in Figure 1, it could precisely distinguish heterozygous mutations and homozygous/hemizygous mutations by capillary electrophoresis (Figure 2). To confirm the accuracy of the SNPscan assay, 30 samples were blindly analysed using PCR amplification and DNA sequencing (Supplementary Table S2). Comparatively speaking to direct DNA sequencing, the SNPscan assay was 100% sensitive [95% confidence interval (CI): 94.87–100%] and 100% specific (95% CI: 87.66–100%), without any cross-reactivity for the identification of *G6PD* mutations. Additionally, the SNPscan assay could precisely distinguish double mutations, such as Canton/Viangchan, Gaohe/Kaiping and Canton/Kaiping. The created approach is dependable for identifying *G6PD* mutations, according to all of the evidence, detailed data are shown in Supplementary Table S3.

3.2 Mutation spectrum of G6PD deficiency

Fifteen *G6PD* mutations were identified by the SNPcan assay in the Baise population (Table 1). Among the 709 G6PD-deficient people, 544

TABLE 1 Frequency of all *G6PD*-positive mutations and predicted consequences before and after amino acid changes.

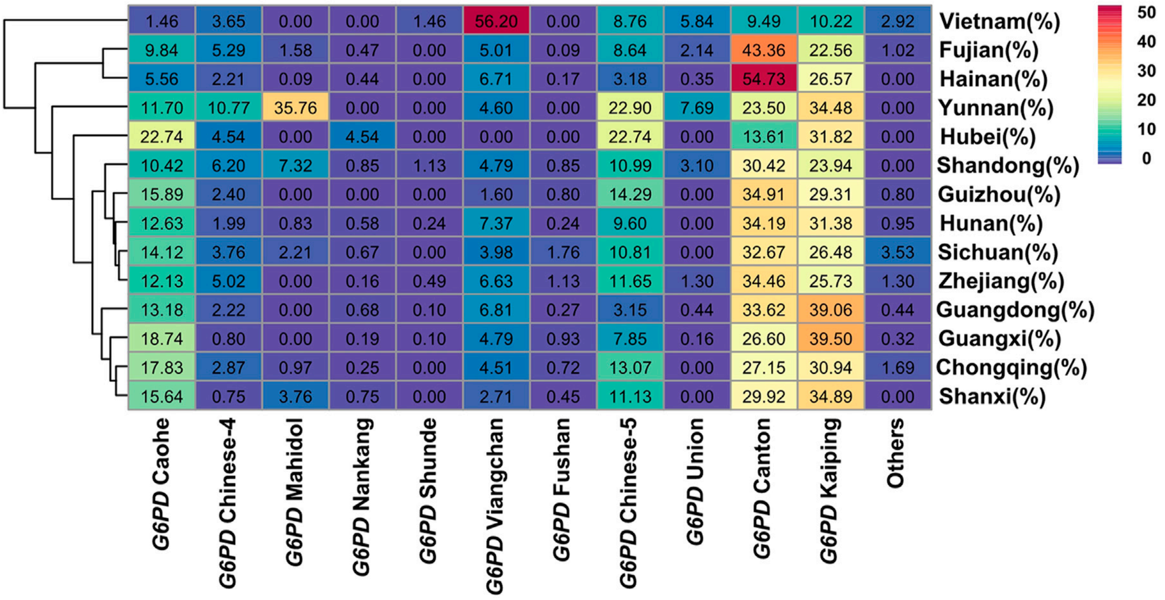
Name	Mutation	Protein	PolyPhen-2	PROVEAN	SIFT	FoldX (stability)	PI	Total (*n*)	Frequency (%)
Gaohe	c.95 A>G	p.His32Arg	PROBABLY DAMAGING	Deleterious	Tolerated	-.583907	6.19	90	17.25
Songklanagarind	c.196 T>A	p.Phe66Ile	BENIGN	Neutral	Tolerated	.58045	6.10	2	.38
NR	c.274 C>T	p.Pro92Ser	BENIGN	Neutral	Tolerated	1.66913	6.10	1	.19
Chinese-4	c.392 G>T	p.Gly131Val	PROBABLY DAMAGING	Deleterious	Damaging	29.6132	6.10	3	.58
Valladolid	c.406 C>T	p.Arg136Cys	PROBABLY DAMAGING	Deleterious	Damaging	2.53579	5.98	11	2.11
Mahidol	c.487 G>A	p.Gly163Ser	POSSIBLY DAMAGING	Deleterious	Damaging	7.96808	6.10	2	.38
Miaoli	c.519 C>T	p.Phe173Leu	PROBABLY DAMAGING	Deleterious	Damaging	1.35032	6.10	7	1.34
Shunde	c.592 C>T	p.Arg198Cys	PROBABLY DAMAGING	Deleterious	Damaging	4.70525	5.99	3	.58
Nanning	c.703 C>T	p.Leu235Pro	PROBABLY DAMAGING	Deleterious	Damaging	6.46278	6.10	2	.38
Viangchan	c.871 G>A	p.Val291Met	PROBABLY DAMAGING	Deleterious	Damaging	-1.19782	6.10	24	4.61
Fushan	c.1004 C>A	p.Ala335Asp	POSSIBLY DAMAGING	Neutral	Damaging	1.44222	6.10	8	1.54
Chinese-5	c.1024 C>T	p.Leu342Phe	BENIGN	Neutral	Tolerated	3.94837	6.10	45	8.64
Union	c.1360 C>T	p.Arg454Cys	PROBABLY DAMAGING	Deleterious	Damaging	2.33769	5.98	1	.19
Canton	c.1376 G>T	p.Arg459Leu	PROBABLY DAMAGING	Deleterious	Tolerated	-.424977	5.99	137	26.30
Kaiping	c.1388 G>T	p.Arg463His	PROBABLY DAMAGING	Deleterious	Damaging	.798808	6.02	185	35.51

NR: class not reported.

TABLE 2 The 709 samples were classified by ethnicity.

Name	Mutation	Zhuang (n, %)	Han (n, %)	Yao (n, %)	Buyi (n, %)	Mulao (n, %)	Total (n, %)
Caohe	c.95 A>G	76 (18.45)	9 (10.34)	0	4 (25.00)	1 (100)	90 (17.25)
Songklanagarind	c.196 T>A	1 (.24)	1 (1.15)	0	0	0	2 (.38)
NR	c.274 C>T	0	1 (1.15)	0	0	0	1 (.19)
Chinese-4	c.392 G>T	2 (.49)	1 (1.15)	0	0	0	3 (.58)
Valladolid	c.406 C>T	9 (2.18)	2 (2.30)	0	0	0	11 (2.11)
Mahidol	c.487 G>A	1 (.24)	1 (1.15)	0	0	0	2 (.38)
Miaoli	c.519 T>G	3 (.73)	2 (2.30)	0	2 (12.5)	0	7 (1.34)
Shunde	c.592 C>T	2 (.49)	1 (1.15)	0	0	0	3 (.58)
Nanning	c.703 C>T	1 (.24)	1 (1.15)	0	0	0	2 (.38)
Viangchan	c.871 G>A	19 (4.61)	5 (5.75)	0	0	0	24 (4.61)
Fushan	c.1004 C>A	7 (1.70)	0	0	1 (6.25)	0	8 (1.54)
Chinese-5	c.1024 C>T	32 (7.77)	10 (11.49)	2 (40)	1 (6.25)	0	45 (8.64)
Union	c.1360 C>T	0	1 (1.15)	0	0	0	1 (.19)
Canton	c.1376 G>T	113 (27.43)	17 (19.54)	3 (60)	4 (25.00)	0	137 (26.30)
Kaiping	c.1388 G>T	146 (35.44)	35 (40.23)	0	4 (25.00)	0	185 (35.51)
	Total	412 (100)	87 (100)	5 (100)	16 (100)	1 (100)	521 (100)

(277 females and 267 males) had at least one mutation in the *G6PD* gene. Among the 277 females with mutated G6PD deficiency, we identified 22 (3.10%) homozygotes and 255 heterozygotes, including 73 compound heterozygotes. The mutations of *G6PD* Kaiping, *G6PD* Canton and *G6PD* Gaohe were the three dominant mutations with an overall frequency of higher than 79.06%, followed by *G6PD* Chinese-5, *G6PD* Viangchan and *G6PD* Valladolid, with a frequency of 2.11% as a minimum, respectively. The number and frequency of various mutations are presented in Table 1.



	G6PD Caohe	*G6PD* Chinese-4	*G6PD* Mahidol	*G6PD* Nankang	*G6PD* Shunde	*G6PD* Viangchan	*G6PD* Fushan	*G6PD* Chinese-5	*G6PD* Union	*G6PD* Canton	*G6PD* Kaiping	Others
Vietnam(%)	1.46	3.65	0.00	0.00	1.46	56.20	0.00	8.76	5.84	9.49	10.22	2.92
Fujian(%)	9.84	5.29	1.58	0.47	0.00	5.01	0.09	8.64	2.14	43.36	22.56	1.02
Hainan(%)	5.56	2.21	0.09	0.44	0.00	6.71	0.17	3.18	0.35	54.73	26.57	0.00
Yunnan(%)	11.70	10.77	35.76	0.00	0.00	4.60	0.00	22.90	7.69	23.50	34.48	0.00
Hubei(%)	22.74	4.54	0.00	4.54	0.00	0.00	0.00	22.74	0.00	13.61	31.82	0.00
Shandong(%)	10.42	6.20	7.32	0.85	1.13	4.79	0.85	10.99	3.10	30.42	23.94	0.00
Guizhou(%)	15.89	2.40	0.00	0.00	0.00	1.60	0.80	14.29	0.00	34.91	29.31	0.80
Hunan(%)	12.63	1.99	0.83	0.58	0.24	7.37	0.24	9.60	0.00	34.19	31.38	0.95
Sichuan(%)	14.12	3.76	2.21	0.67	0.00	3.98	1.76	10.81	0.00	32.67	26.48	3.53
Zhejiang(%)	12.13	5.02	0.00	0.16	0.49	6.63	1.13	11.65	1.30	34.46	25.73	1.30
Guangdong(%)	13.18	2.22	0.00	0.68	0.10	6.81	0.27	3.15	0.44	33.62	39.06	0.44
Guangxi(%)	18.74	0.80	0.00	0.19	0.10	4.79	0.93	7.85	0.16	26.60	39.50	0.32
Chongqing(%)	17.83	2.87	0.97	0.25	0.00	4.51	0.72	13.07	0.00	27.15	30.94	1.69
Shanxi(%)	15.64	0.75	3.76	0.75	0.00	2.71	0.45	11.13	0.00	29.92	34.89	0.00

FIGURE 3
Heatmap of G6PD-deficient allele frequency distributions for Baise populations and others cities populations. Red represents the highest G6PD-deficient allele frequency, while purple represents the lowest.

These 709 samples were classified by ethnicity. There were 412 allele mutations and 15 variants in the Zhuang nationality, most of which were *G6PD* Kaiping, *G6PD* Canton and *G6PD* Gaohe, accounting for 79.08% (Table 2). However, there were only 87 (16.70%) allele mutations in the Han nationality. In addition, in order to examine the association of the major G6PD-deficient alleles in Chinese people and Southern Asian populations, data from our research or other studies were further analysed (Liu et al., 2020; Zheng et al., 2020). As shown in Figure 3, the frequencies of different G6PD-deficient alleles in different regions were plotted on a heatmap. The color of each block varies with the corresponding frequency. Purple represented the lowest allele frequency on the color scale, which went up to red for the greatest allele frequency. Obviously, Four G6PD-deficient alleles (Canton, Kaiping, Gaohe and Chinese-5) were present in relatively high frequencies in Chinese people, whereas *G6PD* Viangchan and *G6PD* Kaiping were prevalent in Southern Asian populations (Vietnam populations).

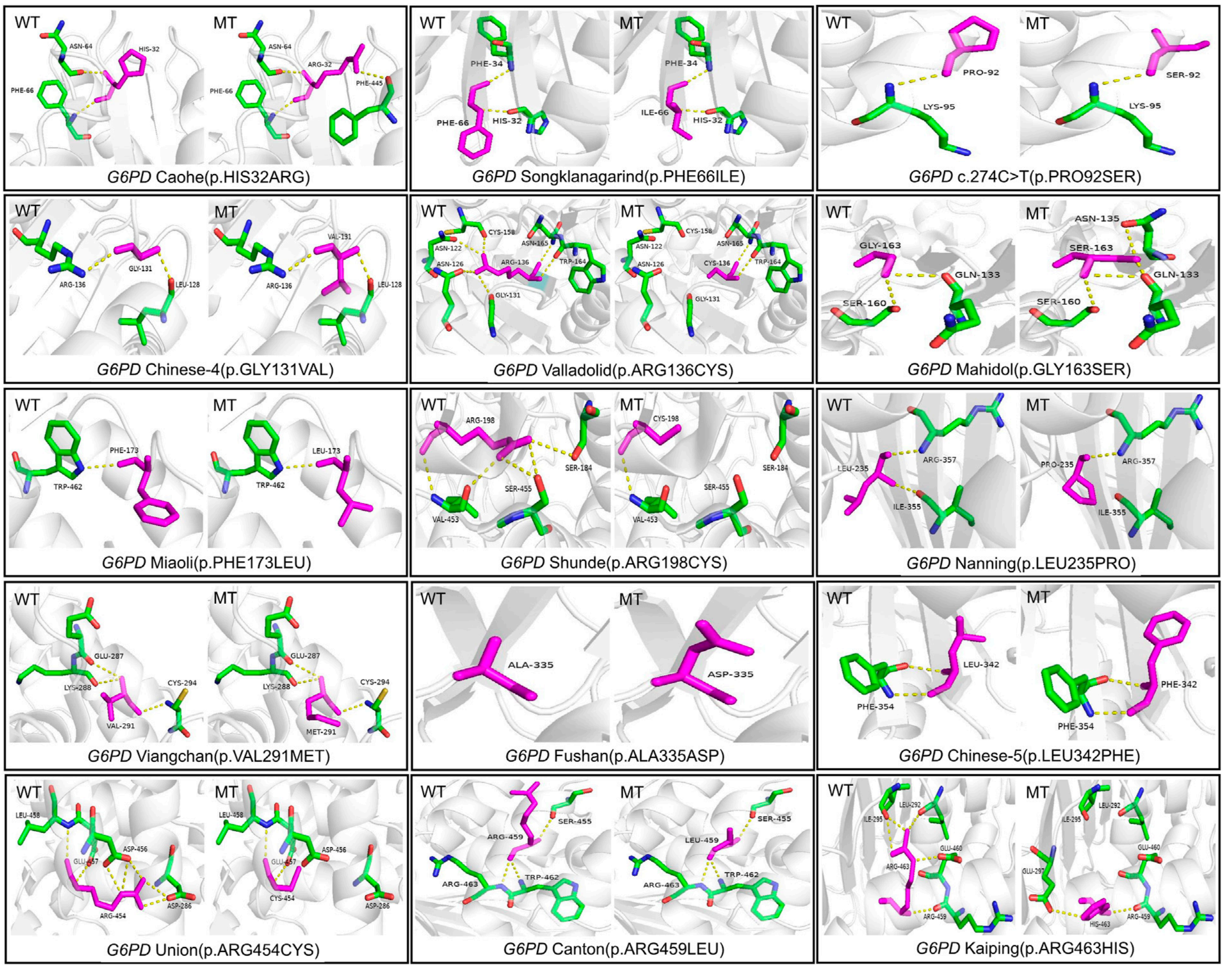
3.3 Effect of mutations on disease manifestation

Tools from bioinformatics were used to forecast how changing an amino acid might affect how a protein function (Table 1). According to PolyPhen2.0, all variants were identified as potentially damaging (prediction score close to 1) except for *G6PD* Songklanagarind, *G6PD* c.274C>T and *G6PD* Chinese-5), similar to the results predicted by PROVEAN (except *G6PD* Fushan). However, SIFT predicted that five missense mutations (*G6PD* Gaohe, *G6PD* Songklanagarind, *G6PD* c.274C>T, *G6PD* Chinese-5, and *G6PD* Canton) could be tolerated, and the rest were damaging. FoldX was used to predict changes in the protein stability of *G6PD* (Table 1), and three variants (*G6PD* Gaohe, *G6PD* Viangchan and *G6PD* Canton) were found to increase the stability of the G6PD protein, while other missense variants were predicted to destabilize the G6PD protein. Additionally, Table 1 provides an overview of the expected pI values for each of the 15 *G6PD* variations. The changes in protein structure and polar bonds before and after *G6PD* mutation are shown in Figure 4.

4 Discussion

In this study, we looked studied the distribution of different *G6PD* gene variants, the prevalence of G6PD deficiency, and the relationship between genotypes and phenotypes related to enzyme function in Baise, Guangxi Zhuang Autonomous Region. The results showed that six of the most prevalent mutations were *G6PD* Kaiping, *G6PD* Canton, *G6PD* Gaohe, *G6PD* Chinese-4, *G6PD* Viangchan and *G6PD* Chinese-5, accounting for more than 60% of G6PD-deficient alleles. This result is consistent with LinZou's research (Liu et al., 2020). The sexes and different sorts of mutation patterns affected how G6PD activities were distributed (Driscoll and Migeon, 1990). These findings present a more precise and thorough characterization of G6PD deficiency in Baise, Guangxi.

The prevalence of G6PD deficiency varies widely by region in China, with northern China having a relatively lower prevalence than southern China. G6PD deficiency was present in 2.1% of China's population overall (He et al., 2020), and over 35 different *G6PD* gene mutations were known, with *G6PD* Kaiping and *G6PD* Canton predominating in earlier investigations (Liu et al., 2020; Jiang et al., 2006). Africa, Asia, southern Europe, the Middle East, Southeast Asia, and Mediterranean nations have the highest prevalence rates, according to reports (He et al., 2020; Liu et al., 2020). In India, in various population groups, it was



G6PD Caohe(p.HIS32ARG)

G6PD Songklanagarind(p.PHE66ILE)

G6PD c.274C>T(p.PRO92SER)

G6PD Chinese-4(p.GLY131VAL)

G6PD Valladolid(p.ARG136CYS)

G6PD Mahidol(p.GLY163SER)

G6PD Miaoli(p.PHE173LEU)

G6PD Shunde(p.ARG198CYS)

G6PD Nanning(p.LEU235PRO)

G6PD Viangchan(p.VAL291MET)

G6PD Fushan(p.ALA335ASP)

G6PD Chinese-5(p.LEU342PHE)

G6PD Union(p.ARG454CYS)

G6PD Canton(p.ARG459LEU)

G6PD Kaiping(p.ARG463HIS)

FIGURE 4
Changes in protein structure and polar bonds before and after *G6PD* mutation.

discovered that 8.5% of people have G6PD deficiencies (Saravu et al., 2016). The prevalence rate varies between the tribal groupings, ranging from 2.3% to 27.0%, with an overall incidence of 7.7% (Mukherjee et al., 2015). In contrast to southern India, where it is continuously low except in the states of Andhra Pradesh and Tamil Nadu, the frequency of the G6PD-deficient allele is higher in northern and western India (Devendra et al., 2020). In Indian caste groupings, *G6PD* Mediterranean was discovered to be the most prevalent variation (Devendra et al., 2020; Sukumar et al., 2004). However, *G6PD* Kaiping was found to be the most common variant in China (Lin et al., 2018; Liu et al., 2020; Yan et al., 2010). In Southeast Asia, G6PD deficiency is diverse, as previously demonstrated by epidemiological and molecular research (Louicharoen and Nuchprayoon, 2005). In Thais, Laotians, Cambodians, and Malaysian Malays, *G6PD* Viangchan appears to be the most prevalent form (Ainoon et al., 2003; Iwai et al., 2001; Louicharoen and Nuchprayoon, 2005; Nuchprayoon et al., 2002), while the most prevalent form of G6PD in the population of Myanmar is Mahidol (Matsuoka et al., 2004). In the current research, a total of 15 harmful mutations were found, which were dominated by *G6PD* Kaiping and *G6PD* Canton, accounting for approximately 42% of all G6PD-deficient alleles. However, it is lower than previous research results (84.1%, 75.3%) in the Guangxi population (Fu et al., 2018; Yan et al., 2006). This can be because we only collected a small number of samples or because geographical disparities exist. In addition, we are a region with a high prevalence of thalassemia, moreover, hemolysis and anemia are common (Lin et al., 2015). Medication, hemolysis, and anemia can affect the detection of G6PD activity (Nuinoon et al., 2022; Pfeffer et al., 2022). In our study, these may be one of the reasons that the 165 samples with no detection of any of the 33 common mutations. However, they were detected with G6PD activity deficiency. Certainly, the other reason is that they may have rare mutations (besides 33 common mutations).

G6PD was first described by Carson in 1956 (ALVING et al., 1956). Its clinical manifestations include fulminant hemolysis, severe hyperbilirubinemia, and kernicterus, which contribute to neonatal neurological injury and risk of death (He et al., 2020; Kaplan et al., 2015; Liu et al., 2020). This condition may be brought on by infections, specific foods (such as fava beans), oxidizing medicines, and/or specific herbal therapies (Liu et al., 2020). To date, after a newborn's screening results in a positive result, the most effective treatment for this illness is to prevent hemolysis by avoiding some oxidative stressors (Liu et al., 2020). Therefore, the general survey of G6PD deficiency, early detection and early prevention are important measures to prevent and treat the disease. There are three common measures to prevent the disease, the most important is to avoid accidental ingestion of fava beans (Reading et al., 2016); secondly, avoid taking anti-malarial drugs (primaquine, chloroquine, malaria quinine, pentaquine and adipine), sulfones (thiazole sulfone, aminophene sulfone), sulfonamides (sulfamethoxazole, sulfadimethoxine, sulfapyridine and salazosulfapyridine) and antipyretics (acetazolamide and acetanilide) and so on (http://www.g6pd.org) (Chu and Freedom, 2019; Reading et al., 2016). Finally, when the patient has an infection (viral hepatitis, influenza, pneumonia, typhoid), which should immediately seek medical help to avoid hemolysis.

Today's G6PD deficiency diagnosis primarily uses the enzyme activity detection assay, and the main diagnosis used to avoid oxidative hemolysis cannot be other than a phenotypic test, especially in women; however, there is an added value in *G6PD* genotyping, different sorts of mutations can result in various classes of variations and exhibit various symptoms (Beutler et al., 2002; WHO, 2022). So, to establish a certain diagnosis of G6PD insufficiency, genotyping of *G6PD* mutations is beneficial (Jiang et al., 2006). In addition, the analysis of *G6PD* genotypes contributes to the study of molecular biology and genetic characterization of human populations (Hamali, 2021; Lee et al., 2022). Aside from this, the genotyping of G6PD deficiency also has a significant impact on the field's understanding of the disorder (Li et al., 2008). The SNPscan assay used in the study covered 33 common mutations in the Chinese population and could identify more than 95% of G6PD deficiencies. Based on the detection of SNP loci, SNPscan technology can simultaneously type multiple SNP loci in one detection process (Yu et al., 2021). Numerous investigations have shown that it has good accuracy, sensitivity, and specificity and is cost-effective and high-throughput (Du et al., 2014; Yin et al., 2014; Zhang et al., 2016). Compared with the direct sequencing method, it saves more tedious operations in the experimental process, can detect multiple sites in multiple samples at the same time, and reduces the cost (Zhang et al., 2016). Compared with the gene chip method, SNPscan technology has more detection sites, so it can be flexibly designed for known target gene mutation sites and achieves high throughput (Chen et al., 2012; Duan et al., 2017; Hu et al., 2015; Zhang et al., 2016). In addition, we investigated a general comparison of costs associated with these different techniques and found that the SNPscan technique has the lowest cost (SNPscan technology: $14.26/sample, direct sequencing method: $20.97/sample, gene chip method: $69.93/sample). Therefore, a trustworthy, quick, and affordable method for identifying *G6PD* point mutations would be beneficial to patients, their families, the doctors who treat them, and the testing labs.

Data availability statement

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

Ethics statement

This protocol was approved by the ethics committee of the Affiliated Hospital of Youjiang Medical University for Nationalities, and written informed consent was obtained from all individuals.

Author contributions

All authors contributed to the study's conception and design. ML, JW, and CW designed the study. WC, LC, and LH collected the samples and entered the data. WW, GX, and WH analysed and interpreted the data. HW, HH, and WH conducted the laboratory work (*G6PD* genotyping and analysis of G6PD enzyme activity). LL, WC, and YZ make the figures and tables. ML, HW, and YL wrote the paper. All authors critically reviewed the paper and approved the final version of the paper for submission.

Funding

This work was partially supported by Key Scientific Research Projects of Guangdong Provincial Department of Education, Grant/Award Number: 2018A030307074; Natural Science Foundation of Guangdong Province, Grant/Award Number: 2018A030307074; National Natural Science Foundation of China, Contract/Grant Number: 81760615; Scientific Research Project of Chaozhou Health Bureau, Grant/Award Number: 2021013.

Acknowledgments

We would like to thank all teammates for contributing to this work.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1000290/full#supplementary-material

References

Ainoon, O., Yu, Y. H., Amir, M. A., Boo, N. Y., Cheong, S. K., and Hamidah, N. H. (2003). Glucose-6-phosphate dehydrogenase (G6PD) variants in Malaysian Malays. *Hum. Mutat.* 21 (1), 101. doi:10.1002/humu.9103

Alving, A. S., Carson, P. E., Flanagan, C. L., and Ickes, C. E. (1956). Enzymatic deficiency in primaquine-sensitive erythrocytes. *Science* 124 (3220), 484–485. doi:10.1126/science.124.3220.484-a

Beutler, E., Gelbart, T., and Miller, W. (2002). Severe jaundice in a patient with a previously undescribed glucose-6-phosphate dehydrogenase (g6pd) mutation and gilbert syndrome. *Blood Cells Mol. Dis.* 28 (2), 104–107.

Boonyuen, U., Songdej, D., Tanyaratsrisakul, S., Phuanukoonnon, S., Chamchoy, K., Praoparotai, A., et al. (2021). Glucose-6-phosphate dehydrogenase mutations in malaria endemic area of Thailand by multiplexed high-resolution melting curve analysis. *Malar. J.* 20 (1), 194. doi:10.1186/s12936-021-03731-0

Chen, X., Li, S., Yang, Y., Yang, X., Liu, Y., Liu, Y., et al. (2012). Genome-wide association study validation identifies novel loci for atherosclerotic cardiovascular disease. *J. Thromb. Haemost.* 10 (8), 1508–1514. doi:10.1111/j.1538-7836.2012.04815.x

Chu, C. S., and Freedman, D. O. (2019). Tafenoquine and G6PD: A primer for clinicians. *J. Travel Med.* 26 (4), taz023. doi:10.1093/jtm/taz023

Devendra, R., Gupta, V., Biradar, S. S., Bhat, P., Hegde, S., Hoti, S. L., et al. (2020). G6PD A-is the major cause of G6PD deficiency among the Siddis of Karnataka, India. *Ann. Hum. Biol.* 47 (1), 55–58. doi:10.1080/03014460.2019.1699954

Dombrowski, J. G., Souza, R. M., Curry, J., Hinton, L., Silva, N., Grignard, L., et al. (2017). G6PD deficiency alleles in a malaria-endemic region in the Western Brazilian Amazon. *Malar. J.* 16 (1), 253. doi:10.1186/s12936-017-1889-6

Driscoll, D. J., and Migeon, B. R. (1990). Sex difference in methylation of single-copy genes in human meiotic germ cells: implications for X chromosome inactivation, parental imprinting, and origin of CpG mutations. *Somat. Cell. Mol. Genet.* 16 (3), 267–282. doi:10.1007/BF01233363

Du, W., Cheng, J., Ding, H., Jiang, Z., Guo, Y., and Yuan, H. (2014). A rapid method for simultaneous multi-gene mutation screening in children with nonsyndromic hearing loss. *Genomics* 104 (4), 264–270. doi:10.1016/j.ygeno.2014.07.009

Duan, S. H., Ma, J. L., Yang, X. L., and Guo, Y. F. (2017). Simultaneous multi-gene mutation screening using SNPscan in patients from ethnic minorities with nonsyndromic hearing-impairment in Northwest China. *Mol. Med. Rep.* 16 (5), 6722–6728. doi:10.3892/mmr.2017.7431

Filosa, S., Giacometti, N., Wangwei, C., De Mattia, D., Pagnini, D., Alfinito, F., et al. (1996). Somatic-cell selection is a major determinant of the blood-cell phenotype in heterozygotes for glucose-6-phosphate dehydrogenase mutations causing severe enzyme deficiency. *Am. J. Hum. Genet.* 59 (4), 887–895.

Fu, C., Luo, S., Li, Q., Xie, B., Yang, Q., Geng, G., et al. (2018). Newborn screening of glucose-6-phosphate dehydrogenase deficiency in Guangxi, China: Determination of optimal cutoff value to identify heterozygous female neonates. *Sci. Rep.* 8 (1), 833. doi:10.1038/s41598-017-17667-6

Gómez-Manzo, S., Marcial-Quino, J., Vanoye-Carlo, A., Serrano-Posada, H., Ortega-Cuellar, D., González-Valdez, A., et al. (2016). Glucose-6-phosphate dehydrogenase: Update and analysis of new mutations around the world. *Int. J. Mol. Sci.* 17 (12), 2069. doi:10.3390/ijms17122069

Hamali, H. A. (2021). Glucose-6-phosphate dehydrogenase deficiency: An overview of the prevalence and genetic variants in Saudi Arabia. *Hemoglobin* 45 (5), 287–295. doi:10.1080/03630269.2022.2034644

He, Y., Zhang, Y., Chen, X., Wang, Q., Ling, L., and Xu, Y. (2020). Glucose-6-phosphate dehydrogenase deficiency in the han Chinese population: Molecular characterization and genotype-phenotype association throughout an activity distribution. *Sci. Rep.* 10 (1), 17106. doi:10.1038/s41598-020-74200-y

Howell, R. R. (2006). Advisory committee on heritable disorders and genetic diseases in newborns and children. *Ment. Retard. Dev. Disabil. Res. Rev.* 12 (4), 313–315. doi:10.1002/mrdd.20126

Hu, R., Lin, M., Ye, J., Zheng, B. P., Jiang, L. X., Zhu, J. J., et al. (2015). Molecular epidemiological investigation of G6PD deficiency by a gene chip among Chinese Hakka of southern Jiangxi province. *Int. J. Clin. Exp. Pathol.* 8 (11), 15013. doi:10.1097/01.hs9.0000566368.48943.78

Iwai, K., Hirono, A., Matsuoka, H., Kawamoto, F., Horie, T., Lin, K., et al. (2001). Distribution of glucose-6-phosphate dehydrogenase mutations in Southeast Asia. *Hum. Genet.* 108 (6), 445–449. doi:10.1007/s004390100527

Ji-Guang, D., Shui-Lan, Y. U., Nong-Zhiand Yi-Chao, Y. (2017). Analysis of inspection certification results on malaria elimination in Baise City. *Zhongguo Xue Xi Chong Bing Fang. Zhi Za Zhi* 29 (4), 512–514. doi:10.16250/j.32.1374.2017017

Jiang, W., Yu, G., Liu, P., Geng, Q., Chen, L., Lin, Q., et al. (2006). Structure and function of glucose-6-phosphate dehydrogenase-deficient variants in Chinese population. *Hum. Genet.* 119 (5), 463–478. doi:10.1007/s00439-005-0126-5

Kaplan, M., Hammerman, C., and Bhutani, V. K. (2015). Parental education and the WHO neonatal G-6-PD screening program: A quarter century later. *J. Perinatol.* 35 (10), 779–784. doi:10.1038/jp.2015.77

Kumar, R., Singh, M., Mahapatra, S., Chaurasia, S., Tripathi, M. K., Oommen, J., et al. (2020). Fine mapping of glucose 6 phosphate dehydrogenase (G6PD) deficiency in a rural malaria area of south west odisha using the clinical, hematological and molecular approach. *Med. J. Hematol. Infect. Dis.* 12 (1), e2020015. doi:10.4084/MJHID.2020.015

Lee, H. Y., Ithnin, A., Azma, R. Z., Othman, A., Salvador, A., and Cheah, F. C. (2022). Glucose-6-Phosphate dehydrogenase deficiency and neonatal hyperbilirubinemia: Insights on pathophysiology, diagnosis, and gene variants in disease heterogeneity. *Front. Pediatr.* 10, 875877. doi:10.3389/fped.2022.875877

Li, L., Zhou, Y. Q., Xiao, Q. Z., Yan, T. Z., and Xu, X. M. (2008). Development and evaluation of a reverse dot blot assay for the simultaneous detection of six common Chinese G6PD mutations and one polymorphism. *Blood Cells Mol. Dis.* 41 (1), 17–21. doi:10.1016/j.bcmd.2008.01.007

Liang, X. Y., Chen, J. T., Ma, Y. B., Huang, H. Y., Xie, D. D., Monte-Nguba, S. M., et al. (2020). Evidence of positively selected G6PD A-allele reduces risk of Plasmodium falciparum infection in African population on Bioko Island. *Mol. Genet. Genom. Med.* 8 (2), e1061. doi:10.1002/mgg3.1061

Lin, F., Lou, Z., Xing, S., Zhang, L., and Yang, L. (2018). The gene spectrum of glucose-6-phosphate dehydrogenase (g6pd) deficiency in guangdong province, China. *Gene* 678, 312–317. doi:10.1016/j.gene.2018.07.068

Lin, M., Yang, L. Y., Xie, D. D., Chen, J. T., Nguba, S. M., Ehapo, C. S., et al. (2015). G6PD deficiency and hemoglobinopathies: Molecular epidemiological characteristics and healthy effects on malaria endemic bioko island, Equatorial Guinea. *PLoS One* 10 (4), e0123991. doi:10.1371/journal.pone.0123991

Liu, Z., Yu, C., Li, Q., Cai, R., Qu, Y., Wang, W., et al. (2020). Chinese newborn screening for the incidence of G6PD deficiency and variant of G6PD gene from 2013 to 2017. *Hum. Mutat.* 41 (1), 212–221. doi:10.1002/humu.23911

Louicharoen, C., and Nuchprayoon, I. (2005). G6PD Viangchan (871G>A) is the most common G6PD-deficient variant in the Cambodian population. *J. Hum. Genet.* 50 (9), 448–452. doi:10.1007/s10038-005-0276-2

Maloukh, L., Kumarappan, A., El-Din, E. H., Al-Kamali, F., Gomma, F., Akhondi, A., et al. (2021). Development of allelic discrimination assay to detect Mediterranean G6PD mutation and its linked inheritance with normal vision and/colorblindness loci for 4 generations among Egyptian and Emirati families. *Saudi J. Biol. Sci.* 28 (9), 5028–5033. doi:10.1016/j.sjbs.2021.05.014

Matsuoka, H., Wang, J., Hirai, M., Arai, M., Yoshida, S., Kobayashi, T., et al. (2004). Glucose-6-phosphate dehydrogenase (g6pd) mutations in Myanmar: g6pd mahidol (487g> a is the most common variant in the Myanmar population. *J. Hum. Genet.* 49 (10), 544–547. doi:10.1007/s10038-004-0187-7

Mukherjee, M. B., Colah, R. B., Martin, S., and Ghosh, K. (2015). Glucose-6-phosphate dehydrogenase (G6PD) deficiency among tribal populations of India - country scenario. *Indian J. Med. Res.* 141 (5), 516–520. doi:10.4103/0971-5916.159499

Nuchprayoon, I., Sanpavat, S., and Nuchprayoon, S. (2002). Glucose-6-phosphate dehydrogenase (g6pd) mutations in Thailand: g6pd viangchan (871g> is the most common deficiency variant in the Thai population. *Hum. Mutat.* 19 (2), 185. doi:10.1002/humu.9010

Nuinoon, M., Krithong, R., Pramtong, S., Sasuk, P., Ngeaiad, C., Chaimusik, S., et al. (2022). Prevalence of g6pd deficiency and g6pd variants amongst the southern Thai population. *PeerJ* 10, e14208. doi:10.7717/peerj.14208

Pan, M., Lin, M., Yang, L., Wu, J., Zhan, X., Zhao, Y., et al. (2013). Glucose-6-phosphate dehydrogenase (G6PD) gene mutations detection by improved high-resolution DNA melting assay. *Mol. Biol. Rep.* 40 (4), 3073–3082. doi:10.1007/s11033-012-2381-6

Pfeffer, D. A., Satyagraha, A. W., Sadhewa, A., Alam, M. S., Bancone, G., Boum, Y. N., et al. (2022). Genetic variants of glucose-6-phosphate dehydrogenase and their associated enzyme activity: A systematic review and meta-analysis. *Pathogens* 11 (9), 1045. doi:10.3390/pathogens11091045

Reading, N. S., Sirdah, M. M., Shubair, M. E., Nelson, B. E., Al-Kahlout, M. S., Al-Tayeb, J. M., et al. (2016). Favism, the commonest form of severe hemolytic anemia in Palestinian children, varies in severity with three different variants of G6PD deficiency within the same community. *Blood Cells Mol. Dis.* 60, 58–64. doi:10.1016/j.bcmd.2016.07.001

Saravu, K., Kumar, R., Ashok, H., Kundapura, P., Kamath, V., Kamath, A., et al. (2016). Therapeutic assessment of chloroquine-primaquine combined regimen in adult cohort of plasmodium vivax malaria from primary care centres in southwestern India. *PLoS One* 11 (6), e0157666. doi:10.1371/journal.pone.0157666

Sukumar, S., Mukherjee, M. B., Colah, R. B., and Mohanty, D. (2004). Molecular basis of G6PD deficiency in India. *Blood Cells Mol. Dis.* 33 (2), 141–145. doi:10.1016/j.bcmd.2004.06.003

Tian, P. L., Zhou, B. Y., Zhao, W. Z., Zheng, L. X., Ye, J. L., Wang, B. X., et al. (2013). Identification of glucose-6-phosphate dehydrogenase gene variants in Guangdong populations. *Zhonghua Xue Ye Xue Za Zhi* 34 (8), 719–721. doi:10.3760/cma.j.issn.0253-2727.2013.08.017

Wei, J., Zheng, L., Liu, S., Yin, J., Wang, L., Wang, X., et al. (2013). MiR-196a2 rs11614913 T > C polymorphism and risk of esophageal cancer in a Chinese population. *Hum. Immunol.* 74 (9), 1199–1205. doi:10.1016/j.humimm.2013.06.012

Who (2022). *Technical consultation to review the classification of glucose-6-phosphate dehydrogenase (g6pd)* Geneva: World Health Organ.

Wisnumurti, D. A., Sribudiani, Y., Porsch, R. M., Maskoen, A. M., Rahayuningsih, S. E., Asni, E. K., et al. (2019). G6PD genetic variations in neonatal Hyperbilirubinemia in

Indonesian Deutromalay population. *BMC Pediatr.* 19 (1), 506. doi:10.1186/s12887-019-1882-z

Yan, J. B., Xu, H. P., Xiong, C., Ren, Z. R., Tian, G. L., Zeng, F., et al. (2010). Rapid and reliable detection of glucose-6-phosphate dehydrogenase (G6PD) gene mutations in Han Chinese using high-resolution melting analysis. *J. Mol. Diagn.* 12 (3), 305–311. doi:10.2353/jmoldx.2010.090104

Yan, T., Cai, R., Mo, O., Zhu, D., Ouyang, H., Huang, L., et al. (2006). Incidence and complete molecular characterization of glucose-6-phosphate dehydrogenase deficiency in the Guangxi Zhuang autonomous region of southern China: Description of four novel mutations. *Haematologica* 91 (10), 1321. doi:10.3389/fgene.2022.994015

Yang, H., Wang, Q., Zheng, L., Zhan, X. F., Lin, M., Lin, F., et al. (2015). Incidence and molecular characterization of Glucose-6-Phosphate Dehydrogenase deficiency among neonates for newborn screening in Chaozhou, China. *Int. J. Lab. Hematol.* 37 (3), 410–419. doi:10.1111/ijlh.12303

Yin, J., Wang, L., Shi, Y., Shao, A., Tang, W., Wang, X., et al. (2014). Interleukin 17A rs4711998 A>G polymorphism was associated with a decreased risk of esophageal cancer in a Chinese population. *Dis. Esophagus.* 27 (1), 87–92. doi:10.1111/dote.12045

Yu, H., Hu, W., Lin, C., Xu, L., Liu, H., Luo, L., et al. (2021). Polymorphisms analysis for association between ADIPO signaling pathway and genetic susceptibility to T2DM in Chinese han population. *Adipocyte* 10 (1), 463–474. doi:10.1080/21623945.2021.1978728

Zhang, F., Xiao, Y., Xu, L., Zhang, X., Zhang, G., Li, J., et al. (2016). Mutation analysis of the common deafness genes in patients with nonsyndromic hearing loss in linyi by SNPscan assay. *Biomed. Res. Int.* 2016, 1302914. doi:10.1155/2016/1302914

Zheng, Y., Wang, J., Liang, X., Huang, H., Ma, Y., Lin, L., et al. (2020). Epidemiology, evolutionary origin, and malaria-induced positive selection effects of G6PD-deficient alleles in Chinese populations. *Mol. Genet. Genom. Med.* 8 (12), e1540. doi:10.1002/mgg3.1540

Zhong, Z., Wu, H., Li, B., Li, C., Liu, Z., Yang, M., et al. (2018). Analysis of glucose-6-phosphate dehydrogenase genetic polymorphism in the Hakka population in southern China. *Med. Sci. Monit.* 24, 7316–7321. doi:10.12659/MSM.908402

TYPE Original Research
PUBLISHED 06 March 2023
DOI 10.3389/fgene.2023.1014014

Conditioning on parental mating types can reduce necessary assumptions for Mendelian randomization

Keisuke Ejima[1,2,3], Nianjun Liu[1], Luis Miguel Mestre[1], Gustavo de los Campos[4,5,6] and David B. Allison[1]*

[1]Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN, United States, [2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore, [3]Department of Global Health Policy, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, [4]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, United States, [5]Department of Statistics and Probability, Michigan State University, East Lansing, MI, United States, [6]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

EDITED BY
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

REVIEWED BY
Adrienne Tin,
University of Mississippi Medical Center,
United States
Karl Skorecki,
Bar-Ilan University, Israel

*CORRESPONDENCE
David B. Allison,
✉ allison@iu.edu

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

RECEIVED 08 August 2022
ACCEPTED 17 February 2023
PUBLISHED 06 March 2023

CITATION
Ejima K, Liu N, Mestre LM, de los Campos G and Allison DB (2023), Conditioning on parental mating types can reduce necessary assumptions for Mendelian randomization.
Front. Genet. 14:1014014.
doi: 10.3389/fgene.2023.1014014

Mendelian randomization (MR) has become a common tool used in epidemiological studies. However, when confounding variables are correlated with the instrumental variable (in this case, a genetic/variant/marker), the estimation can remain biased even with MR. We propose conditioning on parental mating types (a function of parental genotypes) in MR to eliminate the need for one set of assumptions, thereby plausibly reducing such bias. We illustrate a situation in which the instrumental variable and confounding variables are correlated using two unlinked diallelic genetic loci: one, an instrumental variable and the other, a confounding variable. Assortative mating or population admixture can create an association between the two unlinked loci, which can violate one of the necessary assumptions for MR. We simulated datasets involving assortative mating and population admixture and analyzed them using three different methods: 1) conventional MR, 2) MR conditioning on parental genotypes, and 3) MR conditioning on parental mating types. We demonstrated that conventional MR leads to type I error rate inflation and biased estimates for cases with assortative mating or population admixtures. In the presence of non-additive effects, MR with an adjustment for parental genotypes only partially reduced the type I error rate inflation and bias. In contrast, conditioning on parental mating types in MR eliminated the type I error inflation and bias under these circumstances. Conditioning on parental mating types is a useful strategy to reduce the burden of assumptions and the potential bias in MR when the correlation between the instrument variable and confounders is due to assortative mating or population stratification but not linkage.

KEYWORDS

Mendelian randomization, genetic epidemiology, causal inference, study design, linkage disequilibrium

Introduction

Randomized experiments, often called randomized controlled trials, are the gold standard for drawing causal inferences. In randomized experiments, observational units (e.g., subjects) are randomly assigned to different levels of the variable being used to assess the causal effect, e.g., the treatment. The randomization process eliminates the influence of potential confounding variables on the exposure variable (e.g., treatment or control). Therefore, we can conclude that the observed difference in outcomes between groups in randomized controlled trials is purely caused by the treatment (barring stochastic variations). However, randomized experiments are not always ethical, feasible, or practical (Sanson-Fisher et al., 2007).

Observational studies do not always yield unbiased estimates of effects because of their lack of random assignment. Of the multiple limitations that these studies have, herein, we will only consider the bias due to confounding.

To mitigate confounding, researchers often include potential confounders in analyses as covariates in regression models or stratify analyses by confounders. Figure 1A depicts a general causal model with an exposure variable (X), an outcome (Y), a confounder (U), and a genetic marker (G), where U is associated with both X and Y and G determines X. The variables X, Y, and U are assumed to be continuous. Causal effects and associations are represented by directional and bidirectional arrows, respectively. If U is observable and is included in the model, the estimate of the effect of X on Y will be unbiased, provided the estimation method does not induce a bias. However, the confounder (U) is not always measurable or known. If U is a set of confounders of the relationship between X and Y and is not appropriately accounted for in the analysis, the estimator of the regression coefficient of Y on X will be biased.

Mendelian randomization (MR) was proposed to address the issue of unmeasured confounders in observational studies (Smith and Ebrahim, 2003; Boutwell and Adams, 2020; Sanderson et al., 2022). MR uses genotypic (G) data from loci that affect the exposure variable (X), do not have a direct effect on the outcome, and are uncorrelated with potential confounders. The most commonly used process of estimation is as follows: 1) X is regressed on G to obtain the predicted value of X, $\hat{X}$; 2) Y is regressed on $\hat{X}$, and then, the estimated coefficient is an unbiased estimator of the effect of X on Y under some assumptions. As a simple and robust approach for causal inference, MR has become common in epidemiological studies during the last few decades.

However, MR rests on three assumptions (Emdin et al., 2017): "1) the genetic variant is associated with the risk factor; 2) the genetic variant is not associated with confounders; and 3) the genetic variant influences the outcome only through the risk factor." In Figure 1A, these assumptions correspond to the following: 1) G and X are associated, 2) there is no association between G and U, and 3) there is no direct effect of G on Y, not through X. If any of the aforementioned three assumptions are violated, the estimated effect is not guaranteed to be unbiased.

Unfortunately, the violation of assumptions, especially the violation of assumption (2), is quite plausible: the genotype (G) can be associated with confounders (U). Even without a direct effect of G on U (or *vice versa*), assortative mating and population stratification can yield associations between them, which violate assumption (2). Furthermore, it is hard to verify this assumption because U includes unmeasurable variables: "The second and third assumptions, however, cannot be empirically proven and require both judgment by the investigators and the performance of various sensitivity analyses" (Emdin et al., 2017). This paper proposes conditioning on parental mating types (defined as a combination of genotypes of parents at a locus used as an instrumental variable (Allison, 1997)) in MR to eliminate the bias in conventional MR, when there is correlation between the instrumental variable and confounding variables. This means that our approach obviates the need for one of the three necessary assumptions in MR.

This paper consists of two parts. First, we demonstrate that the estimation using conventional MR (without conditioning on parental mating types) could lead to biased estimates, when there is a correlation between the instrumental variable and confounding variables due to assortative mating or population stratification. Second, we propose the use of parental mating types in conventional MR and to assess the utility of this approach.

Materials and methods

Mechanisms violating assumptions for MR: Assortative mating and population stratification

First, we define the variables, parameters, and error terms used in the simulation and analyses (summarized in Table 1), with explicit mathematical expressions and causal mechanisms. There are six variables: X, Y, U, G, H, and P. X, Y, and $\bar{U}$ are an exposure variable, an outcome variable, and a confounder, respectively, and all are quantitative traits (thus, continuous variables), such as weight and height. G and H are genotypes defined by SNPs; thus, they are one of the three statuses: $\{AA, Aa, aa\}$ for G and $\{BB, Bb, bb\}$ for H, respectively; $f_A(*)$ and $f_D(*)$ are functions to calculate the additive and dominance effect of a genotype, respectively (f_A counts the number of A [or B] alleles for the genotype; f_D is 1 for a heterozygote and 0 for a homozygote). P is the parental mating type, a combination of genotypes of parents at a locus used as an instrumental variable, and one of the six statuses: $\{AA/AA, AA/Aa, AA/aa, Aa/Aa, Aa/aa, aa/aa\}$. We introduce five indicator functions to compute the genetic effect of parental mating types, $I_{AA/AA}(P), I_{AA/Aa}(P), I_{AA/aa}(P), I_{Aa/Aa}(P), I_{Aa/aa}(P)$, where the function is 1 if P is the same as the subscript of the function and otherwise, 0. The effect of a variable M on a variable N (i.e., the difference in N due to a single unit increase in M) is represented by β_{MN}. It should be noted that the additive effect and the dominance effect of a genotype M on a variable N are represented as $\beta_{M_A N}$ (i.e., difference in N by substituting allele A [or B] for allele a [or b]) and $\beta_{M_D N}$ (i.e., deviance from the average of genotypic values of the two homozygotes), respectively. Furthermore, there are five coefficients to represent the effect of parental mating type P on a variable M using the parental mating type aa/aa as a reference group. Thus, for example, $\beta_{AA/AAM}$ is the unit increase in M for the parental mating type AA/AA compared to the increase in the parental mating type aa/aa. Estimated regression coefficients are distinguished from the causal effect using the

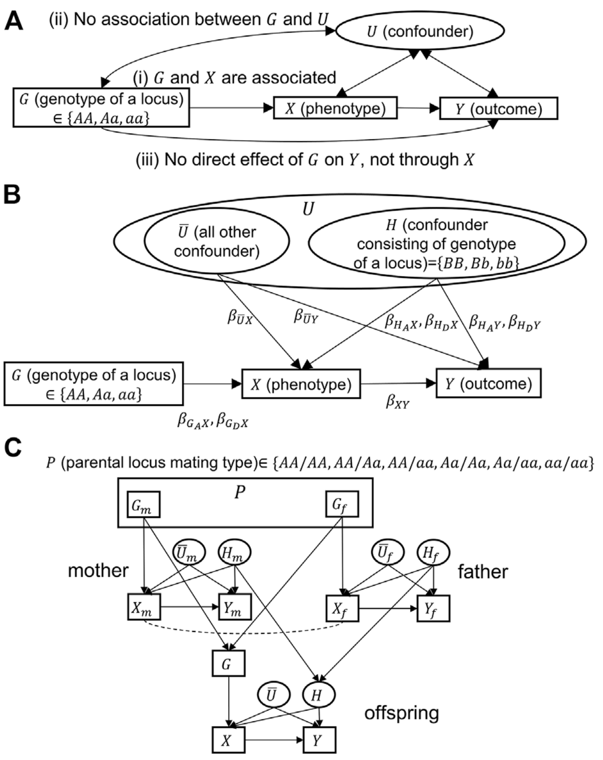

A
(ii) No association between G and U
U (confounder)
(i) G and X are associated
G (genotype of a locus) $\in \{AA, Aa, aa\}$
X (phenotype)
Y (outcome)
(iii) No direct effect of G on Y, not through X
B
U
$\bar{U}$ (all other confounder)
H (confounder consisting of genotype of a locus)$=\{BB, Bb, bb\}$
$\beta_{\bar{U}X}$
$\beta_{\bar{U}Y}$
$\beta_{H_AX}, \beta_{H_DX}$
$\beta_{H_AY}, \beta_{H_DY}$
G (genotype of a locus) $\in \{AA, Aa, aa\}$
X (phenotype)
Y (outcome)
β_{XY}
$\beta_{G_AX}, \beta_{G_DX}$
C
P (parental locus mating type)$\in \{AA/AA, AA/Aa, AA/aa, Aa/Aa, Aa/aa, aa/aa\}$
P
G_m
G_f
$\bar{U}_m$
H_m
$\bar{U}_f$
H_f
mother
father
X_m
Y_m
X_f
Y_f
G
$\bar{U}$
H
offspring
X
Y

FIGURE 1
Causal models for Mendelian randomization. Directional and bidirectional arrows correspond to causal and associational relationships, respectively. βs are regression coefficients. Variables in rectangles and ovals correspond to measurable or unmeasurable variables, respectively. **(A)** Generalized model for Mendelian randomization (MR) with three assumptions. **(B)** Explicit causal model separating confounding variables, U, into the variable consisting of unlinked heritable variants in the nuclear genome, H, and all other confounding variables, $\bar{U}$. **(C)** Explicit causal model for the father−mother−offspring trio. The parental mating type, P, is the combination of parental genotypes (G_m and G_f), which takes one of the six possible values. The dotted line connecting X_m and X_f implies assortative mating of these variables.

following: $\hat{\beta}_{MN}$ is the regression coefficient estimated by regressing N on M.

Figure 1B is a causal model in which the confounding variable set, U, is separated into two sets of variables: one set includes a confounding variable consisting of a genotype on a single biallelic locus (H) with two alleles B and b, and the second set $\bar{U}$ consists of all other confounders. We note that in our scenario, H and the genotype on the biallelic locus, used as an instrumental variable (G), are unlinked. However, G and H could be correlated (i.e., non-linkage disequilibrium), which would violate assumption (2). The following describes two situations, assortative mating and population stratification, which can cause such a non-linkage disequilibrium.

Situation 1: Assortative mating

In human and other animal populations, the choice of a mate does not plausibly occur at random. One may be more likely to mate with another who has specific phenotypes, resulting in non-random or assortative mating (Anonymous, 1903). For example, assortative mating for body mass index (BMI) or body fatness (i.e., individuals with a high BMI or body fatness are more likely to mate with one another, as are individuals with low BMI or body fatness) is widely observed (Allison et al., 1996; Silventoinen et al., 2003; Jackson et al., 2007). We modeled assortative mating as being dependent on the exposure variable X. Mothers and fathers are separately sorted by X, and they are paired according to the order. For this purpose, each parent's genotype, exposure, outcome, and confounders are explicitly modeled. Variables are given with one of the two subscripts, m or f, for either the mother or the father (variables without these subscripts are for an offspring). The model is summarized in Figure 1C.

Briefly, the correlation between G and H is explained as follows: Assortative mating on X (i.e., X_m and X_f) induces associations between G_m and H_f and G_f and H_m, which result in an association between G and H, thus violating the MR assumption (2).

Situation 2: Population stratification

Population stratification occurs and can create genotype–phenotype associations in the absence of linkage or a causal effect of the specific genotype on the specific phenotype, when a population consists of multiple subpopulations (Freedman et al., 2004) and some subpopulations have different allele frequencies and phenotypic distributions. By using the framework given in Figure 1C without assortative mating, we assume two different subpopulations. Therefore, within each subpopulation, three assumptions are held for conventional MR. The difference between the two populations is that they have different allele frequencies. If data from the two subpopulations were analyzed as a single population without accounting for the population substructure, they would yield a spurious association between G and H (because all parental loci [G_m, H_m, G_f, and H_f] are associated), which violates the MR assumption (2).

Correcting the bias in MR: Conditioning on parental mating types

Assuming the aforementioned two situations, the conventional MR estimation procedure can lead to biased estimates because MR assumption (2) is violated. To eliminate the bias, we propose conditioning on the parental mating type P, which is a combination of parental genotypes used for the instrumental variable in MR. The rationale for using P is that both G_m and G_f are located on open (i.e., d connected) paths between genotypes G and H in both situations 1 and 2, and conditioning on P blocks the path. We also follow the approach of using parental *genotypes* instead of *mating types*, as proposed by Hartwig et al. (2018), which is another reference method.

In the following, we show details of three methods: conventional MR, MR conditioning on parental genotypes [a method proposed by Hartwig et al. (2018)], and MR conditioning on parental mating types (which we propose in this study). It should be noted that unmeasurable variables (variables in ovals, given in Figure 1C) do not appear in any of the analyses.

Conventional MR

1) Conventional MR uses the following model: $X = \beta_1 + \beta_{G_AX} f_A(G) + \beta_{G_DX} f_D(G) + \varepsilon_X$, where ε_X is an error

TABLE 1 Summary of variables, functions, and intercepts in regression models.

Parameter	Description
X	Exposure variable
Y	Outcome
G	Genotype of a locus with effects on X $(G \in \{AA, Aa, aa\})$
H	Confounder consisting of a genotype of a locus with effects on X and Y $(H \in \{BB, Bb, bb\})$
$\bar{U}$	All other (non-genetic) confounders (with effects on X and Y)
P	Parental mating type on G $(P \in \{AA/AA, AA/Aa, AA/aa, Aa/Aa, Aa/aa, aa/aa\})$
Function for genetic effects	**Description**
$f_A(M)$	Function to compute the additive effect of genotype M (counting the number of A or B)
$f_D(M)$	Function to compute the dominance effect of genotype M (1 for a heterozygote and 0 for a homozygote)
$I_{AA/AA}(P)$	Indicator function (1 for the parental mating type AA/AA and 0 for the other)
$I_{AA/Aa}(P)$	Indicator function (1 for the parental mating type AA/Aa and 0 for the other)
$I_{AA/aa}(P)$	Indicator function (1 for the parental mating type AA/aa and 0 for the other)
$I_{Aa/Aa}(P)$	Indicator function (1 for the parental mating type Aa/Aa and 0 for the other)
$I_{Aa/aa}(P)$	Indicator function (1 for the parental mating type Aa/aa and 0 for the other)
$I_{aa/aa}(P)$	Indicator function (1 for the parental mating type aa/aa and 0 for the other)
Intercept in the regression model	**Description**
β_1	Genotypic value of aa on X
β_2	Genotypic value of bb on Y when both X and $\bar{U}$ are zero

term. Therefore, the auxiliary regression of X on $f_A(G)$ and $f_D(G)$ is performed to obtain the estimated value of X $(= \hat{X})$: $\hat{X} = \hat{\beta}_1 + \hat{\beta}_{G_A X} f_A(G) + \hat{\beta}_{G_D X} f_D(G)$.

2) Then, the regression of Y on $\hat{X}$ is conducted by assuming the following model with an error term ε_Y: $Y = \beta_2 + \beta_{XY}\hat{X} + \varepsilon_Y$.

MR conditioning on parental genotypes

To correct for the bias in MR, Hartwig et al. (2018) proposed conditioning on parental genotypes. The analysis proceeds as follows:

1) MR conditioning on parental genotypes uses the following model: $X = \beta_1 + \beta_{G_A X} f_A(G) + \beta_{G_D X} f_D(G) + \beta_{G_{mA} X} f_A(G_m) + \beta_{G_{mD} X} f_D(G_m) + \beta_{G_{fA} X} f_A(G_f) + \beta_{G_{fD} X} f_D(G_f) + \varepsilon_X$. Therefore, the auxiliary regression of X on $f_A(G)$ and $f_D(G)$ conditioning on $f_A(G_m)$, $f_D(G_m)$, $f_A(G_f)$, and $f_D(G_f)$ are performed to obtain the estimated value of X $(= \hat{X})$: $\hat{X} = \hat{\beta}_1 + \hat{\beta}_{G_A X} f_A(G) + \hat{\beta}_{G_D X} f_D(G) + \hat{\beta}_{G_{mA} X} f_A(G_m) + \hat{\beta}_{G_{mD} X} f_D(G_m) + \hat{\beta}_{G_{fA} X} f_A(G_f) + \hat{\beta}_{G_{fD} X} f_D(G_f)$.
2) The regression of Y on $\hat{X}$ is conducted assuming the following model: $Y = \beta_2 + \beta_{XY}\hat{X} + \beta_{G_{mA} Y} f_A(G_m) + \beta_{G_{mD} Y} f_D(G_m) + \beta_{G_{fA} Y} f_A(G_f) + \beta_{G_{fD} Y} f_D(G_f) + \varepsilon_Y$.

MR conditioning on parental mating types

Hartwig et al. (2018) assumed an additive model and, thus, used a *parental genotype* as an instrumental variable. However, if the effect of the parental genotype on an offspring's phenotype is non-additive, using a parental mating type, i.e., a combination of parental genotypes taking one of the six possible values (Figure 1C), is more appropriate. The corresponding analysis proceeds as follows:

1) MR conditioning on parental mating types uses the following model: $X = \beta_1 + \beta_{G_A X} f_A(G) + \beta_{G_D X} f_D(G) + \beta_{AA/AAX} I_{AA/AA}(P) + \beta_{AA/AaX} I_{AA/Aa}(P) + \beta_{AA/aaX} I_{AA/aa}(P), + \beta_{Aa/AaX} I_{Aa/Aa}(P) + \beta_{Aa/aaX} I_{Aa/aa}(P) + \varepsilon_X$. Therefore, the auxiliary regression of X on G_1 conditioning on the parental mating type P is performed to obtain the estimated value of X $(= \hat{X})$: $\hat{X} = \hat{\beta}_1 + \hat{\beta}_{G_A X} f_A(G) + \hat{\beta}_{G_D X} f_D(G) + \hat{\beta}_{AA/AAX} I_{AA/AA}(P) + \hat{\beta}_{AA/AaX} I_{AA/Aa}(P) + \hat{\beta}_{AA/aaX} I_{AA/aa}(P), + \hat{\beta}_{Aa/AaX} I_{Aa/Aa}(P) + \hat{\beta}_{Aa/aaX} I_{Aa/aa}(P)$.
2) The regression of Y on $\hat{X}$ is conducted by assuming the following model: $Y = \beta_2 + \beta_{XY}\hat{X} + \hat{\beta}_{AA/AAX} I_{AA/AA}(P) + \hat{\beta}_{AA/AaX} I_{AA/Aa}(P) + \hat{\beta}_{AA/aaX} I_{AA/aa}(P), + \hat{\beta}_{Aa/AaX} I_{Aa/Aa}(P) + \hat{\beta}_{Aa/aaX} I_{Aa/aa}(P) + \varepsilon_Y$.

Simulations

To demonstrate the potential bias when conventional MR is used due to the violation of the MR assumption (2) and the utility of using parental mating types to eliminate the bias, we performed simulations considering assortative mating and population stratification.

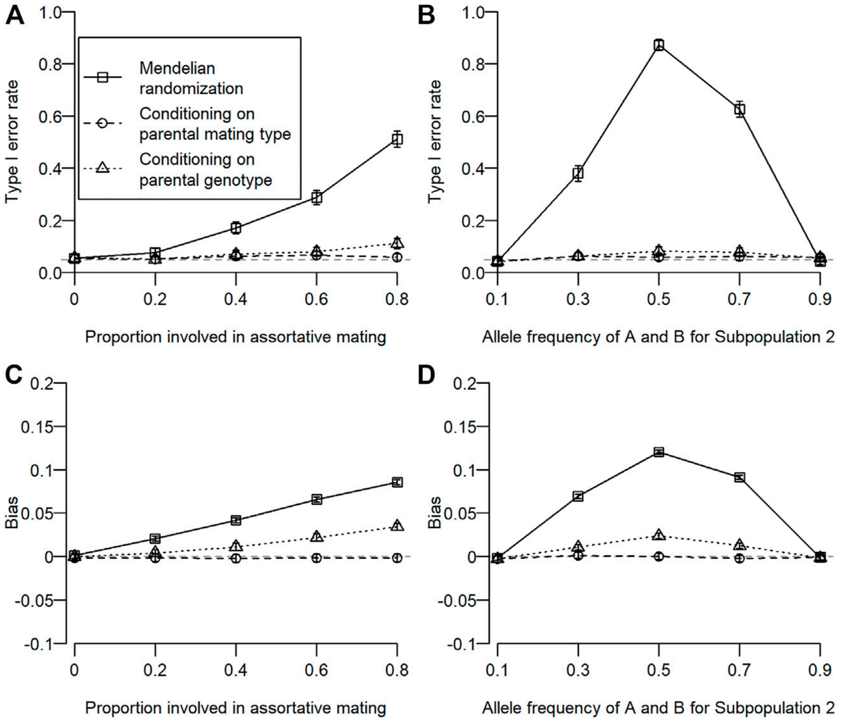


FIGURE 2
Type I error rate and bias of estimated coefficients for three different types of MR. Open squares, open circles, and open triangles correspond to conventional MR, MR conditioning on parental mating types, and MR conditioning on parental genotypes, respectively. For simulation 1, the proportion of the population involved in assortative mating was changed from 0 to 0.8. For simulation 2, allele frequencies of *A* and *B* for subpopulation 2 were varied from 10% to 90%. **(A, B)** Type I error rates for simulations 1 and 2. Gray dotted lines are significance levels (= 0.05). **(C, D)** Bias in the estimated regression coefficient of an offspring's outcome on exposure ($\hat{\beta}_{XY} - \beta_{XY}$) for simulations 1 and 2.

For the simulation of each situation, we created data for 1,000 trio (father–mother–offspring) families (500 trios each for the population for situation 2) for a single simulation and performed three different analyses on each dataset. We repeated the process 1,000 times for each parameter setting. The type I error rate (when $\beta_{XY} = 0$) is defined as the proportion of simulations in which the estimated association between X and Y is statistically significant (false-positive finding). The bias in the estimated coefficient $E[\hat{\beta}_{XY} - \beta_{XY}]$ is also assessed when $\beta_{XY} > 0$. The coefficient β_{XY} was set as 1.0 for bias assessment. The sensitivity of the type I error rate and the bias on the magnitude of the violation of MR assumption (2) were assessed by varying the parameters. The significance level was set as 0.05. The process for generating data and analyses are described in the next section.

Simulation 1: Assortative mating

The following is a step-by-step protocol and parameter setting for the simulation:

1) Allele frequencies of A and B are 10% for each: $Prob(A) = Prob(B) = 0.1, Prob(a) = Prob(b) = 0.9.$ Each parent's genotypes (G_m, H_m, G_f, and H_f) are determined assuming the Hardy–Weinberg equilibrium (Hardy, 1908). It should be noted that G and H are independent.
2) The confounding variables for parents, $\bar{U}_m$ and $\bar{U}_f$, are determined, which follow a bivariate normal distribution: $N(0, 0.1)$.
3) The exposure variables of parents, X_m and X_f, are determined by their genotype and confounding variable: $X_m = \beta_1 + \beta_{G_A X} f_A(G_m) + \beta_{G_D X} f_D(G_m) + \beta_{\bar{U}X}\bar{U}_m + \varepsilon_X$, where $\varepsilon_X \sim N(0, 0.1)$. β_1 is interpreted as the genotypic effect of the genotype aa on X. X_f is determined in the same way as X_m.
4) The outcome of parents, Y_m and Y_f, are determined by their exposure, genotype, and confounding variable: $Y_m = \beta_2 + \beta_{XY}X_m + \beta_{H_A Y} f_A(H_m) + \beta_{H_D Y} f_D(H_m) + \beta_{\bar{U}Y}\bar{U}_m + \varepsilon_Y$, where $\varepsilon_Y \sim N(0, 0.1)$. β_2 is interpreted as the genotypic effect of the genotype bb on Y, when both X and $\bar{U}$ are zero. Y_f is determined in the same way as Y_m.
5) Proportion p is selected from paternal and maternal populations. In the selected population, both parents are sorted separately by the exposure X_m or X_f and are paired according to the order of X_m and X_f. Unselected parents (1-p) are randomly coupled regardless of the values of X and Y.
6) The genotype of the offspring, G and H, are determined by randomly selecting an allele from each parent.
7) The exposure, X, and the outcome, Y, of the offspring are determined by following the same process as for the parents (see 3 and 4).

The sensitivity of the type I error rate and bias was assessed by changing p from 0.0 to 0.8. All effects from $\bar{U}$ to X and Y are assumed to be 1. For genetic effects, we assumed that there is no additive effect ($\beta_{G_A X} = \beta_{H_A X} = \beta_{H_A Y} = 0$), but there is a strong dominance effect ($\beta_{G_D X} = \beta_{H_D X} = \beta_{H_D Y} = 1$) of G and H on any associated variables.

Simulation 2: Population stratification

The simulation setting for simulation 2 is similar to simulation 1 save for a couple of differences: 1) no assortative mating and 2) we assume two populations (i.e., subpopulation 1 and subpopulation 2) with different allele frequencies. Allele frequencies of A and B for subpopulation 1 are 10% each. Otherwise, all simulation settings, including parameter settings, are the same as those in simulation 1. The source of the violation of MR assumption (2) is different allele frequencies. To demonstrate the sensitivity of the type I error rate and bias on the magnitude of the violation of MR assumption (2), allele frequencies of A and B for subpopulation 2 were varied from 10% to 90%. All simulations and analyses were performed using statistical computing software R (version 3.6.1).

Results

The type I error rate for simulation 1 is shown in Figure 2A. Type I error inflation was observed for conventional MR and MR conditioning on parental genotypes, and it increased as the proportion involved in assortative mating increased. Type I error inflation was not observed for MR conditioning on parental mating types. Type I error inflation was mitigated by conditioning on parental genotypes to some extent, which still remained. The type I error rate for simulation 2 is shown in Figure 2B. Type I error inflation was observed for both conventional MR and MR conditioning on parental genotypes but not for MR conditioning on parental mating types. As shown in simulation 1, conditioning on parental genotypes reduced but did not eliminate type I error rate inflation. Interestingly, we observed a large type I error inflation when allele frequencies for the subpopulation were intermediate (0.5). This is because we assumed that homozygous genotypes (i.e., AA, aa and BB, bb) have the same effect on phenotypes (X and Y).

The bias of the estimated coefficient is shown in Figures 2C, D. We observed similar results for the bias in estimation as in type I error rates. When type I error rate inflation was observed, a statistically significant bias was also observed, and magnitudes of type I error rate inflation and absolute bias were positively associated.

Discussion

MR has become a common approach for causal inference in epidemiology, as genetic data become more accessible owing to fast and efficient DNA sequencing technology and as journals and funding bodies encourage data sharing (Levey et al., 2009; Bloom et al., 2014; Loder and Groves, 2015). However, as for most epidemiological approaches, MR has essential assumptions we need to check before performing analysis. Among them, the assumption of no association between genetic variants used in MR and confounders [MR assumption (2)] could be violated or is difficult to check in practice. First, we demonstrated that MR produces inflation in type I error rates and a biased estimation in realistic settings where the assumption is violated. We introduced two plausible situations: assortative mating and population stratification. The sensitivity of type I error rates and estimation bias was assessed by changing parameters relevant to the violation of the MR assumption. As expected, we observed type I error inflation and estimation bias in these realistic settings when conventional MR was used, and such inflations and biases worsened as violations became more severe. They were mitigated by conditioning on parental genotypes to some extent; however, type I error inflation remained. Second, we proposed the use of parental mating types for a valid association inference for these two situations. We successfully confirmed that conditioning on parental mating types solves the problem in both situations.

We noted that we are not the first to propose the idea of considering parental genetic information in an epidemiological study. The idea was originally proposed in testing for linkages in the presence of associations (Allison, 1997). Redden et al. suggested using parental mating types in the inference of genotype–phenotype associations (Redden and Allison, 2006). Later, Liu et al. (2015) extended the idea to testing causal effects of a fetal drive. In this work, they showed the relationship between this idea and MR. In MR, the genetic variant needs to be a causal variant. However, it may be difficult to verify this assumption in practice, if not impossible. Conditioning on parental mating types is one way to identify causal genetic variants, thus relaxing assumptions, specifically assumption (2) of MR (resulting in the strengthening of MR). In the context of MR, Hartwig et al. proposed using parental genotypes in the case of assortative mating, which violates MR assumption (3) (Hartwig et al., 2018). They proposed two methods to integrate parental genotypes in MR analyses. The first method is to adjust conventional MR by parental allele scores, which we used in this study. The second method is to use parental non-transmitted allele scores and the offspring allele score as instrumental variables of parental and offspring exposure variables. They demonstrated that both methods provide unbiased estimates of the exposure–outcome association and avoid type I error inflation even under strong assortative mating conditions. The difference between the study by Hartwig et al. and ours is that we assumed that the locus influencing the outcome (H) also influences the exposure (X). Therefore, their model is considered a special case of ours. Although Hartwig et al. (2018) concluded that only cross-trait assortative mating (between X and Y) yields a bias, we found that same-trait assortative mating (between Xs or between Ys) can also yield a bias due to the heritable confounding variable (H). Furthermore, we found that conditioning of parental genotypes is not enough to control the bias if effects of alleles on phenotypes are non-additive. In our previous work (Liu et al., 2015), we indicated that random mating is not assumed with conditioning on parental mating types. We also explained that it is necessary to condition on parental mating types to achieve randomization, which is the basis for causal inference. Further insights into the rationale for this or other ways of expressing fundamental ideas can be found in the study by Pearl et al. (2016).

We list a few limitations of our approach. One apparent limitation is the data availability. Most genetic epidemiological research studies do not have (or is not designed to collect) parental genetic data (i.e., mother–father–offspring). However, because family trio data collection is considered to be a powerful tool for identifying rare diseases, even outside the context of MR, and owing to technological advancements in gene sequencing, the collection of family trio data may become more common (Infante-Rivard et al., 2009). In a recent study, Young et al. proposed imputing parental genotypes to reduce biases in GWA studies (Young et al., 2022). The imputation strategy presented in this study provides an opportunity to implement methods we proposed here for MR in situations where parental genotypes are not directly available. In this work, we propose that conditioning on parental genetic mating types can reduce assumptions needed for MR. We illustrate this key principle using a simulation study involving one locus with dominance effects. However, the approach we propose is general and does not require dominance effects. Indeed, our approach will also work under an additive model because the additive model is a special case of the more general model we use for conditioning. However, if the mode of action of the locus is strictly additive, conditioning on a parental allele dosage may be enough to reduce the bias. Therefore, in future studies, we plan to assess the superiority of conditioning on parental mating types relative to conditioning on allele dosages. Furthermore, we plan to assess the principle we proposed in a broader range of realistic circumstances. We are, particularly, interested in investigating two situations. The first is to evaluate the performance of the proposed approach in a multi-locus context for models involving epistatic interactions, which seem common (Zhu et al., 2015). The second situation is one where there is a selection bias on the exposure, X. Since X is a collider of $G, H, \text{and} \bar{U}$, if a subpopulation was sampled according to X (people with X higher than the threshold, for example), spurious correlations among $G, H, \text{and} \bar{U}$ might occur. In this case, conditioning on parental genetic mating types can account for the correlation between $G \,\text{and}\, H$ but not for the correlation between $G \,\text{and}\, \bar{U}$ because $\bar{U}$ is not a heritable variable.

However, regardless of the limitations suggested previously, conditioning on parental mating types in MR can strengthen assumptions and help avoid type I error inflation and bias, when a heritable confounding variable is associated with the instrumental variable in MR.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in: Zenodo (doi: 10.5281/zenodo.6972710).

Author contributions

DA conceived the research idea. KE, LM, GC, and NL implemented the simulation and performed the analyses. DA and GC oversaw data analyses. All authors were involved in the writing of the manuscript and gave final approval of submitted and published versions.

Funding

This work was supported in part by the National Institutes of Health (Grant numbers R25HL124208, R25DK099080, and R01AG057703 to DA; Grant numbers R03DE024198 and R03DE025646 to NL), the National Science Foundation (Grant number DMS 2002865 to NL), and the Japan Society for Promotion of Science KAKENHI (Grant number 18K18146 to KE).

Acknowledgments

The authors would like to thank Brian Boutwell (the University of Mississippi) for invaluable comments.

Conflict of Interest

DA and his institutions (Indiana University and the Indiana University Foundation) have received consulting fees, donations, grants, and contracts or promises for the same, from numerous not-for-profit, for-profit (including food, pharmaceutical, litigation, dietary supplement, and other entities), and government organizations with interests in health, causal inference, genetics, and statistics; however, none of these could reasonably be taken to represent a conflict of interest with this statistical methodology paper.

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Allison, D. B., Neale, M. C., Kezis, M. I., Alfonso, V. C., Heshka, S., and Heymsfield, S. B. (1996). Assortative mating for relative weight: Genetic implications. *Behav. Genet.* 26 (2), 103–111. doi:10.1007/BF02359888

Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* 60 (3), 676–690.

Anonymous (1903). Assortative mating in man: A cooperative study. *Biometrika* 2 (4), 481–498. doi:10.1093/biomet/2.4.481

Bloom, T., Ganley, E., and Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS Biol.* 12 (2), e1001797. doi:10.1371/journal.pbio.1001797

Boutwell, B. B., and Adams, C. D. (2020). A research note on mendelian randomization and causal inference in criminology: Promises and considerations. *J. Exp. Criminol.* 18, 171–182. doi:10.1007/s11292-020-09436-9

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *JAMA* 318 (19), 1925–1926. doi:10.1001/jama.2017.17219

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., et al. (2004). Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* 36 (4), 388–393. doi:10.1038/ng1333

Hardy, G. H. (1908). Mendelian propositions in a mixed population. *Science* 28 (706), 49–50. doi:10.1126/science.28.706.49

Hartwig, F. P., Davies, N. M., and Davey Smith, G. (2018). Bias in Mendelian randomization due to assortative mating. *Genet. Epidemiol.* 42 (7), 608–620. doi:10.1002/gepi.22138

Infante-Rivard, C., Mirea, L., and Bull, S. B. (2009). Combining case-control and case-trio data from the same population in genetic association analyses: Overview of approaches and illustration with a candidate gene study. *Am. J. Epidemiol.* 170 (5), 657–664. doi:10.1093/aje/kwp180

Jackson, D. M., Stewart, J., Djafarian, K., and Speakman, J. R. (2007). Assortative mating for obesity. *Am. J. Clin. Nutr.* 86 (2), 316–323. doi:10.1093/ajcn/86.2.316

Levey, A., Stevens, L., Schmid, C., Zhang, Y., Castro, A., Feldman, H., et al. (2009). A new equation to estimate glomerular filtration rate. *Ann. Intern Med.* 150 (9), 604–612. doi:10.7326/0003-4819-150-9-200905050-00006

Liu, N., Archer, E., Srinivasasainagendra, V., and Allison, D. B. (2015). A statistical framework for testing the causal effects of fetal drive. *Front. Genet.* 5, 464. doi:10.3389/fgene.2014.00464

Loder, E., and Groves, T. (2015). The BMJ requires data sharing on request for all trials. *BMJ* 350, h2373. doi:10.1136/bmj.h2373

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. New York, NY: John Wiley & Sons.

Redden, D. T., and Allison, D. B. (2006). The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture. *Behav. Genet.* 36 (5), 678–686. doi:10.1007/s10519-006-9060-0

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., et al. (2022). Mendelian randomization. *Nat. Rev. Methods Prim.* 2 (1), 6–21. doi:10.1038/s43586-021-00092-5

Sanson-Fisher, R. W., Bonevski, B., Green, L. W., and D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am. J. Prev. Med.* 33 (2), 155–161. doi:10.1016/j.amepre.2007.04.007

Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J., and Rose, R. J. (2003). Assortative mating by body height and BMI: Finnish twins and their spouses. *Am. J. Hum. Biol.* 15 (5), 620–627. doi:10.1002/ajhb.10183

Smith, G. D., and Ebrahim, S. (2003). Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32 (1), 1–22. doi:10.1093/ije/dyg070

Young, A. I., Nehzati, S. M., Benonisdottir, S., Okbay, A., Jayashankar, H., Lee, C., et al. (2022). Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* 54 (6), 897–905. doi:10.1038/s41588-022-01085-0

Zhu, Z., Bakshi, A., Vinkhuyzen, A. A., Hemani, G., Lee, S. H., Nolte, I. M., et al. (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am. J. Hum. Genet.* 96 (3), 377–385. doi:10.1016/j.ajhg.2015.01.001

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fgene.2023.1104727

OPEN ACCESS

EDITED BY
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

REVIEWED BY
Yuehua Cui,
Michigan State University, United States
Himel Mallick,
Merck, United States
Zakariya Yahya Algamal,
University of Mosul, Iraq

*CORRESPONDENCE
Swati Biswas,
✉ swati.biswas@utdallas.edu

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

RECEIVED 21 November 2022
ACCEPTED 21 February 2023
PUBLISHED 09 March 2023

CITATION
Sajal IH and Biswas S (2023), Bivariate quantitative Bayesian LASSO for detecting association of rare haplotypes with two correlated continuous phenotypes.
Front. Genet. 14:1104727.
doi: 10.3389/fgene.2023.1104727

Bivariate quantitative Bayesian LASSO for detecting association of rare haplotypes with two correlated continuous phenotypes

Ibrahim Hossain Sajal and Swati Biswas*

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, United States

In genetic association studies, the multivariate analysis of correlated phenotypes offers statistical and biological advantages compared to analyzing one phenotype at a time. The joint analysis utilizes additional information contained in the correlation and avoids multiple testing. It also provides an opportunity to investigate and understand shared genetic mechanisms of multiple phenotypes. Bivariate logistic Bayesian LASSO (LBL) was proposed earlier to detect rare haplotypes associated with two binary phenotypes or one binary and one continuous phenotype jointly. There is currently no haplotype association test available that can handle multiple continuous phenotypes. In this study, by employing the framework of bivariate LBL, we propose bivariate quantitative Bayesian LASSO (QBL) to detect rare haplotypes associated with two continuous phenotypes. Bivariate QBL removes unassociated haplotypes by regularizing the regression coefficients and utilizing a latent variable to model correlation between two phenotypes. We carry out extensive simulations to investigate the performance of bivariate QBL and compare it with that of a standard (univariate) haplotype association test, Haplo.score (applied twice to two phenotypes individually). Bivariate QBL performs better than Haplo.score in all simulations with varying degrees of power gain. We analyze Genetic Analysis Workshop 19 exome sequencing data on systolic and diastolic blood pressures and detect several rare haplotypes associated with the two phenotypes.

KEYWORDS

diastolic blood pressure (DBP), Genetic Analysis Workshop 19, Markov chain Monte Carlo (MCMC), regularization, systolic blood pressure (SBP)

1 Introduction

Information on multiple phenotypes is often collected in health-related studies to obtain a bigger picture of patients' health conditions (Teixeira-Pinto and Normand, 2009). Studies have found variants at numerous genetic loci to be associated with these phenotypes (Solovieff et al., 2013). Sometimes, a genetic variant is associated with more than one phenotype, a phenomenon known as pleiotropy. Recent studies have confirmed the widespread presence of pleiotropy in the human genome, thus showing the underlying common genetic mechanisms of numerous traits (Solovieff et al., 2013; Gratten and Visscher, 2016; Buniello et al., 2019; Watanabe et al., 2019). Investigating and understanding pleiotropy can uncover additional associations, redefine disease classification, and expand our understanding of the genetic basis of complex diseases with wide-ranging implications for healthcare (Hackinger and Zeggini, 2017; Lee et al., 2019; Lee et al., 2021).

The most common way of the testing trait–variant association is to consider one phenotypic trait at a time and test its association with genotypic variants under study. However, such a univariate statistical approach ignores valuable additional information contained in the joint distribution of the phenotypes. Even more importantly, such an approach amounts to a lost opportunity to investigate potential pleiotropy and shared genetic mechanisms. It may also result in a loss of power, especially with multiplicity adjustment, for performing multiple univariate tests. Therefore, considering a multivariate framework to model the phenotypes jointly is appealing from both biological and statistical perspectives.

Several methods have been proposed that utilize a multivariate framework to jointly model multiple correlated phenotypes, including some recent gene-based approaches (Klei et al., 2008; O'Reilly et al., 2012; Van der Sluis et al., 2015; Ray et al., 2016; Hackinger and Zeggini, 2017; Kaakinen et al., 2017; Lee et al., 2017; Ray and Basu, 2017; Deng et al., 2020). However, most of these studies consider single-nucleotide polymorphisms (SNPs) or variants (SNVs) as a genetic unit obtained from genome-wide association studies (GWAS) or next-generation sequencing (NGS) studies. Thus, when rare variants are of interest, one has to rely on SNVs obtained from NGS as rare SNPs are not usually genotyped in GWAS. Yet, most NGS data lack the adequate sample size required for multivariate analysis of correlated phenotypes. Hence, an alternative approach to multiple trait–rare variant association tests that does not necessarily rely on NGS data is warranted.

Haplotype-based tests are powerful alternatives to SNP-based genetic association tests (Bader, 2001; Wang and Lin, 2015). Haplotypes are more biologically meaningful genetic variants as compared to SNPs, which are not inherited independently. Moreover, common SNPs can make up a rare haplotype in a haplotype block, providing avenues to investigate the common disease rare variant (CDRV) hypothesis. Thus, rare variants can also be investigated using GWAS data through haplotype-based tests, allowing the use of data from much larger sample sizes than those of NGS. Several tests have been proposed to investigate the CDRV hypothesis through haplotype-based tests (Guo and Lin, 2009; Li et al., 2010; Li et al., 2011; Biswas and Lin, 2012; Lin et al., 2013), among which logistic Bayesian LASSO (LBL) is a well-studied and powerful method (Biswas and Lin, 2012; Biswas and Papachristou, 2014; Datta and Biswas, 2016; Papachristou and Biswas, 2020). LBL was extended to incorporate gene–environment interactions (Zhang et al., 2017a; Zhang et al., 2017b; Papachristou and Biswas, 2020), data generated using complex sampling designs (Zhang et al., 2017a), and family data (Wang and Lin, 2014; Datta et al., 2018). LBL was also adapted to accommodate two phenotypes, namely, bivariate LBL-2B for binary phenotypes and bivariate LBL-BC for binary and continuous phenotypes (Yuan and Biswas, 2019; Yuan and Biswas, 2021). LBL and its extensions utilize regularization to decrease the unassociated effects close to zero, which, in turn, helps the effect of an associated haplotype, especially if it is a rare one, to stand out. Bivariate LBL-2B and LBL-BC model the dependency between two phenotypes *via* a latent variable. Notably, there is another haplotype-based bivariate genetic association test for correlated quantitative traits; it uses the haplotype trend regression approach (Pei et al., 2009). However, it is only applicable for testing associations with common haplotypes and hence cannot be used for the CDRV hypothesis.

There is no haplotype-based association test currently available that can detect rare haplotypes associated with multiple quantitative phenotypes jointly. To fill this gap, we propose a new method, bivariate quantitative Bayesian LASSO (QBL) to jointly model two correlated continuous phenotypes. We borrow the well-studied framework of bivariate LBL and make appropriate modifications to accommodate quantitative traits. The properties of bivariate QBL are investigated using extensive simulations under various association scenarios, sample sizes, and the number of haplotypes. We also compare its performance to a standard univariate haplotype-based association test, Haplo.score (Schaid et al., 2002). Finally, we apply our proposed method to exome sequencing data from Genetic Analysis Workshop (GAW) 19. We analyze haplotype blocks in several genes of interest (as per literature) and detect rare haplotypes associated with systolic and diastolic blood pressures (SBP and DBP) jointly.

2 Methods

2.1 Likelihood formulation

We closely follow the framework of bivariate LBL-2B and LBL-BC and accordingly the notations used therein. Consider a sample of n subjects with two continuous correlated (standardized) phenotypes denoted by Y_{ic} and $Y_{ic'}$. Let $\boldsymbol{Y_c} = (Y_{1c,}Y_{2c,\ldots,}Y_{nc})$, $\boldsymbol{Y_{c'}} = (Y_{1c'},Y_{2c',\ldots,}Y_{nc'})$, and $\boldsymbol{G} = (G_1, G_2, \ldots, G_n)$, where G_i represents the i^{th} individual's observed genotype on the SNPs, making up the haplotype block under study. Furthermore, let $S(G_i)$ be the set of haplotype pairs compatible with G_i as the haplotype pair for an individual may not be unambiguously determined from the genotype data; Z_{ir} denotes the r^{th} element of $S(G_i)$. We introduce a latent variable u_i to model the marginal dependence between Y_{ic} and $Y_{ic'}$. Let $u_i \sim N(0, \sigma_u^2)$ for all i and $\boldsymbol{u} = (u_1, u_{2,\ldots,}u_n)$. We assume that although Y_{ic} and $Y_{ic'}$ are marginally dependent, they are conditionally independent, given u_i. In other words, the latent variable induces conditional independence between the two correlated outcomes. We also assume that Z_{ir} is independent of u_i. The likelihood can be written as

$$L(\psi) = \prod_{i=1}^{n} \sum_{Z_{ir}\epsilon S(G_i)} P(Y_{ic}, Y_{ic'}, Z_{ir}, u_i)$$

$$\propto \prod_{i=1}^{n} \sum_{Z_{ir}\epsilon S(G_i)} P\left(Y_{ic}, Y_{ic'} \middle| Z_{ir}, u_i\right) P(Z_{ir}, u_i)$$

$$\propto \prod_{i=1}^{n} \sum_{Z_{ir}\epsilon S(G_i)} P(Y_{ic}|Z_{ir}, u_i) P\left(Y_{ic'} \middle| Z_{ir}, u_i\right) P(Z_{ir}) P(u_i),$$

where ψ is the vector of model parameters, which includes regression coefficients, variance parameters, and parameters associated with haplotype frequencies (to be introduced soon). Notably, bivariate QBL does not require specification of the

TABLE 1 Haplotype settings and association scenarios (the effect of target haplotype is shown in boldface).

Setting	Hap	Freq	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
			β_c	$\beta_{c'}$	β_c	$\beta_{c'}$	β_c	$\beta_{c'}$	β_c	$\beta_{c'}$	β_c	$\beta_{c'}$
1	01100	0.300	0	0	0	0	0	0	0	0	0	0
	10100	0.005	0	0	0	0	0	0	0	0	0	0
	11011	0.010	**1**	**1**	**−1**	**−1**	**1**	**−1**	**1**	**0**	**−1**	**0**
	11100	0.155	0	0	0	0	0	0	0	0	0	0
	11111	0.110	0	0	0	0	0	0	0	0	0	0
	10011	0.420	0	0	0	0	0	0	0	0	0	0
2	00111	0.070	0	0	0	0	0	0	0	0	0	0
	01000	0.020	0	0	0	0	0	0	0	0	0	0
	01011	0.050	0	0	0	0	0	0	0	0	0	0
	01101	0.060	0	0	0	0	0	0	0	0	0	0
	01110	0.140	0	0	0	0	0	0	0	0	0	0
	10010	0.080	0	0	0	0	0	0	0	0	0	0
	10100	0.005	0	0	0	0	0	0	0	0	0	0
	11011	0.010	**1**	**1**	**−1**	**−1**	**1**	**−1**	**1**	**0**	**−1**	**0**
	11101	0.090	0	0	0	0	0	0	0	0	0	0
	11110	0.130	0	0	0	0	0	0	0	0	0	0
	11111	0.100	0	0	0	0	0	0	0	0	0	0
	10001	0.245	0	0	0	0	0	0	0	0	0	0

Hap, haplotype; Freq, haplotype frequency.

haplotype pair for an individual (which is typically unknown due to phase ambiguity); rather, it averages over all compatible haplotype pairs for a person to incorporate uncertainty in haplotype pair estimation. Suppose there are m possible haplotypes in the haplotype block and population under study. In the following, we model the probabilities in the aforementioned likelihood in terms of the model parameters (the subscripts i and r are suppressed for simplicity).

2.1.1 Modeling of $P(Y_c|Z,u)$ and $P(Y_{c'}|Z,u)$

A haplotype pair Z consists of two haplotypes denoted as $z_k/z_{k'}$ $(k, k' = 1, 2, \ldots, m)$. Let $\boldsymbol{X_z} = (1, x_1, x_2, \ldots, x_{m-1})$ be a (row) design vector with x_k equal to the number of times z_k appears in the haplotype pair Z; $k = 1, \ldots, m-1$, i.e., $z_k = 0, 1$, or 2. The m^{th} haplotype is assumed to be the baseline without loss of generality. Let $\boldsymbol{\beta_c}$ and $\boldsymbol{\beta_{c'}}$ be the vectors of regression coefficients (including the intercept), i.e., they include the effects of haplotypes on phenotypes Y_c and $Y_{c'}$, respectively. The slope coefficients have the same interpretation as in a usual linear regression model, i.e., the expected change in the quantitative trait if a person carries a copy of a specific haplotype as opposed to the baseline haplotype. As Y_c and $Y_{c'}$ are two continuous phenotypes and u is the latent variable that induces a correlation between them, we use the following linear models: $Y_c = \boldsymbol{X_z\beta_c} + u + \epsilon_c$ and $Y_{c'} = \boldsymbol{X_z\beta_{c'}} + u + \epsilon_{c'}$, where $\epsilon_c \sim N(0, \sigma_c^2)$ and $\epsilon_{c'} \sim N(0, \sigma_{c'}^2)$. We assume $\epsilon_c, \epsilon_{c'}$, and u to be uncorrelated with each other. The marginal correlation coefficient between Y_c and $Y_{c'}$ can be shown to be equal to $\frac{\sigma_u^2}{\sqrt{\sigma_u^2+\sigma_c^2}\sqrt{\sigma_u^2+\sigma_{c'}^2}}$ and, thus, must be non-negative. If the two traits are negatively correlated, then the values for one of them should be multiplied by −1 before applying this method.

2.1.2 Modeling $P(Z)$

We model $P(Z)$ in terms of two sets of parameters: $\boldsymbol{f} = (f_1, f_2, \ldots, f_m)$, denoting the frequencies of m haplotypes in the population, and d, the within-population inbreeding coefficient (Weir, 1996).

For a given haplotype pair $Z = z_k/z_{k'}$

$$P(Z) = P\left(Z = z_k/z_{k'}\middle|\boldsymbol{f}, d\right) = \delta_{kk'} d f_k + (2 - \delta_{kk'})(1-d) f_k f_{k'}$$

where $\delta_{kk'} = 1\,(0)$ if $z_k = z_{k'}$ $(z_k \neq z_{k'})$ and $d \in (-1, 1)$ capture the excess/reduction of homozygosity. The aforementioned expression of $P(Z)$ reduces to the assumption of Hardy–Weinberg equilibrium (HWE) when $d = 0$, while other values of d allow for the Hardy–Weinberg disequilibrium.

2.2 Prior distributions

There are many choices of shrinkage priors to regularize the regression coefficients, such as LASSO, ridge, Student's t-test,

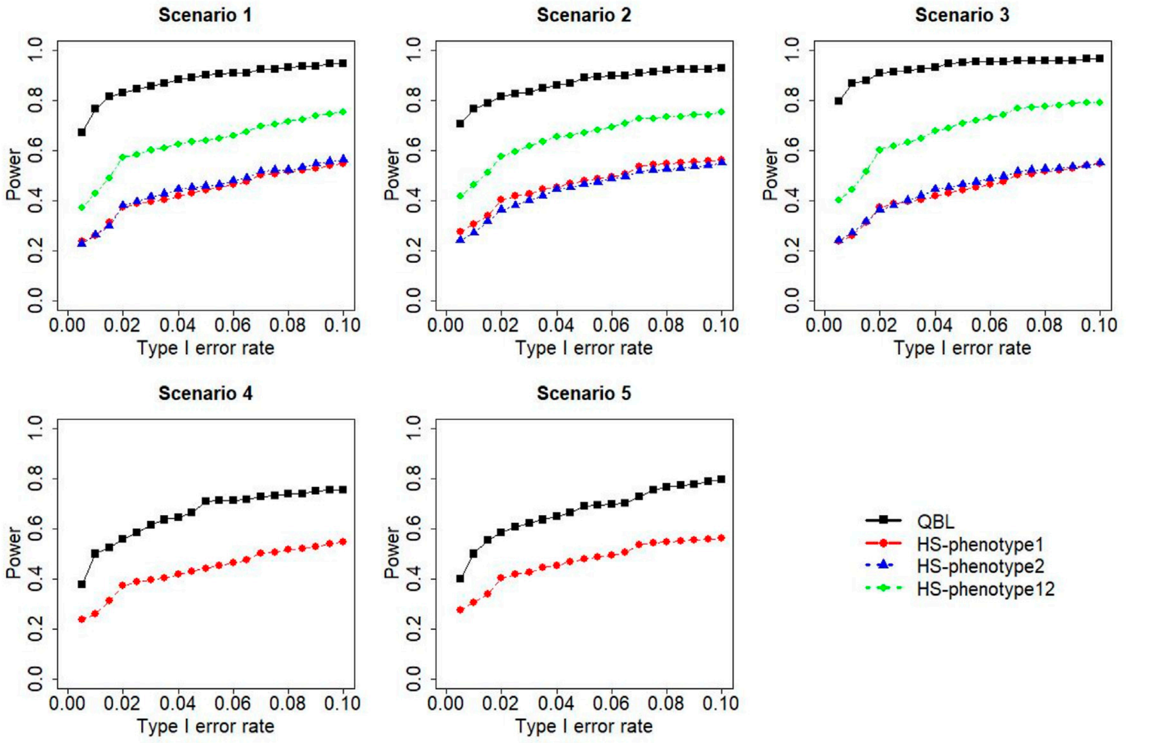


FIGURE 1
Simulation results under sample size 500, setting 1 (six haplotypes), and ρ = 0.1. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

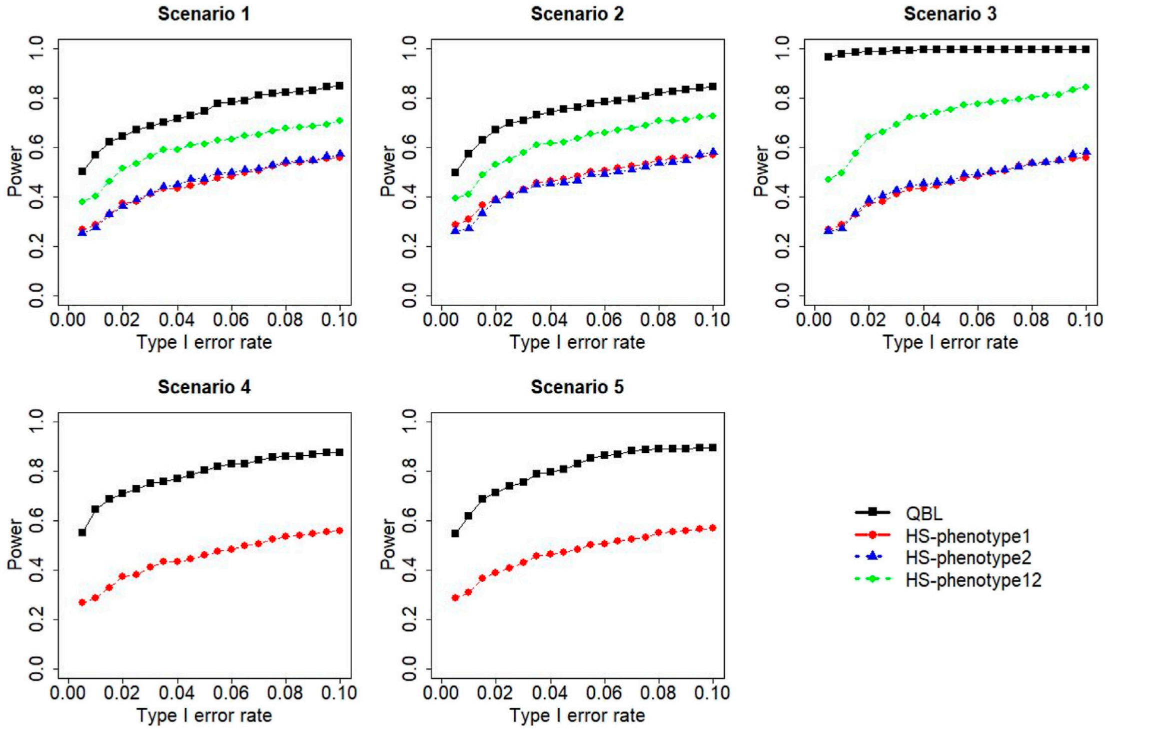


FIGURE 2
Simulation results under sample size 500, setting 1 (six haplotypes), and ρ = 0.5. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

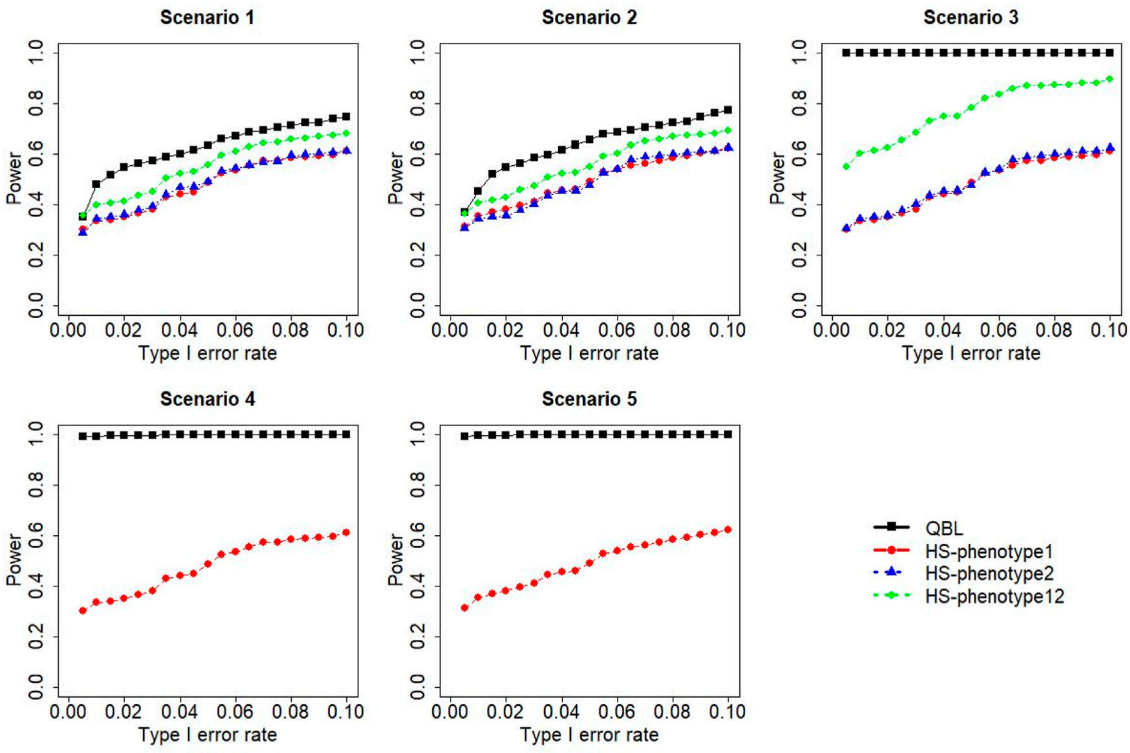


FIGURE 3
Simulation results under sample size 500, setting 1 (six haplotypes), and ρ = 0.9. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

horseshoe, and spike and slab. However, their performances are rather similar when the number of predictors (haplotypes) is smaller than the sample size, as is the case in this study (Van Erp et al., 2019). We choose Bayesian LASSO to regularize the regression coefficients for its ease of implementation, following previous LBL versions. Specifically, the prior for each slope parameter in $\boldsymbol{\beta_c}$ and $\boldsymbol{\beta}_{c'}$ is assigned a double exponential distribution with mean 0 and variance $\frac{2}{\lambda_c^2}$ and $\frac{2}{\lambda_{c'}^2}$, respectively. We use standard normal priors for the intercepts β_{0c} and $\beta_{0c'}$. The amounts of penalty for the slope coefficients are controlled by the hyper-parameters λ_c and $\lambda_{c'}$. We let them follow gamma (a, b) distribution with a = b = 20, following the original LBL method and its extensions (Biswas and Lin, 2012; Yuan and Biswas, 2019; Yuan and Biswas, 2021).

The prior for the frequency vector $\boldsymbol{f}$ is set to be non-informative Dirichlet (1, . . ., 1) consisting of m values. We consider a uniform prior for d. However, given that $P(Z)$, as shown in Section 2.1.2, must always be non-negative, d and $\boldsymbol{f}$ are not independent. In particular, d must be greater than $-\frac{f_k}{1-f_k}$ for all k values. Thus, the prior for d, given $\boldsymbol{f}$, is set to be Uniform$\left(max_k\left\{-\frac{f_k}{1-f_k}\right\}, 1\right)$. We use a weakly informative half-Cauchy prior for σ_u with a fixed hyper-parameter A given by $\pi(\sigma_u) \propto (1 + (\frac{\sigma_u}{A})^2)^{-1}$, where $\sigma_u > 0$, and set A = 10 (Yuan and Biswas, 2019; Yuan and Biswas, 2021). A non-informative uniform prior is used for σ_c^2 and $\sigma_{c'}^2$, whose probability density function is given by $p(\sigma^2) \propto \sigma^{-1}$, where $\sigma^2 > 0$.

2.3 Posterior distributions

The joint posterior distribution of all parameters can be obtained by combining the likelihood and prior distributions as follows:

$$\pi\left(\boldsymbol{\beta}_c, \boldsymbol{\beta}_{c'}, \lambda_b, \lambda_c, \boldsymbol{f}, d, \sigma_u, \sigma_c^2, \sigma_{c'}^2, \boldsymbol{Z} \middle| \boldsymbol{Y}_c, \boldsymbol{Y}_{c'}, \boldsymbol{G}, \boldsymbol{u}\right) \propto$$
$$L(\Psi)\, \pi\left(\boldsymbol{\beta}_c \middle| \lambda_c\right) \pi\left(\boldsymbol{\beta}_{c'} \middle| \lambda_{c'}\right) \pi(\lambda_c)\, \pi(\lambda_{c'})\, \pi\left(d \middle| \boldsymbol{f}\right) \pi(\boldsymbol{f})$$
$$\pi(\sigma_u)\, \pi\left(\sigma_c^2\right) \pi\left(\sigma_{c'}^2\right)$$

where $\boldsymbol{Z}$ consists of all possible haplotype pairs for all n subjects. We use Markov chain Monte Carlo (MCMC) methods to estimate the posterior distributions of all parameters. Details of the MCMC algorithm can be found in Supplementary Appendix A1. Notably, we update the latent variable u at every MCMC iteration, and thus, obtain its posterior distribution.

2.4 Association testing

We use the posterior distributions of regression coefficients for testing the association of haplotypes with the two phenotypes jointly. In particular, to test the association of the j^{th} haplotype with the two continuous phenotypes jointly, the hypotheses are

$$H_0: \left|\beta_{jc}\right| \le \epsilon \text{ and} \left|\beta_{jc'}\right| \le \epsilon \text{ vs } H_a: \left|\beta_{jc}\right| > \epsilon \text{ or} \left|\beta_{jc'}\right| > \epsilon$$

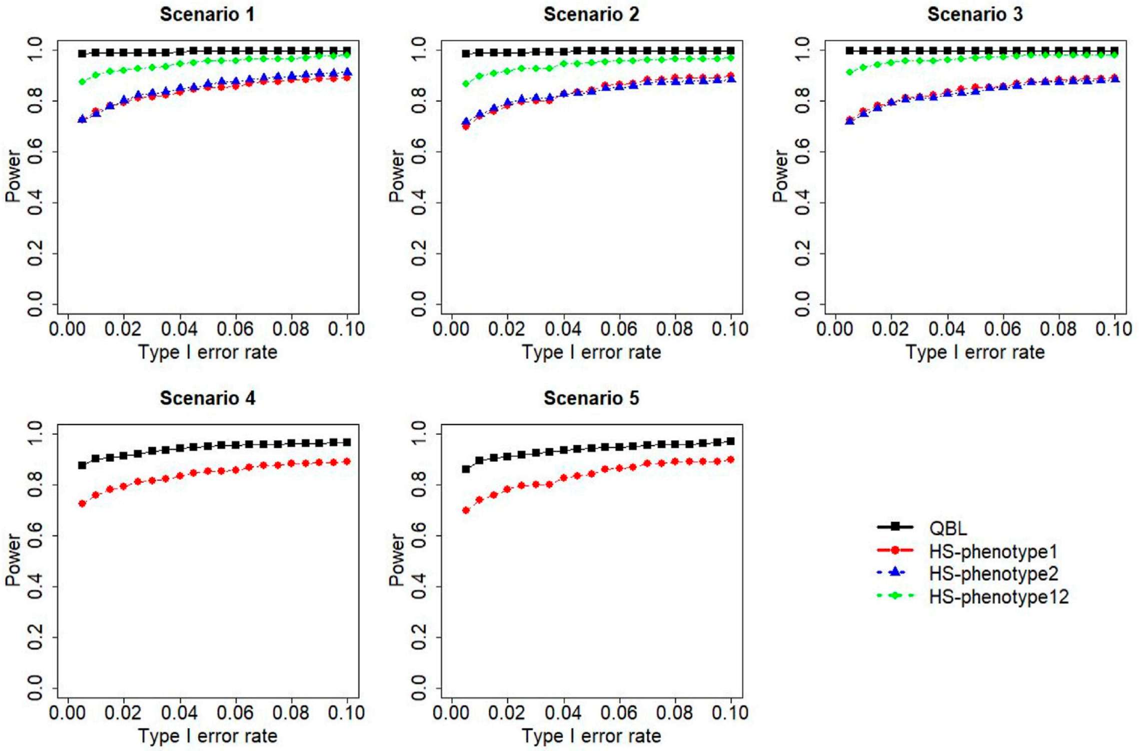


FIGURE 4
Simulation results under sample size 1,000, setting 1 (six haplotypes), and ρ = 0.1. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

where we set ϵ to be 0.1 (Biswas and Lin, 2012; Yuan and Biswas, 2019; Yuan and Biswas, 2021). Notably, the alternate hypothesis corresponds to the association with at least one phenotype.

To carry out this test, we calculated the Bayes factor (BF), which is the ratio of the posterior odds to the prior odds in favor of the alternative hypothesis. The prior odds can be found in Supplementary Appendix A2.

The posterior odds are obtained from the estimated posterior distributions. Once the BF for each haplotype in a block is obtained, their maximum BF is recorded. If this maximum BF exceeds a certain threshold, we conclude that the haplotype block is associated with at least one of the two phenotypes. We calculated the appropriate threshold following Yuan and Biswas (2019) and Yuan and Biswas (2021)—to be described in detail in the Simulation study and Application sections.

We compare the performance of bivariate QBL with a standard haplotype association test, Haplo.score (Schaid et al., 2002). We use the R package Haplo.stats to apply Haplo.score twice to the two continuous phenotypes individually (Sinnwell and Schaid, 2022).

3 Simulation study

3.1 Data generation

We generate data under two haplotype settings and five association scenarios to examine the properties of bivariate QBL and compare with Haplo.score. The two haplotype settings consist of 6 and 12 haplotypes (in a haplotype block under this study), as shown in Table 1. Following the simulation studies conducted previously for investigating univariate and bivariate LBL methods, we formed each haplotype by combining five SNPs (to allow easy comparison across various LBL versions). However, we note that, in principle, bivariate QBL can handle haplotype blocks with a larger number of SNPs at the expense of an increased computational burden (this issue is discussed in the Discussion section). Under each setting, the causal haplotype is 11011, a rare haplotype of frequency 1%. This target haplotype can be associated with one or both phenotype(s) and its effect(s), i.e., the corresponding β coefficient(s) can be positive (risk) or negative (protective). This leads to five association scenarios in total with the non-zero β values (for 11011) chosen to ensure that the power of the proposed method or Haplo.score at type I error rates of 0.5%–10% is in a reasonable range. We assume other haplotypes in the block to be null or non-associated, i.e., their β coefficients are equal to 0.

To generate a haplotype pair for a subject, we use the haplotype frequencies, as shown in Table 1. Using those frequencies and assuming HWE, the probabilities of all possible haplotype pairs can be calculated. Based on those probabilities, we randomly generate one haplotype pair, say Z, for each subject in the sample, which corresponds to a design row vector $\boldsymbol{X}_Z$. After assigning haplotype pairs to all subjects, we generate two continuous phenotypes for each subject using the following bivariate normal (BVN) distribution.

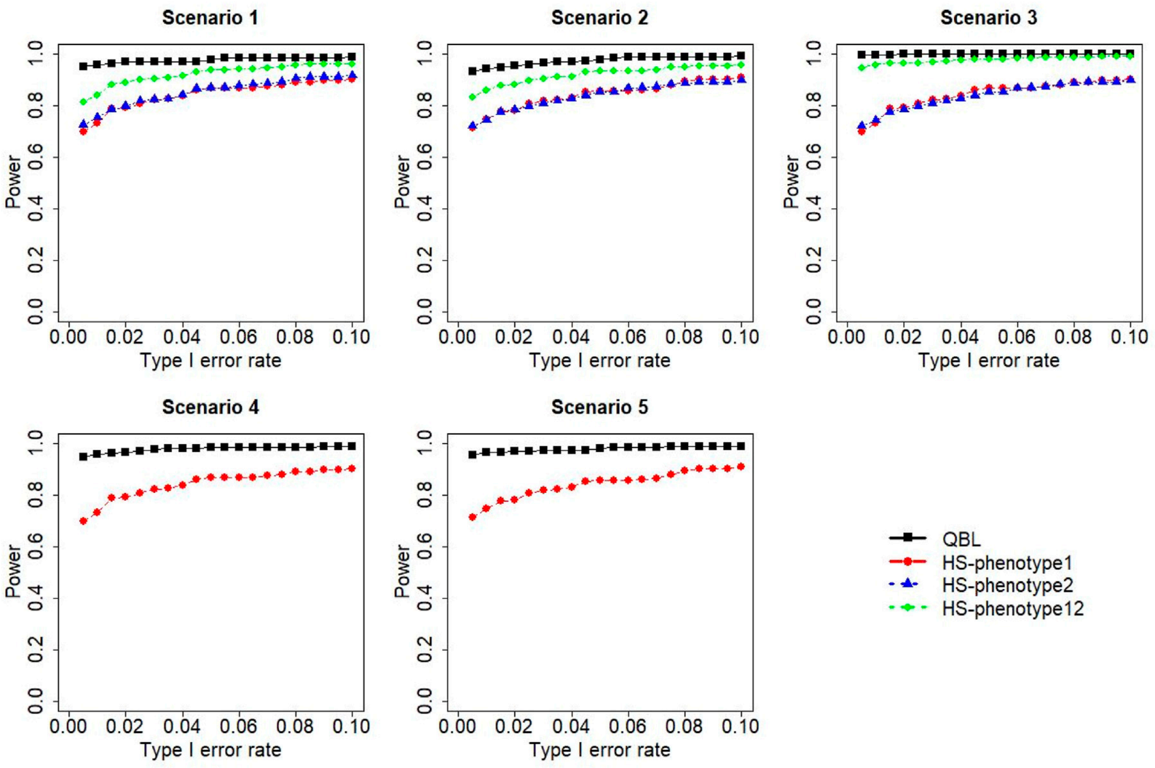


FIGURE 5
Simulation results under sample size 1,000, setting 1 (six haplotypes), and $\rho = 0.5$. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

$$\begin{pmatrix} Y_c \\ Y_{c'} \end{pmatrix} \sim BVN\left(\begin{pmatrix} \boldsymbol{X}_z\boldsymbol{\beta}_c \\ \boldsymbol{X}_z\boldsymbol{\beta}_{c'} \end{pmatrix}, \begin{pmatrix} \sigma_c^2 & \rho\sigma_c\sigma_{c'} \\ \rho\sigma_c\sigma_{c'} & \sigma_{c'}^2 \end{pmatrix}\right),$$

where $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_{c'}$ (excluding intercepts β_{0c} and $\beta_{0c'}$) are as shown in Table 1; $\sigma_c = \sigma_{c'} = 1$ and ρ are varied to be 0.1, 0.5, or 0.9. We set $\beta_{0c} = \beta_{0c'} = 25$.

We generate samples of sizes 500 and 1,000. For each sample size and simulation setup, resulting from a combination of a haplotype setting, a non-null association scenario, and a fixed ρ-value, 500 samples are generated. We also generate the corresponding null scenarios, i.e., for each combination of sample size, haplotype setting, and ρ-value, all βs are set to be equal to 0 and 1,000 samples are generated. To each sample, we apply bivariate QBL to both phenotypes jointly. The MCMC is run for a total of 3,00,000 iterations with 50,000 burn-in to achieve acceptable convergence (Gelman et al., 2003). To declare significance, we use appropriate cutoffs to the resulting BFs. The determination of the cutoffs for both bivariate QBL and Haplo.score is discussed in the following sub-section.

3.2 Calculation of cutoffs

The cutoffs for bivariate QBL are calculated in the following way. For each sample, we obtain one BF value per haplotype. We record the maximum of those BFs. Thus, we obtain 1,000 maximum BF values from the 1,000 null scenario replicates. We sort these 1,000 values in a descending order and obtain the cutoff for a specific type I error rate to be the corresponding percentile. It is to be noted that by taking the maximum overall BF values from a haplotype block, we adjust for multiple testing within that block.

We calculate cutoffs for Haplo.score in a slightly different way because it is applied to each phenotype. For each sample, we obtain two (global) p-values from two Haplo.score analyses. Then, we record the minimum of these two p-values. Similar to bivariate QBL, we obtain 1,000 minimum p-values from the 1,000 null samples. We sort them in an ascending order and obtain the cutoff of Haplo.score for a specific type I error rate by taking the relevant bottom percentile.

Once the cutoffs are obtained in the aforementioned manner, we use these cutoffs to calculate power for the corresponding non-null setups described previously. The type I error rates and power obtained by varying the cutoffs for a p-value (for Haplo.score) and BF (bivariate QBL) are then plotted against each other to obtain receiver operating characteristic (ROC)-type curves. For Haplo.score, the power is shown for detecting associations with at least one of the two phenotypes, as well as with each phenotype separately (in scenarios 1–3, where the target haplotype is associated with both phenotypes).

3.3 Results

The results for settings 1 (six haplotypes) and 2 (12 haplotypes), sample sizes 500 and 1,000, and correlation

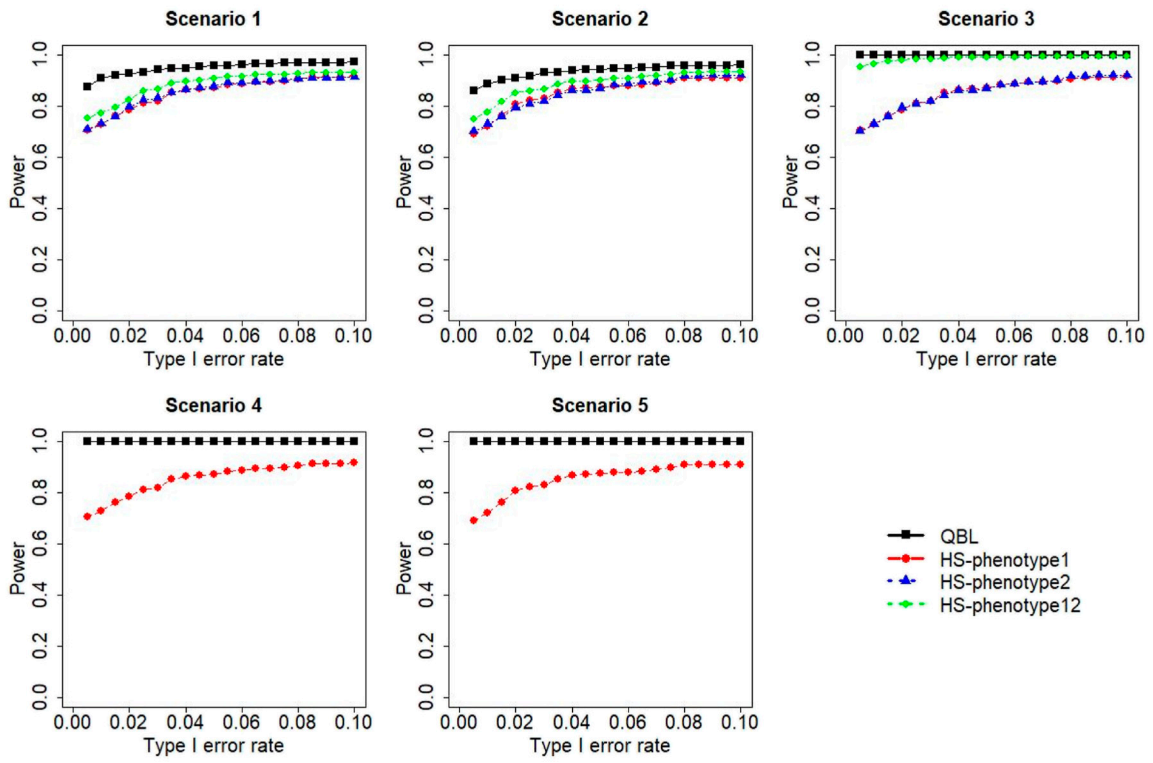


FIGURE 6
Simulation results under sample size 1,000, setting 1 (six haplotypes), and ρ = 0.9. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

coefficients 0.1, 0.5, and 0.9 are shown in Figures 1–12. Notably, bivariate QBL outperforms Haplo.score in all figures even though the margin of difference varies depending on the combination of association scenarios and ρ-values. Bivariate QBL shows the best performance in scenario 3, where the effect sizes of the target haplotype are in opposite directions (one β positive and another β negative). In this scenario, the power of bivariate QBL exceeds Haplo.score by a substantial margin. This margin increases in favor of QBL as the correlation coefficient increases. Bivariate QBL also maintains this superior performance in scenarios 4 and 5, where the target haplotype is unrelated to one phenotype but has a positive (scenario 4) or negative (scenario 5) association with the other phenotypes. Again, the power gain margin of bivariate QBL increases as the correlation between the two phenotypes increases. This outperformance trend can be seen in all combinations of haplotype settings and sample sizes considered in this study.

The performances of bivariate QBL and Haplo.score are the closest in the first two scenarios only when the correlation coefficient is high, i.e., 0.9, as shown in Figures 3, 6, 9. However, Figure 12 shows that even with ρ = 0.9, bivariate QBL is clearly much more powerful than Haplo.score in these two scenarios. Moreover, when the correlation between the two phenotypes is weak or moderate, bivariate QBL outperforms Haplo.score in these scenarios at any combination of haplotype setting and sample sizes.

4 Application to GAW 19 data

We consider two continuous phenotypes, SBP and DBP, available in these data. They are moderately correlated (sample correlation coefficient = 0.55) and likely share a common genetic mechanism (Schillert and Konigorski, 2016). Typically, SBP and DBP are combined to create a single binary phenotype referred to as hypertension. More specifically, clinical thresholds are used for each BP to classify it as high blood pressure (BP); a subject is a case of hypertension if one of them is high (Datta and Biswas, 2016). However, converting a quantitative phenotype to a binary phenotype leads to a loss of information. Furthermore, combining them into one binary phenotype is a lost opportunity to investigate pleiotropy. As bivariate QBL can analyze the two continuous phenotypes jointly, it can potentially provide additional insight into these data.

There are 1,851 subjects in these data after discarding the missing values. Following Yuan and Biswas (2019), we analyze eight genes, namely, *FBN3*, *HRH1*, *INMT*, *MAP4*, *SAT2*, *SHBG*, *ULK4*, and *ZNF280D*. There are 28 SNVs in *FBN3*, 10 in *HRH1*, 18 in *INMT*, 18 in *MAP4*, 7 in *SAT2*, 15 in *SHBG*, 70 in *ULK4*, and 30 in *ZNF280D*. We combine five successive SNVs, starting from the first SNV, and create sliding haplotype blocks covering the whole gene, that is, on each gene, the first haplotype block consists of SNVs 1–5, second block consists of SNVs 2–6, and so on. For example, *ULK4* has 66 haplotype blocks and *MAP4* has 14 blocks.

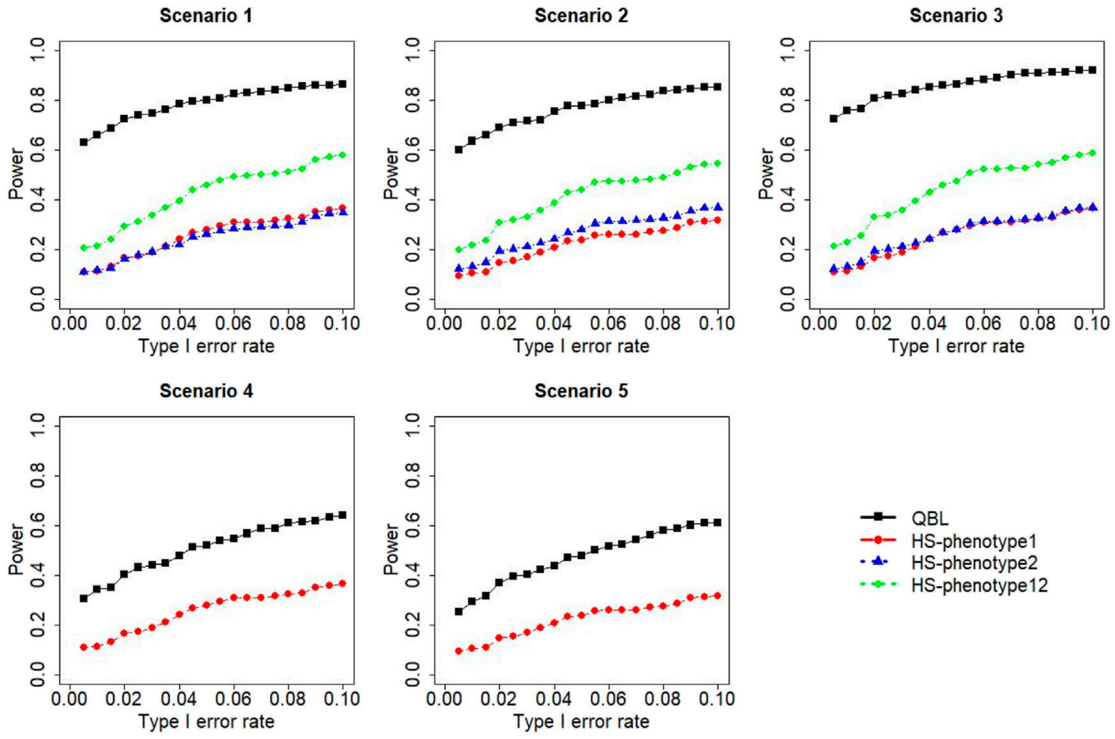


FIGURE 7
Simulation results under sample size 500, setting 2 (12 haplotypes), and ρ = 0.1. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

We apply bivariate QBL to each haplotype block with both phenotypes jointly and Haplo.score to the same haplotype block twice with SBP and DBP separately. We calculate appropriate (and more general purpose) cutoffs for bivariate QBL and Haplo.score based on both simulated data and permutating the GAW19 phenotypes, as described in the following. We simulate 1,200 null samples, following setting 2 of Table 1. To match the GAW19 data more closely, we generate sample sizes of 1,851 with the correlation coefficient (between SBP and DBP) set to 0.55. As GAW19 data are exome sequence and have far more rare haplotypes than those considered in our simulations, we complement 1,200 simulated null samples by GAW19 data with permutated phenotype values. In particular, we permute the phenotypes of all subjects while retaining the pairing between SBP and DBP. Then, we combine the permuted phenotypes with genotypes in the *ULK4* gene to create a null sample. We repeat this process 10 times to obtain 660 (66 × 10) blocks or null samples. Similarly, the permuted phenotypes are also combined with genotypes from *MAP4* gene and repeated 10 times to provide 140 (14 × 10) blocks or null samples. The results from 800 null samples obtained using permutations are combined with those from 1,200 simulated null samples to calculate cutoffs.

The cutoffs based on 2000 null samples are calculated in the same manner, as described in the simulation study section for both bivariate QBL and Haplo.score. The cutoffs for type I error rates of 1% and 2.5% are found to be BFs of 10.91 and 4.65 for bivariate QBL and *p*-values of 0.0004 and 0.0058 for the Haplo.score global test, respectively.

The haplotype blocks found to be significantly associated at a type I error rate of 2.5% using at least one of the methods are shown in Table 2. Bivariate QBL found a larger number of haplotype blocks to be significant, and the findings are consistent with the literature (Datta et al., 2016; Yuan and Biswas, 2019). For example, Haplo.score could not detect the haplotype in *FBN3*, whose $\hat{\beta}$ values for SBP and DBP are in opposite directions. All the haplotype blocks found to be significant using Haplo.score are also detected by bivariate QBL. At the type I error rate of 1%, bivariate QBL identifies all haplotype blocks in *ULK4*, as shown in Table 2, as significant, whereas Haplo.score identifies only one haplotype block (39–43) as significant. Therefore, bivariate QBL appears to perform better than Haplo.score in GAW19 data, which is in agreement with our findings in the simulation study.

5 Discussion

Health-related studies usually collect multiple outcomes to better assess patients' health, understand complex diseases/traits, and inter-connection between them, which, in turn, can help in developing effective prevention and treatment strategies. These outcomes are often correlated and may share a common genetic etiology. A commonly used practice in genetic association studies is to analyze these outcomes in a one-at-a-time manner. Such a univariate approach essentially ignores the additional information contained in the joint distribution of the outcomes. Also, it is a missed chance to investigate the possibility of pleiotropy among

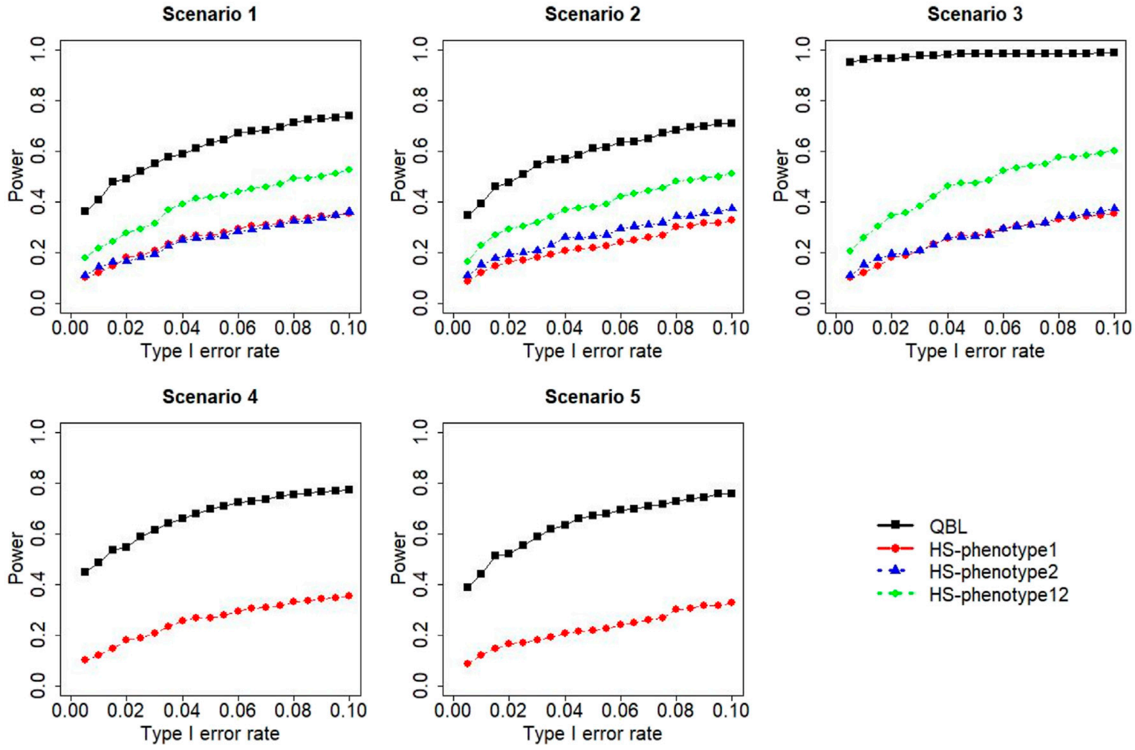


FIGURE 8
Simulation results under sample size 500, setting 2 (12 haplotypes), and ρ = 0.5. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

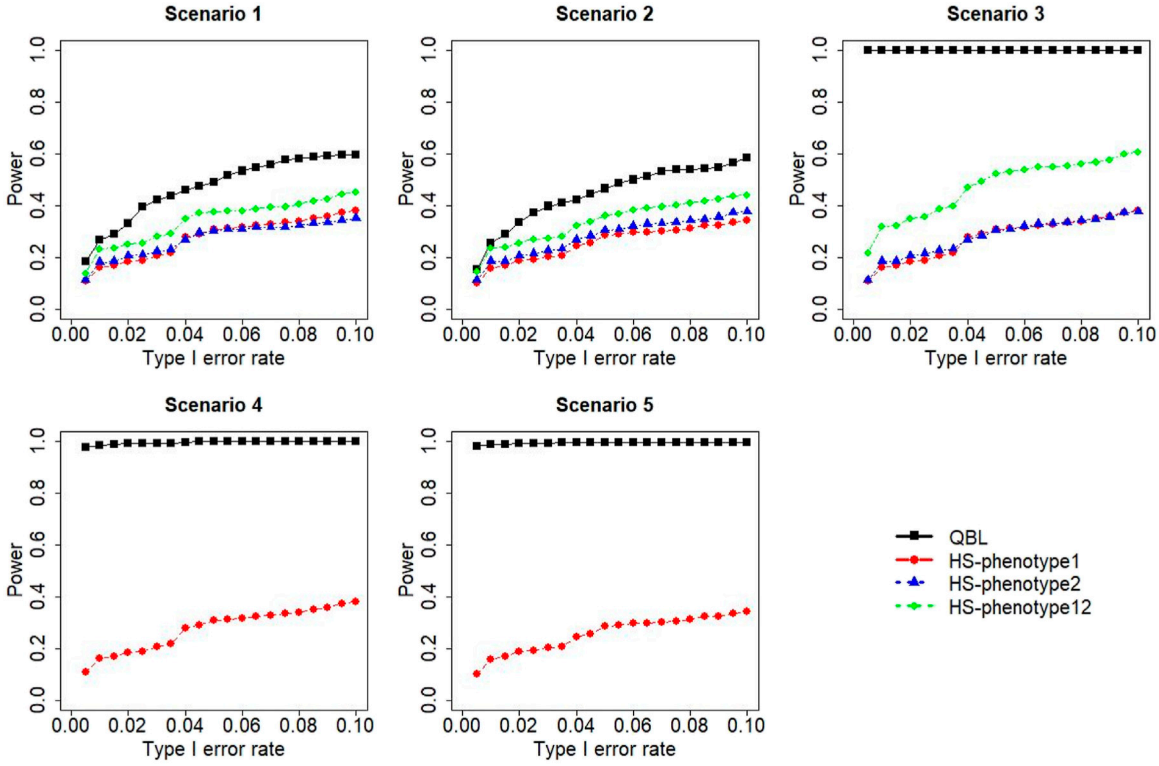


FIGURE 9
Simulation results under sample size 500, setting 2 (12 haplotypes), and ρ = 0.9. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

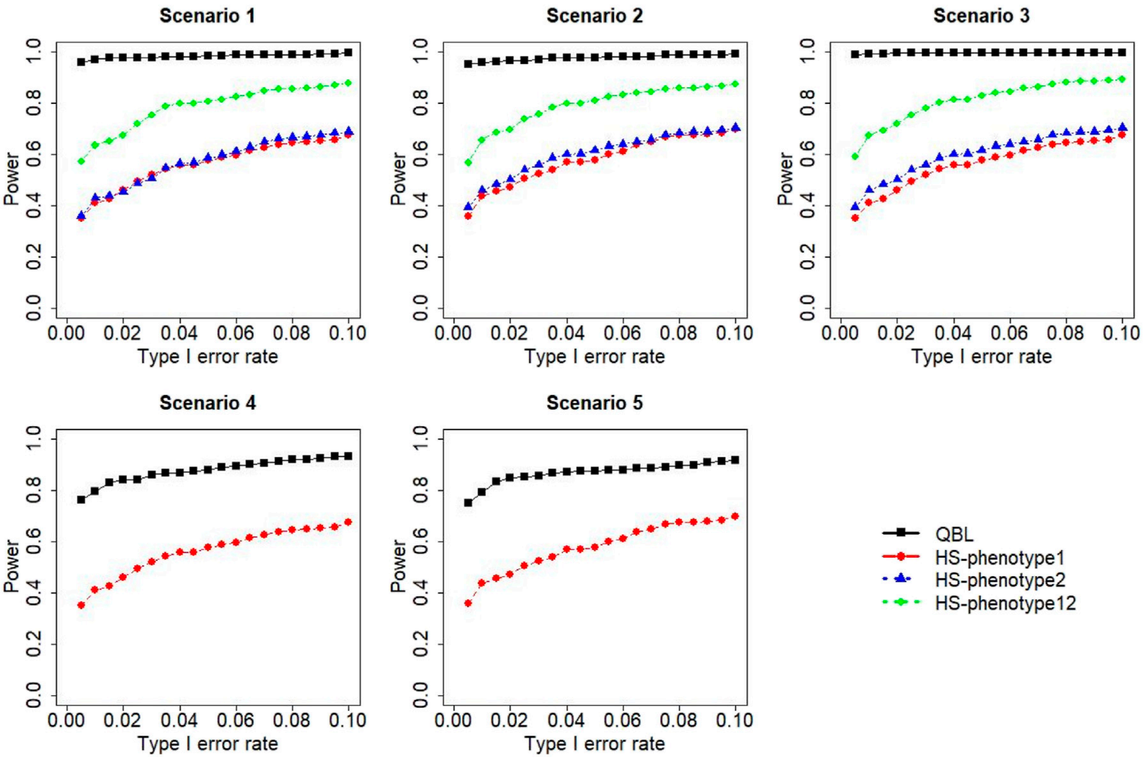


FIGURE 10
Simulation results under sample size 1,000, setting 2 (12 haplotypes), and ρ = 0.1. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12: phenotype 1 or 2.

these outcomes. Therefore, it is statistically and biologically more beneficial to adopt a multivariate approach to analyze the outcomes jointly. Moreover, analyzing haplotypes as genetic variants is advantageous because they are biologically interpretable, and haplotype-based tests can be performed on both NGS and GWAS data. There is no haplotype-based association test available that can detect rare variants associated with multiple continuous phenotypes yet. To fill this void, we propose bivariate QBL to detect the association of two quantitative traits with rare (and common) haplotypes. Our findings from the simulation study show that the method performs better than Haplo.score in all simulation setups that we considered.

Bivariate QBL performs best when the two outcomes have high positive correlation between them, and the target haplotype has discordant effects on the two phenotypes, i.e., one positive β and another negative β. This finding is consistent with the literature (Liu et al., 2009a; Ferreira and Purcell, 2009; Galesloot et al., 2014). In particular, to compare with Galesloot et al. (2014), we note that the first two scenarios in our study (both βs of the same sign) correspond to positive genetic correlation in their terminology, scenario 3 (one positive β and another negative β) corresponds to negative genetic correlation, and scenarios 4 and 5 (one β is 0) correspond to no genetic correlation. In scenarios 3–5, with a negative or zero genetic correlation, bivariate QBL outperforms Haplo.score at any combination of haplotype settings, correlation, and sample sizes, and its power increases as the positive residual correlation (i.e., ρ in our context) increases. Bivariate QBL gains substantial power in these scenarios with increasing residual correlation as it not only avoids the burden of multiple testing but also incorporates the additional information provided by the cross-trait correlation. However, even with type I error rates of less than 1%, bivariate QBL has power close to or practically 1, whereas Haplo.score has a much lower power in these scenarios.

The performance of Haplo.score is close to that of bivariate QBL only when both outcomes are highly correlated and the target haplotype affects both outcomes in the same direction, i.e., scenarios 1 and 2. In these scenarios, the power of bivariate QBL increases as the correlation decreases. In the terminology of Galesloot et al. (2014), this means when both genetic correlation and residual correlation are of the same sign, the power of bivariate QBL decreases as the positive residual correlation increases. This phenomenon of bivariate QBL is also consistent with other multivariate genetic association tests that exist in the literature (Liu et al., 2009a; Ferreira and Purcell, 2009). In practice, it is unlikely that two phenotypes will have a very high correlation. On the other hand, we note that bivariate QBL estimates haplotype frequencies (*f*) jointly with the haplotype effects and other parameters. Haplotype frequencies are estimated very well by bivariate QBL, especially due to the fact that we set the starting values of *f* in the MCMC algorithm to its maximum likelihood estimate (obtained from the hapassoc package) (Burkett et al., 2006; Burkett et al., 2015). Thus, there is practically no impact of

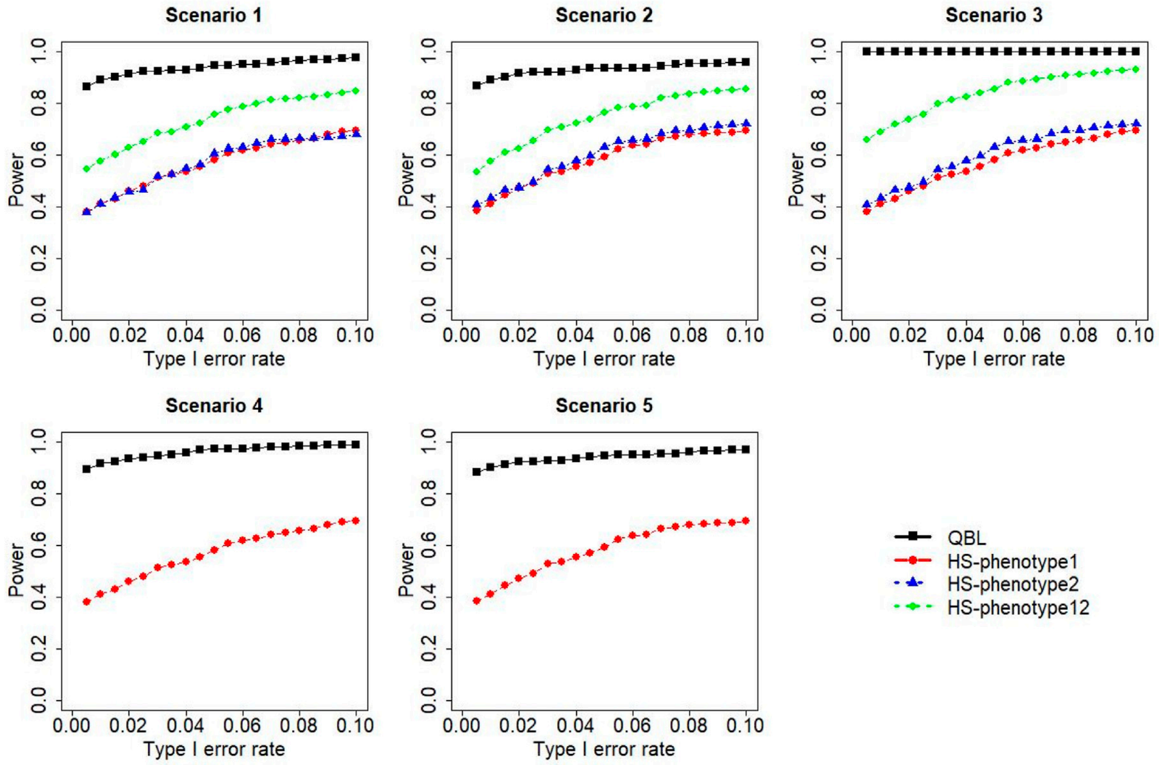


FIGURE 11
Simulation results under sample size 1,000, setting 2 (12 haplotypes), and ρ = 0.5. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12: phenotype 1 or 2.

haplotype frequency estimation on type I error and power of the method.

In GAW19 data, SBP and DBP are moderately correlated (0.55) (Datta et al., 2016; Yuan and Biswas, 2019). As another example, Liu et al. (2009b) observed a correlation between the body mass index and bone mineral density of 0.384 and 0.257, respectively, in two datasets. When there is a weak-to-moderate correlation, bivariate QBL outperforms Haplo.score by a substantial margin. In our GAW19 data application, we detected several rare haplotype blocks to be associated with SBP and DBP jointly. Specifically, nine blocks were detected in *ULK4*, one in *MAP4*, and another in *FBN3*. These results agree with the findings from previous studies (Levy et al., 2009; International Consortium for Blood Pressure Genome-Wide Association Studies Ehret et al., 2011; Ehret and Caulfield, 2013). Notably, the correlation between SBP and DBP is moderate and as per our simulation results, bivariate QBL is far more powerful than Haplo.score in this situation. However, many of those haplotype blocks could not be detected by Haplo.score. This indicates that bivariate QBL can help establish multiple trait–variant associations and identify potential pleiotropic effects for further investigation.

Bivariate QBL has a limitation in terms of computing time. In our simulation study, for a sample size of 500, bivariate QBL takes 86 and 166 s to finish 2,00,000 MCMC iterations for 6 and 12 haplotypes, respectively. This is for a machine with 3.50-GHz Milan processor with 128 cores under the Linux operating system and 256 GB RAM. However, it is faster than both bivariate LBL-2B and LBL-BC. Bivariate QBL can handle a larger number of SNPs in a haplotype at the expense of an increased computational burden. The runtime of bivariate QBL almost doubles when we increase the number of SNPs in a haplotype block from 5 (86 s) to 10 (158 s). Another limitation is that the method can only accommodate two continuous phenotypes at a time. We plan to extend the framework of bivariate QBL (and LBL) to accommodate many correlated continuous and/or binary phenotypes jointly. We also plan to extend the framework to investigate gene–environment interactions and develop a computationally efficient version of this method.

Despite these limitations, we believe bivariate QBL is an important addition to the existing genetic association tests, especially because there is currently no rare haplotype association test available that can analyze two correlated continuous phenotypes jointly.

6 Software

An R package implementing the proposed bivariate QBL method will be made available at https://www.utdallas.edu/~swati.biswas/ and https://github.com/ihsajal/ as part of the existing package LBL.

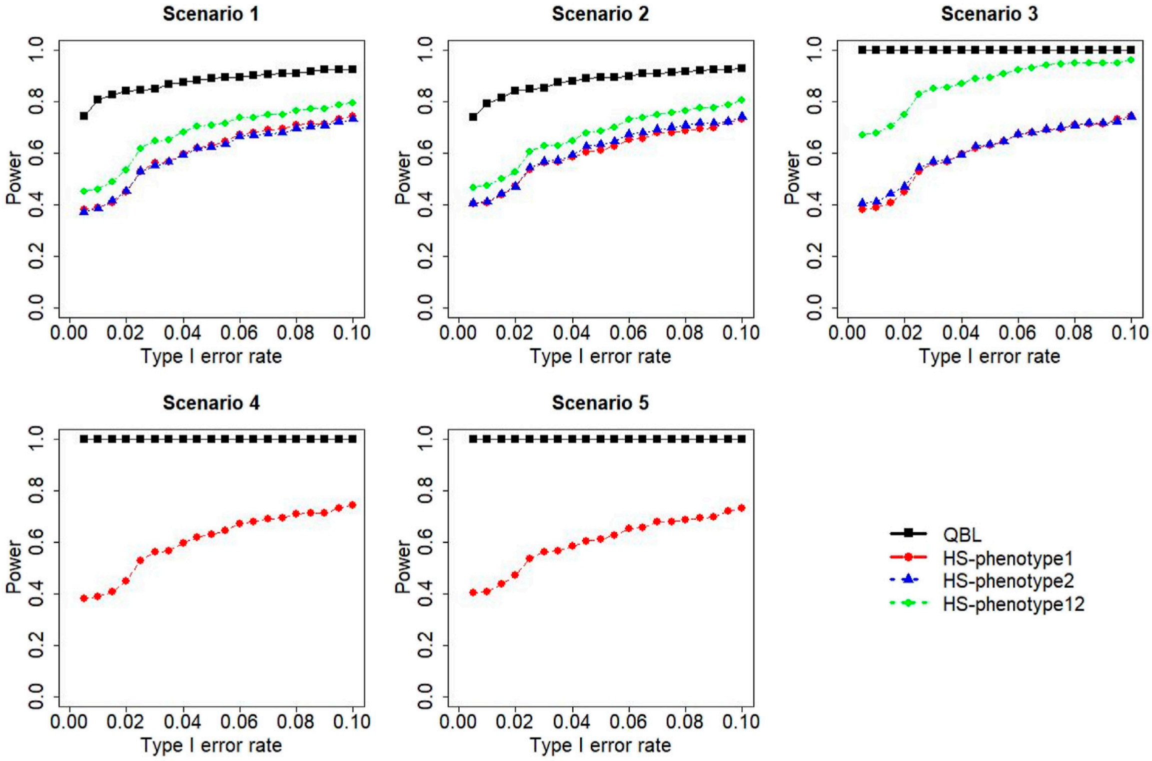


FIGURE 12
Simulation results under sample size 1,000, setting 2 (12 haplotypes), and ρ = 0.9. Scenarios are shown in Table 1. HS, Haplo.score; phenotype12, phenotype 1 or 2.

TABLE 2 Haplotype blocks significant at the 2.5% level on *ULK4*, *MAP4*, and *FBN3* genes using the bivariate QBL or Haplo.score (significant BF or *p*-value is shown in boldface).

				Bivariate QBL			Haplo.score	
Gene	Win	Hap	Freq	β (SBP)	β (DBP)	BF	*p*-value (SBP)	*p*-value (DBP)
ULK4	3–7	h10101	0.0016	1.206	0.824	**14.06**	0.0292	0.0913
ULK4	4–8	h01010	0.0014	1.608	0.747	**54.56**	**0.0056**	0.1308
ULK4	5–9	h10101	0.0014	1.619	0.767	**50.52**	**0.0033**	0.1319
ULK4	6–10	h01010	0.0016	1.211	0.843	**15.67**	**0.0011**	0.0405
ULK4	7–11	h10100	0.0016	1.218	0.849	**16.63**	**0.0007**	0.0335
ULK4	8–12	h01000	0.0016	1.207	0.836	**14.82**	**0.0009**	0.0477
ULK4	9–13	h10000	0.0017	1.209	0.835	**20.66**	**0.0012**	0.0384
ULK4	39–43	h11100	0.0055	0.869	0.666	**41.33**	**0.0001**	0.2726
ULK4	40–44	h11000	0.0052	0.854	0.801	**25.26**	0.0791	0.2656
MAP4	11–15	h10000	0.0043	0.778	1.714	**10.49**	0.0301	0.7634
FBN3	24–28	h00010	0.0014	0.783	−0.54	**10.41**	0.0313	0.2224

Win, window; Hap, haplotype; Freq, haplotype frequency.

Data availability statement

The data analyzed in this study are subject to the following licenses/restrictions: The data are from Genetic Analysis Workshop 19. Participants of the workshop have access to these de-identified data for secondary analysis. Requests to access these datasets should be directed at https://bmcproc.biomedcentral.com/articles/10.1186/s12919-016-0007-z.

Author contributions

SB conceived the study. IS and SB developed the methodology. IS carried out all simulations and data analyses under the supervision of SB. Both authors participated in interpreting the results and writing the manuscript. Both authors approved the final version of the manuscript.

Funding

This work was partially supported by the National Cancer Institute under grant number R03CA171011. Genetic Analysis Workshops are supported by the NIH grant R01 GM031575. The GAW19 whole-genome sequencing data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545.

Acknowledgments

The authors would like to thank the Texas Advanced Computing Center at the University of Texas at Austin for their allocation of computing times.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1104727/full#supplementary-material

References

Bader, J. S. (2001). The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* 2 (1), 11–24. doi:10.1517/14622416.2.1.11

Biswas, S., and Lin, S. (2012). Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics* 68 (2), 587–597. doi:10.1111/j.1541-0420.2011.01680.x

Biswas, S., and Papachristou, C. (2014). Evaluation of logistic Bayesian LASSO for identifying association with rare haplotypes. *BMC Proc.* 8, S54. doi:10.1186/1753-6561-8-S1-S54

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids Res.* 47 (D1), D1005–D1012. doi:10.1093/nar/gky1120

Burkett, K., Graham, J., and McNeney, B. (2006). hapassoc: Software for likelihood inference of trait associations with SNP haplotypes and other attributes. *J. Stat. Softw.* 16, 1–19. doi:10.18637/jss.v016.i02

Burkett, K., McNeney, B., and Burkett, M. K. (2015). Package 'hapassoc'.

Datta, A. S., and Biswas, S. (2016). Comparison of haplotype-based statistical tests for disease association with rare and common variants. *Briefings Bioinforma.* 17 (4), 657–671. doi:10.1093/bib/bbv072

Datta, A. S., Lin, S., and Biswas, S. (2018). A family-based rare haplotype association method for quantitative traits. *Hum. Hered.* 83 (4), 175–195. doi:10.1159/000493543

Datta, A. S., Zhang, Y., Zhang, L., and Biswas, S. (2016). Association of rare haplotypes on ULK4 and MAP4 genes with hypertension. *BMC Proc.* 10 (7), 363–369. doi:10.1186/s12919-016-0057-2

Deng, Y., He, T., Fang, R., Li, S., Cao, H., and Cui, Y. (2020). Genome-wide gene-based multi-trait analysis. *Front. Genet.* 11, 437. doi:10.3389/fgene.2020.00437

International Consortium for Blood Pressure Genome-Wide Association StudiesEhret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Pihur, V., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478 (7367), 103–109. doi:10.1038/nature10405,

Ehret, G. B., and Caulfield, M. J. (2013). Genes for blood pressure: An opportunity to understand hypertension. *Eur. heart J.* 34 (13), 951–961. doi:10.1093/eurheartj/ehs455

Ferreira, M. A., and Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics* 25 (1), 132–133. doi:10.1093/bioinformatics/btn563

Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PloS one* 9 (4), e95923. doi:10.1371/journal.pone.0095923

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian data analysis*. 2 ed. Florida: CRC Press.

Gratten, J., and Visscher, P. M. (2016). Genetic pleiotropy in complex traits and diseases: Implications for genomic medicine. *Genome Med.* 8 (1), 78. doi:10.1186/s13073-016-0332-x

Guo, W., and Lin, S. (2009). Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet. Epidemiol.* 33 (4), 308–316. doi:10.1002/gepi.20382

Hackinger, S., and Zeggini, E. (2017). Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* 7 (11), 170125. doi:10.1098/rsob.170125

Kaakinen, M., Mägi, R., Fischer, K., Heikkinen, J., Järvelin, M.-R., Morris, A. P., et al. (2017). Marv: A tool for genome-wide multi-phenotype analysis of rare variants. *BMC Bioinforma.* 18 (1), 110–118. doi:10.1186/s12859-017-1530-2

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* 32 (1), 9–19. doi:10.1002/gepi.20257

Lee, P. H., Anttila, V., Won, H., Feng, Y.-C. A., Rosenthal, J., Zhu, Z., et al. (2019). Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell.* 179 (7), 1469–1482. doi:10.1016/j.cell.2019.11.020

Lee, P. H., Feng, Y.-C. A., and Smoller, J. W. (2021). Pleiotropy and cross-disorder genetics among psychiatric disorders. *Biol. psychiatry* 89 (1), 20–31. doi:10.1016/j.biopsych.2020.09.026

Lee, S., Won, S., Kim, Y. J., Kim, Y., Consortium, T. D. G., Kim, B. J., et al. (2017). Rare variant association test with multiple phenotypes. *Genet. Epidemiol.* 41 (3), 198–209. doi:10.1002/gepi.22021

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* 41 (6), 677–687. doi:10.1038/ng.384

Li, J., Zhang, K., and Yi, N. (2011). A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies. *Hum. Hered.* 71 (3), 148–160. doi:10.1159/000324841

Li, Y., Byrnes, A. E., and Li, M. (2010). To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *Am. J. Hum. Genet.* 87 (5), 728–735. doi:10.1016/j.ajhg.2010.10.014

Lin, W. Y., Yi, N., Lou, X. Y., Zhi, D., Zhang, K., Gao, G., et al. (2013). Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genet. Epidemiol.* 37 (6), 560–570. doi:10.1002/gepi.21740

Liu, J., Pei, Y., Papasian, C. J., and Deng, H. W. (2009). Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet. Epidemiol.* 33 (3), 217–227. doi:10.1002/gepi.20372

Liu, Y.-Z., Pei, Y.-F., Liu, J.-F., Yang, F., Guo, Y., Zhang, L., et al. (2009). Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing both obesity and osteoporosis phenotypes in males. *PloS one* 4 (8), e6827. doi:10.1371/journal.pone.0006827

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R., et al. (2012). MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. *PloS one* 7 (5), e34861. doi:10.1371/journal.pone.0034861

Papachristou, C., and Biswas, S. (2020). Comparison of haplotype-based tests for detecting gene–environment interactions with rare variants. *Briefings Bioinforma.* 21 (3), 851–862. doi:10.1093/bib/bbz031

Pei, Y. F., Zhang, L., Liu, J., and Deng, H. W. (2009). Multivariate association test using haplotype trend regression. *Ann. Hum. Genet.* 73 (4), 456–464. doi:10.1111/j.1469-1809.2009.00527.x

Ray, D., and Basu, S. (2017). A novel association test for multiple secondary phenotypes from a case-control GWAS. *Genet. Epidemiol.* 41 (5), 413–426. doi:10.1002/gepi.22045

Ray, D., Pankow, J. S., and Basu, S. (2016). Usat: A unified score-based association test for multiple phenotype-genotype analysis. *Genet. Epidemiol.* 40 (1), 20–34. doi:10.1002/gepi.21937

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* 70 (2), 425–434. doi:10.1086/338688

Schillert, A., and Konigorski, S. (2016). Joint analysis of multiple phenotypes: Summary of results and discussions from the genetic analysis workshop 19. *BMC Genet.* 17. doi:10.1186/s12863-015-0317-6

Sinnwell, J. P., and Schaid, D. J. (2022). Package 'haplo. stats'. In (Version 1.8.9) [R Package]. Available at: https://analytictools.mayo.edu/research/haplo-stats/.

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 (7), 483–495. doi:10.1038/nrg3461

Teixeira-Pinto, A., and Normand, S. L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics Med.* 28 (13), 1753–1773. doi:10.1002/sim.3588

Van der Sluis, S., Dolan, C. V., Li, J., Song, Y., Sham, P., Posthuma, D., et al. (2015). Mgas: A powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics* 31 (7), 1007–1015. doi:10.1093/bioinformatics/btu783

Van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *J. Math. Psychol.* 89, 31–50. doi:10.1016/j.jmp.2018.12.004

Wang, M., and Lin, S. (2015). Detecting associations of rare variants with common diseases: Collapsing or haplotyping? *Briefings Bioinforma.* 16 (5), 759–768. doi:10.1093/bib/bbu050

Wang, M., and Lin, S. (2014). FamLBL: Detecting rare haplotype disease association based on common SNPs using case-parent triads. *Bioinformatics* 30 (18), 2611–2618. doi:10.1093/bioinformatics/btu347

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J., et al. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* 51 (9), 1339–1348. doi:10.1038/s41588-019-0481-0

Weir, B. S. (1996). *Genetic data analysis II*. Sunderland, Massachusetts: Sinauer Associates.

Yuan, X., and Biswas, S. (2019). Bivariate logistic Bayesian LASSO for detecting rare haplotype association with two correlated phenotypes. *Genet. Epidemiol.* 43 (8), 996–1017. doi:10.1002/gepi.22258

Yuan, X., and Biswas, S. (2021). Detecting rare haplotype association with two correlated phenotypes of binary and continuous types. *Statistics Med.* 40 (8), 1877–1900. doi:10.1002/sim.8877

Zhang, Y., Hofmann, J. N., Purdue, M. P., Lin, S., and Biswas, S. (2017). Logistic Bayesian LASSO for genetic association analysis of data from complex sampling designs. *J. Hum. Genet.* 62 (9), 819–829. doi:10.1038/jhg.2017.43

Zhang, Y., Lin, S., and Biswas, S. (2017). Detecting rare and common haplotype–environment interaction under uncertainty of gene–environment independence assumption. *Biometrics* 73 (1), 344–355. doi:10.1111/biom.12567

frontiers | Frontiers in Genetics

TYPE Technology and Code
PUBLISHED 06 April 2023
DOI 10.3389/fgene.2023.1088223

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Tao Wang,
Shanghai Jiao Tong University, China
Ang Dong,
Beijing Forestry University, China

*CORRESPONDENCE
Cen Wu,
✉ wucen@ksu.edu

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

RECEIVED 03 November 2022
ACCEPTED 28 February 2023
PUBLISHED 06 April 2023

CITATION
Zhou F, Liu Y, Ren J, Wang W and Wu C (2023), Springer: An R package for bi-level variable selection of high-dimensional longitudinal data.
Front. Genet. 14:1088223.
doi: 10.3389/fgene.2023.1088223

Springer: An R package for bi-level variable selection of high-dimensional longitudinal data

Fei Zhou[1], Yuwen Liu[1], Jie Ren[2], Weiqun Wang[3] and Cen Wu[1]*

[1]Department of Statistics, Kansas State University, Manhattan, KS, United States, [2]Department of Biostatistics and Health Data Sciences, Indiana University School of Medicine, Indianapolis, IN, United States, [3]Department of Food, Nutrition, Dietetics and Health, Kansas State University, Manhattan, KS, United States

In high-dimensional data analysis, the bi-level (or the sparse group) variable selection can simultaneously conduct penalization on the group level and within groups, which has been developed for continuous, binary, and survival responses in the literature. Zhou et al. (2022) (PMID: 35766061) has further extended it under the longitudinal response by proposing a quadratic inference function-based penalization method in gene–environment interaction studies. This study introduces “springer,” an R package implementing the bi-level variable selection within the QIF framework developed in Zhou et al. (2022). In addition, R package “springer” has also implemented the generalized estimating equation-based sparse group penalization method. Alternative methods focusing only on the group level or individual level have also been provided by the package. In this study, we have systematically introduced the longitudinal penalization methods implemented in the “springer” package. We demonstrate the usage of the core and supporting functions, which is followed by the numerical examples and discussions. R package “springer” is available at https://cran.r-project.org/package=springer.

KEYWORDS

bi-level variable selection, gene–environment interaction, repeated measurements, generalized estimating equation, quadratic inference function

1 Introduction

In gene–environment interaction studies, a central task is to detect important G×E interactions that are beyond main G and E effects. Although the main environmental factors are usually preselected and of low dimensionality, in the presence of a large number of G factors, conducting G×E analysis can be performed in the variable selection framework. Recently, Zhou et al. (2021a) surveyed the penalized variable selection methods for interaction analysis, revealing the pivotal role that the sparse group selection played in G×E studies. Specifically, determining whether a genetic factor, such as the gene expression or SNP, is associated with the disease phenotype is equivalent to feature selection on the group level of main G and G×E interactions with respect to that G factor. Further detection of the main and/or interaction effects demands selection within the group. Such bi-level variable selection methods have been extensively studies under continuous, binary, and survival outcomes in G×E studies (Wu et al., 2018a; Ren et al., 2022a; Ren et al., 2022b; Liu et al., 2022).

Zhou et al. (2022a) have further examined the sparse group variable selection for longitudinal studies where measurements on the subjects are repeatedly recorded over a sequence of units, such as time (Verbeke et al., 2014). In general, major competitors for the bi-level selection include LASSO and group LASSO types of regularization methods that only perform variable selection on the individual and group levels, respectively (Wu and Ma, 2015). Zhou et al. (2022a) have also incorporated two alternatives for comparison under the longitudinal response based on the quadratic inference functions (QIFs) (Qu et al., 2000). The sgQIF, gQIF, and iQIF, denoting the penalized QIF methods accommodating sparse group, group-, and individual-level selections, respectively, have been thoroughly examined with different working correlation structures modeling the relatedness among repeated measurements. All these methods have been implemented in R package *springer*.

In this article, we provide a detailed introduction of R package *springer*, which has implemented not only the proposed and alternative regularized QIF methods from Zhou et al. (2022a) but also their counterparts based on the generalized estimating equations (GEEs) (Liang and Zeger, 1986). The GEE, originally proposed by Liang and Zeger (1986), captures the intra-correlation of repeated measurements using their marginal distributions and a working correlation matrix depending on certain nuisance parameters. The QIF has further improved upon GEE *via* bypassing the nuisance parameters, leading to consistent and optimal estimation of regression coefficients even when the working correlation is misspecified (Qu et al., 2000).

GEE and QIF have been the two major frameworks for developing high-dimensional penalization methods, especially under the main effect models. For example, Wang et al. (2012) have proposed a regularized GEE with the SCAD penalty. Cho and Qu (2013) have considered the penalized QIF with penalty functions including LASSO, adaptive LASSO, and SCAD. More recently, the high-dimensional longitudinal interaction models have been developed based on GEE and QIF (Zhou et al., 2019; Zhou et al., 2022a). In terms of statistical software, R package *PGEE*, developed by Inan and Wang (2017), has implemented the penalized GEE methods from Wang et al. (2012). The package *interep* features the mixture of individual- and group-level penalty under the GEE, where selection on the two levels does not overlap and thus is not a sparse group penalty (Zhou et al., 2019; Zhou et al., 2022b).

Package *springer* is among the first of statistical software to systematically implement bi-level, group-level, and individual-level regularization under both GEE and QIF. It focuses on the longitudinal interaction models where the linear G×E interactions have been assumed (Zhou et al., 2021a). The non-linear G×E interactions usually demand the varying coefficient models and their extensions (Wu and Cui, 2013; Wu et al., 2018b; Ren et al., 2020). In longitudinal studies, Wang et al. (2008) and Tang et al. (2013) have developed regularized variable selection based on varying coefficient (VC) models under the least squares and quantile check loss, respectively. They have assumed independence for repeated measurements, so the within-subject correlation has not been incorporated. Chu et al. (2016), on the other hand, have considered the weighted least squares-based VC models, where the weights have been estimated from a marginal non-parametric model to account for intra-cluster interconnections. R package *VariableScreening* has provided the corresponding R codes and examples.

We have made R package *springer* publicly available on CRAN (Zhou et al., 2021b). The core modules of the package have been developed in C++ for fast computation. We organize the rest of the paper as follows. Section 2 provides a summary of bi-level penalization in longitudinal interaction studies. The main and supporting functions in package *springer* are introduced in Section 3. To demonstrate the usage of the package, we present a simulated example in Section 4 and a case study in Section 5. We conclude the article with discussions in Section 6.

2 Materials and methods

2.1 The bi-level model for longitudinal G×E studies

In a typical longitudinal setting with n subjects, the ith subject $(1 \leqslant i \leqslant n)$ is repeatedly measured over t_i time points, which naturally results in t_i repeated measurements that are correlated for the same subject and are assumed to be independent with the measurements taken from other subjects. Then, Y_{ij} denotes the phenotype measured for the ith subject at time point j $(1 \leqslant j \leqslant t_i)$. $G_{ij} = (G_{ij1}, \ldots, G_{ijp})^\top$ and $E_{ij} = (E_{ij1}, \ldots, E_{ijq})^\top$ represent the p-dimensional vector of genetic factors and the q-dimensional vector of environmental factors, respectively. The bi-level G×E model associates the genetic and environmental main effects and their interactions with the repeatedly measured phenotypic response as follows:

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \epsilon_{ij} \\ &= \alpha_{n0} + \sum_{h=1}^{q} \alpha_{nh} E_{ijh} + \sum_{k=1}^{p} \gamma_{nk} G_{ijk} + \sum_{k=1}^{p} \sum_{h=1}^{q} u_{nhk} E_{ijh} G_{ijk} + \epsilon_{ij} \\ &= \alpha_{n0} + \sum_{h=1}^{q} \alpha_{nh} E_{ijh} + \sum_{k=1}^{p} \left(\gamma_{nk} + \sum_{h=1}^{q} u_{nhk} E_{ijh} \right) G_{ijk} + \epsilon_{ij} \\ &= \alpha_{n0} + \sum_{h=1}^{q} \alpha_{nh} E_{ijh} + \sum_{k=1}^{p} \eta_{nk}^\top Z_{ijk} + \epsilon_{ij}, \end{aligned} \tag{1}$$

where α_{n0} is the intercept, and α_{nh}, γ_{nk}, and u_{nhk} denote the regression coefficients of environmental and genetic main effects and their interactions, correspondingly. We also define $\eta_{nk} = (\gamma_{nk}, u_{n1k}, \ldots, u_{nqk})^\top$, and $Z_{ijk} = (G_{ijk}, E_{ij1} G_{ijk}, \ldots, E_{ijq} G_{ijk})^\top$. Z_{ijk} is a $(q + 1)$-dimensional vector representing the main and interaction effects with respect to the kth genetic factor. For $1 \leqslant j \leqslant t_i$, the random error ϵ_{ij} has mean zero and a finite variance. For convenience, the random error ϵ_i is assumed to be multivariate normal as $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{it_i})^\top \sim N_{t_i}(0, \Sigma_i)$, where Σ_i is the covariance matrix corresponding to the ith subject. From now on, we let $t_i = t$. Combined, we can write $\alpha_n = (\alpha_{n1}, \ldots, \alpha_{nq})^\top$, $\eta_n = (\eta_{n1}^\top, \ldots, \eta_{np}^\top)^\top$, and $Z_{ij} = (Z_{ij1}^\top, \ldots, Z_{ijp}^\top)^\top$. The length of the coefficient vector η_n is $p + pq$. Then, model (1) can be equivalently expressed as

$$Y_{ij} = \alpha_{n0} + E_{ij}^\top \alpha_n + Z_{ij}^\top \eta_n + \epsilon_{ij}.$$

The $(1 + q + p + pq)$-dimensional vectors $\beta_n = (\alpha_{n0}, \alpha_n^\top, \eta_n^\top)^\top$ and $W_{ij} = (1, E_{ij}^\top, Z_{ij}^\top)^\top$ are denoted, and a concise form of model (1) is formed as follows:

$$Y_{ij} = W_{ij}^{\top}\beta_n + \epsilon_{ij}.$$

The aforementioned model provides a general formulation under the longitudinal design in which both the response variable and predictors are repeatedly measured. Here, the predictors are G and E main effects and G×E interactions. It still works when only one or neither of the G and E factors are repeatedly measured. In the real data analyzed in Zhou et al. (2022a), both the G and E factors in the interaction study do not vary across time.

2.2 An overview of interaction studies based on GEE and QIF

R package *springer* (Zhou et al., 2021b) includes methods that account for repeated measurements based on the GEE and QIF, respectively. Here, we briefly review the two frameworks for longitudinal interaction studies.

The **generalized estimating equation** has been proposed by Liang and Zeger (1986) to account for intra-cluster correlations using a marginal model by specifying the conditional expectation and variance of each response, Y_{ij}, and the conditional pairwise within-subject association among the vector of repeatedly measured phenotypes. In the longitudinal interaction studies, the marginal expectation of the response is $\mathrm{E}(Y_{ij}) = \mu_{ij} = W_{ij}^{T}\beta_n$, and the conditional variance of Y_{ij} is $\mathrm{Var}(Y_{ij}) = \delta(\mu_{ij})$, where $\delta(\mu_{ij})$ is a known function of the mean μ_{ij}. Then, the score equation for the longitudinal G×E model is defined as

$$\sum_{i=1}^{n}\frac{\partial\mu_i(\beta_n)}{\partial\beta_n}V_i^{-1}(Y_i - \mu_i(\beta_n)) = 0,$$

where $Y_i = (Y_{i1},\ldots,Y_{it})^{\top}$ and the covariance matrix for the intra-subject association V_i is defined as $V_i = A_i^{\frac{1}{2}}R_i(\nu)A_i^{\frac{1}{2}}$. Here, for the ith subject, the diagonal matrix A_i is defined as $A_i = \mathrm{diag}\{\mathrm{Var}(Y_{i1}), \ldots, \mathrm{Var}(Y_{it})\}$, and the "working" correlation matrix $R_i(\nu)$ depends on a finite dimensional parameter vector ν, characterizing the within-subject association. We have $\mu_i(\beta_n) = (\mu_{i1}(\beta_n),\ldots,\mu_{it}(\beta_n))^{\top}$. The ratio term in the aforementioned score equation is equivalent to $W_i = (W_{i1},\ldots,W_{it})^{\top}$. Then, the GEE estimator, $\widehat{\beta}_n$, is the corresponding solution.

The term "working" correlation in GEE is adopted to distinguish $R_i(\nu)$ from the true underlying correlation among intra-subject measurements. Liang and Zeger (1986) have shown that when ν is consistently estimated, the GEE estimator is consistent even if the correlation structure is not correctly specified. However, there is a cost under such misspecification, that is, the GEE estimator is no longer efficient, and ν cannot be consistently estimated.

The **quadratic inference function** overcomes the disadvantage of GEE by avoiding the direct estimation of ν (Qu et al., 2000). It has also been shown that even when the correlation structure is misspecified, the QIF estimator is still optimal. With the bi-level modeling of G×E interactions under the longitudinal response, the inverse of $R(\nu)$ can be calculated by a linear combination of basis matrices within the QIF framework. Specifically, $\mathrm{R}(\nu)^{-1} \approx \sum_{k=1}^{m}c_kB_k$, where B_1 is an identity matrix and B_2, ... , B_m are symmetric basis matrices with unknown coefficients c_1, ... c_m. The specifications of these basis matrices are dependent on the types of working correlation (Qu et al., 2000). The score equations can be rewritten as

$$\sum_{i=1}^{n}W_i^{\top}A_i^{-\frac{1}{2}}(c_1B_1 + \cdots + c_mB_m)A_i^{-\frac{1}{2}}(Y_i - \mu_i(\beta_n)). \tag{2}$$

Accordingly, for the ith subject, we define the extended score vector, $\phi_i(\beta_n)$, for the bi-level G×E model as

$$\phi_i(\beta_n) = \begin{pmatrix} W_i^{\top}A_i^{-\frac{1}{2}}B_1A_i^{-\frac{1}{2}}(Y_i - \mu_i(\beta_n)) \\ \cdot \\ \cdot \\ \cdot \\ W_i^{\top}A_i^{-\frac{1}{2}}B_mA_i^{-\frac{1}{2}}(Y_i - \mu_i(\beta_n)) \end{pmatrix}. \tag{3}$$

We then denote the extended score for all subjects as $\overline{\phi_n}(\beta_n) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(\beta_n)$. The linear combination of all components in $\overline{\phi_n}(\beta_n)$ directly leads to the estimation functions in Eq. 2. The quadratic inference function based on the extended score $\overline{\phi_n}(\beta_n)$ is defined as

$$\mathrm{Q}_n(\beta_n) = \overline{\phi_n}^{\top}(\beta_n)\overline{\Omega_n}(\beta_n)^{-1}\overline{\phi_n}(\beta_n),$$

where the sample covariance matrix of ϕ_i (β_n) is $\overline{\Omega_n}(\beta_n) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(\beta_n)\phi_i(\beta_n)^{\top}$. Minimizing the aforementioned quadratic inference function yields $\widehat{\beta}_n$, i.e., $\widehat{\beta}_n = \arg\min_{\beta_n}\mathrm{Q}_n(\beta_n)$. It should be noted that the minimization does not involve the coefficients c_1, ... c_m in Eq. 2.

2.3 Penalized QIF for the bi-level longitudinal G×E interaction studies

R package *springer* (Zhou et al., 2021b) can perform penalized sparse group variable selection based on both the GEE and QIF framework in order to identify an important subset of main and interaction effects that are associated with the longitudinal phenotype. As QIF is an extension of GEE, we focus on the penalized bi-level QIF in the main text and introduce GEE-based methods in the Supplementary Appendix. The following regularized bi-level QIF has been proposed in Zhou et al. (2022a):

$$U(\beta_n) = Q(\beta_n) + \sum_{k=1}^{p}\rho\left(\|\eta_{nk}\|_{\Sigma_k};\lambda_1,\gamma\right) + \sum_{k=1}^{p}\sum_{h=1}^{q+1}\rho\left(|\eta_{nkh}|;\lambda_2,\gamma\right), \tag{4}$$

where the minimax concave penalty is $\rho(t;\lambda,\gamma) = \lambda\int_0^t(1 - \frac{x}{\gamma\lambda})_+dx$ on $[0,\infty)$ with the tuning parameter λ and regularization parameter γ (Zhang, 2010). The group-level penalty $\rho(\|\eta_{nk}\|_{\Sigma_k};\lambda_1,\gamma)$ is imposed on $\|\eta_{nk}\|_{\Sigma_k}$, which is the empirical norm of η_{nk}, to determine whether the kth SNP has any contribution to the variation in the repeatedly measured phenotype. We define the empirical norm as $\|\eta_{nk}\|_{\Sigma_k} = (\eta_{nk}\Sigma_k\eta_{nk})^{1/2}$ with $\Sigma_k = n^{-1}B_k^{\top}B_k$, where B_k is the subset of the design matrix corresponding to the interactions between the kth genetic factor and all the E factors. If η_{nk} is estimated as a zero vector, the kth SNP is not associated with the phenotypic response. Otherwise, the individual-level penalty $\rho(|\eta_{nkh}|;\ \lambda_2,\ \gamma)$ further selects the main and interaction effects that are associated with the phenotype.

Our choice of the baseline penalty function is the MCP, and the corresponding first derivative function of MCP is defined as $\rho'(t;\lambda,\gamma) = (\lambda - \frac{t}{\gamma})I(0 \le t \le \gamma\lambda)$.

The penalized QIF in (4) is the extension of bi-level variable selection to longitudinal studies, which conducts selections of important groups and individual members within the group simultaneously. It is worth noting that the penalized GEE model proposed by Zhou et al. (2019) does not perform within-group selection. The shrinkage has been imposed on the individual level (G main effect) and group level (G×E interactions) separately. Unlike the model in (4), the terms selected on the individual level in the study by Zhou et al. (2019) are not members of the group. Therefore, it is not the sparse group selection, although in a loose sense, it can be treated as a bi-level variable selection method.

A general form for the objective function of regularization methods is "unpenalized objective function + penalty function" (Wu and Ma, 2015). QIF and GEE are widely adopted unregularized objective functions for repeated measurement studies. LASSO and SCAD have been considered the penalty functions in longitudinal studies, where selection of the main effects are of interest (Wang et al., 2012; Cho and Qu, 2013; Ma et al., 2013). To accommodate more complicated structured sparsity incurred by interaction effects, the shrinkage components in Eq. 4 adopts MCP as the baseline penalty to perform individual- and group-level penalization simultaneously. It is commonly recognized that the structure-specific regularization functions are needed to accommodate different sparsity patterns. For example, to account for strong correlations among predictors, network-based variable selection methods have been developed (Ren et al., 2019; Huang et al., 2021). The penalty functions have been implemented in a diversity of R packages. For example, under generalized linear models, the package *glmnet* has included LASSO and its extensions, such as the ridge penalty and elastic net (Friedman et al., 2010a). R package *regnet* has been developed for network-based penalization under continuous, binary, and survival responses with possible choices on robustness (Ren et al., 2017; Ren et al., 2019). With the longitudinal response, R package *PGEE* has adopted SCAD penalty for penalized GEE to select main effects (Inan and Wang, 2017), and package *interep* has been designed in interaction studies based on MCP (Zhou et al., 2022b).

2.4 The bi-level selection algorithm based on QIF

Optimization of the penalized QIF in (4) demands the Newton–Raphson algorithm that can update $\hat{\beta}_n$ iteratively. Specifically, the estimated coefficient vector $\hat{\beta}_n^{g+1}$ can be obtained based on $\hat{\beta}_n^g$ at the gth iteration as follows:

$$\hat{\beta}_n^{g+1} = \hat{\beta}_n^g + \left[V\left(\hat{\beta}_n^g\right) + nH\left(\hat{\beta}_n^g\right)\right]^{-1}\left[P\left(\hat{\beta}_n^g\right) - nH\left(\hat{\beta}_n^g\right)\hat{\beta}_n^g\right], \qquad (5)$$

where $P(\hat{\beta}_n^g)$ and $V(\hat{\beta}_n^g)$ can be obtained as

$$P\left(\hat{\beta}_n^g\right) = -\frac{\partial Q\left(\hat{\beta}_n^g\right)}{\partial \beta_n} = -2\frac{\partial \overline{\phi_n}}{\partial \beta_n}^{\top}\overline{\Omega_n}^{-1}\overline{\phi_n}\left(\hat{\beta}_n^g\right),$$

and

$$V\left(\hat{\beta}_n^g\right) = \frac{\partial^2 Q\left(\hat{\beta}_n^g\right)}{\partial^2 \beta_n} = 2\frac{\partial \overline{\phi_n}}{\partial \beta_n}^{\top}\overline{\Omega_n}^{-1}\frac{\partial \overline{\phi_n}}{\partial \beta_n}.$$

Moreover, $H(\hat{\beta}_n^g)$ is a diagonal matrix consisting of derivatives of both the individual-and group-level penalty functions, which is defined as

$$\begin{aligned} H\left(\hat{\beta}_n^g\right) = \mathrm{diag}\Bigg(&\underbrace{0,\ldots,0}_{1+q}, \underbrace{\frac{\rho'\left(\|\hat{\eta}_{n1}^g\|_{\Sigma_1}; \sqrt{q+1}\lambda_1, \gamma\right)}{\epsilon + \|\hat{\eta}_{n1}^g\|_{\Sigma_1}}, \ldots, \frac{\rho'\left(\|\hat{\eta}_{n1}^g\|_{\Sigma_1}; \sqrt{q+1}\lambda_1, \gamma\right)}{\epsilon + \|\hat{\eta}_{n1}^g\|_{\Sigma_1}}}_{1+q}, \ldots, \\ &\underbrace{\frac{\rho'\left(\|\hat{\eta}_{np}^g\|_{\Sigma_p}; \sqrt{q+1}\lambda_1, \gamma\right)}{\epsilon + \|\hat{\eta}_{np}^g\|_{\Sigma_p}}, \ldots, \frac{\rho'\left(\|\hat{\eta}_{np}^g\|_{\Sigma_p}; \sqrt{q+1}\lambda_1, \gamma\right)}{\epsilon + \|\hat{\eta}_{np}^g\|_{\Sigma_p}}}_{1+q}\Bigg) + \mathrm{diag}\Bigg(\underbrace{0,\ldots,0}_{1+q}, \\ &\underbrace{\frac{\rho'\left(|\hat{\eta}_{n11}^g|; \lambda_2, \gamma\right)}{\epsilon + |\hat{\eta}_{n11}^g|}, \ldots, \frac{\rho'\left(|\hat{\eta}_{n1\,(q+1)}^g|; \lambda_2, \gamma\right)}{\epsilon + |\hat{\eta}_{n1\,(q+1)}^g|}}_{1+q}, \ldots, \underbrace{\frac{\rho'\left(|\hat{\eta}_{np1}^g|; \lambda_2, \gamma\right)}{\epsilon + |\hat{\eta}_{np1}^g|}, \ldots, \frac{\rho'\left(|\hat{\eta}_{np\,(q+1)}^g|; \lambda_2, \gamma\right)}{\epsilon + |\hat{\eta}_{np\,(q+1)}^g|}}_{1+q}\Bigg), \end{aligned}$$

where the small positive fraction ϵ is set to 10^{-6} to guarantee the numerical stability when the denominator approaches zero. Since the intercept and the environmental factors are not subject to shrinkage selection, the first $(1 + q)$ entries on the main diagonal of the matrix are zero accordingly. With fixed tuning parameters, $\hat{\beta}_n^{g+1}$ is updated iteratively following Eq. 5. The update stops when the convergence criterion has been reached, that is, the difference between the L_1 norm of $\hat{\beta}_n^{g+1}$ and $\hat{\beta}_n^g$ is less than a cutoff (e.g., 0.001). Numerical studies have shown that only a small to moderate number of iterations are required upon convergence (Zhou et al., 2022a).

The sparse group penalty (4) incorporates two tuning parameters, λ_1 and λ_2, to determine the amount of shrinkage on the group and individual level, correspondingly. An additional regularization parameter γ further balances the unbiasedness and convexity of MCP. The performance of the proposed regularized QIF is insensitive under different choices of γ (Zhou et al., 2022a). The best pair of (λ_1, λ_2) can be searched over the two-dimensional grid through K-fold cross-validation. We first split the dataset into K non-overlapping portions of roughly the same size and held out the kth $(k = 1, \ldots, K)$ fold as the testing dataset. The rest of the data are used as training data to fit a regularized QIF by giving a specific pair of (λ_1, λ_2). n_k and n_{-k} denote the index sets of subjects as training and testing samples, respectively. We can compute the prediction error on testing data as

$$\mathrm{PE}_{-k}(\lambda_1, \lambda_2) = \frac{1}{|n_{-k}|}\sum_{i \in n_{-k}}\left(Y_i - \mu_i\left(\hat{\beta}_{n_k}\right)\right)^2,$$

where $|n_{-k}|$ is the size of testing data, and $\hat{\beta}_{n_k}$ is the regularized coefficient obtained using the training data. The computation cycles through each of the K fold for $k = 1, 2. ., K$, yielding the following cross-validation error:

$$\mathrm{CV}(\lambda_1, \lambda_2) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{PE}_{-k}(\lambda_1, \lambda_2). \qquad (6)$$

The cross-validation value with respect to each pair of (λ_1, λ_2) can be retrieved across the entire two-dimensional grid. The optimal pair of tunings is corresponding to the smallest CV value. Details of the algorithm are given as follows:

```
1 The two-dimensional grid of $(\lambda_1, \lambda_2)$ is provided with an appropriate range.
2 Under the fixed $(\lambda_1, \lambda_2)$,
    (a) $\hat{\beta}_n^0$ is initialized using LASSO
    (b) at the $(g + 1)^{\text{th}}$ iteration, $V(\hat{\beta}_n^g), H(\hat{\beta}_n^g), P(\hat{\beta}_n^g)$ is computed and
    (c) $\hat{\beta}_n^{d+1}$ is updated according to Eq. 5.
    (d) The cross-validation error is calculated using Eq. 6.
3 Step 2 is repeated for each pair of $(\lambda_1, \lambda_2)$ until convergence.
4 The optimal $(\lambda_1, \lambda_2)$ is found under the smallest cross-validation error. The corresponding $\hat{\beta}_n$ is reported.
```

The validation approach is a popular alternative of tuning selection to bypass the computational intensity of cross-validation. When the data-generating model is available, the independent testing data with much larger size can be readily generated. Then, the prediction performance of the fitted sparse group PQIF model under (λ_1, λ_2) can be assessed on the testing data directly. On the contrary, in cross-validation, the prediction error can only be obtained after cycling through all the K folds as shown by Equation 6.

3 R package *springer*

Package *springer* includes two core functions, namely, springer and cv.springer. The function springer can fit both GEE- and QIF-based penalization models under longitudinal responses in G×E interaction studies. The function cv.springer computes the prediction error in cross-validation. Moreover, the package also includes supporting functions reformat, penalty, and dmcp, which have been developed by the authors. To speed up computation, we have implemented the Newton–Raphson algorithms in C++. The package is thus dependent on R packages Rcpp and RcppArmadillo (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Sanderson, 2014).

3.1 The core functions

In package *springer*, the R function for computing the penalized estimates under fixed tuning parameters is

springer (clin = NULL,e, g, y, beta0, func, corr, **structure**, lam1, lam2, maxits = 30,tol = 0.001).

The clinical covariates and environmental and genetic factors can be specified by the input arguments clin, e, and g, respectively. This is different from packages conducting feature selection for the main effects, such as *glmnet* and *PGEE*, where the entire design matrix should be used an input (Friedman et al., 2010a; Inan and Wang, 2017). In interaction studies, the design matrix has a much more complicated structure. Our package is user friendly in that users only need to provide the clinical, g, and e factors, and then the function springer will automatically formulate the design matrix tailored for interaction analysis. The clinical covariates are not involved in the interactions with G factors and are not subject to selection. The argument beta0 denotes the initial value of $\hat{\beta}_n^0$, which is used at the first iteration of the Newton–Raphson algorithm. Typical choices of beta0 include the LASSO or ridge estimates under the cross-sectional phenotype measured at one of the time points or the average of the within-subject phenotypic measurements.

The character string argument func specifies one of the two frameworks (GEE and QIF) to be used for regularized estimation. One of the three working correlations from AR-1, exchangeable, and independence can be called through the input argument corr. For example, corr = “exchangeable,” corr = “AR-1,” and corr = “independence” denote exchangeable, AR-1, and independent correlation, respectively. In addition to the bi-level structure, this package has also included sparsity structures on the group and individual level, respectively. To use the bi-level PQIF under the exchangeable working correlation proposed by Zhou et al. (2022a), we need to specify func = ”QIF,” structure = ”bi-level,” and corr = ”exchangeable” at the same time. It is worthwhile noting that the bi-level selection requires two tuning parameters to impose sparsity. When structure = ”group” or structure = ”individual,” only one of the two tuning parameters lam1 and lam2 is needed.

The Newton–Raphson algorithms implemented in the package *springer* proceed in an iterative manner. The input argument maxits provides the maximum number of iterations determined by the users. We can supply the small positive fraction ϵ that is used to ensure the stability of the algorithm through argument tol.

In package *springer*, function cv.springer performs cross-validation based on the regularized coefficients provided by springer. The R code is

cv.springer (clin = NULL,e, g, y, beta0, lambda1, lambda2, nfolds, func, corr, **structure**, maxits = 30,tol = 0.001).

The function cv.springer calls springer to conduct cross-validation over a sequence of tuning parameters and report the corresponding cross-validation error. Therefore, it is not surprising to observe that the two functions share a common group of arguments involving the input of data and specifications on the penalization method used for estimation. Unlike the scalars of lam2 and lam2 in function springer, the arguments lambda1 and lambda2 are user-supplied sequences of tuning parameters. For bi-level selection, cv.springer calculates the prediction error across each pair of tunings determined by lambda1 and lambda2. The number of folds used in cross-validation is specified by nfolds.

3.2 Additional supporting functions

Package *springer* also provides multiple supporting functions in addition to the core functions. As MCP is the baseline penalty adopted in all the penalized variable selection methods implemented in the package, the function dmcp denotes its first-order derivative function used in the formulation under the Newton–Raphson algorithm. The function penalty determines the type of sparse structure (individual-, group-, or bi-level) imposed for variable selection. Both the group- and bi-level penalizations involve the empirical norm $\|\eta_{nk}\|_{\Sigma_k}$. In practice, the form of Σ_k is not unique. For example, Σ_k can be chosen as an identity matrix, and then $\|\eta_{nk}\|_{\Sigma_k}$ reduces to an L_2 norm. While the alternatives might be equally applicable, the default choice of Σ_k in package *Springer* is in the form discussed in Section 2.3.

It is assumed that repeated measurements on the response are given in the wide format with the dimension of 100 by 5, where 100 is the sample size and 5 is the number of time points, then we can use function `reformat` to convert the wide format to long format with dimension 500 by 1. Similarly, the design matrix under sample size 100 and 50 main and interaction effects has a dimensionality of 100 by 50, if they do not vary across time. Then, `reformat` will return a 500 by 51 wide format matrix including the column of intercept. An "id" column will also be generated by `reformat` to show the time points corresponding to 500 columns. Moreover, a simulated dataset, `dat`, is provided to demonstrate the penalized selection in the proposed longitudinal study. We describe more details in the next section.

4 Simulation example

In this section, we demonstrate the fit of bi-level selection using package *Springer* based on simulated datasets. Although model (1) is general in the sense that both the response and predictors are repeatedly measured, it can be reduced to the case where the predictors, consisting of the clinical covariates and environmental and genetic factors, are cross-sectional under the longitudinal response. Model (1) is flexible in which the predictors can have a mixture of cross-sectional and longitudinal measurements. For instance, the repeated measurements are only taken on E factors and not on clinical or G factors.

The motivating dataset for the sparse group variable selection developed in Zhou et al. (2022a) can be retrieved from the Childhood Asthma Management Program (CAMP) in our case study where the clinical, E, and G factors are not repeatedly measured (Childhood Asthma Management Program Research Group, 1999; Childhood Asthma Management Program Research Group Szefler et al., 2000; Covar et al., 2012). Therefore, the current version (version 0.1.7) of package *springer* only accounts for such a case. It is worth noting that technically it is not difficult to extend the package to repeatedly measured predictors because the only difference lies in using time-specific measurements rather than repeating the cross-sectional measurements across all the time points in the estimation procedure. We will discuss potential extensions of the package at the end of this section. In the following simulated example, the longitudinal responses are generated together with cross-sectional predictors. The data-generating function is provided as follows:

```
Data <- function (n,p,k,q)
{
y = matrix (rep (0,n*k),n,k)
sig = matrix (0,p,p)
for (i in 1: p) {
for (j in 1: p) { sig [i,j] = 0.8^abs (i-j) }
}
# Generate genetic factors
g = mvrnorm (n,rep (0,p),sig)
sig0 = matrix (0,q,q)
for (i in 1: q) {
for (j in 1: q) { sig0 [i,j] = 0.8^abs (i-j) }
}
# Generate environmental factors
e = mvrnorm (n,rep (0,q),sig0)
E0 = as.numeric (g [,1]<=0)
E0 = E0+1
e = cbind (E0,e [,-1])
e.out = e
e1 = cbind (rep (1,dim(e)[1]),e)
for (i in 1:p) { e = cbind (e,g [,i]*e1) }
x = scale(e)
ll = 0.3
ul = 0.5
coef = runif (q+25,ll,ul)
mat = x [,c (1:q, (q+1), (q+2), (q+6), (q+4), (2*q+2),
(2*q+3), (2*q+7),
(2*q+5), (3*q+3), (3*q+4), (3*q+8), (3*q+6),
(4*q+4), (4*q+5), (4*q+9), (4*q+7), (5*q+5),
(5*q+6), (5*q+10), (5*q+8), (6*q+6), (6*q+7),
(6*q+11), (6*q+9), (7*q+7))]
for (u in 1:k){ y [,u] = 0.5 + rowSums (coef*mat) }
#Exchangable correlation for repeated measurements
sig1 = matrix (0,k,k)
diag (sig1) = 1
for (i in 1: k) {
for (j in 1: k) { if (j != i){sig1 [i,j] = 0.8} } }
error = mvrnorm (n,rep (0,k),sig1)
y = y + error
dat = list (y = y,x = x,e = e.out, g = g, coef = c (0.5,coef))
return (dat)
}
```

In the aforementioned codes, n, p, and q represent the sample size, dimension of the genetic factors, and environmental factors, respectively. The number of repeated measurements is k. Now, we simulate a dataset with 400 subjects, 100 G factors, and 5 E factors. The number of repeated measurements is set to 5. The correlation coefficient ρ of the compound symmetry working correlation assumed for longitudinal measurements is 0.8. In the data-generating function, `coef` represents the vector of non-zero coefficients, and `mat` is the part of design matrix corresponding to the main and interaction effects associated with non-zero coefficients. With $(n, p, q) = (400, 100, 5)$, `coef` is a vector of length 30, and `mat` is a 400-by-30 matrix. The R code `coef*mat` denotes element-wise multiplication by multiplying the non-zero coefficient to the corresponding main or interaction effects. Therefore, `rowSums(coef*mat)` returns a 400-by-1 vector. The code "0.5 + `rowSums(coef*mat)`" stand for the combined effects from those important main and interaction effects, and the intercept, with 0.5 being the coefficient multiplied to the intercept. We listed the R codes and output in the following section:

```
library (MASS)
library (glmnet)
library (springer)
set.seed (123)
n.train = n = 400
p = 100; k = 5; q = 5
dat.train = Data(n.train,p,k,q)
y.train = dat.train$y
```

```
x.train = dat.train$x
e.train = dat.train$e
g.train = dat.train$g
> dim(y.train)
[1] 400 5
> dim(x.train)
[1] 400 605
> dim(e.train)
[1] 400 5
> dim(g.train)
[1] 400 100
```

In addition, the R codes `dat.train$coef` saves the non-zero coefficients used in the data-generating model. By setting the seed, we can reproduce the data generated through calling the `Data`. A total of 100 genetic factors and 5 environmental factors lead to a total of 605 main and interaction effects, excluding the intercept. We first obtain the initial value of the coefficient vector $\widehat{\beta_0}$ by fitting ridge regression under the univariate response taken from a single time point. Other choices of initial values include fitting ridge regression or LASSO under the average of within-subject measurements, which accommodate the case of unbalanced data, where a proper single point might be difficult to determine. In general, the regularized estimates remain relatively insensitive to different choices of initial value $\widehat{\beta_0}$, as long as $\widehat{\beta_0}$ is reasonable, in other words, not extremely far away from the optimal solution.

```
x.train1 = cbind (data.frame (rep (1,n)),x.train)
x.train1 = data.matrix (x.train1)
lasso.cv = cv.glmnet (x.train1,y.train [,1],alpha = 0,nfolds = 5)
alpha = lasso.cv$lambda.min/2
lasso.fit = glmnet (x.train1,y.train [,1],
family = "gaussian",alpha = 0,nlambda = 100)
beta0 = as.matrix (as.vector (predict (lasso.fit,
s = alpha, type = "coefficients"))[-1])
```

With the initial value obtained previously, we call function `cv.springer` to calculate cross-validation errors corresponding to the pair of tuning parameters (`lambda1` and `lambda2`). The number of fold is 5 by setting `nfolds` to 5 in the following codes. Then, a penalized bi-level QIF model with an independence correlation has been fitted to the simulated data with the optimal tunings. The fitted regression coefficients are saved in `fit.beta`.

```
lambda1 = seq (0.025,0.1,length.out = 5)
lambda2 = seq (1,1.5,length.out = 3)
tunning = cv.springer (clin = NULL, e.train, g.train, y.train, beta0,
lambda1, lambda2, nfolds = 5, func = "QIF",
corr = "independence",structure = "bilevel",
maxits = 30, tol = 0.1)
lam1 = tunning$lam1
lam2 = tunning$lam2
> lam1
[1] 0.0625
> lam2
[1] 1
> tunning$CV
          [,1]      [,2]      [,3]
[1,] 14.873142 15.37916 16.02844
[2,] 12.282850 13.23239 13.81465
[3,] 9.663655 10.62635 11.96531
[4,] 10.133435 11.00219 12.25365
[5,] 11.237012 11.79566 13.17813
fit.beta = springer (clin = NULL, e.train, g.train, y.train, beta0,
func = "QIF",corr = "independence",
structure = "bilevel",lam1,lam2,maxits = 30,tol = 0.1)
```

To assess the model's performance, we will compare the fitted coefficient vector `fit.beta` with the true coefficient vector, which is used to simulate the response variable in `Data`. Since the codes `dat.train$coef` only report the true non-zero coefficient, the resulting vector has a length much less than `fit.beta,` which includes zero coefficient. Therefore, we first retrieve locations of non-zero effects in the coefficient vector used to generate the longitudinal response. In the following codes, `tp`, `tp.main`, and `tp.interaction` represent the locations for all the non-zero effects, that is, the column number of the corresponding effects in the design matrix. Although the coefficients are randomly generated from uniform distributions, the locations of the non-zero effects are fixed. In total, there are 30 non-zero effects, consisting of 5 environmental factors, 7 genetic factors, and 18 gene–environment interactions.

```
## non-zero effects without intercept
tp = c(1:q, (q+1), (q+2), (q+6), (q+4), (2*q+2), (2*q+3), (2*q+7), (2*q+5),
(3*q+3), (3*q+4), (3*q+8), (3*q+6), (4*q+4), (4*q+5), (4*q+9), (4*q+7),
(5*q+5), (5*q+6), (5*q+10), (5*q+8), (6*q+6), (6*q+7), (6*q+11),
(6*q+9), (7*q+7))+1
## non-zero main effects
tp.main = c((q+2), (2*q+3), (3*q+4), (4*q+5), (5*q+6), (6*q+7), (7*q+8))
## non-zero interaction effects
tp.interaction = c((q+2), (q+6), (q+4), (2*q+3), (2*q+7), (2*q+5),
(3*q+4), (3*q+8), (3*q+6), (4*q+5), (4*q+9), (4*q+7), (5*q+6), (5*q+10),
(5*q+8), (6*q+7), (6*q+11), (6*q+9))+1
```

We run the codes in R console to evaluate the accuracy in parameter estimation. The precision in estimating the regression coefficients has been assessed based on `TMSE`, `MSE`, and `NMSE`, respectively. The mean squared error of the fitted coefficient vector `fit.beta` with respect to the true one, denoted as `TMSE`, is defined as

$$\mathrm{TMSE} = \frac{1}{1+p+q+pq}\|\hat{\beta}_n - \beta_n\|,$$

where $\hat{\beta}_n$ corresponds to `fit.beta` and β_n is the true regression coefficient vector used to generate the response in the data-generating function. In this simulation example, there are 100 genetic factors (p = 100) and 5 environmental factors (q = 5), resulting in a coefficient vector of length 606, including the intercept. To observe the estimation accuracy on a finer scale, we further dissect β_n into the component corresponding to `tp` and calculate the mean square error with respect to the counterpart from `fit.beta`, denoted as `MSE`. The mean square error is computed based on the rest of `fit.beta`, and β_n is defined as `NMSE`. The R codes and output are listed as follows:

```
coeff = matrix (fit.beta, length (fit.beta),1)
coeff.train = rep (0,length (coeff))
coeff.train [tp] = dat.train$coef[-1]
TMSE = mean ((coeff-coeff.train)^2)
MSE = mean ((coeff [tp]-coeff.train [tp])^2)
NMSE = mean ((coeff [-tp]-coeff.train [-tp])^2)
> TMSE
[1] 0.003455488
> MSE
[1] 0.06563788
> NMSE
[1] 0.0002168221
```

The `dat.train$coef` only consists of the non-zero coefficients used to generate longitudinal responses in the data-generating model; therefore, its dimension is not the same as `fit.beta` as the estimated regression coefficient vector is sparse and includes zero coefficient, thus having a much larger dimension. In regularized variable selection, the non-zero coefficients from `fit.beta` will not be identical to those in `dat.train$coef` due to the shrinkage estimation in order to achieve variable selection. The aforementioned output shows the estimation errors in terms of `TMSE`, `MSE`, and `NMSE`, respectively. The `NMSE` is much smaller than the `MSE` since it computes the MSE with respect to zero coefficients.

In addition to evaluating the accuracy in parameter estimation, we also examine the performance in identification in terms of number of true- and false-positive effects. Specifically, by comparing the locations of the non-zero components in `fit.beta` and the true coefficient vector used in the data-generating model, we can report the total number of true- and false-positive effects, such as `TP` and `FP`. The identification results have also been summarized for the main genetic effects (`TP1` and `FP1`) and G×E interactions (`TP2` and `FP2`). The locations of important effects saved in `tp` obtained from the chunk of R codes previously also include the environmental main effects that are not subject to selection. When calculating the number of true and false positives in the next section, we only count the effects that are under selection, corresponding to the 7 G factors and 18 G×E interactions. The output is provided in the following section.

```
coeff [abs (coeff) < 0.1] = 0
coeff [1: (1 + q)] = 0
ids = which (coeff != 0)
TP = length (intersect (tp,ids))
res = ids [is.na (pmatch (ids,tp))]
FP = length (res)
coeff1 = rep (0,length (coeff))
coeff1 [1: (1 + q)] = coeff [1: (1 + q)]
for (i in (q+2):length (coeff)) {
if ( i%%(q+1)==1) coeff1 [i]= coeff[i]
}
ids1 = which (coeff1 != 0)
TP1 = length (intersect (tp.main,ids1))
res1 = ids1 [is.na (pmatch (ids1,tp.main))]
FP1 = length (res1)
coeff2 = coeff
coeff2 [1: (1 + q)] = 0
for (i in (q+2):length (coeff)) {
if ( i%%(q+1)==1) coeff2[i] = 0
}
ids2 = which (coeff2 != 0)
TP2 = length (intersect (tp.interaction,ids2))
res2 = ids2 [is.na (pmatch (ids2,tp.interaction))]
FP2 = length (res2)
> TP
[1] 21
> FP
[1] 3
> TP1
[1] 6
> FP1
[1] 0
> TP2
[1] 15
> FP2
[1] 3
```

Results on true and false positives indicate that six out of the seven important main effects have been identified, and 15 out of the 18 interactions used in the data-generating model have been detected. The number of identified false-positive effects is three.

In addition to extensive simulation studies that demonstrate the merit of the proposed sparse group variable selection in longitudinal studies, Zhou et al. (2022a) have also considered scenarios in the presence of missing measurements (Rubin, 1976; Little and Rubin, 2019). Under the pattern of missing completely at random (MCAR), the penalized QIF procedure can still be implemented by using a transformation matrix to accommodate missingness. Such a data-transformation procedure will be incorporated in the release of package *springer* in the near future.

The current version of package *springer* (version 0.1.7) has implemented three working correlation matrices, independence, AR-1, and exchangeable, for individual-, group-, and bi-level variable selection under continuous longitudinal responses in both the GEE and QIF frameworks. The future improvement includes incorporating other working correlations, such as the unstructured working correlation. A question worth exploring is the computational feasibility of unstructured working correlation under QIF as the large number of covariance parameters will potentially lead to much more complicated extended score vectors, incurring prohibitively heavy computational cost for high-dimensional data. We will also consider extensions to discrete responses such as binary, count, and multinomial responses, and longitudinally measured clinical, environmental, and genetic factors, especially after these data are available.

5 Case study

We adopt package *springer* to analyze the high-dimensional longitudinal data from the Childhood Asthma Management Program (Childhood Asthma Management Program Research Group, 1999; Childhood Asthma Management Program Research Group Szefler et al., 2000; Covar et al., 2012). Children with age between 5 and 12 years, who are diagnosed with chronic asthma

TABLE 1 Identified main and interaction effects based on the genes from the Wnt signaling pathway on chromosome 6.

SNP	Gene		Treatment	Age	Gender
rs10948011	TAF8	0	0	0	−0.020
rs33954419	USP49	−0.012	0	0	0
rs12194513	TAF8	0.005	0	0	0
rs205339	MAP3K7	0.016	0	0	0
rs11970772	CCND3	0	0.102	0	0.069
rs1018155	DAAM2	0	0	−0.169	0
rs913574	DAAM2	0	−0.020	0	0
rs13191407	MAP3K7	0	0	−0.009	−0.023
rs2475802	MOCS1	0.095	0	0	0
rs805300	BAG6	−0.110	0	0	0
rs1475114	MOCS1	−0.047	0	0	0
rs1018156	DAAM2	−0.045	0	0	0
rs4607417	CCND3	0	−0.108	0	0
rs284513	MAP3K7	0	0.040	0.075	0.011
rs17812916	RSPO3	0	0.021	0	0.208
rs2077102	BAG6	0	0	−0.266	−0.016
rs3218100	CCND3	0.003	0	0	0
rs2242655	C6orf47	−0.046	0	0	0
rs2493835	TAF8	0.056	0	0	0
rs9491700	RSPO3	0.009	0	0	0
rs3008819	MOCS1	−0.021	0	0	0
rs2255741	PRRC2A	0.066	−0.021	0	0
rs3003931	DAAM2	0.004	0	0	0
rs791048	MAP3K7	0	0.080	0	0
rs9285458	RSPO3	0	0	−0.049	−0.078
rs3008801	DAAM2	−0.072	0	0	0
rs9462082	PPARD	0.026	0	0	0
rs166920	MAP3K7	−0.009	0	0	0
rs1144159	MAP3K7	0.091	0	0	0
rs284512	MAP3K7	0	−0.101	0	0
rs719726	RSPO3	0	−0.028	0.020	0.130
rs6916203	DAAM2	0	0	0	0.010
rs2504097	DAAM2	0	0	0	−0.034
rs4713858	FANCE	0	0	−0.139	0.157
rs1936789	RSPO3	0	−0.030	−0.044	0.072
rs1923084	MAP3K7	0	−0.163	0	0.315
rs9462769	C6orf132	0	0	−0.094	−0.138
rs11759168	DAAM2	0.173	0.027	0	−0.174

(Continued in next column)

TABLE 1 (*Continued*) Identified main and interaction effects based on the genes from the Wnt signaling pathway on chromosome 6.

SNP	Gene		Treatment	Age	Gender
rs707917	ABHD16A	−0.096	0	0.196	0.001
rs9267531	CSNK2B	−0.141	0	0	0
rs9394630	DAAM2	0.116	0	0	0
rs2504790	DAAM2	−0.133	0	0	0
rs2750456	MAP3K7	−0.052	0	0	0
rs3003933	DAAM2	−0.073	0	0	0
rs2984659	MOCS1	0.004	0	0	0
rs282065	MAP3K7	0.076	0	0	0
rs2504805	DAAM2	0	0	0	0.122
rs1046080	PRRC2A	0	0	−0.184	0

have been included in the study and monitored through follow-up visits over 4 years. The response variable is the forced expiratory volume in one second (FEV1), which indicates the amount of air one can expel from the lungs in one second. We focus on FEV1 that has been repeatedly measured during the 12 visits after the application of treatment (budesonide, nedocromil, and Control). For our gene–environment interaction analysis, the G factors are the single nucleotide polymorphisms, and E factors consist of treatment, age, and gender. For the demonstration purpose, we target SNPs based on the genes from chromosome 6 and the Wnt signaling pathway at the same time, resulting in a total of 203 SNPs. Following the NIH guideline, we cannot share the data publicly or disclose them in the R output. The data can be applied from dbGap through the accession number phs000166.v2.p1.

```
# the longitudinal FEV1
> dim(ylong)
[1] 438 12
# environmental factors (treatment, age, gender)
> dim(e)
[1] 438 3
# genetic factos (SNP)
> dim(X)
[1] 438 203
```

Both the environmental and genetic factors are cross-sectional. For example, as shown previously, each of the three E factors is a 438-by-1-column vector, forming a 438-by-3 matrix. We obtained the optimal tuning parameters using function `cv.springer`. One can start the process by defining a grid interval for each tuning parameter. We applied the `cv.springer` function with estimating function type `func = "QIF"` and working correlation matrix type `corr = "exchangeable"` as follows:

```
> library (springer)
> #define input arguments
> lambda1 = seq (0.5,1,length.out = 5)
> lambda2 = seq (3,3.5,length.out = 5)
> #run cross-validation
> tunning = cv.springer (clin = NULL, e, X, ylong, beta0, lambda1,
+ lambda2, nfolds = 5, func = "QIF", corr = "exchangeable",
+ structure = "bilevel", maxits = 30, tol = 0.001)
> #print the results
> print (tuning)
$lam1
[1] 0.5
$lam2
[1] 3
$CV
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.2827513 0.2838438 0.2846629 0.2855799 0.2865723
[2,] 0.2858653 0.2867847 0.2877162 0.2885925 0.2894861
[3,] 0.2884425 0.2897974 0.2906588 0.2916546 0.2925146
[4,] 0.2919309 0.2927759 0.2936686 0.2945191 0.2954699
[5,] 0.2948042 0.2954983 0.2962844 0.2971886 0.2979241
```

The optimal tuning parameters within the range have been selected as 0.5 and 3 for lambda1 and lambda2, respectively. We have then applied the `springer` function to the dataset using the optimal tuning parameters as follows:

```
> #fit the bi-level selection model
> beta = springer (clin = NULL, e, X, ylong, beta0, func = "QIF",
+ corr = "exchangeable", structure = "bilevel", lam1, lam2,
+ maxits = 30, tol = 0.001)
```

The `springer` function returns the estimated coefficients for the intercept, environmental factors, genetic factors, and G×E interactions. We organized the output to show the identified genetic main effects and G×E interactions in Table 1. The selected SNPs and the corresponding genes are listed in the first two columns. The last four columns contain the estimated coefficients of the main effects for each SNP and the corresponding interactions between the SNPs and environmental factors .

6 Discussion

Before the formulation of the bi-level (or sparse group) selection in high-dimensional statistics (Friedman et al., 2010b), the relevant statistical models have already been extensively studied in genetic association studies (Lewis, 2002; Wu et al., 2012), which involve the simultaneous selection of important pathways (or gene sets) and corresponding genes within the pathways (or gene sets) (Schaid et al., 2012; Wu and Cui, 2014; Jiang et al., 2017). For G×E interaction studies, the bi-level selection has served as the umbrella model and led to a wide array of extensions (Zhou et al., 2021a).

Package *springer* cannot be applied directly on the ultra-high-dimensional data (Fan and Lv, 2008), which is essentially due to the limitation of regularization methods. A more viable path is to conduct marginal screening first and then apply regularization methods on a smaller set of features suitable for penalized selection (Jiang et al., 2015; Li et al., 2015; Wu et al., 2019). In fact, such an idea on screening has motivated the migration of joint analyses to marginal penalization in recent G×E studies (Chai et al., 2017; Lu et al., 2021; Wang et al., 2022). It is marginal in the sense that only the main and interaction effects with respect to the same G factor are considered in the model. Thus, marginal penalization is of a parallel nature and suitable for handling the ultra-high-dimensional data. To use our R package conducting marginal regularization on the ultra-high-dimensional longitudinal data, we just need to set the argument `g` in function `springer` to one genetic factor at a time, which will return the regression coefficients for all the clinical and environmental factors and main G and G×E interactions with respect to that G factor. The magnitude of the coefficients corresponding to the effects subject to the selection will be used as the measure for ranking and selecting important effects.

Robust penalization methods have drawn increasing attention in recent years (Freue et al., 2019; Hu et al., 2021; Chen et al., 2022; Sun et al., 2022). In high-dimensional longitudinal studies, incorporation of robustness is more challenging. The corresponding variable selection methods are expected to be insensitive to not only the outliers and data contaminations but also to misspecification of working correlation structure capturing the correlations among repeated measurements. It has been widely recognized that GEE is vulnerable to long-tailed distributions in the response variable, even though it yields consistent estimates when working correlations are misspecified (Qu and Song, 2004). Therefore, the more robust QIF emerges as a powerful alternative for developing variable selection methods. Our R package *springer* can facilitate further understanding of robustness in bi-level selection models.

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. Authorized access should be granted before accessing the data analyzed in the case study. Request to access the data should be sent to Database of Genotype and Phenotype (dbGaP) at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000166.v2.p1 through accession number phs000166.v2.p1.

Author contributions

Conceptualization: FZ, JR, and CW; resources: WW and CW; methodology: FZ, YL, JR, and CW; writing—original draft preparation: FZ, YL, and CW; software: FZ, YL, JR, and CW; data analysis: YL, FZ, and CW; writing—review and editing: all authors; supervision: CW; project administration: CW; funding acquisition: CW and WW.

Acknowledgments

The authors thank the editor and reviewer for their careful review and insightful comments, leading to a significant improvement of this article. This work was partially supported by an Innovative Research Award from the Johnson Cancer Research Center at Kansas State University.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1088223/full#supplementary-material

References

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S. E., et al. (2017). Identifying gene-environment interactions for prognosis using a robust approach. *Econ. statistics* 4, 105–120. doi:10.1016/j.ecosta.2016.10.004

Chen, J., Bie, R., Qin, Y., Li, Y., and Ma, S. (2022). Lq-based robust analytics on ultrahigh and high dimensional data. *Statistics Med.* 41, 5220. doi:10.1002/sim.9563

Childhood Asthma Management Program Research Group (1999). The childhood asthma management program (CAMP): Design, rationale, and methods. Childhood asthma management program research group. *Control. Clin. trials* 20 (1), 91–120.

Childhood Asthma Management Program Research GroupSzefler, S., Weiss, S., Tonascia, J., Adkinson, N. F., Bender, B., et al. (2000). Long-term effects of budesonide or nedocromil in children with asthma. *N. Engl. J. Med.* 343 (15), 1054–1063. doi:10.1056/NEJM200010123431501

Cho, H., and Qu, A. (2013). Model selection for correlated data with diverging number of parameters. *Stat. Sin.* 23 (2), 901–927. doi:10.5705/ss.2011.058

Chu, W., Li, R., and Reimherr, M. (2016). Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *Ann. Appl. statistics* 10 (2), 596–617. doi:10.1214/16-AOAS912

Covar, R. A., Fuhlbrigge, A. L., Williams, P., and Kelly, H. W. (2012). The childhood asthma management program (camp): Contributions to the understanding of therapy and the natural history of childhood asthma. *Curr. Respir. Care Rep.* 1 (4), 243–250. doi:10.1007/s13665-012-0026-9

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

Eddelbuettel, D., and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance C++ linear algebra. *Comput. Statistics Data Analysis* 71, 1054–1063. doi:10.1016/j.csda.2013.02.005

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Berlin, Germany: Springer.

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70 (5), 849–911. doi:10.1111/j.1467-9868.2008.00674.x

Freue, G. V. C., Kepplinger, D., Salibián-Barrera, M., and Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *Ann. Appl. Statistics* 13 (4), 2065–2090. doi:10.1214/19-AOAS1269

Friedman, J., Hastie, T., and Tibshirani, R. (2010). *A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Hu, Z., Zhou, Y., and Tong, T. (2021). Meta-analyzing multiple omics data with robust variable selection. *Front. Genet.* 1029, 656826. doi:10.3389/fgene.2021.656826

Huang, H. H., Peng, X. D., and Liang, Y. (2021). Splsn: An efficient tool for survival analysis and biomarker selection. *Int. J. Intelligent Syst.* 36 (10), 5845–5865. doi:10.1002/int.22532

Inan, G., and Wang, L. (2017). Pgee: An r package for analysis of longitudinal data with high-dimensional covariates. *R J.* 9 (1), 393. doi:10.32614/rj-2017-030

Jiang, L., Liu, J., Zhu, X., Ye, M., Sun, L., Lacaze, X., et al. (2015). 2HiGWAS: A unifying high-dimensional platform to infer the global genetic architecture of trait development. *Briefings Bioinforma.* 16 (6), 905–911. doi:10.1093/bib/bbv002

Jiang, Y., Huang, Y., Du, Y., Zhao, Y., Ren, J., Ma, S., et al. (2017). Identification of prognostic genes and pathways in lung adenocarcinoma using a bayesian approach. *Cancer Inf.* 16, 1176935116684825. doi:10.1177/1176935116684825

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings Bioinforma.* 3 (2), 146–153. doi:10.1093/bib/3.2.146

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. statistics* 9 (2), 640–664. doi:10.1214/15-AOAS808

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73 (1), 13–22. doi:10.1093/biomet/73.1.13

Little, R. J., and Rubin, D. B. (2019). *Statistical analysis with missing data, vol. 793*. New York, NY, USA: John Wiley & Sons.

Liu, M., Zhang, Q., and Ma, S. (2022). A tree-based gene–environment interaction analysis with rare features. *Stat. Analysis Data Min. ASA Data Sci. J.* 15, 648–674. doi:10.1002/sam.11578

Lu, X., Fan, K., Ren, J., and Wu, C. (2021). Identifying gene-environment interactions with robust marginal bayesian variable selection. *Front. Genet.* 12, 667074. doi:10.3389/fgene.2021.667074

Ma, S., Song, Q., and Wang, L. (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli* 19 (1), 252–274. doi:10.3150/11-bej386

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* 87 (4), 823–836. doi:10.1093/biomet/87.4.823

Qu, A., and Song, P. X.-K. (2004). Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika* 91 (2), 447–459. doi:10.1093/biomet/91.2.447

Ren, J., Du, Y., Li, S., Ma, S., Jiang, Y., and Wu, C. (2019). Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genet. Epidemiol.* 43 (3), 276–291. doi:10.1002/gepi.22194

Ren, J., He, T., Li, Y., Liu, S., Du, Y., Jiang, Y., et al. (2017). Network-based regularization for high dimensional snp data in the case–control study of type 2 diabetes. *BMC Genet.* 18 (1), 44–12. doi:10.1186/s12863-017-0495-5

Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., et al. (2020). Semiparametric bayesian variable selection for gene-environment interactions. *Statistics Med.* 39 (5), 617–638. doi:10.1002/sim.8434

Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y., and Wu, C. (2022). Robust bayesian variable selection for gene–environment interactions. *Biometrics.* doi:10.1111/biom.13670

Ren, M., Zhang, S., Ma, S., and Zhang, Q. (2022). Gene–environment interaction identification via penalized robust divergence. *Biometrical J.* 64 (3), 461–480. doi:10.1002/bimj.202000157

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63 (3), 581–592. doi:10.1093/biomet/63.3.581

Schaid, D. J., Sinnwell, J. P., Jenkins, G. D., McDonnell, S. K., Ingle, J. N., Kubo, M., et al. (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet. Epidemiol.* 36 (1), 3–16. doi:10.1002/gepi.20632

Sun, Y., Luo, Z., and Fan, X. (2022). Robust structured heterogeneity analysis approach for high-dimensional data. *Statistics Med.* 41, 3229. doi:10.1002/sim.9414

Tang, Y., Wang, H. J., and Zhu, Z. (2013). Variable selection in quantile varying coefficient models with longitudinal data. *Comput. Statistics Data Analysis* 57 (1), 435–449. doi:10.1016/j.csda.2012.07.015

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Stat. methods Med. Res.* 23 (1), 42–59. doi:10.1177/0962280212445834

Wang, J. H., Wang, K. H., and Chen, Y. H. (2022). Overlapping group screening for detection of gene-environment interactions with application to tcga high-dimensional survival genomic data. *BMC Bioinforma.* 23 (1), 202–219. doi:10.1186/s12859-022-04750-7

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Am. Stat. Assoc.* 103 (484), 1556–1569. doi:10.1198/016214508000000788

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* 68 (2), 353–360. doi:10.1111/j.1541-0420.2011.01678.x

Wu, C., and Cui, Y. (2013). A novel method for identifying nonlinear gene–environment interactions in case–control association studies. *Hum. Genet.* 132 (12), 1413–1425. doi:10.1007/s00439-013-1350-z

Wu, C., and Cui, Y. (2014). Boosting signals in gene-based association studies via efficient snp selection. *Briefings Bioinforma.* 15 (2), 279–291. doi:10.1093/bib/bbs087

Wu, C., Jiang, Y., Ren, J., Cui, Y., and Ma, S. (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics Med.* 37 (3), 437–456. doi:10.1002/sim.7518

Wu, C., Li, S., and Cui, Y. (2012). Genetic association studies: An information content perspective. *Curr. genomics* 13 (7), 566–573. doi:10.2174/138920212803251382

Wu, C., and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings Bioinforma.* 16 (5), 873–883. doi:10.1093/bib/bbu046

Wu, C., Zhong, P.-S., and Cui, Y. (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Stat. Appl. Genet. Mol. Biol.* 17 (2). doi:10.1515/sagmb-2017-0008

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., and Ma, S. (2019). A selective review of multi-level omics data integration using variable selection. *High-throughput* 8 (1), 4. doi:10.3390/ht8010004

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statistics* 38 (2), 894–942. doi:10.1214/09-aos729

Zhou, F., Lu, X., Ren, J., Fan, K., Ma, S., and Wu, C. (2022). Sparse group variable selection for gene–environment interactions in the longitudinal study. *Genet. Epidemiol.* 46 (5-6), 317–340. doi:10.1002/gepi.22461

Zhou, F., Lu, X., Ren, J., and Wu, C. (2021). *Package 'springer': Sparse group variable selection for gene-environment interactions in the longitudinal study*. R package version 0.1.2.

Zhou, F., Ren, J., Li, G., Jiang, Y., Li, X., Wang, W., et al. (2019). Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study. *Genes.* 10 (12), 1002. doi:10.3390/genes10121002

Zhou, F., Ren, J., Liu, Y., Li, X., Wang, W., and Wu, C. (2022). Interep: An r package for high-dimensional interaction analysis of the repeated measurement data. *Genes.* 13 (3), 544. doi:10.3390/genes13030544

Zhou, F., Ren, J., Lu, X., Ma, S., and Wu, C. (2021). Gene–environment interaction: A variable selection perspective. *Methods Mol. Biol.* 2212, 191–223. doi:10.1007/978-1-0716-0947-7_13

TYPE Original Research
PUBLISHED 21 April 2023
DOI 10.3389/fgene.2023.1095050

OPEN ACCESS

EDITED BY
Simon Charles Heath,
Center for Genomic Regulation (CRG),
Spain

REVIEWED BY
Renxi Wang,
Capital Medical University, China
Vincent Salvatore Gallicchio,
Clemson University, United States

*CORRESPONDENCE
Bo Yang,
✉ 18275616479@163.com

RECEIVED 10 November 2022
ACCEPTED 13 April 2023
PUBLISHED 21 April 2023

CITATION
Ai Q and Yang B (2023), Are inflammatory bowel diseases associated with an increased risk of COVID-19 susceptibility and severity? A two-sample Mendelian randomization study.
Front. Genet. 14:1095050.
doi: 10.3389/fgene.2023.1095050

Are inflammatory bowel diseases associated with an increased risk of COVID-19 susceptibility and severity? A two-sample Mendelian randomization study

Qixiong Ai and Bo Yang*

Department of Gastroenterology and Hepatology, Guizhou Aerospace Hospital, Zunyi, Guizhou, China

Background: Due to inconsistent findings in observational studies regarding the relationship between inflammatory bowel disease (IBD), encompassing ulcerative colitis (UC) and Crohn's disease (CD), and COVID-19, our objective is to explore a potential causative correlation between IBD and COVID-19 susceptibility and its severity using a two-sample Mendelian randomization (MR) analysis.

Methods: Using summary data from genome-wide association studies, IBD, including UC and CD, were used as exposure instruments, while COVID-19 susceptibility, hospitalization, and very severe illness were employed as the outcome. The five analysis methods were adopted to evaluate the causal relationship between two diseases, with the inverse variance weighted (IVW) method being the most important. Also, sensitivity analyses were done to make sure that the main results of the MR analyses were reliable.

Results: In the analysis using five methods, all *p*-values were higher than 0.05. There was no association between IBD and COVID-19 susceptibility, hospitalization, and severity in our MR study. The random-effect model was applied due to the existence of heterogeneity. MR-Egger regression revealed no indication of directional pleiotropy, and sensitivity analysis revealed similar relationships.

Conclusion: This MR study found no evidence to support that IBD (which includes UC and CD) increases the risk of COVID-19 susceptibility or severity. Our result needs further confirmation through larger epidemiological studies.

KEYWORDS

inflammatory bowel disease, ulcerative colitis, Crohn 's disease, COVID-19, Mendelian randomization

Introduction

Coronavirus disease 2019 (COVID-19) is a contagious illness that has spread throughout the world and primarily affects the respiratory system (Guan et al., 2020). By 28 October 2022, a total of 626,337,158 patients had been diagnosed, with 6,566,610 confirmed deaths globally (https://covid.who.int/). Until now, the virus has expanded at an increasing rate, and the pandemic has swiftly spread to many nations. Global nations are bearing a significant socioeconomic burden as a result of the isolation and treatment measures implemented in response to COVID-19. As a result, one of the most important ways to prevent COVID-19 right now is to find the likely risk factors for the disease and take preventive measures for people who are at high risk.

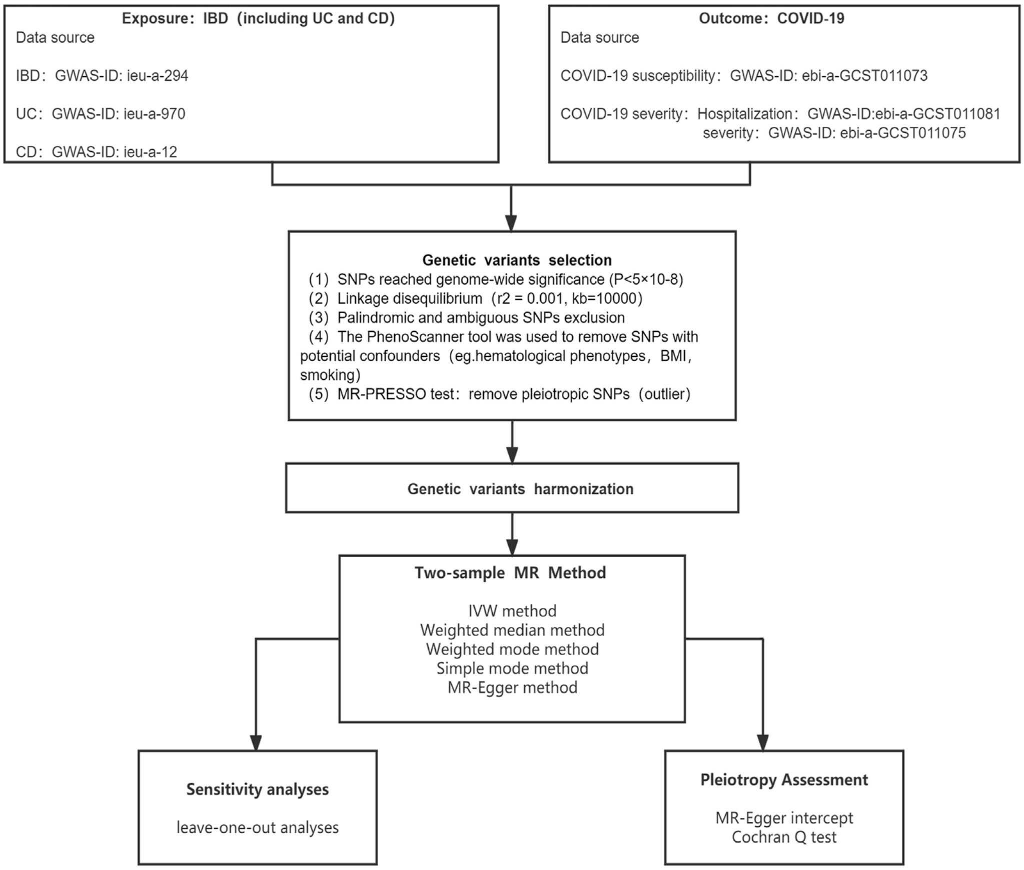

Exposure：IBD（including UC and CD）
Data source
IBD：GWAS-ID: ieu-a-294
UC：GWAS-ID: ieu-a-970
CD：GWAS-ID: ieu-a-12
Outcome：COVID-19
Data source
COVID-19 susceptibility：GWAS-ID: ebi-a-GCST011073
COVID-19 severity：Hospitalization：GWAS-ID:ebi-a-GCST011081
severity：GWAS-ID: ebi-a-GCST011075
Genetic variants selection
(1) SNPs reached genome-wide significance (P<5×10-8)
(2) Linkage disequilibrium（r2 = 0.001, kb=10000）
(3) Palindromic and ambiguous SNPs exclusion
(4) The PhenoScanner tool was used to remove SNPs with potential confounders（eg.hematological phenotypes，BMI，smoking）
(5) MR-PRESSO test：remove pleiotropic SNPs（outlier）
Genetic variants harmonization
Two-sample MR Method
IVW method
Weighted median method
Weighted mode method
Simple mode method
MR-Egger method
Sensitivity analyses
leave-one-out analyses
Pleiotropy Assessment
MR-Egger intercept
Cochran Q test

FIGURE 1
The Mendelian randomization (MR) analysis pipeline of the current study.

Several studies have shown that COVID-19 susceptible risk factors include white blood cells, type 2 diabetes, obesity, and smoking (Leong et al., 2021; Sun et al., 2021; Au Yeung et al., 2022; Cao et al., 2022). Inflammatory bowel disease (IBD) is a group of chronic, non-specific inflammatory diseases affecting the gut, for which the cause is still unknown. The two main types of IBD are Crohn's disease (CD) and ulcerative colitis (UC) (Ananthakrishnan, 2015). There may be an increased risk of infection in IBD patients due to immune system imbalance and prolonged use of immunosuppressive drugs, and COVID-19-related symptoms could potentially exacerbate inflammation in the intestines. (Geremia et al., 2014). According to recent research, immune-mediated IBD may enhance the risk of COVID-19 infection (Derikx et al., 2021). Likewise, additional research revealed that age exceeding 65 years and active IBD were among the factors that correlated with heightened susceptibility to COVID-19. (Bezzio et al., 2020). This might be due to the abnormal intestinal immune response, the infiltration of neutrophils, lymphocytes, and plasma cells into the intestinal mucosa, and the disorder of cytokine secretion that occurs during IBD activity (Cassinotti et al., 2014). On the other hand, contrary studies suggest that IBD patients do not have a higher rate of COVID-19 infection compared to the general population. (Macaluso and Orlando, 2020; Monteleone and Ardizzone, 2020; Popa et al., 2020). Non-etheless, these results are susceptible to confounding variables and reverse causation, which cannot be completely ruled out in observational research. Further investigation is required to identify the link between IBD and the COVID-19 infection and severity.

Mendelian randomization (MR) is based on the assumption that genetic variations are randomly distributed in the population and not associated with confounding factors. It uses genetic variants as instrumental variables (IVs) to explore causality between exposure and outcome. (Burgess et al., 2020). MR is based on the random distribution of gametes during meiosis, which allows it to circumvent the confusion and reverse causation that frequently plague observational studies. (Holmes et al., 2017). The goal of this research was to explore whether there was a link between IBD (including UC and CD) and COVID-19 susceptibility and severity using a two-sample MR analysis.

Methods

Study design

The whole research plan is shown in Figure 1. Specifically, the MR method consists of two primary steps: First, randomizing participants based on IVs; then, evaluating the causal relationships between IBD and COVID-19 outcomes (Davey Smith and Hemani, 2014; Emdin et al., 2017). The IVs must adhere to three essential criteria: 1) that IVs and IBD are tightly associated; 2) that IVs and confounders are unrelated; and 3) that IVs should only impact COVID-19 results via IBD, not through other routes (Davies et al., 2018).

Data source

An important part of running MR analysis was choosing relevant genetic variants. The IEU GWAS database offers users the opportunity to get GWAS summary statistics. The database contains a large number of genetic variants from GWAS summary-level datasets for search or download (Hemani et al., 2018). Consequently, we selected SNPs as IVs for exposures and outcomes using this database. All SNPs and associated summary data were acquired from studies involving solely European populations to mitigate the effects of population stratification. From the Inflammatory Bowel Disease Genetics Consortium, genetic variations linked to IBD were extracted (Liu et al., 2015). IBD was diagnosed using imaging, endoscopic, and histological examinations. We chose SNPs as IVs for IBD, UC, and CD (GWAS ID: ieu-a-294; ieu-a-970; ieu-a-12). From the COVID-19 Host Genetics Initiative round 5, the COVID-19 susceptibility and severity analysis data were collected. We chose SNPs as IVs for COVID-19 susceptibility, hospitalization, and severity (GWAS-ID: ebia-GCST011073; ebia-GCST011081; ebia-GCST011075). Supplementary Table S1 displays detailed characteristics.

All of the information was taken from previously published GWAS summary data that was made accessible to the public. As a result, neither ethical approval nor patient consent were required for the research.

Selection of instrumental variables

The SNPs that were eligible were selected using a variety of quality control techniques. Appropriate SNPs utilized as IVs must be strongly linked to IBD (p < 5E−08). A clumping algorithm (r^2 = 0.001, kb = 10,000) was performed to confirm the independence of SNPs and eliminate linkage disequilibrium (LD). The PhenoScanner database was used to filter out the identified SNPs that were linked to other phenotypes and potentially influencing results. When COVID-19 was identified as the outcome, hematological phenotypes (e.g., platelet count, percentage of neutrophils in granulocytes, and lymphocyte count), type 2 diabetes, obesity, and smoking were identified as confounding variables (Leong et al., 2021; Sun et al., 2021; Au Yeung et al., 2022; Cao et al., 2022). To further evaluate the instrumental value of SNPs, we computed F-statistics, with an F-statistic of more than 10 considered reliable. In addition, the MR-PRESSO test was used to verify whether pleiotropy existed and to manually delete outlier SNPs (p < 0.05). After eliminating these outlier SNPs, the remaining SNPs were used for subsequent MR analysis.

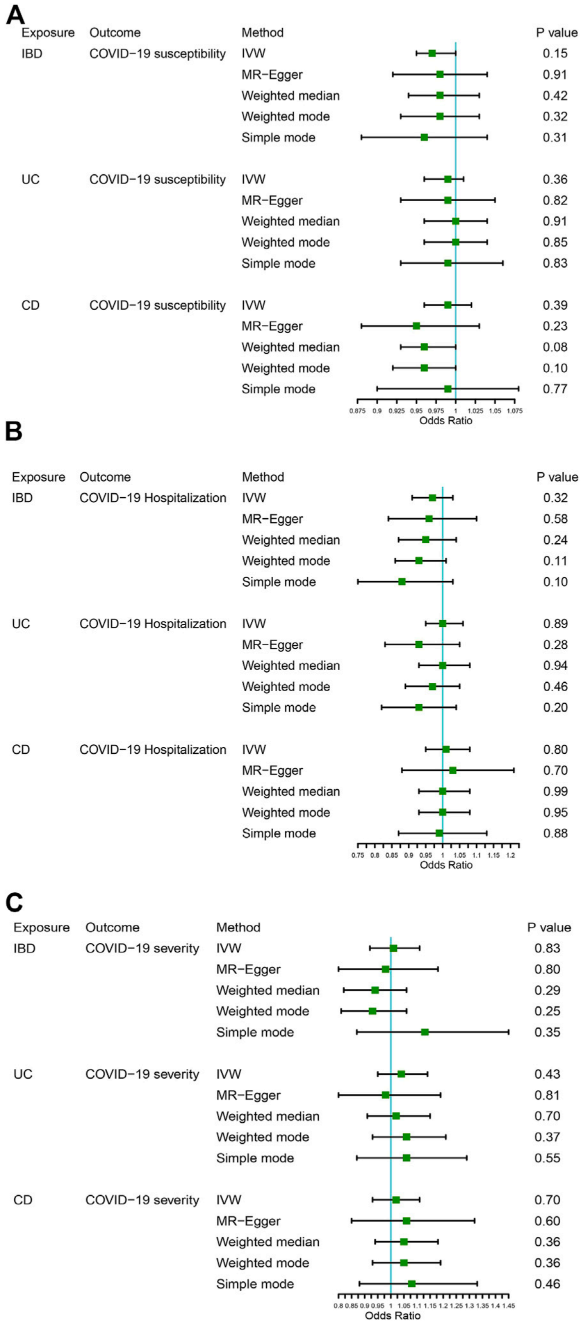


FIGURE 2
The association between IBD (including UC and CD) and COVID-19 susceptibility, hospitalization, and severity. **(A)** COVID-19 susceptibility, **(B)** COVID-19 hospitalization, **(C)** COVID-19 severity.

Mendelian randomization analysis

The effects of IBD on COVID-19 susceptibility and severity were explored using a variety of methods, the most important of which were the inverse variance weighted (IVW), followed by the Mendelian randomization-Egger (MR-Egger), the weighted median, the weighted mode, and the simple mode. The IVW method has the best statistical validity and reliably calculates the causal impact of exposure on the outcome (Burgess et al., 2013).

TABLE 1 Heterogeneity and horizontal pleiotropy analyses results.

Exposure	Outcome	Cochran Q statistic (IVW method)	Heterogeneity *p*-value (IVW method)	MR-Egger	
				Intercept	Intercept *p*-value
IBD	COVID-19 susceptibility	105.9348	0.0005	−0.0024	0.5198
	COVID-19 hospitalization	113.4307	1.02E-04	0.0006	0.9319
	COVID-19 severity	108.1956	0.0003	0.0045	0.6856
UC	COVID-19 susceptibility	43.5409	0.2129	−0.0009	0.8239
	COVID-19 hospitalization	49.2361	0.0859	0.0114	0.1943
	COVID-19 severity	58.1414	0.0147	0.0099	0.4887
CD	COVID-19 susceptibility	71.1531	0.0032	0.0048	0.3533
	COVID-19 hospitalization	80.0273	0.0003	−0.0034	0.7546
	COVID-19 severity	71.8791	0.0027	−0.0062	0.6819

Pleiotropy, heterogeneity, and sensitivity analyses were used to check the quality. The Cochran's Q test and the MR-Egger intercept test were employed to investigate heterogeneity and directed horizontal pleiotropy, respectively, to confirm the reliability of the findings. If there was no evidence of heterogeneity, the fixed-effect model was employed; otherwise, the random-effect model was used. Additionally, we evaluated the consistency and effectiveness of MR findings using the "leave-one-out" method.

Statistical analysis

The statistical analyses were carried out in R version 4.1.3 using the "TwoSampleMR" and "MRPRESSO" packages, respectively (Hemani et al., 2018; Verbanck et al., 2018). There was no heterogeneity across IVs when the Q statistic was $p > 0.05$, but there was heterogeneity when $p < 0.05$. If the MR-Egger regression intercept was not zero and $p < 0.05$, the IV was thought to exhibit horizontal pleiotropy. On the other hand, if $p > 0.05$, the results were considered not to have horizontal pleiotropy. Regarding the correction for multiple testing, we employed a Bonferroni correction to reduce the likelihood of type 1 error, thereby improving the reliability of our results. The Bonferroni-correction (0.0055, 0.05/3 exposures/3 outcomes) was employed to account for the issue of multiple testing. A possible correlation was considered to exist when the *p*-value was less than 0.0055.

Results

Filter instrument variables

After applying stringent exclusion criteria, we included 134, 88, and 122 SNPs as IVs for IBD, UC, and CD, respectively. We found and removed 9 (IBD), 4 (UC), and 7 (CD) palindromic SNPs and 13 (IBD), 10 (UC), and 16 (CD) ambiguous SNPs. Using PhenoScanner, 46 (IBD), 33 (UC), and 52 (CD) SNPs were manually removed. Two SNPs (rs2143178 and rs516246) in IBD, one SNP (rs9611131) in UC, and two SNPs (rs2413583 and rs516246) in CD were excluded based on the MR-PRESSO test. Finally, after strict screening, 66 SNPs (IBD), 40 SNPs (UC), and 45 SNPs (CD) qualified as IVs for the MR analysis. The F-statistics of all three IVs were more than 10 (ranging from 100.2550 to 3614.1141 for IBD; 110.0008 to 1571.1295 for UC; and 114.3503 to 3044.4854 for CD). The characteristics of SNPs for IBD are shown in Supplementary Tables S2–S4.

Association between IBD and COVID-19 susceptibility and severity

Figure 2 displays the results of MR analysis, which demonstrated that none of the five methods had a statistically significant relationship with IBD (including UC and CD) and COVID-19 susceptibility, hospitalization or severity (all $p > 0.0055$). The sensitivity studies, such as the Cochran's Q test and the MR-Egger intercept test, were performed to evaluate the robustness of the aforementioned findings (Table 1). The scatter plots of association estimates between IBD (including UC and CD) and COVID-19 susceptibility, hospitalization, and severity, as well as the MR causal estimates, are shown in Figure 3. The Cochran's Q test, however, revealed heterogeneity between IBD and COVID-19 susceptibility, hospitalization, or severity. As a result, the random-effect model was used in the IVW approach (heterogeneity *p*-value<0.05). Using the MR-Egger intercept test, it was discovered that there was no existence of directional pleiotropy (Intercept *p*-value>0.05). The consistency of the MR impact estimates was further confirmed by the leave-one-out method (Supplementary Figure S1). Forest plots of MR analysis of IBD (including UC and CD) and COVID-19 susceptibility, hospitalization, and severity are displayed in Supplementary

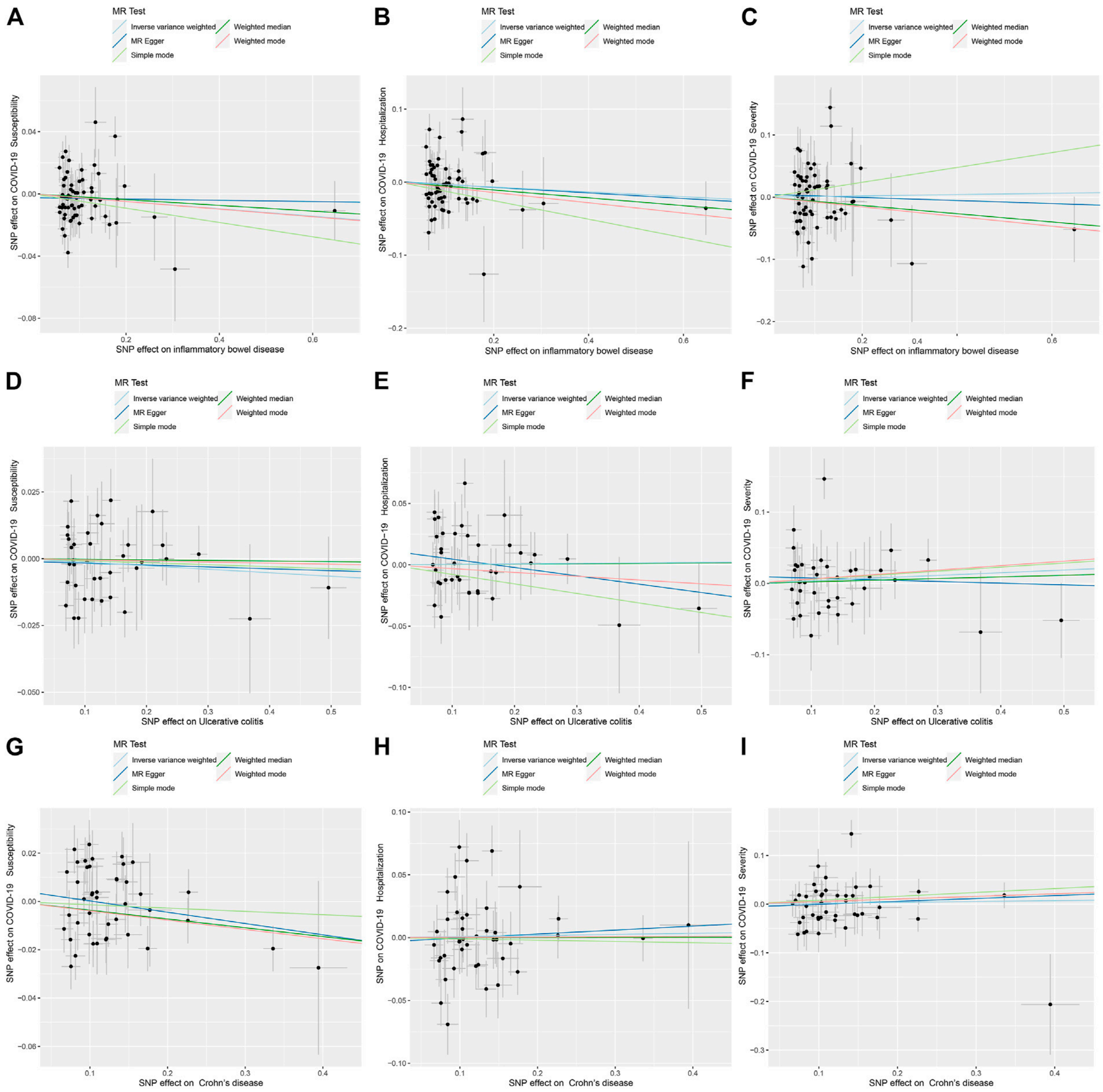


FIGURE 3
Scatter plots for causal SNP effect of IBD (including UC and CD) on COVID-19 susceptibility, hospitalization, and severity. **(A)** Effect of IBD on COVID-19 susceptibility; **(B)** Effect of IBD on COVID-19 hospitalization; **(C)** Effect of IBD on COVID-19 severity; **(D)** Effect of UC on COVID-19 susceptibility; **(E)** Effect of UC on COVID-19 hospitalization; **(F)** Effect of UC on COVID-19 severity; **(G)** Effect of CD on COVID-19 susceptibility; **(H)** Effect of CD on COVID-19 hospitalization; **(I)** Effect of CD on COVID-19 severity.

Figure S2. The funnel plots revealed the heterogeneity in the estimations for each SNP (Supplementary Figure S3).

Discussion

This work used large-scale GWAS data from the IBDGC and the COVID-19 Host Genetics Initiative Round 5 to assess the probable causal connection of IBD with COVID-19 susceptibility and severity. This is the first MR study, to our knowledge, to explore the causal relationship between IBD and COVID-19 risk. The MR analysis revealed that there was insufficient evidence to suggest that IBD (UC and CD) may enhance COVID-19 susceptibility, hospitalization, and severity.

In terms of clinical presentation, IBD and COVID-19 share similar symptoms, such as stomach discomfort, diarrhea, pneumonia, and so on. In a prior study of 709 IBD patients, 53 were shown to be concurrently infected with COVID-19 (Vigano et al., 2020). The researchers found that the proportion of people with IBD and COVID-19 who also suffered from diarrhea was 49%, which was a significant increase above the proportion of people with only IBD. An earlier observational study revealed that patients with IBD are more likely to infect COVID-19, particularly when experiencing active disease or taking immunosuppressive therapy (Bezzio et al., 2020). Another study found that SARS-CoV2, which is widely expressed in the lung and gut, has been shown to be an inflammatory protective factor that is downregulated and upregulated, respectively, in COVID-19 and IBD, suggesting the presence of a coregulatory mechanism (Tao et al., 2022). However, there are no published studies that detail the underlying process. In line with the majority of other studies, our MR analysis showed no robust evidence of a connection between IBD and COVID-19 susceptibility, hospitalization, or severity. A large countrywide cohort study in the Netherlands compared the incidence of COVID-19 in people with IBD to that in the general population and found no statistically significant difference (Derikx et al., 2021). Likewise, the risk of COVID-19 infection was also shown to be the same in both the IBD and non-IBD groups, according to a multi-center network investigation (Singh et al., 2020). In addition, a recent meta-analysis and comprehensive review also showed the same result regarding IBD not increasing COVID-19 morbidity and mortality (Lee et al., 2023). Reverse causality confounding and other biases in observational studies may alter the causal effects of illness exposure on outcomes, resulting in incorrect results. The reported causal links between IBD and COVID-19 outcomes may be messed up in observational research, likely because of confounding factors like hematological phenotypes, type 2 diabetes, body mass index (BMI), and smoking. The MR analysis takes advantage of strong genetic variation in order to produce more reliable evidence for predicting the cause of illness (Davey Smith and Hemani, 2014; Bowden and Holmes, 2019). MR is increasingly used to infer causal relationships between exposures and outcomes. It can now be speculated that the data does not support the causal link between IBD and COVID-19 infection and severity, taking into account both the information already available and the findings of our investigation.

There is no question that our research has certain shortcomings as well. Firstly, since this research only included people of European heritage, the findings cannot be applied to other ethnic groups. Therefore, future MR studies in non-European populations would be valuable to further confirm the causal relationship between IBD and COVID-19 infection and severity. Secondly, genetic instruments may impact outcomes via other confounding variables. Tight exclusion criteria and the PhenoScanner tool can exclude genetic instruments related to confounding factors as much as possible, but this cannot be totally eliminated. Thirdly, in the MR model, we only included a linear impact association between IBD and COVID-19. Additionally, we were unable to investigate the non-linearity of the link between IBD and COVID-19 using the GWAS summary data. Fourth, we failed to perform a stratified analysis based on active or remission periods in IBD due to the limited available datasets. Fifth, as the infectiousness of COVID-19 is a dynamic outcome that is influenced by other confounding factors such as social restrictions, the interpretation and understanding of the study results are more challenging and complex. Hence, considering these limits, evaluating patient hospitalization and severity at the same time may provide more persuasive and reliable results, which can help to comprehensively understand the relationship between IBD and COVID-19. Finally, we discovered that COVID-19 infection risk was not directly influenced by IBD, although the underlying molecular mechanism was still unknown. It was necessary to conduct further functional experiments to verify our conclusions.

Conclusion

Overall, the cause-and-effect connection between IBD and COVID-19 susceptibility and severity were assessed using the two-sample MR method. Based on the findings of this MR investigation, it seems that IBD does not appear to increase the risk of COVID-19 susceptibility, hospitalization, and severity. To validate our findings, we need to do more large-scale epidemiological studies and more research into the biological link between IBD and COVID-19.

Data availability statement

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

Author contributions

The study was created by QA. The manuscript was written by BY. QA and BY contributed to the data analysis, interpretation, and paper correction. The essay was written by all of the writers, and the final version was approved by all of them.

Acknowledgments

We appreciate the efforts of all the researchers that contributed to this overview of the GWAS for IBD and COVID-19.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1095050/full#supplementary-material

References

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* 12 (4), 205–217. doi:10.1038/nrgastro.2015.34

Au Yeung, S. L., Li, A. M., He, B., Kwok, K. O., and Schooling, C. M. (2022). Association of smoking, lung function and COPD in COVID-19 risk: A two-step mendelian randomization study. *Addiction* 117 (7), 2027–2036. doi:10.1111/add.15852

Bezzio, C., Saibeni, S., Variola, A., Allocca, M., Massari, A., Gerardi, V., et al. (2020). Outcomes of COVID-19 in 79 patients with IBD in Italy: An IG-IBD study. *Gut* 69 (7), 1213–1217. doi:10.1136/gutjnl-2020-321411

Bowden, J., and Holmes, M. V. (2019). Meta-analysis and mendelian randomization: A review. *Res. Synth. Methods* 10 (4), 486–496. doi:10.1002/jrsm.1346

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37 (7), 658–665. doi:10.1002/gepi.21758

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat. Commun.* 11 (1), 376. doi:10.1038/s41467-019-14156-4

Cao, H., Baranova, A., Wei, X., Wang, C., and Zhang, F. (2022). Bidirectional causal associations between type 2 diabetes and COVID-19. *J. Med. Virol.* 95, e28100. doi:10.1002/jmv.28100

Cassinotti, A., Sarzi-Puttini, P., Fichera, M., Shoenfeld, Y., de Franchis, R., and Ardizzone, S. (2014). Immunity, autoimmunity and inflammatory bowel disease. *Autoimmun. Rev.* 13 (1), 1–2. doi:10.1016/j.autrev.2013.06.007

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23 (R1), R89–R98. doi:10.1093/hmg/ddu328

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

Derikx, L., Lantinga, M. A., de Jong, D. J., van Dop, W. A., Creemers, R. H., Romkens, T. E. H., et al. (2021). Clinical outcomes of covid-19 in patients with inflammatory bowel disease: A nationwide cohort study. *J. Crohns Colitis* 15 (4), 529–539. doi:10.1093/ecco-jcc/jjaa215

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *JAMA* 318 (19), 1925–1926. doi:10.1001/jama.2017.17219

Geremia, A., Biancheri, P., Allan, P., Corazza, G. R., and Di Sabatino, A. (2014). Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun. Rev.* 13 (1), 3–10. doi:10.1016/j.autrev.2013.06.004

Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* 382 (18), 1708–1720. doi:10.1056/NEJMoa2002032

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

Holmes, M. V., Ala-Korpela, M., and Smith, G. D. (2017). Mendelian randomization in cardiometabolic disease: Challenges in evaluating causality. *Nat. Rev. Cardiol.* 14 (10), 577–590. doi:10.1038/nrcardio.2017.78

Lee, M. H., Li, H. J., Wasuwanich, P., Kim, S. E., Kim, J. Y., Jeong, G. H., et al. (2023). COVID-19 susceptibility and clinical outcomes in inflammatory bowel disease: An updated systematic review and meta-analysis. *Rev. Med. Virol.* 33 (2), e2414. doi:10.1002/rmv.2414

Leong, A., Cole, J. B., Brenner, L. N., Meigs, J. B., Florez, J. C., and Mercader, J. M. (2021). Cardiometabolic risk factors for COVID-19 susceptibility and severity: A mendelian randomization analysis. *PLoS Med.* 18 (3), e1003553. doi:10.1371/journal.pmed.1003553

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47 (9), 979–986. doi:10.1038/ng.3359

Macaluso, F. S., and Orlando, A. (2020). COVID-19 in patients with inflammatory bowel disease: A systematic review of clinical data. *Dig. Liver Dis.* 52 (11), 1222–1227. doi:10.1016/j.dld.2020.09.002

Monteleone, G., and Ardizzone, S. (2020). Are patients with inflammatory bowel disease at increased risk for covid-19 infection? *J. Crohns Colitis* 14 (9), 1334–1336. doi:10.1093/ecco-jcc/jjaa061

Popa, I. V., Diculescu, M., Mihai, C., Cijevschi-Prelipcean, C., and Burlacu, A. (2020). COVID-19 and inflammatory bowel diseases: Risk assessment, shared molecular pathways, and therapeutic challenges. *Gastroenterol. Res. Pract.* 2020, 1918035. doi:10.1155/2020/1918035

Singh, S., Khan, A., Chowdhry, M., Bilal, M., Kochhar, G. S., and Clarke, K. (2020). Risk of severe coronavirus disease 2019 in patients with inflammatory bowel disease in the United States: A multicenter research network study. *Gastroenterology* 159 (4), 1575–1578. doi:10.1053/j.gastro.2020.06.003

Sun, Y., Zhou, J., and Ye, K. (2021). White blood cells and severe COVID-19: A mendelian randomization study. *J. Pers. Med.* 11 (3), 195. doi:10.3390/jpm11030195

Tao, S. S., Wang, X. Y., Yang, X. K., Liu, Y. C., Fu, Z. Y., Zhang, L. Z., et al. (2022). COVID-19 and inflammatory bowel disease crosstalk: From emerging association to clinical proposal. *J. Med. Virol.* 94 (12), 5640–5652. doi:10.1002/jmv.28067

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50 (5), 693–698. doi:10.1038/s41588-018-0099-7

Vigano, C., Massironi, S., Pirola, L., Cristoferi, L., Fichera, M., Bravo, M., et al. (2020). COVID-19 in patients with inflammatory bowel disease: A single-center observational study in northern Italy. *Inflamm. Bowel Dis.* 26 (11), e138–e139. doi:10.1093/ibd/izaa244

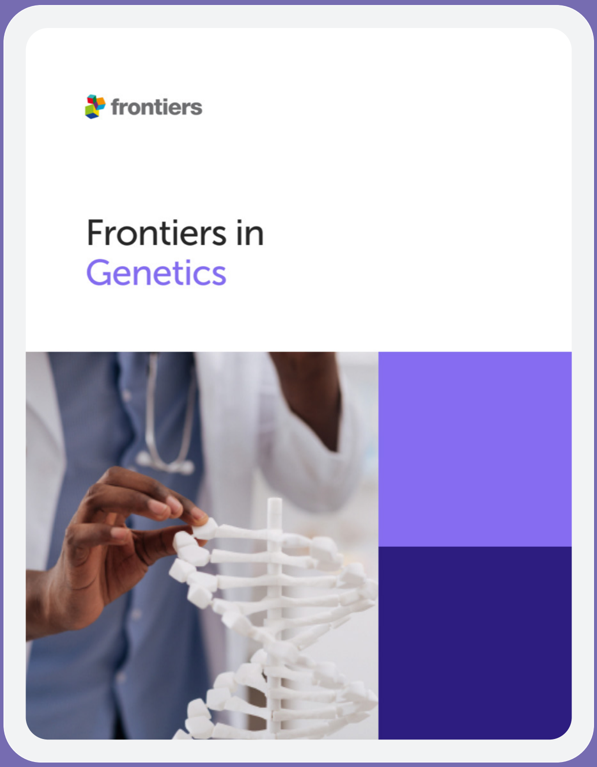
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact