STATISTICAL METHODS, COMPUTING, AND RESOURCES FOR GENOME-WIDE ASSOCIATION STUDIES

EDITED BY: Riyan Cheng, Lide Han, Hailan Liu, Min Zhang and Guolian Kang
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-212-0
DOI 10.3389/978-2-88971-212-0

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

STATISTICAL METHODS, COMPUTING, AND RESOURCES FOR GENOME-WIDE ASSOCIATION STUDIES

Topic Editors:
Riyan Cheng, University of California, United States
Lide Han, Vanderbilt University Medical Center, United States
Hailan Liu, Maize Research Institute of Sichuan Agricultural University, China
Min Zhang, Purdue University, United States
Guolian Kang, St. Jude Children's Research Hospital, United States

Citation: Cheng, R., Han, L., Liu, H., Zhang, M., Kang, G., eds. (2021). Statistical Methods, Computing, and Resources for Genome-Wide Association Studies. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-212-0

Table of Contents

frontiers in Genetics

EDITORIAL
published: 05 August 2021
doi: 10.3389/fgene.2021.714894

Editorial: Statistical Methods, Computing and Resources for Genome-Wide Association Studies

Hailan Liu[1], *Lide Han*[2], *Guolian Kang*[3], *Min Zhang*[4] *and Riyan Cheng*[5]*

[1] Maize Research Institute, Sichuan Agricultural University, Chengdu, China, [2] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, United States, [3] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, United States, [4] Department of Statistics, Purdue University, West Lafayette, IN, United States, [5] Department of Psychiatry, University of California, San Diego, San Diego, CA, United States

Keywords: genome-wide association study, multi-trait analysis, nested association mapping, pleiotropy, imprinted QTL

Editorial on the Research Topic

Statistical Methods, Computing and Resources for Genome-Wide Association Studies

Thanks to the recent advances in genotyping technologies, genome-wide association studies (GWAS) have been an established approach to identifying genetic variants that influence certain characteristics of economic or scientific interest in plants, animals and humans. Applications of GWAS cover a wide range of areas in genetics and have enhanced our understanding of the genetic mechanisms in diseases, physiological or behavioral traits and have generated promises in agriculture, medicine and wildlife conservation. Despite great success, GWAS remains challenged by statistical modeling and computing. This collection of twelve articles presents a variety of interesting scientific problems and novel approaches in GWAS.

Nested association mapping (NAM) is a technique for dissecting the genetic architecture of complex traits in crops. It is designed for NAM-specific populations by taking the advantages of linkage analysis and association mapping while avoiding their disadvantages. Bu et al. developed a multi-locus association mapping model for the analysis of data from multiple families in the NAM population. A notable feature of their method lies in its ability to deal with genetic heterogeneity due to subpopulations, and therefore their approach improves statistical power for quantitative trait locus (QTL) detection and accuracy of QTL effect estimation. In real data analyses, they found that their method identified most QTLs that were detected by linkage analyses of single-family datasets and was also able to disclose some new QTLs with small effects. Three of the 12 articles in this collection are concerned with multi-trait analysis. Multi-trait analysis has been of interest in GWAS due to its potential gain in statistical power and its ability for formal hypothesis testing of biological importance such as pleiotropy vs. linkage. As a common practice, multiple traits are analyzed separately and markers are scanned one at a time. Deng et al. argued that the former does not take advantage of correlations between traits and thus can be a limitation on statistical power, and the latter ignores complex interactions between genomic variants. Therefore, they proposed a genome-wide gene-based multi-trait method to overcome these limitations. Technically, they adopted kernel-based testing to evaluate the joint effect of multiple variants in a gene and proposed an omnibus test strategy to integrate the test results. They demonstrated that their method achieved excellent power with reasonable control of type I error rates. Lin et al. discussed the interpretation of the results from multi-trait analyses. They introduced a bioinformatic tool, MetaPha, which implements a meta-analysis approach by constructing multivariate analysis from univariate GWAS results and then decomposing multivariate associations into multiple ones that

Edited by:
Chaeyoung Lee,
Soongsil University, South Korea

Reviewed by:
Carsten Carlberg,
University of Eastern Finland, Finland

***Correspondence:**
Riyan Cheng
ric025@ucsd.edu

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

Received: *26 May 2021*
Accepted: *15 June 2021*
Published: *05 August 2021*

Citation:
Liu H, Han L, Kang G, Zhang M and Cheng R (2021) Editorial: Statistical Methods, Computing and Resources for Genome-Wide Association Studies. Front. Genet. 12:714894. doi: 10.3389/fgene.2021.714894

facilitate interpretation. They validated their method using lipid data from the Global Lipids Genetics Consortium and found that only three to five central traits of the twenty-one traits they studied were needed at the majority of the loci of their interest. Fernandes et al. focused on a biologically interesting question, linkage vs. pleiotropy, which can be tested by either multivariate or univariate approaches. Using simulation studies, they found that neither of these two approaches alone delivered a satisfactory result, and thus suggest that multivariate and univariate GWAS should be complementary rather than competing. For those interested in mapping imprinted QTLs, Zheng et al. proposed a special type of immortalized F2 population and correspondingly two methods for mapping imprinted QTL and demonstrated the merits of their proposed population and methods in mapping precision. If your research is related to Barley, you may be interested in Li et al. who studied a few traits in Qingke Barley. Other interests include heritability estimation (Xu et al. and Frouin et al.), disease studies using machine learning techniques (Fan et al. and Li et al.), and Mendelian randomization (Xu et al.).

Lastly, we would like to introduce Zhu et al. who reviewed statistical methods for identifying trait-relevant tissues and cell types. Genome-wide association studies (GWAS) have reported numerous quantitative trait loci (QTL). However, few reported QTL have been validated while the ultimate goal of GWAS is to help understand the biological mechanisms of the trait-QTL association. Some researchers have developed statistical methods to integrate genomic information with functional annotations, gene expression data, and gene network information into GWAS, and aim to identify relevant tissue and cell types. Zhu et al. extensively reviewed ten of these methods.

To conclude, this volume highlights new insights and fascinating perspectives in statistical methods, computing, and resources for GWAS. We hope the collection will stimulate more developments in this important topic as biotechnology continues to evolve.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

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 in Genetics

METHODS
published: 15 May 2020
doi: 10.3389/fgene.2020.00431

MetaPhat: Detecting and Decomposing Multivariate Associations From Univariate Genome-Wide Association Statistics

Jake Lin[1], Rubina Tabassum[1], Samuli Ripatti[1,2,3] and Matti Pirinen[1,2,4]**

[1] Institute for Molecular Medicine Finland FIMM, Helsinki Institute of Life Science HiLIFE, University of Helsinki, Helsinki, Finland, [2] Public Health, University of Helsinki, Helsinki, Finland, [3] Broad Institute, Massachusetts Institute of Technology, Harvard University, Cambridge, MA, United States, [4] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

OPEN ACCESS

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

Reviewed by:
Jaeyoon Chung,
Boston University, United States
Wenjian Bi,
University of Michigan, United States

***Correspondence:**
Jake Lin
jake.lin@helsinki.fi
Matti Pirinen
matti.pirinen@helsinki.fi

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

Received: *04 February 2020*
Accepted: *07 April 2020*
Published: *15 May 2020*

Citation:
Lin J, Tabassum R, Ripatti S and Pirinen M (2020) MetaPhat: Detecting and Decomposing Multivariate Associations From Univariate Genome-Wide Association Statistics. Front. Genet. 11:431. doi: 10.3389/fgene.2020.00431

Background: Multivariate testing tools that integrate multiple genome-wide association studies (GWAS) have become important as the number of phenotypes gathered from study cohorts and biobanks has increased. While these tools have been shown to boost statistical power considerably over univariate tests, an important remaining challenge is to interpret which traits are driving the multivariate association and which traits are just passengers with minor contributions to the genotype-phenotypes association statistic.

Results: We introduce MetaPhat, a novel bioinformatics tool to conduct GWAS of multiple correlated traits using univariate GWAS results and to decompose multivariate associations into sets of central traits based on intuitive trace plots that visualize Bayesian Information Criterion (BIC) and *P*-value statistics of multivariate association models. We validate MetaPhat with Global Lipids Genetics Consortium GWAS results, and we apply MetaPhat to univariate GWAS results for 21 heritable and correlated polyunsaturated lipid species from 2,045 Finnish samples, detecting seven independent loci associated with a cluster of lipid species. In most cases, we are able to decompose these multivariate associations to only three to five central traits out of all 21 traits included in the analyses. We release MetaPhat as an open source tool written in Python with built-in support for multi-processing, quality control, clumping and intuitive visualizations using the R software.

Conclusion: MetaPhat efficiently decomposes associations between multivariate phenotypes and genetic variants into smaller sets of central traits and improves the interpretation and specificity of genome-phenome associations. MetaPhat is freely available under the MIT license at: https://sourceforge.net/projects/meta-pheno-association-tracer.

Keywords: multivariate analysis, genotype phenotype correlation studies, feature selection, Bayesian information criteria, visualilzation, canonical correlation, multivariate GWAS, pheno- and genotypes

INTRODUCTION

Genome-wide association studies (GWAS) of common diseases and complex traits in large population cohorts have linked thousands of genetic variants to individual phenotypes. In emerging biobank studies as well as in some disease specific collections have focused on, for example, Type 2 diabetes (T2D) (Mahajan et al., 2018) or coronary artery disease (CAD) (Ripatti et al., 2016), multiple related quantitative traits are simultaneously available for genetic association studies. The statistical power in these discovery efforts can be boosted considerably by multivariate tests, which have become more practical through recent implementations that require only univariate summary statistics, such as MultiPhen (O'Reilly et al., 2012), TATES (van der Sluis et al., 2013), CONFIT (Gai and Eskin, 2018), MTAG (Turley et al., 2018), MTAR (Guo and Wu, 2019), and metaCCA (Cichonska et al., 2016). The merits of many of these methods are further discussed by Chung et al. (2019). Concretely, canonical correlation analysis (CCA) (Hotelling, 1936) is the direct extension of the correlation coefficient to identify linear associations between two sets of variables, and it has been successfully applied also to GWAS (Inouye et al., 2012). Moreover, metaCCA extended CCA to work directly from GWAS summary statistics (effect size estimates and standard errors) of related traits and studies. However, a remaining challenge is to interpret which traits are driving the multivariate association and which traits are just passengers contributing little to the association statistic. A successful identification of a subset of central traits for each associated variant can lead to new biological insights in studies of disease progression and heterogeneity. To address this important task, we have introduced MetaPhat (Meta-Phenotype Association Tracer), a novel method to efficiently and systematically:

1. identify and annotate significant variants via multivariate GWAS from univariate summary statistics using metaCCA;
2. perform decomposition by systematically tracing the traits of highest and lowest statistical importance to identify subsets of central traits at each associated variant;
3. plot the traces of trait decompositions and cluster the variants based on the ranking of the importance of traits.

MATERIALS AND METHODS

Workflow

MetaPhat requires as input a set of related GWAS summary statistics from correlated traits. The program implements efficient multi-trait genome-wide association testing, identification of significant associations, and systematic tracing of trait subsets to identify the central traits that consist of a statistically optimal set of traits together with a set of driver traits. A workflow is shown in **Figure 1**. In steps one to three, genome-wide significant variants [P < 5e-8, the established genome-wide threshold in the field (Sherry et al., 2001; Pe'er et al., 2008)] were identified and were clumped into independent groups that are subsequently represented by the lead variant of each group (i.e., the variant with the smallest P-value). By default, two lead variants were defined as independent if their distance is higher than 1 million base pairs. At step four, we carried out the decompositions of multivariate association by starting from model with all K traits and removing one trait at a time until only one trait remains. We proceeded via two different strategies that we named the *highest trace* and the *lowest trace*. More specifically, starting from the model with all K traits, we tested all unique combinations of (K-1) traits to find the subset with the highest CCA statistic (lowest P-value) that we assigned to the highest trace and the subset with the lowest CCA statistic (highest P-value) that we assigned to the lowest trace. We continued both traces iteratively until only a single trait remained by always choosing the subset with the highest CCA statistic on the highest trace and the subset with the lowest CCA statistic on the lowest trace. Intuitively, at each step, the trait dropped on the highest trace was the trait that was best replaceable by the other traits in the model with respect to the genetic association considered. Analogously, at each step, the trait dropped on the lowest trace was the trait that was most irreplaceable by the other traits in the model with respect to the genetic association considered. Altogether, we evaluated K^2 subsets out of all possible 2^K subsets while building these two traces. Base pair distances, GWAS P-value thresholds, and other program parameters could be updated using command-line arguments.

We used the two traces to identify central traits that are primarily responsible for the association with the variant as explained next.

Evaluating Models

We used two quantities to evaluate models: CCA P-values and Bayesian Information Criterion (BIC; Schwarz, 1978). P-values allowed us to compare each association to the established "genome-wide significance threshold" of 5e-8 (Pe'er et al., 2008). By using the lowest trace, we could identify those traits without which the multivariate P-value is no longer genome-wide significant by simply collecting the traits that have been removed from the full model when the P-value on the lowest trace is first time above 5e-8. We call these traits the driver traits since they drive the association in the sense that without them the association does not anymore reach genome-wide significance and hence would not have been reported as a discovery in a GWAS. This definition of driver traits is based on a fixed P-value threshold, which is an established practice in the field, but does not claim any statistical optimality properties in terms of model comparison. Hence, to more rigorously compare models with different dimensionalities, we used BIC, which approximates the negative marginal likelihood of the model and thus penalizes for the model dimension (Schwarz, 1978). A lower BIC value suggests a statistically better description of the data. A subset of traits with minimum BIC would thus be the model of choice. We defined the optimal subset as the subset with the lowest BIC among all subsets on the highest trace and all subsets on the inverted lowest trace. The inverted lowest trace aggregates the traits that have been dropped on the lowest trace, and, in particular, includes the set of the driver traits as one of its subsets.

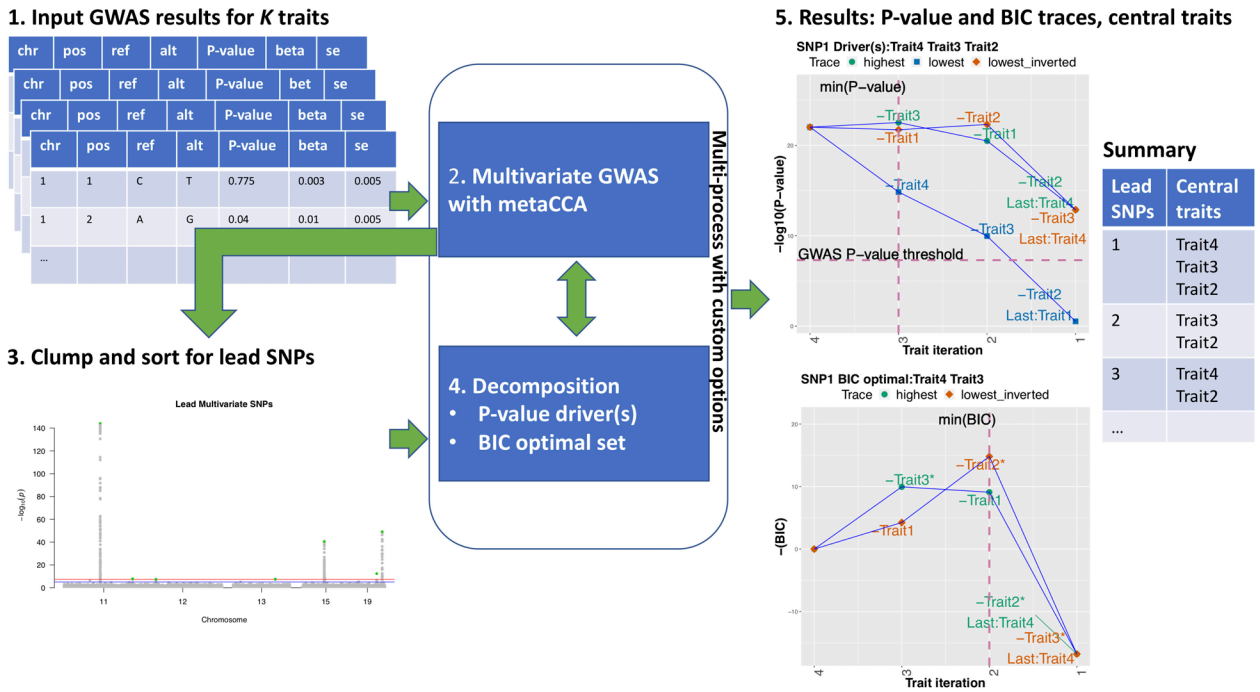

FIGURE 1 | MetaPhat workflow **1.** GWAS results for K traits are accepted as input. **2.** After quality control and filtering, a multivariate GWAS is performed on the full model with all *K* traits using metaCCA via efficient multi-processing and chunking to reduce computation time. **3.** Lead SNPs are detected and sorted based on the leading canonical correlation/*P*-value and then clumped based on a user-specified window size. Custom variants can be added. **4.** Decomposition of chosen variants is performed through highest and lowest traces to find an optimal subset with a minimum BIC and driver traits based on the established *P*-value threshold. **5.** MetaPhat results include trace plots for *P*-values and BIC, univariate association statistics plots for all lead SNPs, cluster maps (shown in **Figure 2**), and a summary table listing central traits (union of drivers and optimal subset).

Subsequently, we defined the central traits as the union of traits from the drivers and optimal BIC subset. MetaPhat traces and terms are summarized in **Table 1**.

Computing *P*-Values and BIC From GWAS Summary Statistics

metaCCA outputs the first canonical correlation r_1 between the genetic variant x and the set of k traits $y_1,\ldots,y_k$ and computes the corresponding *P*-value (Clarke et al., 2011; Cichonska et al., 2016). In this case, the first canonical correlation r_1 equals to the maximum correlation between the variant and any linear combination of the traits and hence is equal to the square root of the variance explained R^2 from the linear regression of x on $y_1,\ldots,y_k$. In general, the expression for BIC is

$$\mathrm{BIC} = \log(n)\,k - 2\widehat{\log lk}$$

where n is the sample size, k is the number of parameters (here traits), and $\widehat{\log lk}$ is the maximized log-likelihood. Next, we have shown how to use metaCCA output r_1 to derive BIC from the maximized likelihood of the linear model written as a function of $R^2 = {r_1}^2$.

Consider a linear model between a (mean-centered) variant x and (mean-centered) traits $\boldsymbol{y} = (y_1,\ldots,y_k)^T$.

$$x = y^T\beta + \varepsilon = y_1\beta_1 + \cdots + y_k\beta_k + \varepsilon,\ \varepsilon \sim N\left(0, \sigma^2\right),$$

where we do not include the intercept parameter as its maximum likelihood estimate (MLE) is zero after mean-centering. The log-likelihood function is

$$\mathrm{loglk}(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{\left(x - y^T\beta\right)^T\left(x - y^T\beta\right)}{2\sigma^2},$$

and MLEs are

$$\hat{\beta} = \left(y^T y\right)^{-1} y^T x \text{and} \hat{\sigma^2} = \frac{1}{n}\left(\left(x - y^T\hat{\beta}\right)^T\left(x - y^T\hat{\beta}\right)\right).$$

Thus, the log-likelihood at maximum is

$$\widehat{\log lk} = \log lk(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\widehat{\sigma^2}\right)$$

$$-\frac{(x - y^T\hat{\beta})^T(x - y^T\hat{\beta})}{\widehat{\sigma^2}} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\widehat{\sigma^2}\right) - \frac{n}{2}$$

$$R^2 = x^T x - \frac{(x - y^T\hat{\beta})^T(x - y^T\hat{\beta})}{x^T x} = 1 - \frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}},$$

that is, $$\frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}} = 1 - R^2 \text{ where } \widehat{\sigma_0^2} = \mathrm{var}(x).$$

TABLE 1 | MetaPhat terminology.

Highest trace	Starting from the full model of K traits, we tested all unique combinations of (K-1) traits to find the subset with the highest CCA statistic (lowest *P*-value), and we iterated until $K = 2$. The goal was to drop most replaceable traits first.
Lowest trace	Starting from the full model of K traits, we tested all unique combinations of (K-1) traits to find the subset with the lowest CCA statistic (highest *P*-value), and we iterated until $K = 2$. The goal was to drop most irreplaceable traits first.
Inverted trace	Aggregates the traits that have been dropped on the lowest trace. The goal was to include the driver sets into the search space for the *optimal set*.
Drivers/driver traits	The traits that have been dropped on the lowest trace at the step where the multivariate *P*-value was for the first time no longer genome-wide significant. Interpretation: traits that make the multivariate association statistically significant.
Optimal set	The subset of traits that has the lowest BIC among subsets across all three traces. Interpretation: the set that is a statistically optimal description of the multivariate association.
Central traits	Union of drivers and optimal set. Interpretation: includes the important traits of the multivariate association.

Hence, the logarithm of the likelihood ratio between the MLE and the null model can be written as

$$\log \mathrm{LR} = \widehat{\log \mathrm{lk}} - \log \mathrm{lk}(0, \widehat{\sigma_0^2}) = -\frac{n}{2}\log\frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}} = 1 - \frac{n}{2}\log(1 - R^2).$$

Hence, we have that, for an additive constant $c = -2\mathrm{loglk}\left(0, \hat{\sigma_0^2}\right)$,

$$\mathrm{BIC} = k\log(n) - 2\left(\widehat{logl\hat{k}}\right) = k\log(n) + n\log\left(1 - R^2\right) + c,$$

which is possible to compute directly from the metaCCA output for models with at least two traits up to an additive constant *c*. Since *c* does not depend on the model dimension, we can ignore it in the BIC calculation, when we are only interested in the differences in BIC between models.

Finally, for a single-trait model, R^2 can be computed directly from the univariate GWAS summary statistics as

$$R^2 = \frac{1}{(1 + n/z^2)} \text{where } z = \frac{\text{GWAS effect}}{\text{standard error}},$$

which can be plugged in the BIC formula above to yield BIC for the single-trait model.

Implementation and Output

MetaPhat is written in Python (compatible for 2.7 and 3+) and requires R (3.4+) for plotting. The command-line based program has been tested on multiple operating systems and cloud images. Library requirements and command options are further described in **Supplementary Table S1**, and test data are accessible from the project page: https://sourceforge.net/projects/meta-pheno-association-tracer.

MetaPhat outputs tabular text files and several plots. A summary result file contains, for each chosen variant, the driver traits and the optimal subset with their *P*-value and BIC statistics. For each variant, trace plots using *P*-values and BIC are generated, showing the highest trace, the lowest trace and the inverted lowest trace. In addition, the univariate *P*-values and directions of effects for each trait are also plotted. The estimated phenotype correlation matrix, clustered heatmaps of trait importance for the chosen variants and a similarity between variants using trait rankings on the lowest trace are produced. Optionally, intermediate statistics during the decomposition can be plotted to get a more detailed view of the decomposition process.

Materials

Our lipidomics data set consisted of the univariate GWAS results of 21 correlated lipid species with polyunsaturated fatty acids that were reported to exhibit high heritability (Tabassum et al., 2019) and showed high correlation (**Supplementary Figure S2**). These results originated from 2,045 Finnish subjects with imputed genotypes available at ∼8.5 million SNPs. The arbitrarily assigned lipid species identifiers along with their class names and fatty acid chemical properties are listed in **Table 2A**. To further validate MetaPhat, we processed summary statistics from four basic lipids [high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides (TG), and total cholesterol (TC)] conducted by the Global Lipids Genetics Consortium (GLGC) (Willer et al., 2013; Zhu et al., 2018), and these are listed in **Table 2B**. With the GLGC data set our aim was to compare MetaPhat results with univariate results reported by GLGC for all variants reported to be significantly associated with two or more traits by GLGC.

RESULTS

Using the lipidomics data sets with GWAS summary statistics from the 21 polyunsaturated lipids, MetaPhat found seven independent lead variants after clumping the 415 variants exceeding the standard GWAS *P*-value threshold of 5e-8 within a window of 1 Mb. **Table 3** lists these variants along with their gene annotation, multivariate *P*-value, and central traits. MetaPhat has strongly reduced the multivariate association for all seven variants into smaller and more specific groups of central traits.

We considered in more detail rs7412, which is a missense variant in the *APOE* gene and is known for its effect on LDL, as reported, for example, in the GLGC analysis (Willer et al., 2013). With the lipidomics data, this variant would not have been identified from any of the 21 univariate GWAS as the smallest univariate *P*-value was 1.1e-4 (trait PCO23, **Supplementary Figure S3.6**). On contrary, the multivariate GWAS by MetaPhat clearly highlighted this variant associated with the multivariate lipidomics (P = 4.2e-13) and further determined that the

TABLE 2 | Lipid traits used in MetaPhat analysis.

(A) PLASMA LIPIDOMICS

Identifier	Lipid class	Lipid species	QC'd variants	HDL corr.	LDL corr.	TG corr.
CE14	Cholesteryl ester	*CE*(20 : 4; 0)	8,711,715	0.032	0.464	0.251
CE15	Cholesteryl ester	*CE*(20 : 5; 0)	8,711,715	0.067	0.396	0.188
CE17	Cholesteryl ester	*CE*(22 : 6; 0)	8,711,665	0.107	0.394	0.107
LPC8	Lysophospatidylcholines	*LPC*(20 : 4; 0)	8,710,151	0.011	− 0.124	− 0.083
LPC9	Lysophospatidylcholines	*LPC*(22 : 6; 0)	8,694,250	0.114	− 0.015	− 0.118
LPE5	Lysophosphatidylethanolamine	*LPE*(20 : 4; 0)	8,710,162	0.077	− 0.077	0.073
LPE6	Lysophosphatidylethanolamine	*LPE*(22 : 6; 0)	8,711,037	0.235	0.005	0.041
PC17	Phosphatidylcholine	*PC*(16 : 0; 0 − 20 : 4; 0)	8,711,715	0.120	0.115	0.361
PC18	Phosphatidylcholine	*PC*(16 : 0; 0 − 20 : 5; 0)	8,711,533	0.126	0.196	0.248
PC29	Phosphatidylcholine	*PC*(17 : 0; 0 − 20 : 4; 0)	8,704,982	0.113	0.138	0.250
PC36	Phosphatidylcholine	*PC*(18 : 0; 0 − 20 : 4; 0)	8,711,715	0.033	0.190	0.336
PC37	Phosphatidylcholine	*PC*(18 : 0; 0 − 20 : 5; 0)	8,751,062	0.061	0.242	0.243
PC46	Phosphatidylcholine	*PC*(18 : 1; 0 − 20 : 4; 0)	8,711,715	0.240	0.105	0.214
PC21	Phosphatidylcholine	*PC*(16 : 0; 0 − 22 : 6; 0)	8,711,715	0.154	0.204	0.219
PCO7	Phosphatidylcholine-ether	*PC* − *O*(16 : 0; 0 − 20 : 4; 0)	8,711,715	0.081	0.194	0.076
PCO23	Phosphatidylcholine-ether	*PC* − *O*(18 : 0; 0 − 20 : 4; 0)	8,711,560	0.187	0.115	− 0.154
PCO29	Phosphatidylcholine-ether	*PC* − *O*(18 : 1; 0 − 20 : 4; 0)	8,710,292	0.198	0.115	− 0.086
PE7	Phosphatidylethanolamine	*PE*(18 : 0; 0 − 20 : 4; 0)	8,707,361	− 0.027	0.028	0.585
PEO3	Phosphatidylethanolamine-ether	*PE* − *O*(16 : 1; 0 − 20 : 4; 0)	8,706,846	0.083	0.198	0.154
PEO11	Phosphatidylethanolamine-ether	*PE* − *O*(18 : 2; 0 − 20 : 4; 0)	8,693,147	0.148	0.238	0.099
PI9	Phosphatidylinositol	*PI*(18 : 0; 0 − 20 : 4; 0)	8,711,715	− 0.026	0.231	0.460

(A) *Polyunsaturated lipid species with acyl chains- C20:4 (14 lipids), C20:5 (3 lipids), and C22:6 (4 lipids) measured for 2,045 individuals (Tabassum et al., 2019). After quality control (QC), a total of 8,576,290 variants were available for all 21 traits. Correlations to basic lipids HDL, LDL, and TG are also shown.* ***(B)*** *Four basic lipids from GLGC (Willer et al., 2013). After quality control, a total of 2,267,285 variants were available for all four traits.*

(B) GLGC LIPIDS

Identifier	Lipid class	QC'd variants	Sample size
HDL	High-density lipoprotein cholesterol	2,343,025	95,129
LDL	Low-density lipoprotein cholesterol	2,271,091	90,421
TC	Total cholesterol	2,341,292	95,537
TG	Triglycerides	2,286,633	91,598

association was driven by CE14 and PCO23 (*P*-value after excluding these driver traits is 1.8e-06). The BIC-optimal subset for this variant extended the drivers by one additional trait and included CE14, PC36, and PCO23, which form the central traits. The trace plots for rs7412 are shown in **Figure 2A** (*P*-values for defining driver traits) and **Figure 2B** (BIC for defining optimal subset).

Variants rs66505542 near *BUD13* and rs261290 near *ALDH1A2* both have only one driver trait (PI9 for *BUD13* and PE7 for *ALDH1A2*) and three or five central traits (**Table 3**). Earlier, the *APOA1* variant rs964184 within 100 kb of rs66505542 has been reported to be associated with TG (lead trait, P = 7.0e-224), TC, HDL, and LDL in GLGC data and rs66505542 itself with several cell phenotypes (platelet count, red cell distribution width, sum of eosinophil and basophil counts) in the GWAS catalog, while rs261290 has been reported to be associated with HDL (lead trait, P = 1.0e-188), TC, and TG in GLGC data (mapped to *LIPC* gene) and with HDL in the GWAS catalog.

A very different picture emerges for rs174567 near *FADS1/2* since its 18 central traits show its wide effects across the lipidomics traits studied here. Previously reported *FADS1/2* associations are with all lipid traits (TG lead trait, P = 7.0e-38) in GLGC data and with metabolite measurements and gallstones in the GWAS catalog.

Trait importance map that clusters each variant based on the lowest trace is shown in **Figure 2C** and the similarity of the variants as measured by rank correlation of the traits on the lowest trace is shown in **Figure 2D**. The trace plots for the other six variants than rs7412 are shown in **Supplementary Figure S1**.

Validation and Global Lipids Genetics Consortium

We processed the Global Lipids Genetics Consortium (GLGC) GWAS study for four plasma lipids (HDL, LDL, TC, and TG, as listed in **Table 2B**). These correlated traits along with large sample sizes and available summary files are suitable for MetaPhat GWAS and decomposition. We focused on the 13 variants reported by GLGC to have associations with three or more lipid traits (**Supplementary Tables S2** and **S3** from Willer et al., 2013). In **Table 4**, we validated that at 12 out of the 13 variants the same associations are confirmed by MetaPhat's central traits. The only discordance was at rs6831256 (*DOK7*) where we found TC and TG as central traits compared to previously reported univariate associations with TC, TG, and LDL. As TC and LDL are highly correlated, it is understandable

TABLE 3 | MetaPhat results of the 7 lead variants from the multivariate analyses of the lipidomics data.

Variant/Gene	Samples missing	*P*-value all traits	Driver trait(s)	*P*-value without drivers	BIC optimal subset	*P*-value BIC optimal subset	Central traits
*rs174567/FADS2	1.3%	2.40e–145	PC36, CE14, PC17, LPC8, PEO11, PEO3, LPE5, PC21, PC46, PC29, CE15, PC37, PC18, PCO7, PCO29, PCO23, PI9, PE7	1.95e–05	CE15, LPC8, PC17, PC21, PC36, PC46, PE7, PEO11, PI9	2.10e–146	PC36, CE14, PC17, LPC8, PEO11, PEO3, LPE5, PC21, PC46, PC29, CE15, PC37, PC18, PCO7, PCO29, PCO23, PI9, PE7
*rs66505542/BUD13	0.1%	1.55e–08	PI9	3.39e–04	PI9, LPC9, PC36	3.27e–12	PI9, LPC9, PC36
rs146327691/SLCO1A2_UTR	1.2%	4.27e–08	LPE5	1.91e–06	LPE5, LPC9, LPE6, PE7	5.60e–11	LPE5, LPC9, LPE6, PE7
rs188167837/ENSG00000200733_UTR_13KB	1.0%	2.95e–08	PC17	7.59e–05	PC17, CE14, CE17, PC21	4.64e–09	PC17, CE14, CE17, PC21
*rs261290/ALDH1A2	0.6%	2.51e–40	PE7	2.04e–07	PE7, CE15, PC17, PCO29, PI9	1.37e–46	PE7, CE15, PC17, PCO29, PI9
*rs7412/APOE	0%	4.17e–13	CE14, PCO23	1.82e–06	CE14, PCO23, PC36	5.79e–18	CE14, PCO23, PC36
rs8736/MBOAT7	23.6%	9.12e–50	PI9	5.89e–02	PI9, LPE6, PC36, PE7	1.25e–81	PI9, LPE6, PC17

*The lipid trait class names and acyl chain properties are listed in **Table 2A**. *Variant region reported as significant for basic lipids by GLGC (Willer et al., 2013).*

that the smaller dimension of the set TC, TG, may in some analyses be preferred over the set that also includes LDL. In **Supplementary Table S2**, we further report high concordance between our central traits and GLGC variants found associated with two or more standard lipids.

Performance

For computing the test statistic, MetaPhat uses metaCCA that, for a single SNP, has previously been shown to reliably estimate the results of standard CCA applied to individual level data (canoncorr function in Matlab) (Cichonska et al., 2016). Additionally, we also empirically validated MetaPhat multivariate findings with GLGC results.

MetaPhat considerably cuts down the computational demands of comprehensive subset testing. With K traits, there are 2^K-1 non-empty subsets that have quickly become infeasible to systematically assess, while MetaPhat only considers about K^2 models. For example, in our example with $K = 21$ traits, the gain in performance is about 4,700-fold compared to the complete subset testing. To further increase performance and usability, we have implemented flexibility for multi-thread processing to enable high performance and memory efficiency. On a moderate Google cloud image (16 vCPUs, 8 GB), the complete MetaPhat workflow for our lipidomics analysis, containing 21 lipids and 8.5 million SNPs, was completed in less than 2.5 h (143 min). Using 10 processors and 9 gigabytes of memory, the GLGC job with the four basic lipids and 2.4 million imputed SNPs completed in 24 min. MetaPhat also allows decomposition and plotting of custom SNPs. For example, the custom analysis of the 13 GLGC variants associated with three or more traits, shown in **Table 4**, was run again on existing GLGC MetaPhat results, and decomposition and plotting took only 2 min. We note that the run time could be longer on shared servers but also substantially shorter using more powerful dedicated cloud images.

DISCUSSION

It is expected that a particular genetic variant may affect only a subset of related biomarkers that are risk factors of complex disorders, such as T2D or coronary heart disease. We implemented MetaPhat to systematically decompose and visualize statistically significant multivariate genome-phenome associations into a smaller group of central traits, based only on univariate GWAS summary statistics. We are not aware of comparable software to MetaPhat that would automatically carry out multivariate GWAS and identify central traits for the associations from summary statistics. ASSET (Bhattacharjee et al., 2012) aims to find the best trait subsets within a pool of multiple studies and has been applied particularly for case-control studies. MTAG (Turley et al., 2018) can be applied to GWAS results of multiple related traits and overlapping samples, but its aim is to improve the accuracy of the univariate effect sizes by using the information from correlated traits rather than decomposing the multivariate association to individual traits.

In our results from an analysis of 21 lipidomics traits, we demonstrated that the *APOE* association (rs7412) benefited from multivariate testing (driven by CE14 and PCO23 traits), as the univariate *P*-value was insignificant ($P >$ 1e-4) across all 21 GWAS traits (shown in **Supplementary Figure S3.6**), but multivariate *P*-value was low ($P <$ 5e-13). This variant is known to have a strong effect on LDL, and **Table 2** shows that CE14 has the highest correlation with LDL (0.464). The other two

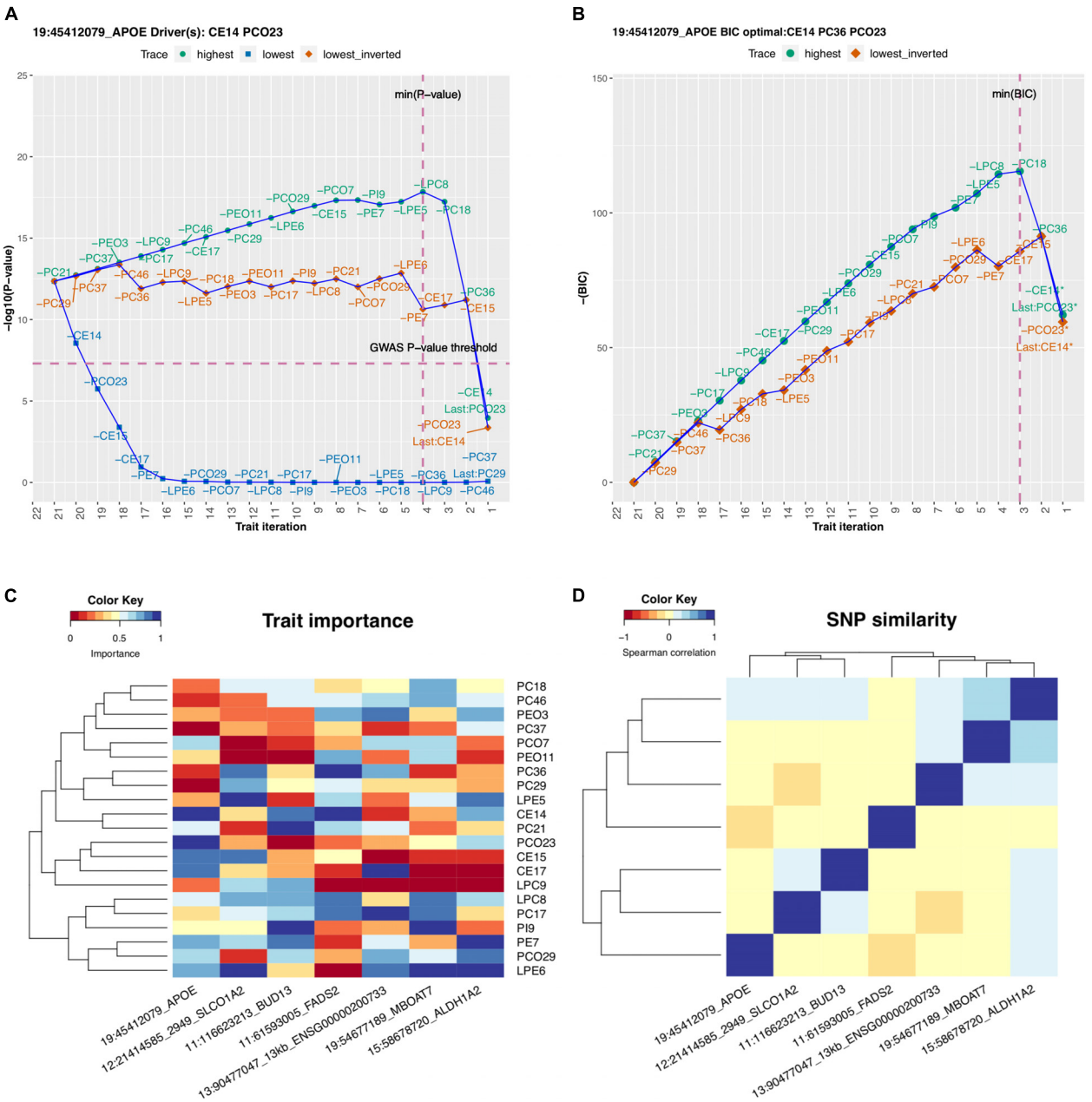

FIGURE 2 | MetaPhat results using multivariate lipidomics data. **(A)** Trace plot of rs7412 identifies CE14 and PCO23 as the driver traits. **(B)** CE14, PC36, and PCO23 form the optimal subset as defined by minimum BIC (highest negative BIC). **(C)** Trait importance map of each SNP is the rank on the lowest trace where the rankings are transformed to the range of 0 and 1 values, with darker blue shades representing the most important traits of the multivariate association. **(D)** SNP similarity based on the rank correlation on the lowest trace.

central traits of this variant, PCO23 and PC36, did not have any correlation to basic lipids larger than 0.20 in absolute magnitude.

Table 3 lists the multivariate results including which four of these seven variants were previously reported by GLGC as associated with at least one of the four basic lipids. The other three variants also have some nearby variants that have been reported in the GWAS catalog (Buniello et al., 2019). First, rs8736 in *MBOAT7* has been previously reported to be associated with human blood metabolites (Shin et al., 2014) as well as alcohol related cirrhosis of the liver (Buch et al., 2015). Second, variants in the region of rs146327691, near the *SLCO1A2* gene, have been previously reported for response to serum metabolites (Krumsiek et al., 2012) and, interestingly, also for response to statins (Ho et al., 2006; Carr et al., 2019). Lastly, variants in the region of rs188167837 have been previously identified to be associated with nasopharyngeal carcinoma (Su et al., 2013). Additionally,

TABLE 4 | MetaPhat detection of driver and optimal lipid sets for 13 variants reported to be associated with at least three lipids by GLGC (12).

Gene	Variant Chr:Pos	GLGC associated lipids	GLGC lead *P*-value	MetaPhat all traits *P*-value	MetaPhat driver(s)	Without driver (s) *P*-value	BIC optimal set	Central traits
		HDL lead						
PIGV-NR0B2	rs12748152 chr1:27138393	HDL LDL TG	1e–15	2.8e–23	HDL LDL TG	3.0e–06	HDL LDL	HDL LDL TG
PPP1R3B	rs9987289 chr8:9183358	HDL LDL TC	2e–41	1.6e–76	HDL TC LDL	1.0e–04	HDL LDL	HDL LDL TC
LIPC (ALDH1A2)	rs1532085 chr15:58683366	HDL TC TG	1e–188	0	HDL TC TG	6.4e–01	HDL TC TG	HDL TC TG
CETP	rs3764261 chr16:56993324	ALL	1e–769	0	ALL	NA	HDL LDL TG	ALL
		LDL lead						
MIR148A	rs4722551 7:25991826	LDL TG TC	4e–14	2.5e–24	TG LDL TC	2.0e–02	LDL TG	LDL TG TC
APOE	rs4420638 19:45422946	ALL	2e–178	6.3e–210	ALL	NA	LDL HDL TC	ALL
		TC lead						
TIMD4	rs6882076 5:156390297	TC LDL TG	5e–41	1.3e–49	TG TC LDL	6.9e–01	TC TG	TC LDL TG
CILP2	rs10401969 19:19407718	TC TG LDL	4e–77	1.3e–138	TG TC LDL	1.0e–01	TC TG	TC TG LDL
		TG lead						
LRPAP1 (DOK7)	rs6831256 4:3473139	TG TC LDL	2e–12	6.3e–16	TG TC	1.0e–07	TG TC	TG TC
ANGPTL3	rs2131925 1:63025942	TG LDL TC	3e–74	7.8e–157	TG LDL TC	9.5e–05	TG TC HDL	ALL
TRIB1	rs2954029 8:126490972	ALL	1e–107	1.6e–148	ALL	NA	TG TC LDL	ALL
FADS123	rs174546 11:61569830	ALL	7e–38	1.3e–104	ALL	NA	ALL	ALL
APOA1	rs964184 11:116648917	ALL	7e–224	7.9e–264	ALL	NA	TG TC	ALL

We confirmed that the vast majority of the MetaPhat central traits are either the same or a subset of the reported GLGC associated lipids (11/13 for driver traits, 12/13 for BIC).

MetaPhat decomposed most variants to substantially smaller sets of central traits than the full set of 21 traits, which can provide new biological insight regarding the variants identified. On the other hand, the essential role of *FADS2* gene region in regulating unsaturation in fatty acids was clearly reflected in MetaPhat results, as we observed as many as 18 central traits at the lead variant. Provided that the exact mechanistic roles of polyunsaturated lipids toward heart disease (Teslovich et al., 2010; Malovini et al., 2016; Pizzini et al., 2017) are under active investigation, our findings warrant further evaluation. We further confirmed good concordance (60/67, **Supplementary Table S2**) with MetaPhat central traits with respect to the earlier reported GLGC associations with two or more standard lipids, and excellent concordance (12/13) with the associations with three or more standard lipids.

MetaPhat optimal subsets are derived from the minimum BIC score representing the model that best describes the data when we account for both the model fit and the model dimension. Qualitatively BIC statistic is similar to the widely-used AIC (Akaike, 1973) statistic, but BIC quantitatively differs from AIC by favoring smaller dimensions, which also improves the interpretation of the optimal models. As intuitively expected, and as seen in **Table 3**, the driver traits tend to be members of the optimal set although they do not always agree, since the driver traits are defined by a GWAS-specific criterion of *P*-value threshold 5e-8, which does not need to coincide with the optimal subset chosen by a more statistically justified BIC criterion.

Our software implements flexible parameters for custom multi-thread chunking to enable high performance, genome-wide, multi-trait meta-analysis while integrating metaCCA for multivariate testing followed by systematic decomposition of traits. Thus, a limitation of MetaPhat is that it relies on metaCCA, but other multivariate GWAS algorithms could also be used provided that these methods can work with univariate GWAS results as inputs and produce suitable metrics that can be used to derive the model comparison statistics. With regard to false positives, we used the standard GWAS cutoff (P = 5e-8), as carried out only a single multivariate GWAS to pick the lead variants. This cutoff can be adjusted according to the preferences of the users. MetaPhat also optionally allows the running of metaCCA+ (Cichonska et al., 2016) shown to protect against false positives via shrinkage that adds robustness to the analysis.

Finally, we remind the reader that MetaPhat decompositions are sequential, dropping one trait at a time, and hence are not guaranteed to produce the globally optimal subset. Additionally, for highly correlated traits, such as LDL and total cholesterol, the choice of which one is dropped first may not be completely robust to small changes in data.

The ability of MetaPhat to identify and visualize central traits will also be valuable in supporting efforts and pipelines (Fatumo et al., 2019) comparing results between univariate and multivariate associations as well as in studies that aim to increase specificity of multi-trait associations. We also expect

that the multi-phenotype clustering results of MetaPhat can assist researchers investigating disease subtypes.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

MP, SR, and JL conceived the project. MP developed the theory. JL implemented and tested the method. RT assisted with the testing. JL and MP draft the manuscript. All authors provided critical feedbacks and important contributions to the final manuscript.

FUNDING

This work was supported by the Academy of Finland (Grant Nos. 288509, 312076, 319181, and 325999).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00431/full#supplementary-material

REFERENCES

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, eds B. N. Petrov and F. Csáki (Budapest: Akadémiai Kiadó), 267–281.

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., et al. (2012). A subset- based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* 90, 821–835. doi: 10.1016/j.ajhg.2012.03.015

Buch, S., Stickel, F., Trépo, E., Way, M., Herrmann, A., Nischalke, H. D., et al. (2015). A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. *Nat. Genet.* 47, 1443–1448. doi: 10.1038/ng.3417

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

Carr, D. F., Francis, B., Jorgensen, A., Zhang, E., Chinoy, H., Heckbert, S. R., et al. (2019). Genomewide association study of statin-induced myopathy in patients recruited using the UK clinical practice research datalink. *Clin. Pharmacol. Ther.* 106, 1353–1361. doi: 10.1002/cpt.1557

Chung, J., Jun, G. R., and Dupuis, J. (2019). Comparison of methods for multivariate gene-based association tests for complex diseases using common variants. *Eur. J. Hum. Genet.* 27, 811–823. doi: 10.1038/s41431-018-0327-8

Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimäki, T., et al. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* 32, 1981–1989. doi: 10.1093/bioinformatics/btw052

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., Zondervan, K. T., et al. (2011). Basic statistical analysis in genetic case-control studies. *Nat. Protoc.* 6, 121–133. doi: 10.1038/nprot.2010.182

Fatumo, S., Carstensen, T., Nashiru, O., Gurdasani, D., Sandhu, M., and Kaleebu, P. (2019). Complimentary methods for multivariate genome-wide association study identify new susceptibility genes for blood cell traits. *Front. Genet.* 10:334. doi: 10.3389/fgene.2019.00334

Gai, L., and Eskin, E. (2018). Finding associated variants in genome-wide association studies on multiple traits. *Bioinformatics* 34, i467–i474. doi: 10.1093/bioinformatics/bty249

Guo, B., and Wu, B. (2019). Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. *Bioinformatics* 35, 2251–2257. doi: 10.1093/bioinformatics/bty961

Ho, R. H., Tirona, R. G., Leake, B., and Glaeser, H. (2006). Drug and bile acid transporters in rosuvastatin hepatic uptake: function, expression, and pharmacogenetics. *Gastroenterology* 130, 1793–1806. doi: 10.1053/j.gastro.2006.02.034

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi: 10.1093/biomet/28.3-4.321

Inouye, M., Ripatti, S., Kettunen, J., Leo-Pekka, L., Oksala, N., Laurila, P., et al. (2012). Novel loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet.* 8:e1002907. doi: 10.1371/journal.pgen.1002907

Krumsiek, J., Suhre, K., Evans, A. M., Matthew, W., Robert, P. M., Michael, V., et al. (2012). Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.* 8:e1003005. doi: 10.1371/journal.pgen.1003005

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* 50, 1505–1513. doi: 10.1038/s41588-018-0241-6

Malovini, A., Bellazzi, R., Napolitano, C., and Guffanti, G. (2016). Multivariate methods for genetic variants selection and risk prediction in cardiovascular Diseases. *Front. Cardiovasc. Med.* 3:17. doi: 10.3389/fcvm.2016.00017

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Federico, C. F., Paul, C., Marjo-Riitta, E., et al. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* 7:e34861. doi: 10.1371/journal.pone.0034861

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* 32, 381–385. doi: 10.1002/gepi.20303

Pizzini, A., Lunger, L., Demetz, E., Hilbe, R., Weiss, G., Ebenbichler, C., et al. (2017). The role of omega-3 fatty acids in reverse cholesterol transport: a review. *Nutrients* 9:1099. doi: 10.3390/nu9101099

Ripatti, P., Rämö, J. T., Söderlund, S., Surakka, I., Matikainen, N., Pirinen, M., et al. (2016). The contribution of GWAS loci in familial dyslipidemias. *PLoS Genet.* 12:e1006078. doi: 10.1371/journal.pgen.1006078

Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. Statist.* 6, 461–464. doi: 10.1214/aos/1176344136

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550. doi: 10.1038/ng.2982

Su, W. H., Yao, J., Shugart, Y., Chang, K. P., Tsang, N. M., Tse, K. P., et al. (2013). How genome-wide SNP-SNP interactions relate to nasopharyngeal carcinoma susceptibility. *PLoS One* 8:e83034. doi: 10.1371/journal.pone.0083034

Tabassum, R., Rämö, J. T., Ripatti, P., Jukka, T. K., Mitja, K., Karjalainen, J., et al. (2019). Genetic architecture of human plasma lipidome and its link to cardiovascular disease. *Nat. Commun.* 10:4329. doi: 10.1038/s41467-019-11954-8

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

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

van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet.* 9:e1003235. doi: 10.1371/journal.pgen.1003235

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283. doi: 10.1038/ng.2797

Zhu, Z., Anttila, V., Smoller, J. W., and Lee, P. H. (2018). Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies. *PLoS One* 13:e0193256. doi: 10.1371/journal.pone.0193256

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 May 2020
doi: 10.3389/fgene.2020.00437

Genome-Wide Gene-Based Multi-Trait Analysis

*Yamin Deng[1], Tao He[2], Ruiling Fang[1], Shaoyu Li[3], Hongyan Cao[1] and Yuehua Cui[4]**

[1] Division of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, China, [2] Department of Mathematics, San Francisco State University, San Francisco, CA, United States, [3] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, United States, [4] Department of Statistics and Probability, Michigan State University, East Lansing, MI, United States

Genome-wide association studies focusing on a single phenotype have been broadly conducted to identify genetic variants associated with a complex disease. The commonly applied single variant analysis is limited by failing to consider the complex interactions between variants, which motivated the development of association analyses focusing on genes or gene sets. Moreover, when multiple correlated phenotypes are available, methods based on a multi-trait analysis can improve the association power. However, most currently available multi-trait analyses are single variant-based analyses; thus have limited power when disease variants function as a group in a gene or a gene set. In this work, we propose a genome-wide gene-based multi-trait analysis method by considering genes as testing units. For a given phenotype, we adopt a rapid and powerful kernel-based testing method which can evaluate the joint effect of multiple variants within a gene. The joint effect, either linear or nonlinear, is captured through kernel functions. Given a series of candidate kernel functions, we propose an omnibus test strategy to integrate the test results based on different candidate kernels. A *p*-value combination method is then applied to integrate dependent *p*-values to assess the association between a gene and multiple correlated phenotypes. Simulation studies show a reasonable type I error control and an excellent power of the proposed method compared to its counterparts. We further show the utility of the method by applying it to two data sets: the Human Liver Cohort and the Alzheimer Disease Neuroimaging Initiative data set, and novel genes are identified. Our method has broad applications in other fields in which the interest is to evaluate the joint effect (linear or nonlinear) of a set of variants.

Keywords: gene-based association, kernel function, multi-trait, nonlinear effect, *p*-value combination

OPEN ACCESS

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

Reviewed by:
Ming Li,
Indiana University, United States
Qing Lu,
University of Florida, United States

***Correspondence:**
Yuehua Cui
cuiy@msu.edu

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

Received: *24 February 2020*
Accepted: *08 April 2020*
Published: *19 May 2020*

Citation:
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

INTRODUCTION

Methods on genome-wide association studies (GWAS) are mostly focused on single variant (e.g., single nucleotide polymorphism, SNP) analysis with a single phenotype, the so-called single-variant single-trait analysis. Increasing evidence shows that pleiotropy, the effect of one gene on multiple phenotypes (often correlated), plays a pivotal role in many complex traits (Stearns, 2010; Schifano et al., 2013). For example, cognitive ability is often assessed in many domains such as memory, intelligence, language, and visual–spatial function (Yang and Wang, 2012). Instead of analyzing one trait at a time, we can take the correlated structure of multiple phenotypes into account and analyze them in a multi-trait analysis. As a complementary approach, such type of analysis can not only

gain association power by aggregating multiple weak signals (He et al., 2013; Schifano et al., 2013; Wang, 2014) but also lead to better understanding of disease etiology by detecting genetic variants with pleiotropic effects (Amos and Laing, 1993; Jiang and Zeng, 1995; Schifano et al., 2013).

For a multi-trait analysis, one commonly applied method is the one-way multivariate analysis of variance (MANOVA) (Bilodeau, 2013). Unfortunately, most multi-trait data do not satisfy the multivariate normal assumption for MANOVA, hence greatly limiting its applicability. Other methods are developed based on the idea of dimension reduction. For example, a multivariate response can be summarized into a univariate score using principal component (PC) analysis, based on which traditional univariate association methods can be applied (e.g., Zhang et al., 2012). As the first PC contains the most information about multiple phenotypes, this can change the test between a SNP and multiple phenotypes into a univariate test of association between a SNP and the first PC. The downside for this analysis is the lack of interpretability. Methods focusing on summary statistics have gained much popularity recently since the individual-level data are typically unavailable (e.g., Kim et al., 2015; Turley et al., 2018). However, such methods are largely undermined if the published GWAS summary statistics have limited accuracy. In addition, the marginal SNP effect is usually quite small in many complex diseases, and many identified SNPs have limited biological interpretation, for example, SNPs identified in non-coding regions.

These limitations motivated the development of gene- or pathway-based association analysis aimed at improving the statistical power and gaining novel insight into disease etiology (Wang et al., 2007; Cui et al., 2008; Liu et al., 2010). Firstly, the gene- or pathway-based analysis can largely alleviate the multiple testing burden by more than 10 or 100 folds. Secondly, due to allelic heterogeneity, most diseases are associated with a set of SNPs at different loci, making it hard to replicate the results based on a single-SNP analysis (Neale and Sham, 2004). In this case, a gene- or pathway-based analysis may provide additional insight to reveal the functional mechanism of complex diseases (Wang et al., 2010). Unlike the heterogeneity of a single locus, the biological function of genes is more consistent across populations, which enhances the likelihood of replication (Neale and Sham, 2004; Wang et al., 2010).

Most reports in the literature on multi-trait analysis are focused on a single-variant analysis, which shares the same limitation as described for the single-trait GWAS. Although methods for gene-based analysis focusing on a single trait have been developed, multi-trait analysis focusing on genes or gene sets is largely under-developed. There is a pressing need to develop a gene-based method for a multi-trait analysis.

In a gene-based single-trait analysis, the kernel-based testing (KBT) method is gaining much popularity recently due to its power and flexibility in capturing potential nonlinear effects (Kwee et al., 2008; Mukhopadhyay et al., 2010; Wu et al., 2010; Li and Cui, 2012; Lin et al., 2013; Marceau et al., 2015; Wei and Lu, 2017). The power of the KBT methods depends on the choice of kernel functions which measure the similarity between individuals across multiple genetic variants in a gene. When the underlying true disease function is unknown, this limits the applicability of the KBT methods since the choice of the kernel function needs to be determined. Given a series of candidate kernel functions under the KBT framework, a common method is to choose the kernel function leading to the smallest p-value. This idea, however, could inflate the type 1 error rate due to the greedy process of kernel selection. We recently proposed a nonparametric KBT testing procedure which relaxes the distributional assumption required in most KBT methods (He et al., 2019). The asymptotic distribution of the test statistics approximately follows a normal distribution when the number of SNP variants in a gene set, p, is large. In fact, the normal approximation works well under a large p setting. Given a series of candidate kernel functions, we provided an analytical procedure to evaluate the p-value of the maximum statistics.

Based on empirical studies, the approximation method could be underperformed when p is relatively small. In this work, we borrowed the same idea but relaxed the large p assumption required for the normal approximation and proposed an omnibus testing procedure when multiple candidate kernels are available. Obtaining a p-value needs almost negligible computation and can be extremely fast. When extending the method to a multi-trait analysis, we adopted a Fisher p-value combination (FPC) method with correlated dependent variables, as proposed by Yang et al. (2016). The FPC provides an alternative approach for multi-trait analysis by integrating the single-trait analysis results. The proposed Omnibus Multi-trait Gene-based Association (OMGA) analysis can capture linear or nonlinear effects without kernel selection and is computationally efficient.

We conduct extensive simulation studies to evaluate the type I error control and power and further compare it with its counterparts. We demonstrate the performance of our proposed method through two real data applications of the Human Liver Cohort (HLC) study and the Alzheimer Disease Neuroimaging Initiative (ADNI) study. The results tell which genes are specific to a single phenotype or contributed to a common genetic construction of multiple phenotypes. Our OMGA method enriches the literature of genome-wide gene-based multi-trait association analysis and has broad applications in other fields where the interest is to evaluate the joint effect (linear or nonlinear) of a set of variants.

STATISTICAL METHODS

Gene-Based Association Test Based on a Single Trait

The Model

To model the association between a gene and a quantitative trait, we consider the following semiparametric model (He et al., 2019),

$$Y_i = \mu + \alpha^T W_i + h(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots\ldots, n, \qquad (1)$$

where Y_i is the response variable for the i-th individual, n is the sample size, α is the effect corresponding to $W_i = (W_{i1}, W_{i2}, \ldots\ldots W_{iH})^T$, a vector of H-dimensional covariates containing variables such as age and gender,

$x_i = \left(x_{i1}, \ldots\ldots x_{ip}\right)^T$ is a vector of a p-dimensional SNP set in a given gene where p can be large, $h(\cdot)$ is an unknown function that captures the joint effect of multiple variants in a given SNP set, and ε_i is the random error with mean 0 and variance σ^2. Here, we relax the error distribution assumption for the error term which does not have to follow a normal distribution.

Following model (1), assessing the effect of multiple variants in a given SNP set (e.g., a gene) is equivalent to test the hypotheses H_0: $h(\cdot) = 0$, while adjusting for the effects of covariates. Wu et al. (2011) proposed a kernel-based test by considering the joint effect of multiple SNPs in a given set and showed great power compared to a multiple-regression approach. In Wu et al. (2011), the function $h(\cdot)$ is modeled as a random effect and $h(\cdot) \sim N\left(0, \tau^2 K\right)$ where τ^2 is the variance and K is a kernel matrix which measures the similarity between individuals across multiple SNP variants. However, the normality assumption on $h(\cdot)$ limits its power when this assumption is violated. To relax this assumption, He et al. (2019) proposed a U-statistic defined as:

$$T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K\left(X_i, X_j\right)\left(Y_i - \hat{Y}_i\right)\left(Y_j - \hat{Y}_j\right)/\hat{\sigma}^2,$$

where $\hat{Y}$ and $\hat{\sigma}^2$ are sample estimates under the null model $Y_i = \mu + \alpha^T W_i + \varepsilon_i$; $K\left(X_i, X_j\right) = \frac{K_\theta\left(X_i, X_j\right)}{\sqrt{E\{K_\theta(X_i,X_i)\}E\{K_\theta(X_j,X_j)\}}}$ is the normalized kernel for kernel $K_\theta\left(X_i, X_j\right)$. In practice, the choice of kernel function for $K_\theta\left(X_i, X_j\right)$ depends on the underlying relationship between SNPs and the disease response. For example, a linear kernel is applied if the relationship between multiple SNP variants and the disease response is linear, and a Gaussian or polynomial kernel can be applied if a nonlinear relationship between multiple SNPs and the disease response is assumed. Several widely used kernel functions include the linear kernel $K_\theta\left(X_i, X_j\right) = X_i^T X_j/\theta$, IBS kernel for discrete SNP genotype data $K_\theta\left(X_i, X_j\right) = \sum_{k=1}^{p}\left(2 - \left|X_{ik} - X_{jk}\right|\right)/2p$, and Gaussian kernel $K_\theta\left(X_i, X_j\right) = \exp\left(-||X_i - X_j||^2/\theta\right)$. These kernels will be our candidate kernels in the simulation and real data analysis.

Let $\tilde{W}_{n\times(L+1)} = [1_n, W_{n\times L}]$ and $\mathrm{A} = \tilde{W}\left(\tilde{W}^T\tilde{W}\right)^{-1}\tilde{W}^T$. Then, we have $\hat{\sigma}^2 = Y^T(I - A)Y/(n - L - 1)$ and $\hat{Y} = \mathrm{AY}$. Following the Eigen-decomposition, $K\left(X_i, X_j\right) = \sum_{m=1}^{\infty}\lambda_m\phi_m(X_i)\phi_m\left(X_j\right)$ where λ_m is the eigenvalues and $\phi_m(\cdot)$ is the orthonormal eigenvectors of the kernel K. For any positive integer k, let $V_k = \sum_{m=1}^{\infty}\lambda_m^k$. Then, under the null hypothesis of no association, the asymptotic distribution of the test statistic T_n follows a chi-square distribution, i.e.:

$$n\,T_n/V_1 \xrightarrow{d} \sum_{m=1}^{\infty}\lambda_{k,m}\left(x_{1,m}^2 - 1\right),$$

where $x_{1,m}^2$ are independent chi-square distributions with one degree of freedom. Then, we can apply a Satterthwaite approximation to the mixture of chi-squares by a scaled chi-square distribution $\hat{a}\chi_{\hat{g}}^2/\hat{V}_1 - 1$, where $\hat{g} = \hat{V}_1/\hat{a}$, $\hat{a} = \hat{\sigma}_{T_n}^2/2\hat{V}_1$, and $\hat{V}_1 = n^{-1}\mathrm{tr}(\boldsymbol{HK})$ is a consistent estimator of V_1 with $\boldsymbol{H} = \boldsymbol{I} - n^{-1}\boldsymbol{J}$ as a projection matrix. Then, an asymptotic α-level test rejects the null if

$$(nT_n + \hat{V}_1)/\hat{a} > \chi_{\hat{g},1-\alpha}^2,$$

where $\chi_{(\hat{g},1-\alpha)}^2$ is the (1 – α)th quantile of a chi-square distribution with $\hat{g}$ degrees of freedom. Following He et al. (2019), $\sigma_{T_n}^2$ can be estimated by

$$1/n^2\left(2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n}\right)tr(\boldsymbol{HKHK}) - \left(\frac{2}{n} + \frac{\hat{\Delta}}{n}\right)tr^2(\mathbf{HK}) + \hat{\Delta}\,tr(\boldsymbol{B}\circ\boldsymbol{B}),$$

where $\circ$ represents the Hadamard product, $\hat{\Delta} = n^{-1}\sum_{i=1}^{n}\left[\frac{(Y)_i - \bar{Y}_n}{\hat{\sigma}}\right]^4 - 3$, and $\boldsymbol{B} = \mathbf{HK}\boldsymbol{H}$. Then, the p-value of T_n can be obtained.

An Omnibus Test With Multiple Candidate Kernels

The method described above works for a given kernel function. There are various kernel functions available to use. For example, if a linear relationship is assumed, then one can apply a linear kernel, while a Gaussian kernel can be applied when potential nonlinear relationship exists. Thus, the power of the proposed test statistic largely depends on the choice of the kernel function. If the optimal kernel function that captures the underlying true relationship cannot be determined, the testing power will suffer. In practice, the true relationship is generally unknown, so does the choice of the kernel function.

To overcome the issue of selecting the optional kernel function, we propose an omnibus test strategy in this work. Given a set of L candidate kernels denoted by $K_1(\cdot,\cdot), K_2(\cdot,\cdot), \cdots, K_L(\cdot,\cdot),$, we can apply the proposed method and get the corresponding p-value denoted by p_1, p_2,…p_L. These L kernel functions can come from a wide range of choices, such as the linear kernel, the Gaussian kernel, and the polynomial kernel. Then, we can transform the L p-values by a Cauchy transformation and combine the transformed p-values to form a new statistic (Liu et al., 2019),

$$T_O = \frac{1}{L}\sum_{j=1}^{L}\tan\left\{\left(0.5 - p_j\right)\pi\right\}.$$

If p_j comes from the null hypothesis, the transformation $\tan\left\{\left(0.5 - p_i\right)\pi\right\}$ follows a Cauchy distribution. Then, the p-value of T_O can be approximated by

$$p\text{-value} \approx 0.5 - \{\arctan(T_O)\}/\pi$$

This Cauchy combination method performs similarly as the minimum p-value method. In addition, it works well under different correlation structures. Thus, when the underlying true relationship is unknown, if the choice of the kernel function is rich enough, we can always achieve good power regardless

of the underlying disease gene action mode. More importantly, this method is computationally fast and robust to different dependence structures between p-values (Liu and Xie, 2019).

Gene-Based Association Test With Multiple Traits

When multiple correlated traits are available, it is more powerful to analyze them together to find the disease–gene association. One way to do so is to perform a multivariate analysis by treating multiple traits as a multivariate response. Generally speaking, it is much easier to conduct a univariate association test than a multivariate association test. Suppose there are a total of d quantitative traits. For a given gene, we can get d gene-level p-values, denoted by $p_1, p_2, \ldots p_d$. Since these d traits are generally correlated and the p-values are obtained based on the same gene, these p-values are typically correlated. To obtain a gene-based p-value for multiple traits, one simple way is to do a p-value combination. Unfortunately, the aforementioned Cauchy combination method does not work well in many cases since it functions like a minimum p-value approach, and this is not the intention for multi-trait analysis.

When the d p-values are independent, the Fisher combination method defined as $T = -2\sum_{j=1}^{d}\log\left(p_j\right)$ follows a chi-square distribution with $2d$ degrees of freedom (Littell and Folks, 1971). For correlated traits, this method cannot be directly applied to find the association between one gene and multiple traits. In fact, the statistic T is a sum of correlated chi-square statistics which can be approximated by a scaled chi-square distribution δx_τ^2 or a gamma distribution with a scale parameter of 2δ and a shape parameter of $\tau/2$ under the null hypothesis (Yang et al., 2016). Let $E\left(T\right) = \mu$ and $\mathrm{Var}\left(T\right) = \sigma^2$. Then, δ and τ can be computed as $\delta = \sigma^2/2\mu$ and $\tau = 2\mu^2/\sigma^2$. Here we adopt the method proposed by Yang et al. (2016) to combine the d-dependent p-values. The variance σ^2 can be calculated as

$$\begin{aligned}\sigma^2 &= Var\left[T\right] = Var\left\{-2\sum_{j=1}^{d}\log\left(p_j\right)\right\}\\ &= \sum_{j=1}^{d} Var\left\{-2\log\left(p_j\right)\right\} + \sum_{j\neq k} cov\left(-2\log\left(p_j\right), -2\log\left(p_k\right)\right)\\ &= 4d + \sum_{j\neq k} cov\left(-2\log\left(p_j\right), -2\log\left(p_k\right)\right)\end{aligned}$$

Let $\delta_{jk} = cov\left\{-2\log\left(p_j\right), -2\log\left(p_k\right)\right\}$. Yang et al. (2016) proposed a method to estimate δ_{jk} based on which we can estimate σ^2 [please refer to Yang et al. (2016) for the technical details of estimating σ^2 and μ]. An R package implementing the method can be found at https://github.com/jjyang2019/FisherCombinationStat. Then, based on the estimators of μ and σ^2 for the gamma distribution parameters, the overall testing p-value of T can be calculated as

$$p\text{-value} = 1 - \Gamma\left(\mu^2/\sigma^2, \sigma^2/\mu\right).$$

The number of the gene-level test is much smaller than the number of the SNP-level test. After obtaining the gene-level p-values, multiple testing adjustment such as FDR can be applied to claim the significance of a gene.

SIMULATION STUDIES

Simulation Design

To evaluate the statistical power and the type 1 error rate of the proposed method, we conducted extensive simulation studies to compare the proposed method (OMGA) with some existing methods. Specifically, we compared with the method of multivariate multiple linear regression (RMMLR) proposed by Basu et al. (2013) and the MANOVA method. RMMLR was developed based on multivariate regression and transformed the phenotype and genotype data to achieve a rapid gene-based genome-wide association test for multiple traits. The R package that implements the method, termed as RMMLR, is available at GitHub: https://github.com/SAONLIB/RMMLR. For the MANOVA analysis, the association between each SNP in a gene and multi-trait is implemented with the MANOVA function in R. The minimum p-value in a gene is recorded as the gene-level p-value.

The genetic data were simulated to mimic the real structure of a gene through the software EpiSIM (Shang et al., 2013). The software package of EpiSIM can be downloaded at https://sourceforge.net/projects/episimsimulator/. We simulated correlated quantitative phenotypes with the following model:

$$Y_i = 0.02Z_{i1} + 0.6Z_{i2} + h\left(X_i\right) + \epsilon_i, \quad i = 1, \ldots\ldots, n,$$

where $\epsilon_i = \left(\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{id}\right)^T$ is a d-dim random error vector generated from a multivariate normal distribution with mean 0 and covariance Σ; $Y_i = \left(Y_{i1}, Y_{i2}, \cdots, Y_{id}\right)^T$ is a d-dim-dependent trait vector; $Z_{i1} \sim N\left(2, 1\right)$ and $Z_{i2} \sim Ber\left(0.6\right)$ are two independent covariates; $X_i = \left(X_{i1}, X_{i2}, \cdots, X_{ip}\right)^T$ is a p-dim SNP genotype vector in a gene. Under all scenarios, we simulated genes with different dimensions, i.e., $p = 50$ and $p = 100$, and with different sample sizes, namely, $n = 100$, 200, and 400. For the number of traits, we assumed $d = 5$. The correlation between traits was assumed to be $p = 0.3$ and 0.8, with the purpose to evaluate the impact of correlation on the testing power. In each scenario, we applied 1,000 simulation replications.

We assessed the type 1 error rates under the null hypothesis [i.e., $h(\cdot) = 0$] by the proportion of results that incorrectly rejected the null hypothesis. To evaluate the power, we set up four different scenarios for the $h(\cdot)$ function and recorded the proportion of results that rejected the null hypothesis. Under scenario A, we assumed that $h\left(x\right) = 0.2\left(x_1 - x_6\right) + \cos\left(x_6\right)\exp\left(-x_6^2/4\right)$, where the 1st and 6th SNP have a main effect with different directions and the 6th SNP also has a nonlinear effect on the five response traits. Under scenario B, we assumed that $h\left(x\right) = 0.3x_2 + 0.6x_4 - 0.07x_8$.

To mimic the situation where a large number of SNPs influence the traits, we assumed the following model:

$$h\left(x\right) = c_M\sum_{k\in S_M}\alpha_k x_k + c_N\sum_{k,k'\in S_N}\beta_{kk'}x_k x_{k'}$$

where S_M consists of a predefined set of 10 SNPs with main effect, and S_N contains a set of 30 SNP pairs with interactions. Both $\{\alpha_k, k \in S_M\}$ and $\{\beta_{kk'}, (k, k') \in S_N\}$ were generated from a uniform distribution with Unif (0, 0.02), and were fixed for all simulation replicates once generated. Under scenario C, we set $C_M = 0.02$ and $C_N = 1.8$, which gave a combination of weak main effect and relatively strong interaction effect. Under scenario D, we set $C_M = 3.8$, and $C_N = 0$, with a pure main effect model. The four scenarios with their corresponding mean functions are summarized here:

Scenario A: $h(x) = 0.2(x_1 - x_6) + \cos(x_6)\exp(-x_6^2/4)$
Nonlinear effect
Scenario B: $h(x) = 0.3x_2 + 0.6x_4 - 0.07x_8$
Linear effect
Scenario C: $h(x) = 0.02\sum_{k\in S_M}\alpha_k x_k + 1.8\sum_{k,k'\in S_N}\beta_{kk'}x_k x_{k'}$
Weak main but strong interaction effects
Scenario D: $h(x) = 3.8\sum_{k\in S_M}\alpha_k x_k$
Pure main effects

Simulation Results

Table 1 displays the empirical type 1 error rate of different methods under different settings, from which we conclude that the three methods maintained reasonable type 1 error rate control in most settings.

The power simulation results for the case with $p = 0.3$ are shown in **Figure 1**. Under different scenarios, the power of the three methods all increases as the sample size increases. Among the three methods, MANOVA performs the worst in most cases. Although the power decreases as the SNP dimension increases for all the three methods, the power decrease is more dramatic for RMMLR and MANOVA compared to that for OMGA. This indicates the relative advantage of the proposed method against the other two when the data dimension is high. The result clearly shows that the proposed omnibus test outperforms the other two methods under different scenarios since it can better capture the potential nonlinear effect of variants within a gene by applying a nonparametric KBT procedure with different kernel choices.

TABLE 1 | The type 1 error rate of different methods under different settings.

Data dimension	Sample size (n)	Correlation (ρ)	OMGA	RMMLR	MANONA
$p = 50$	100	0.3	0.059	0.037	0.052
		0.8	0.045	0.052	0.041
	200	0.3	0.050	0.061	0.038
		0.8	0.048	0.049	0.032
	400	0.3	0.048	0.064	0.052
		0.8	0.051	0.061	0.061
$p = 100$	100	0.3	0.044	0.052	0.046
		0.8	0.049	0.038	0.044
	200	0.3	0.061	0.041	0.046
		0.8	0.041	0.067	0.043
	400	0.3	0.051	0.057	0.035
		0.8	0.047	0.050	0.037

Figure 2 shows the empirical testing power of the three methods with $p = 0.8$. Compared with the $p = 0.3$ case, the power of RMMLR and MANONA decreased, while our proposed method can still maintain a comparable power as the $p = 0.3$ case. Note that the MANOVA method implemented here uses a minimum p-value approach among multiple SNPs to denote a gene-level p-value. The simulation result echoes the work of Basu and Pan (2011), in which the minimum p-value method performs the worst among the three methods that the authors compared in their simulation study.

In summary, the simulation results clearly demonstrate that the proposed omnibus test method can maintain a reasonable type I error control while having better power than the other two methods under different scenarios. This is because the proposed omnibus testing method can efficiently capture a linear or a nonlinear relationship between multiple variants in a gene and multiple phenotypes. In practice, the underlying true disease–gene relationship is never known. This makes our proposed omnibus test method particularly attractive in real application since it does not put any model assumption. As long as the choice of kernel functions is rich enough, the omnibus test can achieve its power advantage against the other methods which only function well under the desired model assumption.

REAL DATA ANALYSIS

Case One: The Human Liver Cohort Data Analysis

To demonstrate the power and the applicability of our approach, we applied the proposed method OMGA together with RMMLR and MANONA to a HLC study data set, which can be downloaded from https://www.synapse.org/#!Synapse:syn4499. The HLC study aims to explore the genetic architecture of gene expressions in human liver. There are a total nine phenotypes of P450 enzymes (CYP1A2, 2B6, 2C8, 2A6, 2C9, 2D6, 2C19, 2E1, and 3A4) from unrelated liver samples of Caucasian individuals. The samples were removed if their genotype and phenotype information were missed, and the final data included in our study contained 170 individuals. DNAs were genotyped by the Illumina 650Y SNP and Affymetrix 500K SNP genotyping arrays. SNPs with a minor allele frequency (MAF) less than 5% were removed. The total number of SNPs that remained was 312,082, which were further mapped into 11,579 genes using tools from the NCBI website ftp://ftp.ncbi.nih.gov/snp/.

The cytochrome P450s compose a superfamily of monooxygenases which are critical for anabolic and catabolic metabolism in almost all living organisms (Nelson et al., 1996; Aguiar et al., 2005; Plant, 2007). With its importance in physiology and drug metabolism in human, the regulatory mechanisms and genetic variations of P450 enzyme have been extensively studied. As there is a relatively close relationship among the CYP family enzymes, a joint analysis of multiple P450 enzyme traits and gene association can potentially lead to the identification of novel genes. Based on a hierarchical clustering analysis, we focused on six enzyme activity traits,

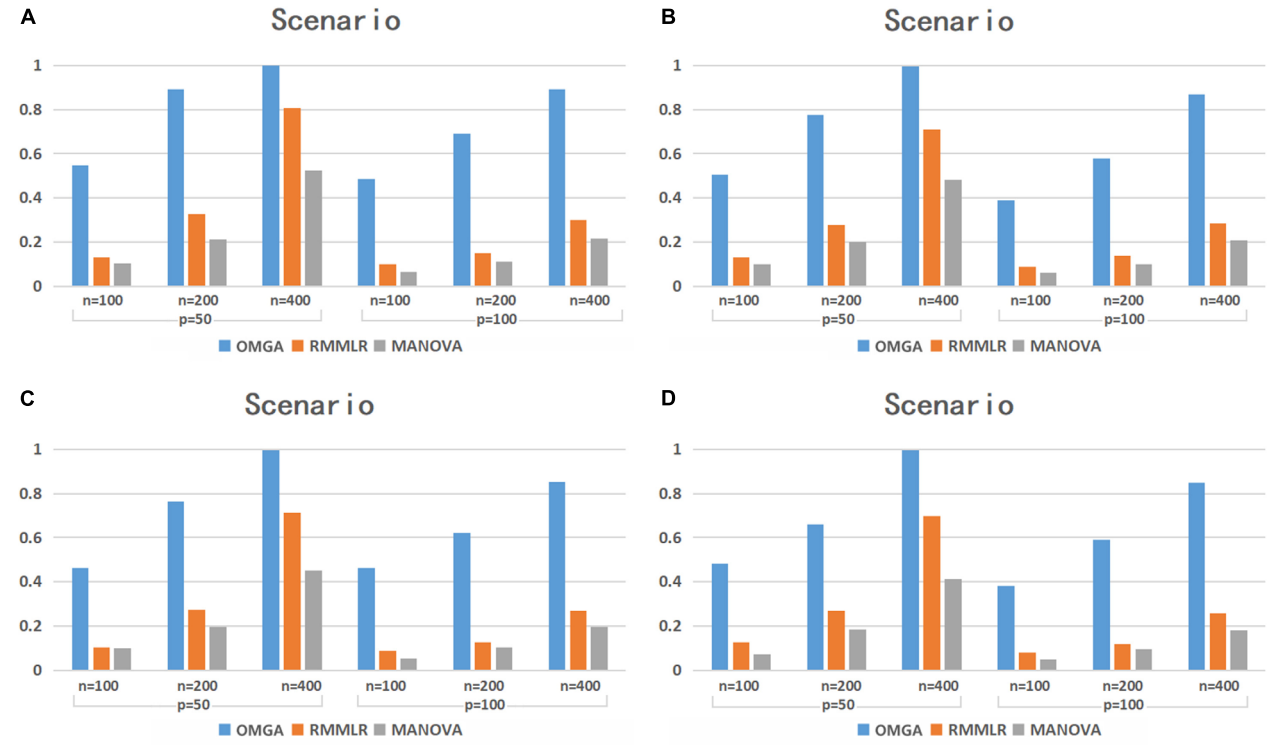

FIGURE 1 | The testing power of different methods under the four scenarios with $p = 0.3$.

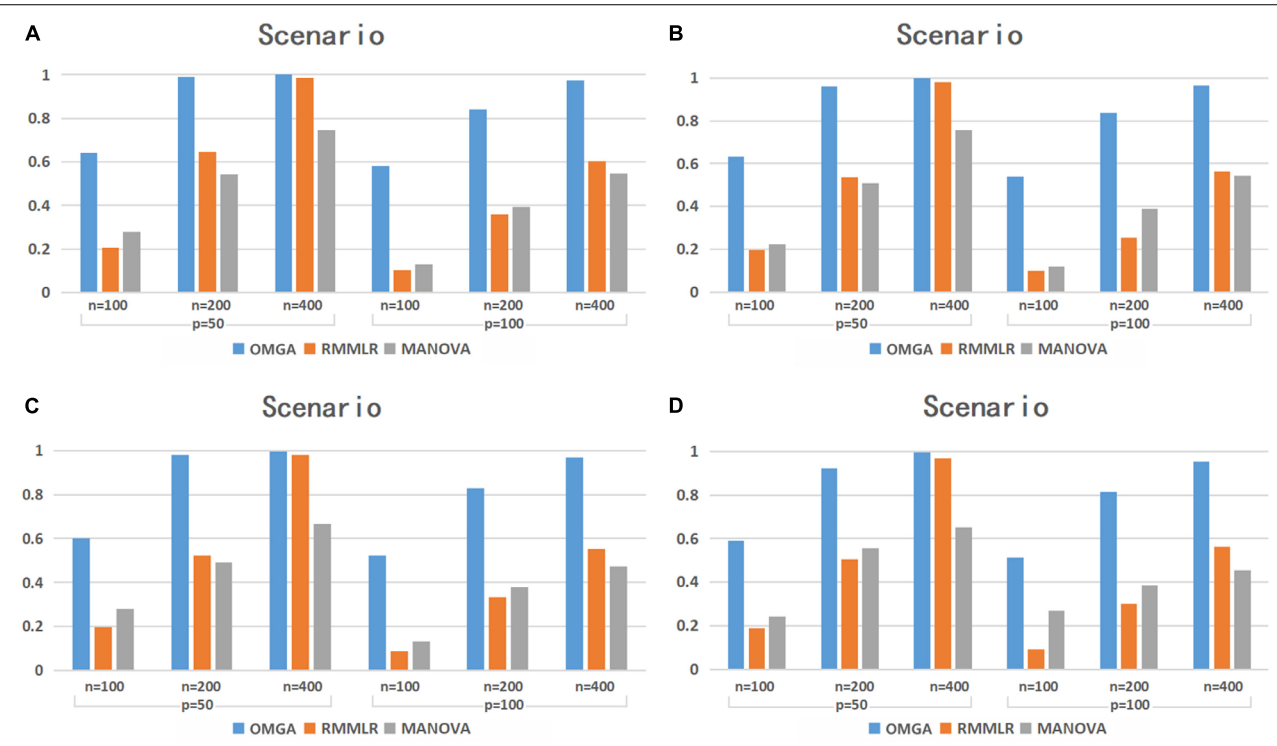

FIGURE 2 | The testing power of different methods under the four scenarios with $p = 0.8$.

namely, CYP1A2, CYP3A4T, CYP2C8, CYP2B6, CYP2C9, and CYP2A6, as the response variables since they show a moderate correlation (see **Supplementary Figure S1**). We included age and gender as covariates in the analysis and log-transformed the six response variables.

For each individual trait, we first conducted a marginal gene-based single-trait analysis with the omnibus KBT. Then, we integrated the *p*-values for the six traits and applied the *p*-value combination method to get a gene-based multi-trait *p*-value. In the multi-trait analysis, we also applied the RMMLR and the MANONA methods. The *Q*–*Q* plot of the single-trait analysis is shown in **Supplementary Figure S2** and no *p*-value inflation was observed. **Figure 3** shows the *Q*–*Q* plot of the multi-trait analysis.

If we use the genome-wide gene-level Bonferroni correction, the threshold to claim a significant gene level significance is 4.3×10^{-6}. This leads to no significant genes in our analysis. Here, we only listed a few top genes with *p*-value less than 6×10^{-5} as suggestive significance. In the single-trait analysis, the top genes for each trait are *HAUS8* and *IRS12* for CYP1A2, *TRAPPC10* for CYP3A4T, *TARID* and *FUNDC2* for CYP2C9, and *PAPLN* for CYP2A6. No genes pass the suggested threshold for trait CYP2B6 and CYP2C8 (see **Supplementary Table S1** for a detailed list of associated genes for each trait and the corresponding *p*-values). For the multi-trait analysis, we listed in **Table 2** the results of the top genes along with the results by RMMLR and MANOVA. Among the four genes, *TARID*, *TRAPPC10*, and *HAUS8* were also in the list of single-trait analysis. Gene *ATAD3C* is not shown in the top list of the single-trait analysis. This may be due to the low power of the single-trait analysis. If we ignore the correlation information among the six enzyme traits and only focus on a single-trait analysis, we may miss some discoveries.

TABLE 2 | List of top genes and the *p*-values with different methods in the Human Liver Cohort study.

Gene name	Number of single nucleotide polymorphisms	Chr	OMGA	RMMLR	MANONA
TARID	80	6	1.11E−05	0.1227	0.1048
TRAPPC10	58	21	1.29E−05	0.0072	0.1003
HAUS8	42	19	4.22E−05	0.0425	0.1022
ATAD3C	150	1	5.53E−05	0.0789	0.0926

For the top four genes by OMGA, the *p*-values by RMMLR and MANOVA are all quite large. This could be due to the potential complex functional relationship between the genes and the traits. RMMLR and MANOVA were not designed to capture those complex relationships.

Empirical evidence supports some of the identified genes. For example, gene *ATAD3C* has been reported in literature to be associated with aldosterone metabolism and P450 enzyme (Chu et al., 2017). Gene *TARID* participates in liver cell metabolism (Yuan et al., 2016). Gene *TRAPPC10* is associated with the toxic effect of octylphenol on the expression of genes in the liver (Li et al., 2014).

Case Two: The Alzheimer Disease Neuroimaging Initiative Data Analysis

We also applied the developed OMGA method to the ADNI data set which can be accessed at http://adni.loni.usc.edu/. From the ADNI1 and ADNI2 studies, we selected 490 samples with complete genetic and phenotypic information. We deleted SNPs with MAF < 0.05 or those that could not pass the Hardy–Weinberg equilibrium test. This ended up with 620,901 SNPs.

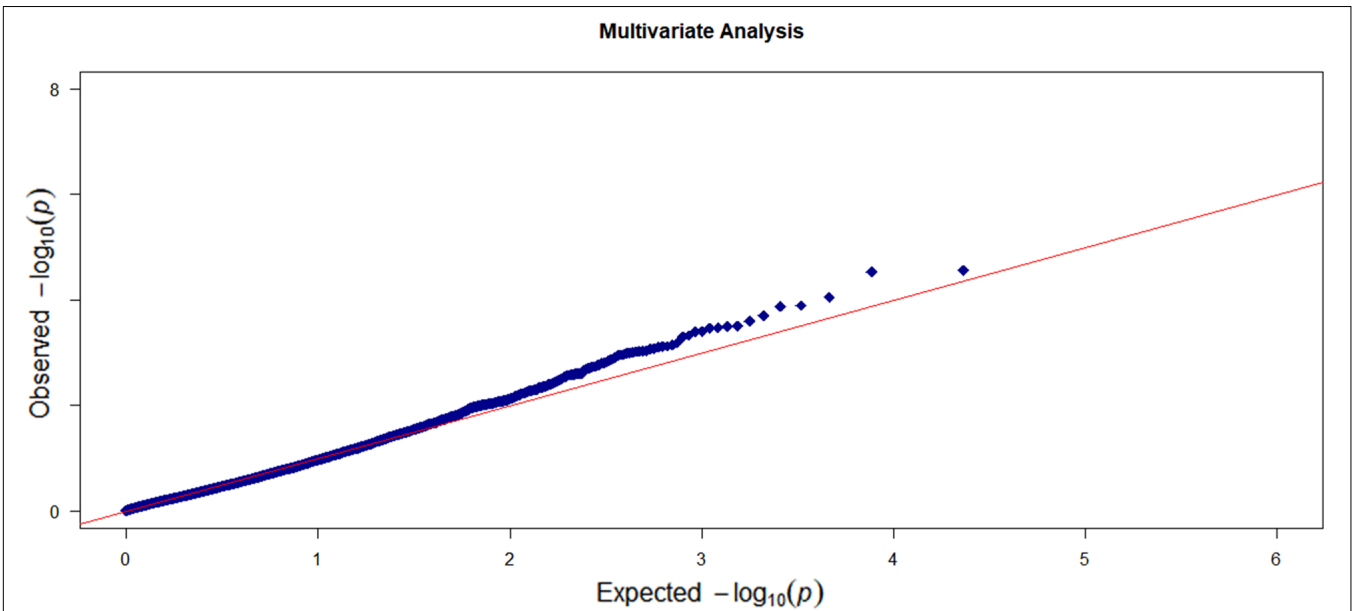

FIGURE 3 | The *Q*–*Q* plot of the observed –log10 (*p*-value) versus the expected –log10 (*p*-value) for the six enzyme traits based on the multi-trait analysis.

We included SNPs within 20 kb upstream and downstream of each gene and mapped them to 22,890 genes according to human genome version GRCh38.

Alzheimer's disease (AD) is a central nervous system degenerative disease with insidious onset and chronic progress and has affected over 5.5 million Americans, especially among the elderly over the age of 65 years. ADNI provides pre-calculated volumes of five cortical regions including entorhinal, hippocampus, ventricles, midtemp, and fusiform. Brain atrophy is a typical clinical symptom among AD patients (Ferrarini et al., 2006). Studies have pointed out that the volumes in the different cortical regions show different rates of decline and are functionally related to AD. For example, the hippocampus region helps humans to deal with memory sounds, long-term learning, and taste and is a sensitive early indicator of AD (Mu and Gage, 2011). The loss in the entorhinal region is highly correlated with the severity of AD and the loss is obvious even in mild AD patients (Juottonen et al., 1998). Similarly, the volumes in the regions fusiform and midtemp also slightly decrease in AD patients (Thambisetty et al., 2011). This motivates us to take the volumes of the five cortical regions as a multi-trait response and to identify which genes are associated with the volume variation in the different brain regions.

We first conducted the marginal single-trait analysis with the proposed gene-based omnibus kernel testing approach. We log-transformed the volumes of the five cortical regions and took the age, education level, gender, and *APOE4* alleles as the covariates. The *Q*–*Q* plot of the gene-based single-trait analysis is shown in **Supplementary Figure S3**. No sign of *p*-value inflation was observed. Also, there is no strong indication of significant signals either. Then, we carried out the multi-trait analysis which can more accurately reflect the brain atrophy in AD patients. We also applied MANOVA and RMMLR methods for multi-trait analysis. The *Q*–*Q* plot of the multi-trait analysis results by OMGA is shown in **Figure 4**. There is no significant indication of *p*-value inflation.

Again no significant genes were identified based on the genome-wide gene-level Bonferroni threshold. Here, we listed the top 12 genes based on a suggestive threshold of 5×10^{-5} in **Table 3**. From the single-trait analysis, we found eight, 10, 10, five, and six genes associated with the regions entorhinal, ventricles, hippocampus, fusiform, and midtemp, respectively (see **Supplementary Table S2** for a detailed list of the genes). Two genes (*SNORA30* and *TLR4*) that were not in the single-trait analysis list but showed up in the multi-trait analysis list are highlighted in bold font in **Table 3**. Compared to RMMLR and MANOVA analyses, the *p*-values by OMGA are uniformly smaller, indicating the power of OMGA by taking both linear and nonlinear effect into consideration.

For the 12 genes associated with multi-trait of brain atrophy in AD patients, some of them have been reported in the literature. For example, gene *RBM45*, known as the RNA-binding motif protein 45 or developmentally regulated RNA-binding protein-1 (*Drbp1*), has been shown to be associated with the degenerative neurological changes in AD patients (Eck et al., 2018). Gene *UPK1B* has been shown to be cooperated with *CD9* and *CD81* and is directly involved in the pathological process of AD (De Strooper and Wakabayashi, 2011; Orre et al., 2014; Wężyk and Żekanowski, 2017). Mutation in gene *TLR4* reduces microglial activation, increases Aβ deposits, and exacerbates cognitive deficits in a mouse model of AD (Song et al., 2011). A study showed that polymorphisms in gene *TLR4* and *CD14* were closely related to AD (Balistreri et al., 2008). Others reported the increasing expressions of *TLR2* and *TLR4* on the peripheral blood mononuclear cells of AD patients

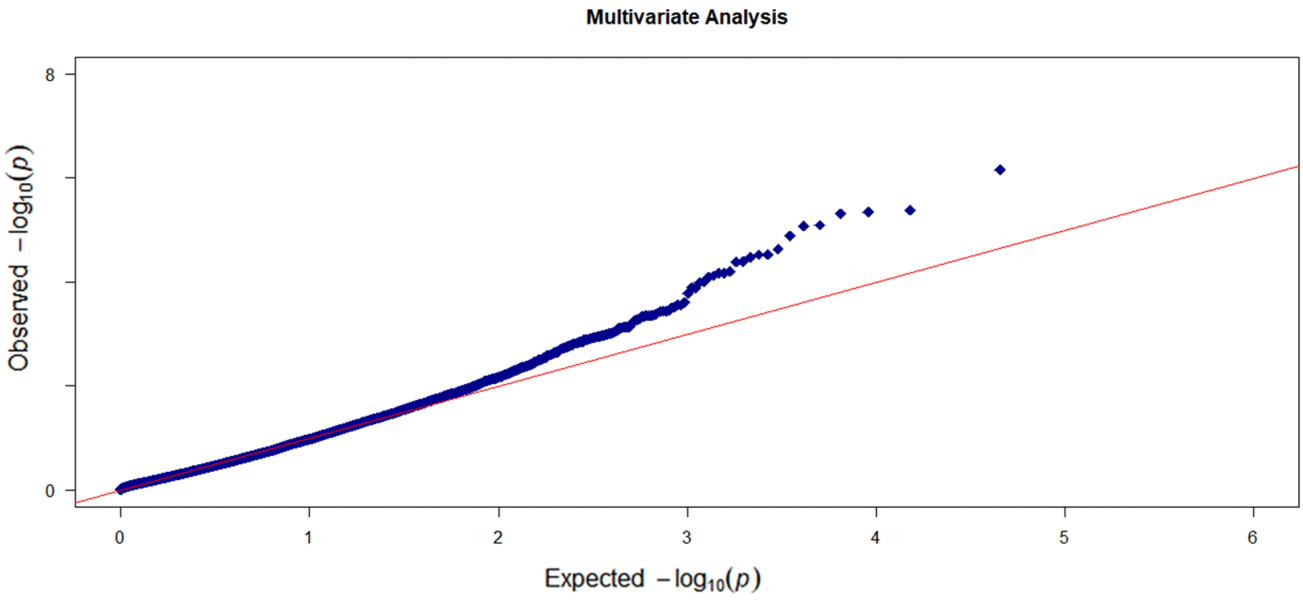

FIGURE 4 | The *Q*–*Q* plots of the observed –log10 (*p*-value) versus the expected –log10 (*p*-value) for the five cortical regions based on the multi-trait analysis.

TABLE 3 | List of top genes and the *p*-values with different methods in the Alzheimer Disease Neuroimaging Initiative study.

Gene name	Number of single nucleotide polymorphisms	Chr	OMGA	RMMLR	MANONA
TMEM26-AS1	731	10	3.45E−06	0.0004	0.2572
TPRG1-AS2	320	3	6.60E−06	0.0238	0.4595
ST3GAL4	2,457	11	8.37E−06	0.1373	0.0165
LMNTD1	89	12	9.64E−06	0.6580	0.1698
OR4F5	2,234	1	1.03E−05	0.1887	0.1364
MIR6723	170	14	1.83E−05	0.5421	0.2648
RBM45	468	2	2.25E−05	0.0017	0.0077
ADAMTS7P1	1,444	15	2.29E−05	0.0003	0.3606
SNORA30	200	16	2.30E−05	0.0213	0.0093
TLR4	153	9	3.45E−05	0.0015	0.1364
C5orf46	663	5	3.69E−05	0.1254	0.0232
UPK1B	772	3	4.10E−05	0.1855	0.0036

(Zhang et al., 2012). These empirical evidences support the results of the analysis.

DISCUSSION

Increasing evidence has shown that, for correlated phenotypes, multi-trait analysis can significantly increase the power of association analysis (e.g., He et al., 2013; Schifano et al., 2013; Wang, 2014). Given that genes are functional units in most living organisms, we proposed a rapid and powerful gene-based multi-trait analysis method. Our method is developed under the KBT framework without specific error distribution assumptions. It possesses a few advantages over existing methods. First, the method achieves fast calculation speed and decreases the computational burden for high-dimensional data. A testing *p*-value can be quickly computed with the asymptotic results, making the method computationally attractive. Second, it can capture a potential nonlinear effect within genes by using a nonparametric KBT procedure. By incorporating different kernel functions, potential linear or nonlinear genetic effects can be captured and tested. When a given series of candidate kernel functions is available, the omnibus testing procedure is robust against misspecification of kernel functions. Moreover, it is built upon the Cauchy transformation and is computationally fast (Liu and Xie, 2019). Thus, the proposed method enjoys both theoretical rigor and computational efficiency and can be widely used in gene-based analysis.

We conducted extensive simulation studies to evaluate the type I error control and the power of the proposed method. The results show that the proposed OMGA method can maintain a reasonable type 1 error rate and achieve great power compared to other popular methods such as MANOVA and RMMLR. Furthermore, the omnibus testing procedure incorporating different kernels performs as well as if the underlying true genetic function is correctly specified. Thus, the method is safe to apply in real applications regardless of the underlying disease function, making the method practically attractive.

For multi-trait analysis, there are two different frameworks proposed. One is to jointly model multiple traits as a multivariate response and further assess their association with SNP variants. This framework can directly take correlation information into consideration. Methods for such type of multi-trait analysis include the RMMLR and the MANOVA methods as discussed in this work and many others (e.g., Maity et al., 2012]. Another framework is to conduct a single-trait disease–gene association test and then combine *p*-values to assess the joint association. The method developed by Yang et al. (2016) falls into this category. Nevertheless, methods to combining *p*-values have to take the correlation information into consideration. Otherwise, the results can be biased. Ideally, the first framework should be preferable since it models multiple traits simultaneously in one joint model. On the other hand, the second framework has its advantages. For example, it can be computationally less expensive and ease theoretical evaluations. Especially with the proposed method in this work, the second framework can be a better choice since the asymptotic evaluation of the joint association statistics can be theoretically challenging or may not even be feasible.

Our method can be easily applied to a genome-wide pathway-based multi-trait analysis. It is known that genes usually do not work alone. For example, cellular pathways and complex molecular networks are often more directly involved in the progression and the susceptibility of diseases. Thus, a pathway-based analysis can shed light on the mechanics of complex diseases. On the other hand, the current study only focused on quantitative multivariate phenotypes. It can be extended to qualitative response variables or a combination of qualitative and quantitative phenotypes. However, the extension is non-trivial and will be studied in our future investigation. The R code that implements the method can be found in GitHub at https://github.com/yamin-19/OMGA.

DATA AVAILABILITY STATEMENT

The HLC dataset can be downloaded at https://www.synapse.org/#!Synapse:syn4499. The ADNI dataset can be accessed through http://adni.loni.usc.edu/.

AUTHOR CONTRIBUTIONS

YD implemented the method and drafted the manuscript. TH derived the kernel testing method. RF, SL, and HC were involved in the simulation and data analysis. YC conceived the idea, designed the study, and drafted the manuscript. All the authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00437/full#supplementary-material

REFERENCES

Aguiar, M., Masse, R., and Gibbs, B. F. (2005). Regulation of cytochrome P450 by posttranslational modification. *Drug Metab. Rev.* 37, 379–404. doi: 10.1081/dmr-200046136

Amos, C., and Laing, A. J. G. E. (1993). A comparison of univariate and multivariate tests for genetic linkage. *Genetic Epidemiol.* 10, 671–676. doi: 10.1002/gepi.1370100657

Balistreri, C., Grimaldi, M., Chiappelli, M., Licastro, F., Castiglia, L., Listì, F., et al. (2008). Association between the polymorphisms of TLR4 and CD14 genes and Alzheimer's disease. *Curr. Pharm. Design* 14, 2672–2677. doi: 10.2174/138161208786264089

Basu, S., and Pan, W. J. G. E. (2011). Comparison of statistical tests for disease association with rare variants. *Gen. Epidemiol.* 35, 606–619. doi: 10.1002/gepi.20609

Basu, S., Zhang, Y., Ray, D., Miller, M. B., Iacono, W. G., and McGue, M. J. H. H. (2013). A rapid gene-based genome-wide association test with multivariate traits. *Hum. Hered.* 76, 53–63. doi: 10.1159/000356016

Bilodeau, M. (2013). *Analysis of Variance, Multivariate (MANOVA)*. Hoboken, NJ: John Wiley & Sons.

Chu, C., Zhao, C., Zhang, Z., Wang, M., Zhang, Z., Yang, A., et al. (2017). Transcriptome analysis of primary aldosteronism in adrenal glands and controls. 10, 10009–10018.

Cui, Y., Kang, G., Sun, K., Qian, M., Romero, R., and Fu, W. J. G. (2008). Gene-centric genomewide association study via entropy. *Genetics* 179, 637–650. doi: 10.1534/genetics.107.082370

De Strooper, B., and Wakabayashi, T. (2011). *Extracellular Targets for Alzheimer's Disease. Google Patents. US20110008350A1*.

Eck, A. G., Lopez, K. J., and Henderson, J. O. J. J. (2018). RNA-binding motif protein 45 (Rbm45)/developmentally regulated RNA-binding protein-1 (Drbp1): association with neurodegenerative disorders. *J. Stud. Res.* 7, 33–37.

Ferrarini, L., Palm, W. M., Olofsen, H., van Buchem, M. A., Reiber, J. H., and Admiraal-Behloul, F. (2006). Shape differences of the brain ventricles in Alzheimer's disease. *Neuroimage* 32, 1060–1069. doi: 10.1016/j.neuroimage.2006.05.048

He, Q., Avery, C. L., and Lin, D. Y. J. G. E. (2013). A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic Epidemiol.* 37, 759–767. doi: 10.1002/gepi.21759

He, T., Li, S., Zhong, P. S., and Cui, Y. (2019). An optimal kernel-based U-statistic method for quantitative gene-set association analysis. *Gen. Epidemiol.* 43, 137–149. doi: 10.1002/gepi.22170

Jiang, C., and Zeng, Z.-B. J. G. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140, 1111–1127.

Juottonen, K., Laakso, M., Insausti, R., Lehtovirta, M., Pitkänen, A., Partanen, K., et al. (1998). Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease. *Neurobiol. Aging* 19, 15–22. doi: 10.1016/s0197-4580(98)00007-4

Kim, J., Bai, Y., and Pan, W. J. G. E. (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. *Gen. Epidemiol.* 39, 651–663. doi: 10.1002/gepi.21931

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. J. (2008). A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Gen.* 82, 386–397. doi: 10.1016/j.ajhg.2007.10.010

Li, S., and Cui, Y. (2012). Gene-centric gene-gene interactions: a model-based kernel machine method. *Ann. Appl. Stat.* 6, 1134–1161. doi: 10.1214/12-aoas545

Li, X.-Y., Xiao, N., and Zhang, Y.-H. J. E. (2014). Toxic effects of octylphenol on the expression of genes in liver identified by suppression subtractive hybridization of Rana chensinensis. *Ecotoxicology* 23, 1–10. doi: 10.1007/s10646-013-1144-z

Lin, X., Lee, S., Christiani, D. C., and Lin, X. J. B. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 14, 667–681. doi: 10.1093/biostatistics/kxt006

Littell, R. C., and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Am. Stat. Assoc.* 66, 802–806. doi: 10.1080/01621459.1971.10482347

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

Liu, W., and Xie, J. (2019). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* 114, 384–392.

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

Maity, A., Sullivan, P. F., and Tzeng, J. Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.* 36, 686–695. doi: 10.1002/gepi.21663

Marceau, R., Lu, W., Holloway, S., Sale, M. M., Worrall, B. B., Williams, S. R., et al. (2015). A fast multiple-kernel method with applications to detect gene-environment interaction. *Gen. Epidemiol.* 39, 456–468. doi: 10.1002/gepi.21909

Mu, Y., and Gage, F. H. (2011). Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol. Neurodeg.* 6:85. doi: 10.1186/1750-1326-6-85

Mukhopadhyay, I., Feingold, E., Weeks, D. E., and Thalamuthu, A. J. G. E. T. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Gen. Epidemiol.* 34, 213–221.

Neale, B. M., and Sham, P. C. J. T. (2004). The future of association studies: gene-based analysis and replication. *Am. J. Hum Gen.* 75, 353–362. doi: 10.1086/423901

Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., et al. (1996). P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Eur. PMC* 6, 1–42. doi: 10.1097/00008571-199602000-00002

Orre, M., Kamphuis, W., Osborn, L. M., Jansen, A. H., Kooijman, L., Bossers, K., et al. (2014). Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction. *Neurobiol. Aging* 35, 2746–2760. doi: 10.1016/j.neurobiolaging.2014.06.004

Plant, N. (2007). The human cytochrome P450 sub-family: transcriptional regulation, inter-individual variation and interaction networks. *Biochim. Biophys. Acta Gen. Sub.* 1770, 478–488. doi: 10.1016/j.bbagen.2006.09.024

Schifano, E. D., Li, L., Christiani, D. C., and Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *Am. J. Hum. Gen.* 92, 744–759. doi: 10.1016/j.ajhg.2013.04.004

Shang, J., Zhang, J., Lei, X., Zhao, W., and Dong, Y. J. G. (2013). EpiSIM: simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis. *Genes Genomics* 35, 305–316. doi: 10.1007/s13258-013-0081-9

Song, M., Jin, J., Lim, J.-E., Kou, J., Pattanayak, A., Rehman, J. A., et al. (2011). TLR4 mutation reduces microglial activation, increases Aβ deposits and exacerbates cognitive deficits in a mouse model of Alzheimer's disease. *J. Neuroinflam.* 8:92. doi: 10.1186/1742-2094-8-92

Stearns, F. W. J. G. (2010). One hundred years of pleiotropy: a retrospective. *Genetics* 186, 767–773. doi: 10.1534/genetics.110.122549

Thambisetty, M., Simmons, A., Hye, A., Campbell, J., Westman, E., Zhang, Y., et al. (2011). Plasma biomarkers of brain atrophy in Alzheimer's disease. *PLoS One* 6:e28527.

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

Wang, K. (2014). Testing genetic association by regressing genotype over multiple phenotypes. *PLoS One* 9:e106918. doi: 10.1371/journal.pone.0106918

Wang, K., Li, M., and Bucan, M. J. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Gen.* 81, 1278–1283. doi: 10.1086/522374

Wang, K., Li, M., and Hakonarson, H. J. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11:843. doi: 10.1038/nrg2884

Wei, C., and Lu, Q. (2017). A generalized association test based on U statistics. *Bioinformatics* 33, 1963–1971. doi: 10.1093/bioinformatics/btx103

Wężyk, M. M., and Żekanowski, C. J. (2017). "Presenilins interactome in Alzheimer disease and pathological ageing," in *Senescence: Physiology or Pathology*, Vol. 95, eds J. Dorszewska and W. Kozubski (London: InTechOpen).

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., et al. (2010). Powerful SNP set analysis for case-control genome wide association studies. *Am. J. Hum. Genet.* 86, 929–942. doi: 10.1016/j.ajhg.2010.05.002

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

Yang, J. J., Li, J., Williams, L. K., and Buu, A. (2016). An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinform.* 17:19.

Yang, Q., and Wang, Y. J. J. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *J. Probab. Stat.* 2012: 652569.

Yuan, S.-X., Zhang, J., Xu, Q.-G., Yang, Y., and Zhou, W. (2016). Long noncoding RNA, the methylation of genomic elements and their emerging crosstalk in hepatocellular carcinoma. *Cancer Lett.* 379, 239–244. doi: 10.1016/j.canlet.2015.08.008

Zhang, F., Guo, X., Wu, S., Han, J., Liu, Y., Shen, H., et al. (2012). Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. *PLoS One* 7:e53320. doi: 10.1371/journal.pone.0053320

Zhang, W., Wang, L.-Z., Yu, J.-T., Chi, Z.-F., and Tan, L. J. (2012). Increased expressions of TLR2 and TLR4 on peripheral blood mononuclear cells from patients with Alzheimer's disease. *J. Neurol. Sci.* 315, 67–71. doi: 10.1016/j.jns.2011.11.032

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 June 2020
doi: 10.3389/fgene.2020.00620

An Analysis Regarding the Association Between the *ISLR* Gene and Gastric Carcinogenesis

Shu Li[1]*, *Wei Zhao*[2] and *Manyi Sun*[3]

[1] Department of Gastroenterology and Hepatology, Tianjin Medical University General Hospital, Tianjin, China, [2] General Data Technology Co., Ltd., Tianjin, China, [3] Department of Gastroenterology, Tianjin Union Medical Center, Tianjin, China

For datasets of gastric cancer collected by TCGA (The Cancer Genome Atlas) and GEO (Gene Expression Omnibus) repositories, we applied a bioinformatics approach to obtain expression data for the *ISLR* (immunoglobulin superfamily containing leucine-rich repeat) gene, which is highly expressed in gastric cancer tissues and closely associated with clinical prognosis. Although we did not observe an overall association of *ISLR* mutation, high expression or copy number variation with survival, hypomethylation of four methylated sites (assessed by the probes cg05195566, cg17258195, cg09664357, and cg07297039) of *ISLR* was negatively correlated with high expression levels of *ISLR* and was associated with poor clinical prognosis. In addition, we detected a correlation between *ISLR* expression and the infiltration levels of several immune cells, especially $CD8^+$ T cells, macrophages and dendritic cells. We also identified a series of genes that were positively and negatively correlated with *ISLR* expression based on the TCGA-STAD, GSE13861, and GSE29272 datasets. Principal component analysis and random forest analysis were employed to further screen for six hub genes, including *ISLR*, *COL1A2*, *CDH11*, *SPARC*, *COL3A1*, and *COL1A1*, which exhibited a good ability to differentiate between tumor and normal samples. GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway and gene set enrichment analysis data also suggested a potential relationship between *ISLR* gene expression and epithelial-mesenchymal transition (EMT). *ISLR* expression was negatively correlated with sensitivity to PX-12 and NSC632839. Taken together, these results show that the *ISLR* gene is involved in gastric carcinogenesis, and the underlying molecular mechanisms may include DNA methylation, EMT, and immune cell infiltration.

Keywords: *ISLR*, expression, methylation, immune cell infiltration, gastric cancer

OPEN ACCESS

Edited by:
Min Zhang,
Purdue University, United States

Reviewed by:
Jian Li,
Tulane University, United States
Jun Li,
University of Notre Dame, United States

***Correspondence:**
Shu Li
tj_lishu@163.com

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

Received: *12 December 2019*
Accepted: *22 May 2020*
Published: *16 June 2020*

Citation:
Li S, Zhao W and Sun M (2020) An Analysis Regarding the Association Between the ISLR Gene and Gastric Carcinogenesis. Front. Genet. 11:620. doi: 10.3389/fgene.2020.00620

INTRODUCTION

The Cancer Genome Atlas (TCGA), a publicly funded project, archives multiple types of genomic data from various types of cancer, including gene expression, mutation, copy number variation (CNV), genome methylation, and clinical data (Cancer Genome Atlas Research Network, 2014; Tomczak et al., 2015; Wang et al., 2016). In addition, GEO (Gene Expression Omnibus) molecular datasets also offer many clinical cancer-related gene expression data (Barrett et al., 2013; Clough and Barrett, 2016). The complicated pathogenesis of gastric cancer involves multiple clinical prognosis-associated

oncogenes. Previously, based on the datasets of gastric cancer within TCGA and GEO, we identified the *ISLR* (immunoglobulin superfamily containing leucine-rich repeat) gene by means of principal component analysis (PCA) and random forest analysis (data not shown), which showed a high expression level in gastric cancer tissues and was closely linked to clinical prognosis. The present study attempted to investigate the possible oncogenic roles of the *ISLR* gene in the pathogenesis and prognosis of gastric cancer.

The human *ISLR* gene is situated on human chromosome 15q23-q24 (Nagasawa et al., 1997). The human ISLR protein, a member of the Ig superfamily, contains a leucine-rich repeat (LRR) with conserved flanking sequences and a C2-type immunoglobulin (Ig)-like domain (Nagasawa et al., 1997). The ISLR protein has been reported to be involved in some biological events, such as cell replicative senescence of human dermal fibroblasts (Yoon et al., 2004), embryo development (Homma et al., 2009), and Gaucher disease (Lugowska et al., 2019). However, no study has mainly investigated the potential functional relationship between the *ISLR* gene and cancer events thus far.

In the current study, we elucidated the underlying molecular mechanisms of the *ISLR* gene in gastric carcinogenesis from the perspectives of genetic mutation, copy number variation, DNA methylation, immune cell infiltration, expression correlation, pathway enrichment and drug sensitivity for the first time.

MATERIALS AND METHODS

Expression Analysis

We first investigated the expression level of the *ISLR* gene between gastric cancer and negative controls samples within the TCGA-STAD (The Cancer Genome Atlas stomach adenocarcinoma) cohort and the GTEx (Genotype Tissue Expression) database using the online tool GEPIA2[1] (Tang et al., 2019). A log2 (FC) (fold change) cutoff = 1, a *P*-value cutoff = 0.01, and a jitter size = 4 were set. Log_2 [TPM (transcripts per million) + 1] values were used for log-scale. Gene expression data were visualized by the "boxplot" function of the R language (for the cancer and control samples) or the "vioplot" R package [for the pathological stage (stages I, II, III, and IV)]. Then, we obtained the expression dataset of "Chen et al. (2003)," which contains a total of 11 diffuse gastric adenocarcinoma and 24 normal control samples, by means of Oncomine[2]. The log2 (median-centered intensity) data were visualized by GraphPad Prism software, version 5.01 (San Diego).

Furthermore, we utilized the "GEOquery" R package to obtain the available expression and group datasets in GSE13861 and GSE29272. The difference in expression of the *ISLR* gene between gastric cancer cases and normal controls was analyzed by the t.test function of the compare_means () and visualized by the ggviolin () function of the "ggpubr" R package. We then used the wilcox.test function of the compare_means () with the setting of "paired = TRUE" to analyze the difference in expression between the gastric tumor tissues and adjacent normal tissues and displayed the results using the ggdotchart () function of the "ggpubr" R package. R language software [R-3.6.1, 64-bit][3] was used.

Survival Curve Analysis

We conducted OS (overall survival) and DFS (disease-free survival) analyses of gastric cancer cases in the TCGA-STAD cohort according to the expression status of the *ISLR* gene through GEPIA2. A group cutoff of "quartile" was set, and the Kaplan–Meier curve was plotted. We also pooled the gastric cancer cases in the GSE14210, GSE15459, GSE22377, GSE29272, GSE51105, and GSE62254 datasets for the OS, FP (first progression), and PPS (post progression survival) analyses using the Kaplan–Meier plotter tool (Szasz et al., 2016). The automatically selected best cutoff was used. We considered clinical factors including sex (female or male), pathologic stage (stages 1∼4, T2∼4, N0∼3, M0∼1), HER2 status (negative or positive), Lauren classification (intestinal, diffuse, or mixed), differentiation (poor, moderate, or well), and treatment (surgery alone, 5-Fu-based adjuvant or other adjuvant). Furthermore, we employed the Coxph (Cox proportional hazard) model to determine the correlation between *ISLR* expression and the clinical prognosis of gastric cancer cases in TCGA-STAD through the web-based tool TIMER (Tumor Immune Estimation Resource) (Li et al., 2016, 2017). Clinical factors, including age, sex, race, stage, and tumor purity, were included in the Coxph model.

Genetic Alteration Analysis

The alteration frequency of the *ISLR* gene in several studies of gastric cancer, including the TCGA pub (2014), PanCan 2018 (Ellrott et al., 2018; Gao et al., 2018; Hoadley et al., 2018; Liu J. et al., 2018; Sanchez-Vega et al., 2018; Taylor et al., 2018; Bhandari et al., 2019), TCGA cohort, Pfizer and UHK (Wang et al., 2014), UHK (Wang et al., 2011), and U Tokyo (Kakiuchi et al., 2014) studies, was analyzed via the cBioPortal database[4]. We provided data of genomic alteration type, mutation site profile, OS and D/PFS (disease/progression-free survival) analyses. In addition, we generated a MEXPRESS plot (Koch et al., 2015, 2019) to analyze the CNV types of the *ISLR* gene. The correlation between CNV and the expression level of *ISLR* was also analyzed by Pearson's test. The overall survival analysis according to the CNV status of the *ISLR* gene (masked CNV $\geq$ or < -0.019) was performed through UCSC Xena[5]. The log-rank test was performed.

DNA Methylation Analysis

We analyzed the methylation status of *ISLR* DNA in the gastric cases in the TCGA-STAD cohort through MEXPRESS (Koch et al., 2015, 2019). Pearson's test was used to determine the correlation between methylation

[1]http://gepia2.cancer-pku.cn/#analysis

[2]https://www.oncomine.org/resource/main.html

[3]https://www.r-project.org/

[4]https://www.cbioportal.org/

[5]https://xenabrowser.net/

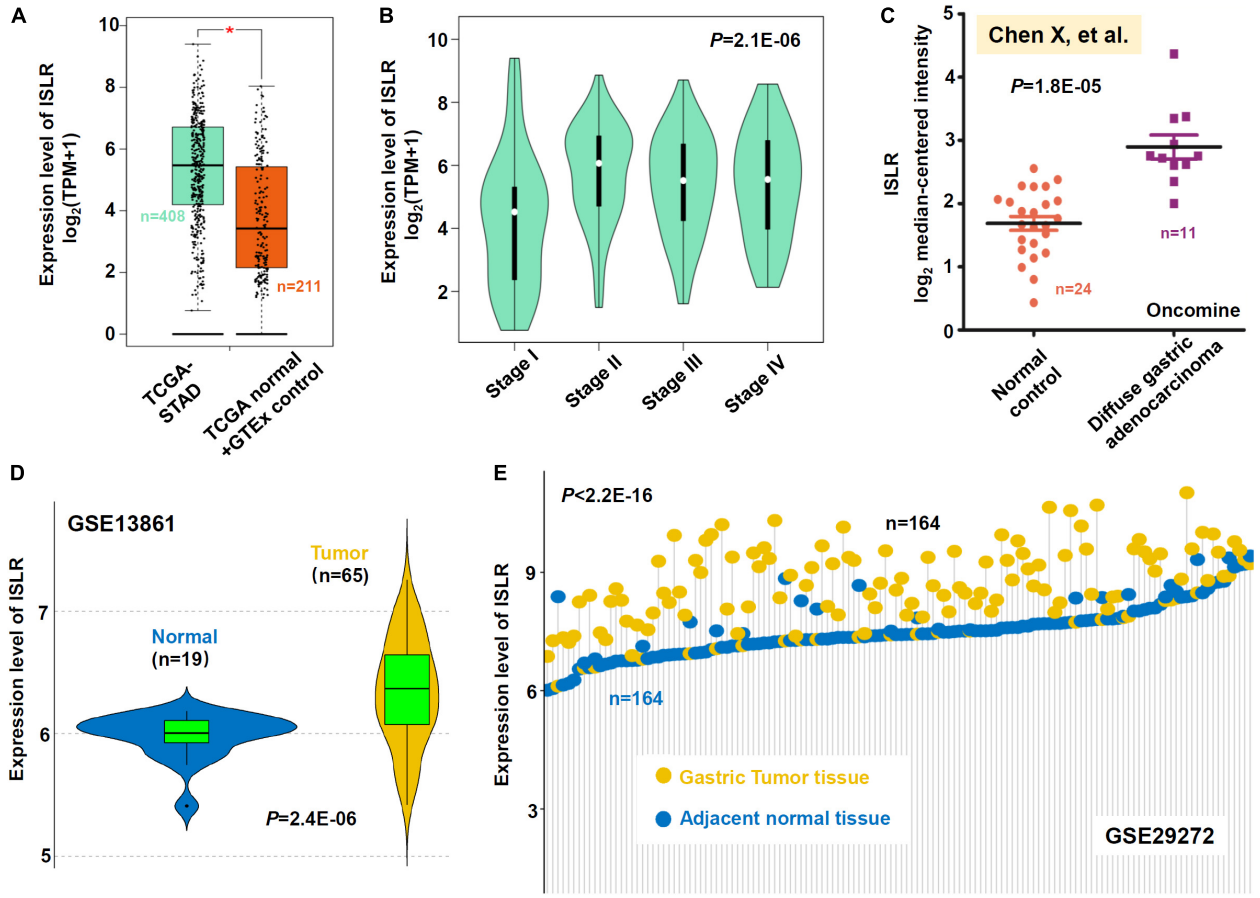

FIGURE 1 | Expression analysis of the *ISLR* gene between gastric cancer and negative control samples. **(A)** Expression level of *ISLR* between gastric cancer tissues in the TCGA-STAD cohort (n = 408) and control tissues (n = 211). Normal tissues in both the TCGA and GTEx databases were included as negative controls. *$P < 0.05$. **(B)** The expression level of *ISLR* among different pathologic stages (stage I, II, III, and IV) was also analyzed through GEPIA2. We also compared the expression level of *ISLR* between diffuse gastric adenocarcinoma cases and normal controls from the studies of "Chen et al. (2003)" through the Oncomine database **(C)**. The differences in *ISLR* gene expression between gastric cancer cases and normal controls in GSE13861 **(D)** and GSE29272 **(E)** were analyzed as well. Violin plots were used for GSE13861. Adjacent normal tissues were used for GSE29272, and the data were visualized by a dot plot.

and the expression level of the *ISLR* gene. We determined correlation coefficients (R) and Benjamini–Hochberg-adjusted *P*-values regarding different methylation probes, such as cg05195566, cg15480336, cg02077702, and cg16926502. The waterfall plot of the methylation level of *the ISLR* gene and Kaplan–Meir plots of the relationship between *ISLR* DNA hypermethylation/hypomethylation and cancer survival were generated with the MethSurv tool (Modhukur et al., 2018).

Immune Cell Infiltration Analysis

We used GEPIA2 to perform pairwise gene correlation analysis between *ISLR* expression and the signatures of the following immune cells: macrophages, TAMs (tumor-associated macrophages), dendritic cells, monocytes, NK (natural killer) cells; mast cells, neutrophils, eosinophils, basophils, B cells, Th1 cells, Th2 cells, Th17 cells, $CD8^+$ T cells, Tfh (follicular helper T) cells, resting Treg cells, effector Treg cells, and exhausted T cells. Then, based on a TIMER2 approach, we calculated immune infiltration estimations for TCGA-STAD samples with the TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC algorithms. A heatmap with the purity-adjusted Spearman's rho value was obtained by the "pheatmap" R package. Specific scatter plots were provided. In addition, the correlation between *ISLR* SNVs and the level of infiltrating immune cells, including dendritic cells, neutrophils, $CD8^+$ T cells, $CD4^+$ T cells, B cells, and macrophages, was also investigated by the TIMER tool.

ISLR-Correlated Gene Cluster Analysis

We utilized the "TCGAbiolinks" R package to download the gene expression and clinical information data of TCGA-STAD cohorts from the TCGA database. Log_2 [FPKM (Fragments per Kilobase Million) + 1] values were used for log-scale. The 25/75% quartile cutoff of *ISLR* expression in three datasets, including TCGA-STAD, GSE13861, and GSE29272, was used to define high and low groups of *ISLR* expression. We then analyzed the *ISLR*-correlated genes through the "limma" R package. The positively or negatively correlated significant genes

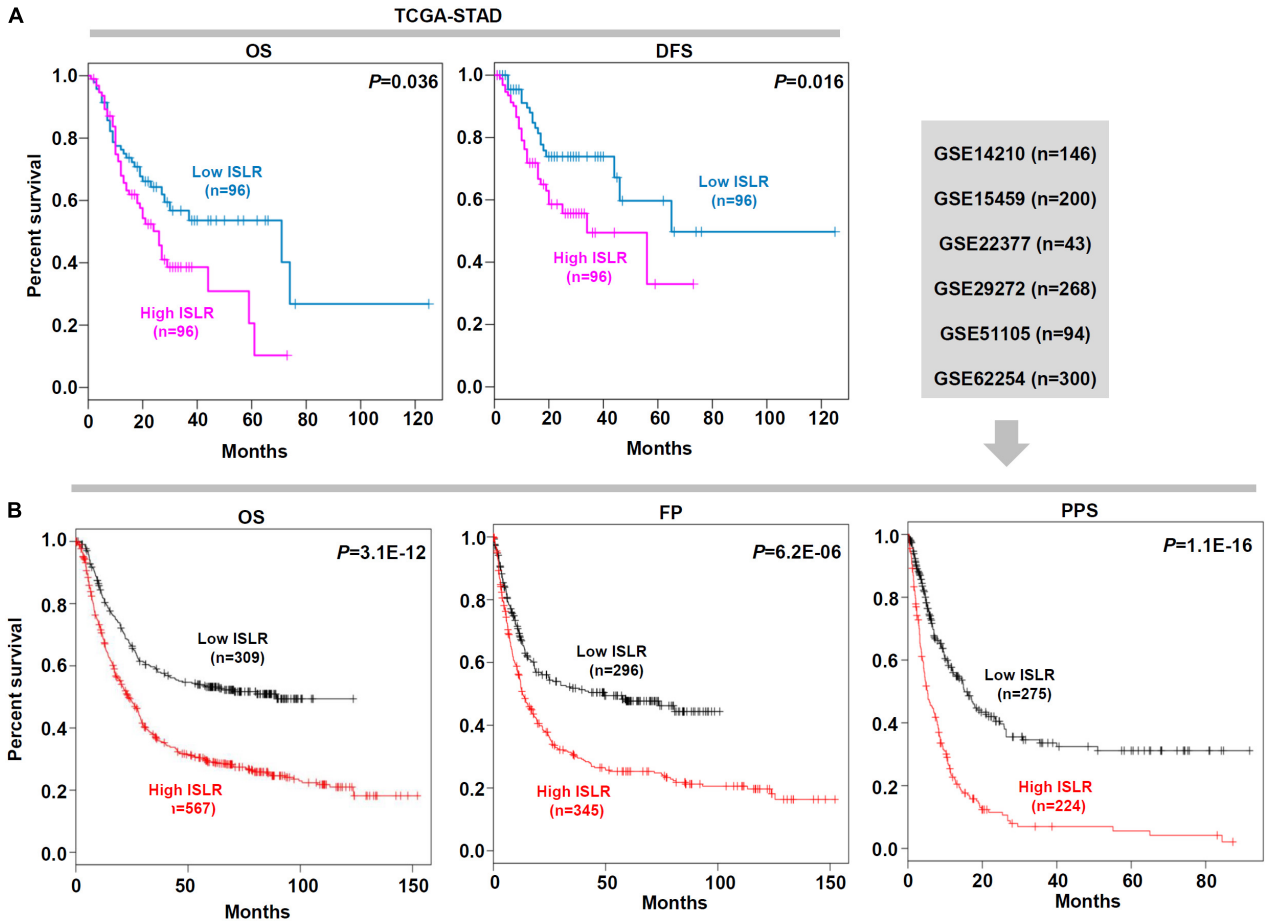

FIGURE 2 | Survival curve analysis of the *ISLR* gene for gastric cancer cases. **(A)** Overall survival (OS) and disease-free survival (DFS) analyses according to the expression level of the *ISLR* gene were performed using gastric cancer cases in the TCGA-STAD cohort. **(B)** Based on the data of gastric cancer cases in GSE14210, GSE15459, GSE22377, GSE29272, GSE51105, and GSE62254, we also performed OS, first progression (FP), and post progression survival (PPS) analyses through Kaplan–Meier plotter.

were visualized by the "ggplot2" R package. The "VennDiagram" R package was used to identify the common genes among TCGA-STAD, GSE13861, and GSE29272. Furthermore, the "clusterProfiler" and "pathview" R packages were used for the GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analyses. The data were visualized by the functions cnetplot () and dotplot (). The GOCircle and chord plots using extracellular matrix-associated terms were visualized by the "GOplot" R package.

In addition, we performed *ISLR*-correlated GSEA (gene set enrichment analysis) and pathway activation/inhibition analyses through a LinkedOmics approach (Vasaikar et al., 2018). The following cutoffs were used: simulations = 500, minimum number of genes = 3, and rank criteria = FDR (false discovery rate). The pathway activity module presents the difference in *ISLR* expression between pathway activity groups (activation and inhibition) defined by pathway scores. The pathway activity module presents the difference in gene expression between pathway activity groups (activation and inhibition) defined by pathway scores.

Principal Component Analysis

Based on the above common differentially expressed genes, we used the prcomp () function for principal component analysis (PCA) to classify the normal and tumor sections in the TCGA-STAD, GSE13861, and GSE29272 datasets. A scree plot was obtained by the plot () function, and a three-dimensional map [principal component 1 (PC1), PC2, and PC3] was drawn using the "scatterplot3d" package.

After using the "VennDiagram" R package, common hub genes among TCGA-STAD, GSE13861, and GSE29272 were identified. Then, the cor () function and "corrplot" R package were used for the Spearman correlation analysis of these hub genes. The scatter plots were then obtained by the "ggpubr" R package. The "factoextra" R package was utilized to show the principal component weight and to generate two-dimensional contribution maps of common hub genes.

Random Forest Analysis

Based on the above hub genes, we used the "randomForest" package (ntree = 500) to perform random forest modeling. The

TABLE 1 | Correlation of ISLR expression and the overall survival of gastric cancer patients in the GEO cohort (Kaplan–Meier plotter).

Factor	Group	Sample size	HR	95% CI	*logRank_P*
Gender	Female	244	2.4	1.57–3.68	**3.2E-05**
	Male	567	2.02	1.58–2.59	**1.2E-08**
Stage	Stage 1	69	2.7	0.9–8.07	0.065
	Stage 2	145	2.04	1.11–3.74	**0.019**
	Stage 3	319	2.72	1.84–4.03	**1.9E-07**
	Stage 4	152	1.82	1.22–2.71	**0.0029**
Stage T	T2	253	1.7	1.12–2.6	**0.013**
	T3	208	1.93	1.33–2.95	**6.0E-04**
	T4	39	2.66	1.06–6.68	**0.032**
Stage N	N0	76	2.95	1.13–7.69	**0.021**
	N1	232	2.74	1.78–4.21	**1.8E-06**
	N2	129	3.15	1.88–5.27	**4.9E-06**
	N3	76	2.08	1.18–3.67	**0.0097**
	N1 + 2 + 3	437	2.22	1.69–2.90	**2.9E-09**
Stage M	M0	459	2.67	1.44–4.96	**0.0012**
	M1	58	2.12	1.59–2.82	**1.6E-07**
HER2	Negative	641	2.09	1.61–2.72	**1.3E-08**
	Positive	425	1.74	1.26–2.39	**0.00058**
Lauren classification	Intestinal	336	2.85	2.00–4.08	**1.7E-09**
	Diffuse	248	1.94	1.36–2.77	**0.00018**
	Mixed	33	3.31	1.17–9.42	**0.018**
Differentiation	Poor	166	1.31	0.86–2.01	0.21
	Moderate	67	1.75	0.92–3.34	0.086
	Well	32	5.97	2.25–15.85	**6.1E-05**
Treatment	Surgery alone	393	1.59	1.19–2.13	**0.0017**
	5-Fu based adjuvant	158	0.63	0.44–0.91	**0.013**
	Other adjuvant	80	3.13	1.30–7.52	**0.0072**

HR, hazard ratio; CI, confidence interval; HER2, Erb-B2 Receptor Tyrosine Kinase 2. Bold values mean P < 0.05.

MDSplot () function was used to obtain a multidimensional scale. The mean decrease accuracy and mean decrease Gini values were calculated by the ggdotchart () function in the "ggpubr" package. Using the "pROC" package, the receiver operating characteristic (ROC) curve was plotted, and the area under the ROC curve (AUC) value was calculated.

TABLE 2 | Correlation of ISLR expression and the clinical prognosis of gastric cancer patients in the TCGA-STAD cohort (Cox proportional hazard model).

Factor	HR	95% CI_up	95% CI_down	*Cox_P*
ISLR	1.161	1.026	1.314	**0.018**
Purity	0.638	0.304	1.338	0.234
Age	1.032	1.011	1.052	**0.002**
Gender (male)	1.133	0.754	1.702	0.549
Race (Black)	1.619	0.657	3.992	0.295
Race (White)	1.109	0.681	1.806	0.679
Clinical stage2	1.471	0.679	3.188	0.328
Clinical stage3	2.428	1.190	4.954	**0.015**
Clinical stage4	3.813	1.422	10.223	**0.008**

HR, hazard ratio; CI, confidence interval. Bold values mean P < 0.05.

Drug Sensitivity Analysis

The correlation between *ISLR* and sensitivity to small molecules and/or drugs was investigated using the GSCALite tool (Liu C. J. et al., 2018). Drug sensitivity and gene expression profiling data of cancer cell lines in the Cancer Therapeutics Response Portal (CTRP) were integrated for investigation (Rees et al., 2016; Liu C. J. et al., 2018). The correlation of *ISLR* gene expression with the small molecule/drug sensitivity (half-inhibitory concentration, IC50) was determined through a Spearman correlation analysis.

RESULTS

Expression Analysis Data

First, the difference in *ISLR* gene expression between gastric cancer tissues and negative control tissues was measured. A total of 408 gastric cancer tissue samples in the TCGA-STAD cohort were included, and the adjacent tissues within TCGA-STAD and normal tissues in the GTEx database were included as negative controls (n = 211). As shown in **Figure 1A**, there was high expression of *ISLR* in the gastric cancer tumor samples ($^*P < 0.05$) compared with the controls. We further analyzed the difference in *ISLR* gene expression among different pathological stages of gastric cancer cases in the TCGA-STAD cohort and

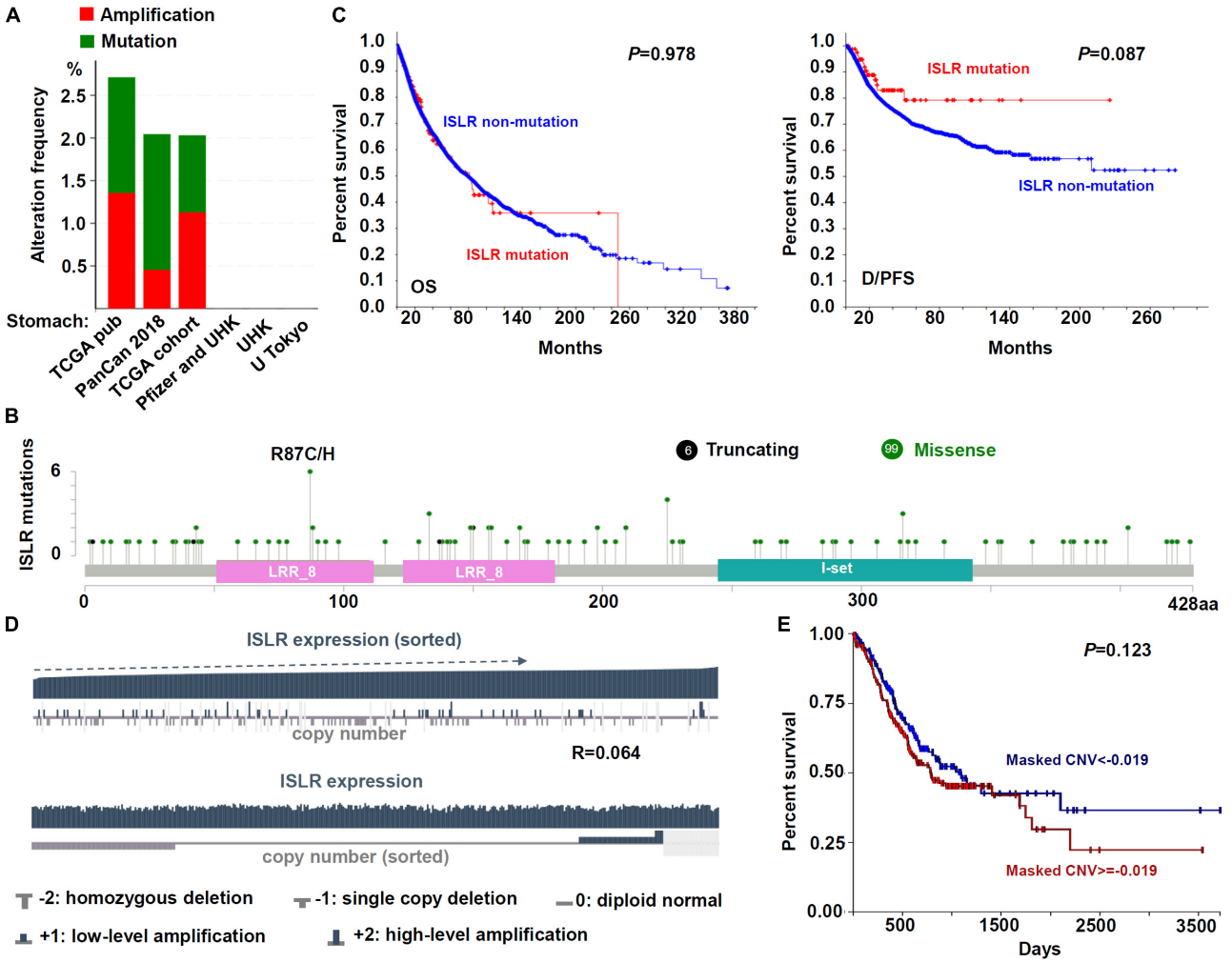

FIGURE 3 | Mutation and CNV analyses of the *ISLR* gene. **(A)** The alteration frequency of the *ISLR* gene for the cases in several studies as analyzed through cBioPortal. **(B)** The mutation site profile of the *ISLR* gene is shown. **(C)** OS and D/PFS analyses according to the mutation status of the *ISLR* gene were performed. **(D)** Correlation between copy number variation and expression of *ISLR*. **(E)** OS analysis according to the CNV status of the *ISLR* gene was performed.

identified a positive correlation (**Figure 1B**, P = 2.1E-06). Then, based on the dataset reported by Chen et al. (2003), we observed that the expression level of the *ISLR* gene in 11 diffuse gastric adenocarcinoma cases was higher than that in 24 normal controls (**Figure 1C**, P = 1.8E-05). A similar expression difference between tumor and normal samples was detected in the GSE13861 dataset (**Figure 1D**, P = 2.4E-06). Moreover, we observed an obvious high expression level of *ISLR* in 164 gastric tumor tissues compared with 164 adjacent normal tissues within the GSE29272 dataset (**Figure 1E**, P < 2.2E-16). Collectively, these results indicated that the expression level of the *ISLR* gene in gastric cancer cases was higher than that in negative controls, which suggests the potential role of the *ISLR* gene in the etiology of gastric cancer.

Survival Curve Analysis Data

Next, we explored the correlation between *ISLR* expression patterns and clinical prognosis for gastric cancer cases in the TCGA-STAD cohort. As shown in **Figure 2A**, we observed lower rates of overall survival (P = 0.036) and disease-free survival (P = 0.016) in the high *ISLR* expression group than in the low *ISLR* expression group. We also pooled a total of six GSEA datasets for the clinical prognosis analyses. As shown in **Figure 2B**, there were lower overall survival (P = 3.1E-12), first progression (P = 6.2E-06), and post progression survival (P = 1.1E-16) rates in the *ISLR* high expression group than in the low expression group. Additionally, we fully considered the effect of different clinical factors (e.g., sex, pathologic stage, HER2 status, Lauren classification, differentiation and treatment) during the above analyses. Survival curve analyses were carried out when grouping the samples by the different clinical factors. As shown in **Table 1**, there was a relationship between high *ISLR* expression and poor overall survival (hazard ratio, HR > 1, P < 0.05) in most subgroups but not in the subgroups with poor (P = 0.21) or moderate (P = 0.086) differentiation, or stage 1 disease (P = 0.065). Surprisingly, for the 158 gastric cancer cases treated with 5-Fu-based adjuvant therapy, a high level of

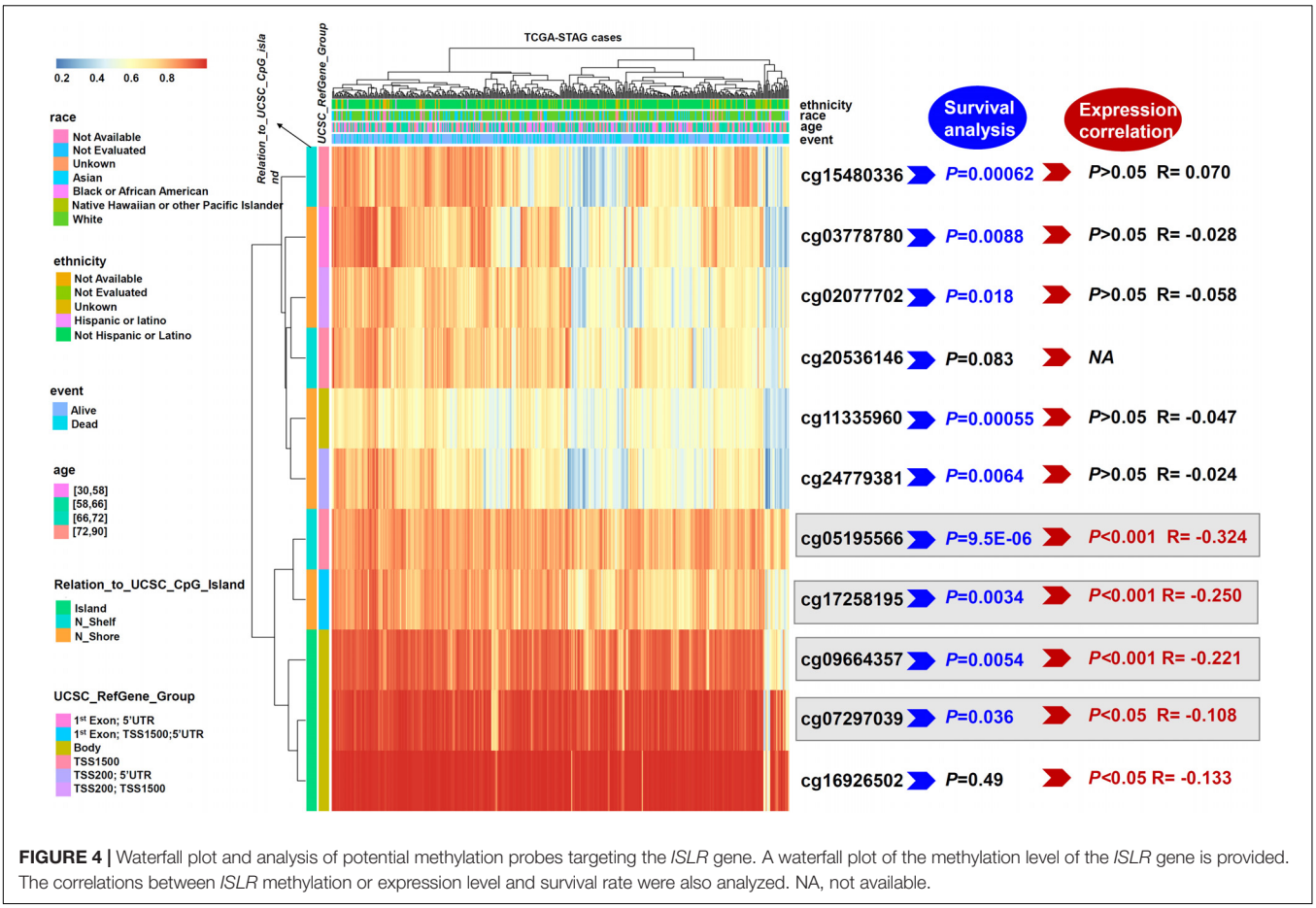

FIGURE 4 | Waterfall plot and analysis of potential methylation probes targeting the *ISLR* gene. A waterfall plot of the methylation level of the *ISLR* gene is provided. The correlations between *ISLR* methylation or expression level and survival rate were also analyzed. NA, not available.

ISLR expression was linked to a better clinical prognosis than a low level of *ISLR* expression (**Table 1**, HR = 0.63, $P = 0.013$), indicating a possible connection of *ISLR* expression with drug sensitivity. We observed similar results in the correlation analysis of *ISLR* expression and first progression and post-progression survival (**Supplementary Tables S1, S2**). Moreover, we included the factors of tumor purity, age, sex, race, clinical stage, and *ISLR* expression in a Cox proportional hazard model and obtained a statistical correlation between high *ISLR* expression and poor clinical prognosis (**Table 2**, $P = 0.018$). These findings offer evidence regarding the relationship between *ISLR* expression and clinical outcomes. This led us to perform a more in-depth molecular mechanism study.

Genetic Alteration Analysis Data

We attempted to study the potential mechanism of the *ISLR* gene in the pathogenesis of gastric cancer in terms of gene mutation and copy number variation. As shown in **Figure 3A**, we detected the mutation frequency in six groups of gastric cancer cases through the cBioPortal database. There was a low mutation rate (∼2%) of *ISLR* in the cases in the TCGA-STAD, TCGA pub (2014), and PanCan 2018 (Ellrott et al., 2018; Gao et al., 2018; Hoadley et al., 2018; Liu J. et al., 2018; Sanchez-Vega et al., 2018; Taylor et al., 2018; Bhandari et al., 2019) cohorts and no mutation in gastric cancer cases in the Pfizer and UHK (Wang et al., 2014), UHK (Wang et al., 2011), and U Tokyo (Kakiuchi et al., 2014) cohorts. The type and location of specific mutations, with the most frequent missense mutation being R87C/H ($n = 6$), are shown in **Figure 3B**. Additionally, we did not observe a statistically significant correlation between *ISLR* gene mutation and the OS rate (**Figure 3C**, $P = 0.978$) or the D/PFS rate (**Figure 3C**, $P = 0.087$).

Next, we investigated the CNV status of the *ISLR* gene. As shown in **Figure 3D**, the *ISLR* gene mainly exhibited two kinds of CNVs, namely, single copy deletion and low-level amplification. However, there was no statistically significant association between *ISLR* CNV and gene expression (**Figure 3D**, $R = 0.064$) or the overall survival rate of gastric cancer cases (**Figure 3E**, $P = 0.123$). These results suggested that *ISLR* gene mutation and copy number variation may not affect gastric tumorigenesis.

DNA Methylation Analysis Data

Next, we aimed to investigate whether the *ISLR* gene was closely linked to *ISLR* DNA methylation. Based on methylation data from TCGA-STAD, we observed that the methylation values from four methylation probes, cg05195566, cg17258195, cg09664357, and cg07297039, were negatively correlated with the expression level of the *ISLR* gene (**Figure 4**, $P < 0.05$). **Supplementary Figure S1** presents the specific information of methylation

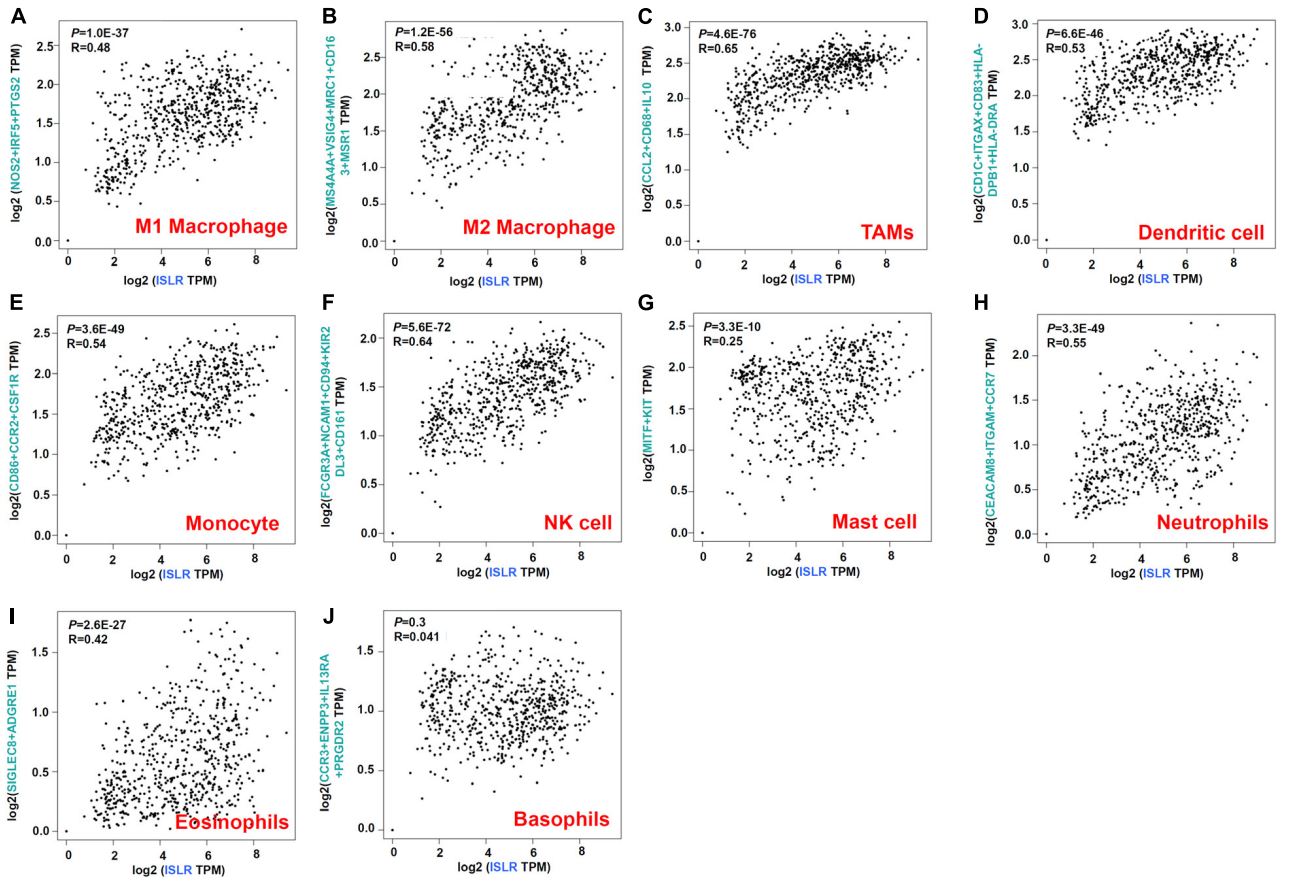

FIGURE 5 | Correlation between *ISLR* expression and markers of immune cells as analyzed through GEPIA2. **(A)** M1 macrophages; **(B)** M2 macrophages; **(C)** TAMs; **(D)** dendritic cells; **(E)** monocytes; **(F)** NK cells; **(G)** mast cells; **(H)** neutrophils; **(I)** eosinophils; **(J)** basophils.

probe sites and the correlation results of *ISLR* gene expression with methylation level. Additionally, some methylation probes showed a correlation between *ISLR* hypomethylation and poor overall survival in gastric cancer (**Figure 4** and **Supplementary Figure S2**, cg05195566, P = 9.5E-06; cg17258195, P = 0.0034; cg09664357, P = 0.0054; cg07297039, P = 0.036).

Immune Cell Infiltration Analysis Data

Herein, we sought to explore possible molecular mechanisms through immune cell infiltration during the etiology of gastric cancer. First, through GEPIA2, we analyzed the association between *ISLR* gene expression and immune cell infiltration status. As shown in **Figure 5**, we observed a positive correlation between *ISLR* expression and the marker genes of M1 macrophages (R = 0.48, P = 1.0E-37), M2 macrophages (R = 0.58, P = 1.2E-56), TAMs (R = 0.65, P = 4.6E-76), dendritic cells (R = 0.53, P = 6.6E-46), monocytes (R = 0.54, P = 3.6E-49), NK cells (R = 0.64, P = 5.6E-72), mast cells (R = 0.25, P = 3.3E-10), neutrophils (R = 0.55, P = 3.3E-49), and eosinophils (R = 0.42, P = 2.6E-27) but not between *ISLR* expression and basophils (R = 0.041, P = 0.3). We observed similar results for the different types of T and B cells, such as Tfh cells (R = 0.56, P = 2.7E-52) (R = 0.6, P = 4.0E-61), and exhausted T cells (**Supplementary Figure S3**).

Then, we utilized the TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, and EPIC algorithms for further immune infiltration estimations. As shown in **Figure 6**, we observed a relatively obvious correlation between *ISLR* expression and the immune infiltration levels of $CD8^+$ T cells, monocytes, macrophages (especially the M2 type), activated mast cells and dendritic cells when adjusted by tumor purity.

Additionally, we detected the correlation between *ISLR* CNV and the overall infiltration level of immune cells (**Supplementary Figure S4**). The copy deletion type of *ISLR* CNV was correlated with the infiltration level of dendritic cells, neutrophils, $CD8^+$ T cells, $CD4^+$ T cells, B cells, and macrophages (all $P < 0.05$), while the low-level amplification CNV was only associated with the infiltration level of dendritic cells ($P < 0.001$), neutrophils ($P < 0.01$), and $CD8^+$ T cells ($P < 0.001$).

Cluster Analysis Data

Based on the "limma" R package, we obtained genes positively or negatively correlated with *ISLR* among three datasets: TCGA-STAD, GSE13861, and GSE29272 (**Figure 7A**). Then, we performed intersection analysis and identified 134 common positively correlated genes and 8 common negatively correlated

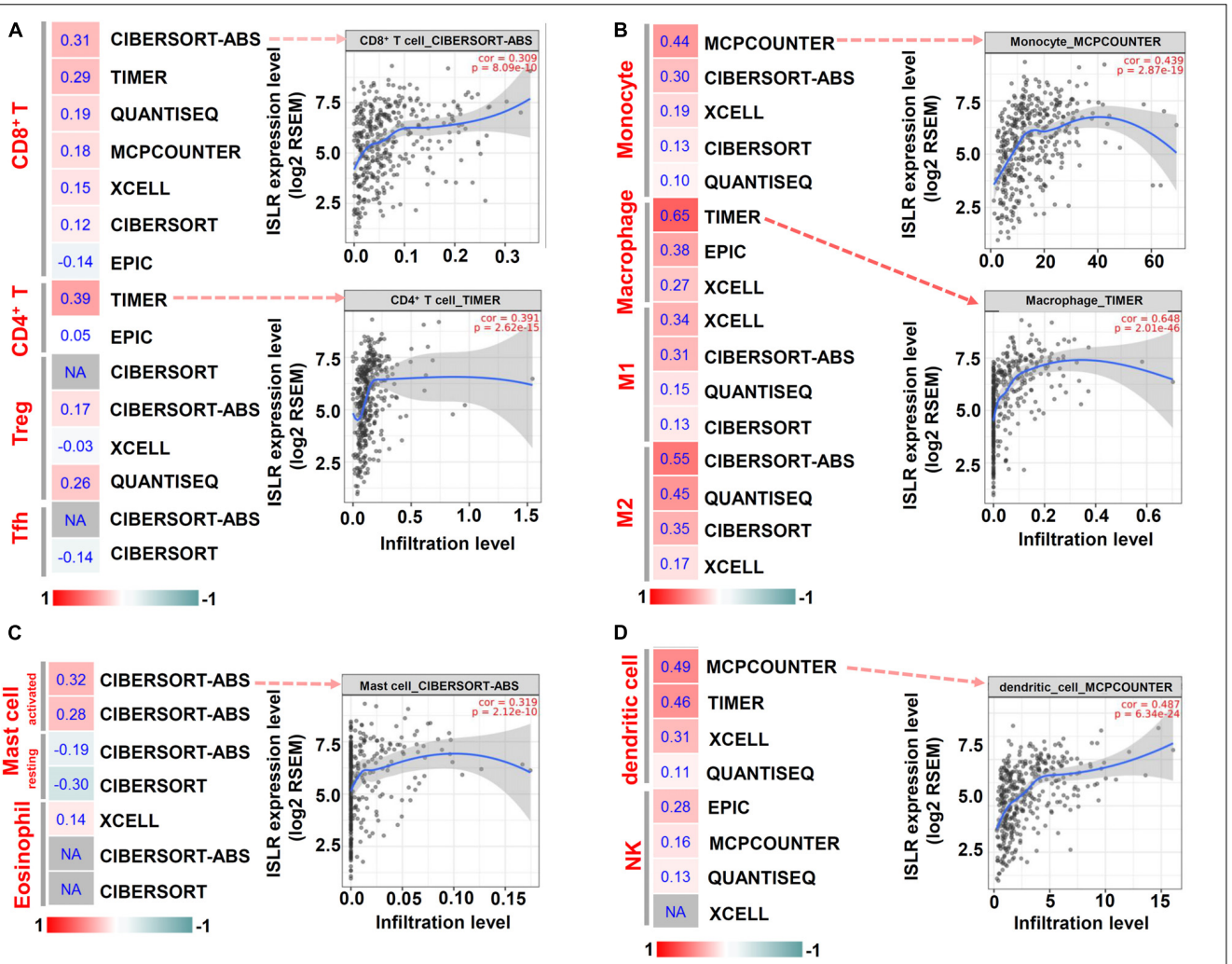

FIGURE 6 | Correlation between *ISLR* expression and the infiltration level of immune cells. The TIMER, CIBERSORT, CIBERSORT.ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC algorithms were applied for the immune infiltration estimations of CD8$^+$ T cells, CD4$^+$ T cells, Treg cells, and Tfh cells **(A)**; Monocytes and macrophages (M1, M2 type) **(B)**; mast cells and eosinophils **(C)**; and dendritic cells and NK cells **(D)**. A heatmap with the purity-adjusted Spearman's rho value and specific scatter plots are provided as examples. "NA" means a lack of an association between *ISLR* expression and the infiltration level of immune cells.

genes (**Figure 7B**). Then, we performed GO enrichment analyses. We observed extracellular matrix-associated terms, such as extracellular structure organization and extracellular matrix structural constituents, in the GO_biological_process (**Figure 7C**), GO_cellular_component (**Supplementary Figure S5A**), and GO_molecular_function (**Supplementary Figure S5B**) categories. Then, we displayed the extracellular matrix-associated terms in GOCircle (**Figure 7D**) and chord (**Figure 7E**) plots.

KEGG analysis data identified the ECM-receptor interaction (**Supplementary Figure S6**). GSEA data also showed the extracellular matrix (ECM)-associated gene sets, including extracellular structure organization, extracellular matrix structural constituents, ECM-receptor interaction, miRNA targets in ECM and membrane receptors (**Supplementary Figure S7**). Based on the GSCALite pathway score analysis, we further observed activation of the epithelial-mesenchymal transition (EMT) pathway (**Supplementary Figure S8**).

PCA and Random Forest Analysis Data

To further identify *ISLR*-correlated hub genes for the differentiation of tumor from normal samples, we performed PCA. As shown in **Figures 8A–C**, we used PC1, PC2 and PC3 to distinguish normal from tumor samples in the three datasets: TCGA-STAD, GSE13861, and GSE29272. Then, we conducted an intersection analysis and identified six hub genes, *ISLR*, *COL1A2* (collagen type I alpha 2 chain), *CDH11* (cadherin 11), *SPARC* (secreted protein acidic and cysteine-rich), *COL3A1* (collagen type III alpha 1 chain), and *COL1A1* (collagen type I alpha 1 chain) (**Figure 8D**). There were strong positive correlations of expression among these genes, and all correlation coefficients were greater than 0.8 (**Figure 8E**). **Figure 8F** presents the correlation between *ISLR* and *COL1A2* gene expression in the

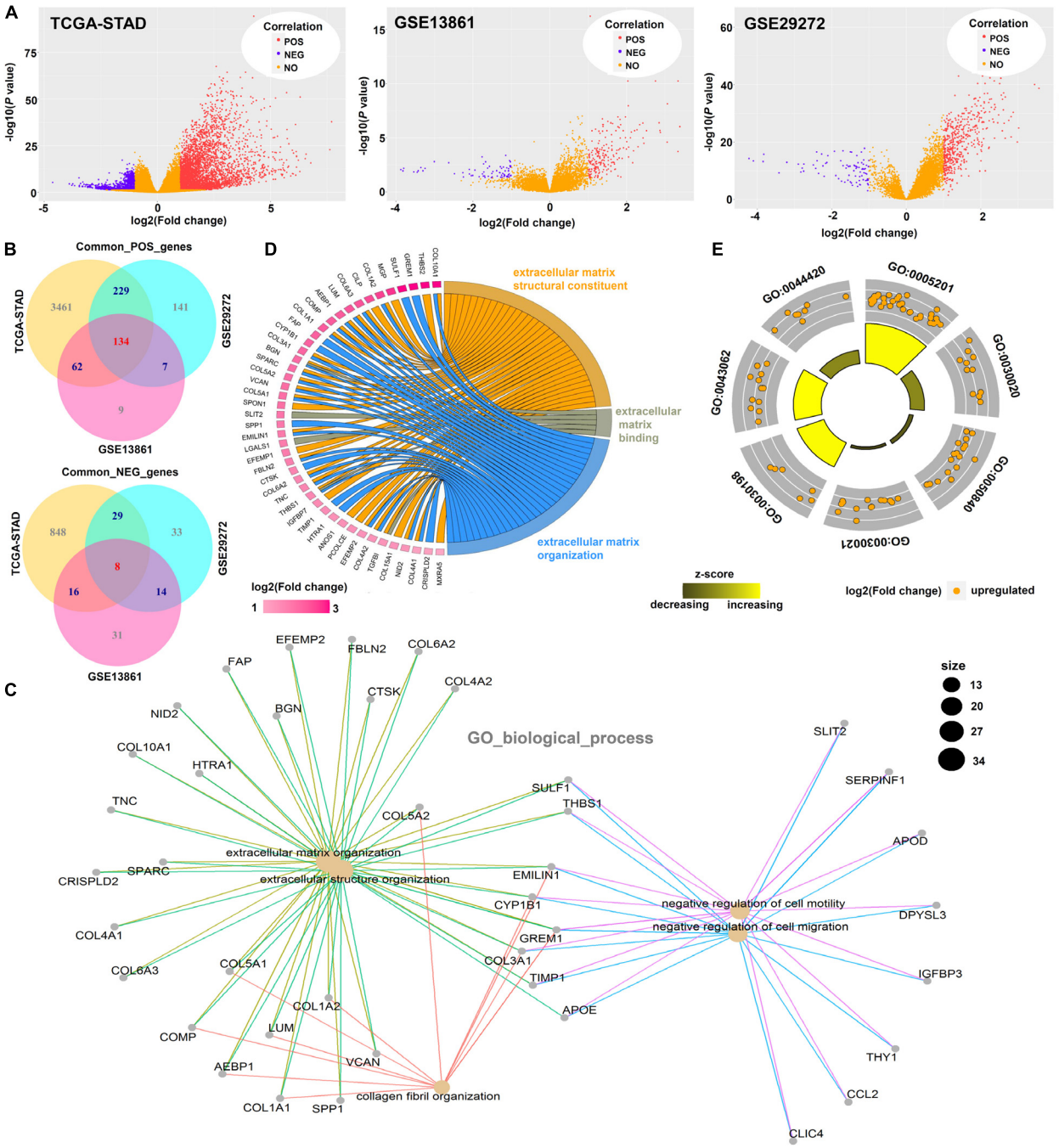

FIGURE 7 | Cluster analysis of significant *ISLR*-correlated genes. **(A)** After expression difference analyses between the ISLR-high and -low expression groups were performed with the "limma" R package on the TCGA-STAD, GSE13861, and GSE29272 datasets, volcano plots were constructed. POS, genes positively correlated with *ISLR*; NEG, genes negatively correlated with *ISLR*; NO, genes not correlated with *ISLR*. **(B)** The "VennDiagram" R package was used for the common genes positively (POS) and negatively (NEG) correlated with ISLR. **(C)** The "clusterProfiler" R package was used for the GO_biological_process enrichment analysis. The GOCircle **(D)** and chord plots **(E)** using extracellular matrix-associated terms were visualized by the "GOplot" R package.

three datasets (R = 0.91, P < 2.2E-16). **Figures 8G–I** further shows the contribution of these hub genes to PC1 and PC2.

Subsequently, we carried out a random forest analysis based on these six hub genes. The multidimensional scale plot in **Figure 9A** suggests the effective differentiation of normal from tumor samples in the TCGA-STAD cohort. **Figure 9B** shows the mean decrease accuracy and mean decrease Gini data. The AUC value of 0.869 indicated high classification accuracy (**Figure 9C**). Similar results were observed in the GSE16831 (**Figures 9D–F**) and GSE29272 (**Figures 9G–I**) datasets.

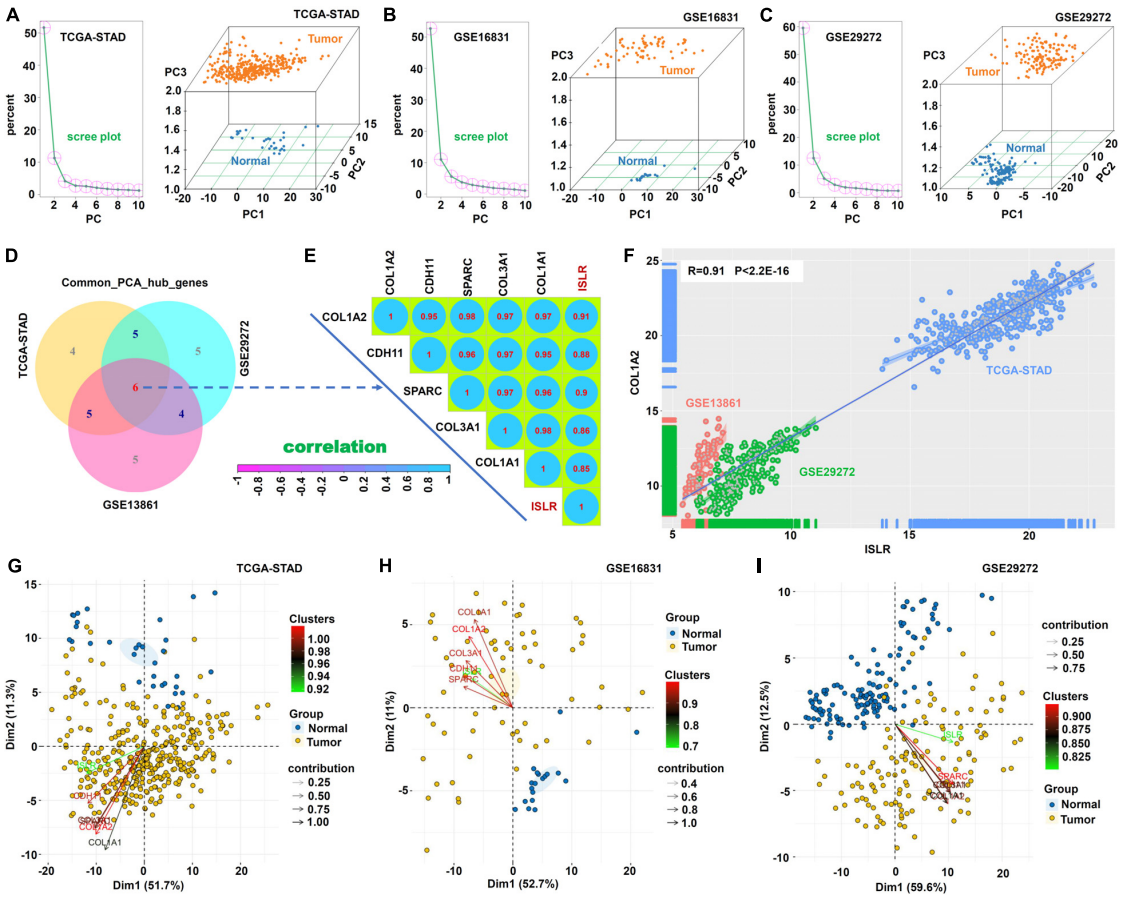

FIGURE 8 | Principal component analysis. Based on the common_POS_genes and common_NEG_genes, the prcomp () function was used for PCA to classify the normal and tumor sections in TCGA-STAD **(A)**, GSE13861 **(B)**, and GSE29272 **(C)**. A scree plot and a three-dimensional map (PC1, PC2, and PC3) are provided. **(D)** The "VennDiagram" R package was used for the common_PCA_hub_genes. **(E)** The cor () function and "corrplot" R package were used for the Spearman correlation analysis of these hub genes. Correlation coefficients are shown. **(F)** The scatter plot for the correlation between *ISLR* and *COL1A2* gene expression is provided. **(G–I)** The "factoextra" R package was utilized to show the principal component weight and two-dimensional contribution maps of common hub genes.

Drug Sensitivity Analysis Data

Finally, based on the CTRP database, we conducted a small molecule/drug sensitivity (IC50) evaluation and further detected that the expression of *the ISLR* gene was negatively related to sensitivity to PX-12 and NSC632839 (**Supplementary Figure S9**).

DISCUSSION

Based on the available datasets of gastric cancer cases collected by TCGA and GEO, for the first time, we found a statistical correlation between high expression of the *ISLR* gene and poor overall survival, disease-free survival, first progression, and post-progression survival. There were significant differences in *ISLR* expression among different pathological stages (stages 1–4). When gastric cancer samples were divided by clinical information, a positive correlation between *ISLR* expression and gastric cancer prognosis existed in most subgroups, such as subgroups based on different Lauren classifications (intestinal or diffuse). Notably, we only observed a correlation between *ISLR* gene expression and OS in the well-differentiated subgroup but not in the poorly or moderately differentiated subgroup. In addition, we detected a positive effect of *ISLR* expression on survival in the pathological stage 3 subgroup but not the stage 1, 2, or 4 subgroups. These results implied that the prognostic ability of high *ISLR* gene expression may increase with tumor differentiation or pathological grade.

Upon integrated analysis, we observed that high expression of the *ISLR* gene showed a correlation with low sensitivity to PX-12 (an irreversible inhibitor of thioredoxin-1) (Metcalfe et al., 2016) and NSC632839 (a non-selective isopeptidase inhibitor) (Nicholson et al., 2008), indicating that high *ISLR* gene expression may be associated with chemoresistance in gastric cancer. Unexpectedly, during the survival analysis of gastric cancer patients treated with a 5-Fu-based adjuvant, high expression of the *ISLR* gene was linked to a better prognosis than low expression of the *ISLR* gene. It is possible that 5-Fu treatment interferes with the expression of the *ISLR* gene in patients, which leads to changes in survival outcomes. Additionally, although we did not detect a correlation between *ISLR* gene

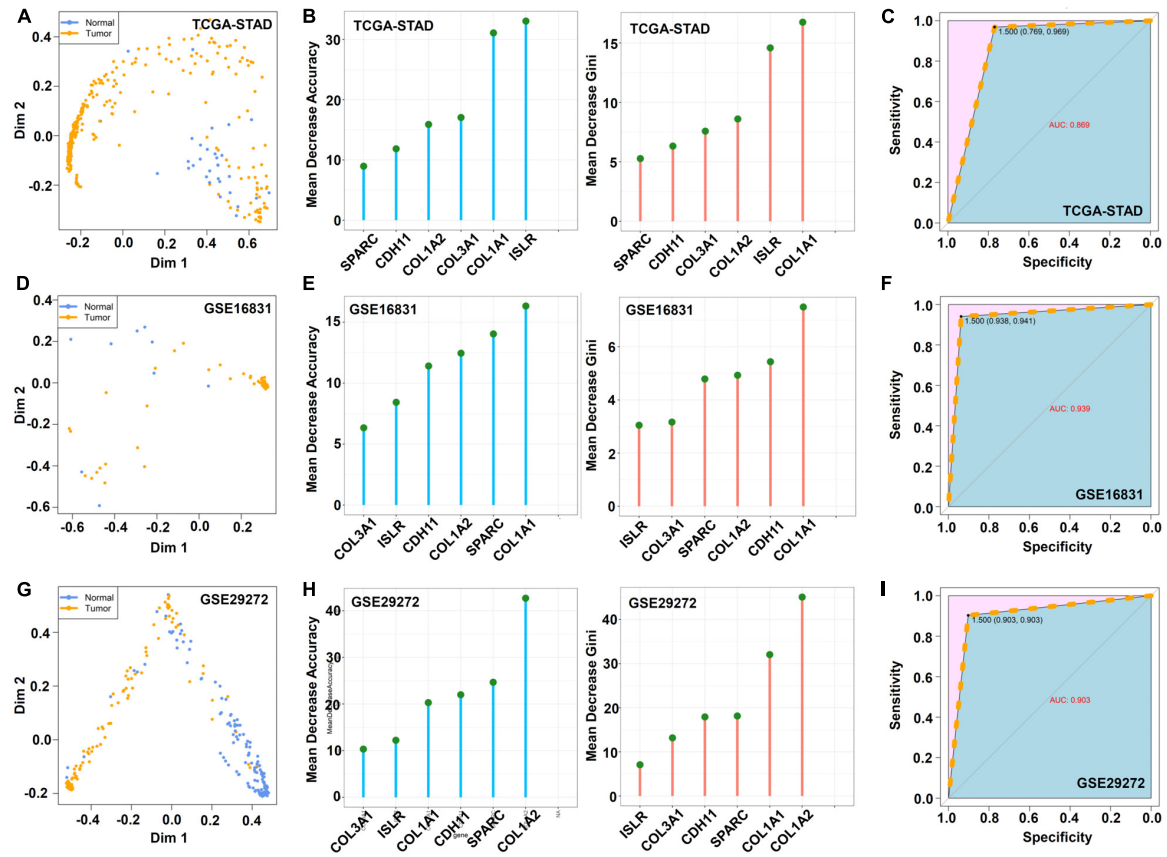

FIGURE 9 | Random forest analysis. Based on the above hub genes, the "randomForest" package (ntree = 500) was utilized for the random forest modeling analysis. The MDSplot () function was used to obtain a multidimensional scale **(A,D,G)**. The data of mean decrease accuracy and mean decrease Gini were visualized by the ggdotchart () function in the "ggpubr" package **(B,E,H)**. The receiver operating characteristic (ROC) curve was plotted by the "pROC" package, and the area under the ROC curve (AUC) value was calculated **(C,F,I)**.

expression and 5-Fu drug sensitivity through our preliminary assessment, more clinical gastric cancer samples specifically under the treatment of 5-Fu and comprehensive analysis are needed to validate the relationship between *ISLR* expression and 5-Fu chemotherapy resistance.

DNA methylation status is closely associated with the carcinogenesis or drug resistance of gastric cancer (Tahara and Arisawa, 2015; Choi et al., 2017). Although we failed to detect a correlation between *ISLR* gene mutations or CNVs and the clinical prognosis of gastric cancer, the hypomethylation status of several sites within *ISLR* (cg05195566, cg17258195, cg09664357, and cg0729703) was linked to high expression of *ISLR* and clinically poor survival outcomes. We noted that the cg05195566 and cg17258195 sites are situated in the promoter region, while cg09664357 and cg0729703 are outside the promoter region. It is worthwhile to further investigate how methylation of different sites within *ISLR* affects the expression level and survival outcomes of gastric cancer patients.

Considering the structure of the ISLR protein as a member of the Ig superfamily (Nagasawa et al., 1997) and the functional links between immune infiltration and gastric cancer (Kim et al., 2016; Liu et al., 2016; Pan et al., 2019; Xu et al., 2019), we first investigated the correlation between *ISLR* gene expression and macrophage, neutrophil, dendritic cell, B cell, T cell and other immune cell infiltration levels based on gene expression correlations and the TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC algorithms. The results were adjusted for tumor purity. We observed a positive correlation between *ISLR* gene expression and several immune cells, especially $CD8^+$ T cells, macrophages and dendritic cells. We also detected a correlation between *ISLR* CNV and immune infiltration. These results indicated that the tumor microenvironment may be key in the complex molecular mechanism by which *the ISLR* gene affects carcinogenesis of gastric cancer.

The extracellular matrix (ECM) and epithelial-mesenchymal transition (EMT) have been reported to be associated with the invasion and migration of gastric cancer (Lukaszewicz-Zajac et al., 2011; Peng et al., 2014; Huang et al., 2015). After pooling the *ISLR* expression-associated genes, we detected significantly enriched ECM-related pathways, including miRNA targets in the ECM and membrane receptors. Our PCA and random forest analysis further identified six extracellular matrix-associated hub

genes, which were able to distinguish between gastric cancer and normal control samples. We also found that *ISLR* gene expression was associated with the activation of the EMT pathway. Considering the connection between miRNAs and EMT in gastric cancer (Bure et al., 2019), we performed GSCALite mRNA-miRNA regulation network analysis to identify several potential *ISLR*-binding miRNAs, including hsa-miR16-5p, hsa-miR-3116, hsa-mir-934, hsa-miR98-5p, and hsa-miR-339-5p (data not shown). It is meaningful to evaluate the relationship between *ISLR* expression and EMT from the perspective of miRNA and to investigate the mechanism underlying the progression of gastric cancer. In addition, the chief aim of our research was only to examine the potential mechanism by which the *ISLR* gene participates in gastric carcinogenesis. It should be noted that *ISLR* does not show specificity for gastric cancer tissue (data not shown), and its role in other cancers cannot be ruled out. Most likely, ISLR works as an effective prognostic marker during gastric carcinogenesis because it forms functional protein-protein and protein-nucleic acid complexes.

CONCLUSION

After our bioinformatics and biostatistics analyses of gastric cancer cases within the TCGA and GEO cohorts, high *ISLR* expression was identified as a potential prognostic biomarker for gastric cancer. DNA hypomethylation of *ISLR* is linked to high expression of the *ISLR* gene and overall clinical prognosis. *ISLR* expression was also correlated with the infiltration of several immune cells (e.g., $CD8^+$ T cells, macrophages and dendritic cells), EMT pathway activity and sensitivity to PX-12 and NSC632839. Our findings are of great significance for conducting ISLR-based cell or animal experimental validation.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

SL conceived and designed the study and drafted the manuscript. SL and WZ performed the *ISLR* expression, survival curve analysis, and genetic alteration analysis. SL and MS performed the DNA methylation, immune cell infiltration analysis, *ISLR*-correlated pathway, PCA, random forest, and drug sensitivity analysis. All authors reviewed the manuscript before submission and approved the final version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00620/full#supplementary-material

FIGURE S1 | Correlation between methylation status and expression of *ISLR* in gastric cases in the TCGA-STAD cohort. Detailed information on the methylation probe is provided. Pearson correlation coefficients (R) and Benjamini–Hochberg-adjusted *P*-values (*$P < 0.05$, ***$P < 0.001$) for the comparison are shown as well.

FIGURE S2 | Correlation between the methylation status of *ISLR* DNA and the survival rate of gastric patients in the TCGA-STAD cohort. A total of eleven methylation probes, including cg15480336 **(A)**, cg03778780 **(B)**, cg02077702 **(C)**, cg20536146 **(D)**, cg11335960 **(E)**, cg24779381 **(F)**, cg05195566 **(G)**, cg177258195 **(H)**, cg09664357 **(I)**, cg07297039 **(J)**, and cg16926502 **(K)**, were used.

FIGURE S3 | Correlation between *ISLR* and markers of T or B cells as analyzed through GEPIA2. **(A)** B cells; **(B)** Th1 cells; **(C)** Th2 cells; **(D)** Th17 cells; **(E)** $CD8^+$ T cells; **(F)** Tfh cells; **(G)** resting Treg cells; **(H)** effector Treg cells; **(I)** exhausted T cells.

FIGURE S4 | Correlation between *ISLR* CNV and the infiltration level of immune cells as analyzed through TIMER. **(A)** dendritic cells; **(B)** neutrophils; **(C)** $CD8^+$ T cells; **(D)** $CD4^+$ T cells; **(E)** B cells; **(F)** macrophages. (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

FIGURE S5 | GO_biological_process and GO_molecular_function enrichment analysis. The "clusterProfiler" R package was used for the GO_biological_process **(A)** and GO_molecular_function **(B)** enrichment analyses.

FIGURE S6 | KEGG enrichment analysis. The "pathview" R package was used for the KEGG enrichment analysis.

FIGURE S7 | GSEA. A LinkedOmics approach was utilized for the *ISLR*-correlated GSEA profiles. Four extracellular matrix (ECM)-associated gene sets, including genes involved in extracellular structural organization **(A)**, extracellular matrix structural constituents **(B)**, ECM-receptor interactions **(C)**, and miRNA targets in the ECM and membrane receptors **(D)**, were identified.

FIGURE S8 | Correlation between *ISLR* expression and pathway activation or inhibition.

FIGURE S9 | Correlation between *ISLR* expression and small molecule/drug sensitivity (IC50).

TABLE S1 | Correlation of *ISLR* expression and the first progression status of gastric cancer patients in the GEO cohort (Kaplan–Meier plotter).

TABLE S2 | Correlation of *ISLR* expression and the PPS of gastric cancer patients in the GEO cohort (Kaplan–Meier plotter).

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bhandari, V., Hoey, C., Liu, L. Y., Lalonde, E., Ray, J., Livingstone, J., et al. (2019). Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* 51, 308–318. doi: 10.1038/s41588-018-0318-2

Bure, I. V., Nemtsova, M. V., and Zaletaev, D. V. (2019). Roles of E-cadherin and noncoding RNAs in the epithelial-mesenchymal transition and progression in gastric cancer. *Int. J. Mol. Sci.* 20:2870. doi: 10.3390/ijms20122870

Cancer Genome Atlas Research Network, (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Chen, X., Leung, S. Y., Yuen, S. T., Chu, K. M., Ji, J., Li, R., et al. (2003). Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell* 14, 3208–3215. doi: 10.1091/mbc.e02-12-0833

Choi, S. J., Jung, S. W., Huh, S., Chung, Y. S., Cho, H., and Kang, H. (2017). Alteration of DNA methylation in gastric cancer with chemotherapy. *J. Microbiol. Biotechnol.* 27, 1367–1378. doi: 10.4014/jmb.1704.04035

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* 6, 271–281.e7. doi: 10.1016/j.cels.2018.03.002

Gao, Q., Liang, W. W., Foltz, S. M., Mutharasu, G., Jayasinghe, R. G., Cao, S., et al. (2018). Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* 23, 227–238.e3. doi: 10.1016/j.celrep.2018.03.050

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e6. doi: 10.1016/j.cell.2018.03.022

Homma, S., Shimada, T., Hikake, T., and Yaginuma, H. (2009). Expression pattern of LRR and Ig domain-containing protein (LRRIG protein) in the early mouse embryo. *Gene Expr. Patterns* 9, 1–26. doi: 10.1016/j.gep.2008.09.004

Huang, L., Wu, R. L., and Xu, A. M. (2015). Epithelial-mesenchymal transition in gastric cancer. *Am. J. Transl. Res.* 7, 2141–2158.

Kakiuchi, M., Nishizawa, T., Ueda, H., Gotoh, K., Tanaka, A., Hayashi, A., et al. (2014). Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. *Nat. Genet.* 46, 583–587. doi: 10.1038/ng.2984

Kim, J. W., Nam, K. H., Ahn, S. H., Park, D. J., Kim, H. H., Kim, S. H., et al. (2016). Prognostic implications of immunosuppressive protein expression in tumors as well as immune cell infiltration within the tumor microenvironment in gastric cancer. *Gastric Cancer* 19, 42–52. doi: 10.1007/s10120-014-0440-5

Koch, A., De Meyer, T., Jeschke, J., and Van Criekinge, W. (2015). MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics* 16:636. doi: 10.1186/s12864-015-1847-z

Koch, A., Jeschke, J., Van Criekinge, W., van Engeland, M., and De Meyer, T. (2019). MEXPRESS update 2019. *Nucleic Acids Res.* 47, W561–W565. doi: 10.1093/nar/gkz445

Li, B., Severson, E., Pignon, J. C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17:174.

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, e108–e110. doi: 10.1158/0008-5472.can-17-0307

Liu, C. J., Hu, F. F., Xia, M. X., Han, L., Zhang, Q., and Guo, A. Y. (2018). GSCALite: a web server for gene set cancer analysis. *Bioinformatics* 34, 3771–3772. doi: 10.1093/bioinformatics/bty411

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, 400–416.e11. doi: 10.1016/j.cell.2018.02.052

Liu, J. Y., Yang, X. J., Geng, X. F., Huang, C. Q., Yu, Y., and Li, Y. (2016). Prognostic significance of tumor-associated macrophages density in gastric cancer: a systemic review and meta-analysis. *Minerva Med.* 107, 314–321.

Lugowska, A., Hetmanczyk-Sawicka, K., Iwanicka-Nowicka, R., Fogtman, A., Ciesla, J., Purzycka-Olewiecka, J. K., et al. (2019). Gene expression profile in patients with Gaucher disease indicates activation of inflammatory processes. *Sci. Rep.* 9:6060.

Lukaszewicz-Zajac, M., Mroczko, B., and Szmitkowski, M. (2011). Gastric cancer - The role of matrix metalloproteinases in tumor progression. *Clin. Chim. Acta* 412, 1725–1730. doi: 10.1016/j.cca.2011.06.003

Metcalfe, C., Ramasubramoni, A., Pula, G., Harper, M. T., Mundell, S. J., and Coxon, C. H. (2016). Thioredoxin inhibitors attenuate platelet function and thrombus formation. *PLoS One* 11:e0163006. doi: 10.1371/journal.pone.0163006

Modhukur, V., Iljasenko, T., Metsalu, T., Lokk, K., Laisk-Podar, T., and Vilo, J. (2018). MethSurv: a web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics* 10, 277–288. doi: 10.2217/epi-2017-0118

Nagasawa, A., Kubota, R., Imamura, Y., Nagamine, K., Wang, Y., Asakawa, S., et al. (1997). Cloning of the cDNA for a new member of the immunoglobulin superfamily (ISLR) containing leucine-rich repeat (LRR). *Genomics* 44, 273–279. doi: 10.1006/geno.1997.4889

Nicholson, B., Leach, C. A., Goldenberg, S. J., Francis, D. M., Kodrasov, M. P., Tian, X., et al. (2008). Characterization of ubiquitin and ubiquitin-like-protein isopeptidase activities. *Protein Sci.* 17, 1035–1043. doi: 10.1110/ps.083450408

Pan, J. H., Zhou, H., Cooper, L., Huang, J. L., Zhu, S. B., Zhao, X. X., et al. (2019). LAYN is a prognostic biomarker and correlated with immune infiltrates in gastric and colon cancers. *Front. Immunol.* 10:6. doi: 10.3389/fimmu.2019.00006

Peng, Z., Wang, C. X., Fang, E. H., Wang, G. B., and Tong, Q. (2014). Role of epithelial-mesenchymal transition in gastric cancer initiation and progression. *World J. Gastroenterol.* 20, 5403–5410. doi: 10.3748/wjg.v20.i18.5403

Rees, M. G., Seashore-Ludlow, B., Cheah, J. H., Adams, D. J., Price, E. V., Gill, S., et al. (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* 12, 109–116. doi: 10.1038/nchembio.1986

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321–337.e10. doi: 10.1016/j.cell.2018.03.035

Szasz, A. M., Lanczky, A., Nagy, A., Forster, S., Hark, K., Green, J. E., et al. (2016). Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* 7, 49322–49333. doi: 10.18632/oncotarget.10337

Tahara, T., and Arisawa, T. (2015). DNA methylation as a molecular biomarker in gastric cancer. *Epigenomics* 7, 475–486. doi: 10.2217/epi.15.4

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* 47, W556–W560. doi: 10.1093/nar/gkz430

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., et al. (2018). Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* 33, 676–689.e3. doi: 10.1016/j.ccell.2018.03.007

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963. doi: 10.1093/nar/gkx1090

Wang, K., Kan, J., Yuen, S. T., Shi, S. T., Chu, K. M., Law, S., et al. (2011). Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat. Genet.* 43, 1219–1223. doi: 10.1038/ng.982

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H., Shi, S. T., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* 46, 573–582. doi: 10.1038/ng.2983

Wang, Z., Jensen, M. A., and Zenklusen, J. C. (2016). A practical guide to the cancer genome atlas (TCGA). *Methods Mol. Biol.* 1418, 111–141. doi: 10.1007/978-1-4939-3578-9_6

Xu, J., Yu, Y., He, X., Niu, N., Li, X., Zhang, R., et al. (2019). Tumor-associated macrophages induce invasion and poor prognosis in human gastric cancer in a cyclooxygenase-2/MMP9-dependent manner. *Am. J. Transl. Res.* 11, 6040–6054.

Yoon, I. K., Kim, H. K., Kim, Y. K., Song, I. H., Kim, W., Kim, S., et al. (2004). Exploration of replicative senescence-associated genes in human dermal fibroblasts by cDNA microarray technology. *Exp. Gerontol.* 39, 1369–1378. doi: 10.1016/j.exger.2004.07.002

Conflict of Interest: WZ was employed by the General Data Technology Co., Ltd, Tianjin.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 July 2020
doi: 10.3389/fgene.2020.00603

Causal Effects of Overall and Abdominal Obesity on Insulin Resistance and the Risk of Type 2 Diabetes Mellitus: A Two-Sample Mendelian Randomization Study

Hua Xu[1,2†], *Chuandi Jin*[3†] *and Qingbo Guan*[1,2*]

[1] Department of Endocrinology, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, China, [2] Shandong Clinical Medical Center of Endocrinology and Metabolism, Institute of Endocrinology and Metabolism, Shandong Academy of Clinical Medicine, Jinan, China, [3] Institute for Medical Dataology, Shandong University, Jinan, China

OPEN ACCESS

Edited by:
Lide Han,
Vanderbilt University Medical Center, United States

Reviewed by:
Dan Zhou,
Vanderbilt University Medical Center, United States
Ping Zeng,
Xuzhou Medical University, China

***Correspondence:**
Qingbo Guan
doctorguanqingbo@163.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: *26 January 2020*
Accepted: *18 May 2020*
Published: *02 July 2020*

Citation:
Xu H, Jin C and Guan Q (2020) Causal Effects of Overall and Abdominal Obesity on Insulin Resistance and the Risk of Type 2 Diabetes Mellitus: A Two-Sample Mendelian Randomization Study. Front. Genet. 11:603. doi: 10.3389/fgene.2020.00603

Overall and abdominal obesity were significantly associated with insulin resistance and type 2 diabetes mellitus (T2DM) risk in observational studies, though these associations cannot avoid the bias induced by confounding effects and reverse causation. This study aimed to test whether these associations are causal, and it compared the causal effects of overall and abdominal obesity on T2DM risk and glycemic traits by using a two-sample Mendelian randomization (MR) design. Based on summary-level statistics from genome-wide association studies, the instrumental variables for body mass index (BMI), waist-to-hip ratio (WHR), and WHR adjusted for BMI (WHRadjBMI) were extracted, and the horizontal pleiotropy was analyzed using MR–Egger regression and the MR–pleiotropy residual sum and outlier (PRESSO) method. Thereafter, by using the conventional MR method, the inverse-variance weighted method was applied to assess the causal effect of BMI, WHR, and WHRadjBMI on T2DM risk, Homeostatic model assessment of insulin resistance (HOMA-IR), fasting insulin, fasting glucose, and Hemoglobin A1c (HbA1c). A series of sensitivity analyses, including the multivariable MR (diastolic blood pressure, systolic blood pressure, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol as covariates), MR–Egger regression, weighted median, MR–PRESSO, and leave-one-out method, were conducted to test the robustness of the results from the conventional MR. Despite the existence of horizontal pleiotropy, consistent results were found in the conventional MR results and sensitivity analyses, except for the association between BMI and fasting glucose, and WHRadjBMI and fasting glucose. Each one standard deviation higher BMI was associated with an increased T2DM risk [odds ratio (OR): 2.741; 95% confidence interval (CI): 2.421–3.104], higher HbA1c [1.054; 1.04–1.068], fasting insulin [1.202; 1.173–1.231], and HOMA-IR [1.221; 1.187–1.255], similar to findings for causal effect of WHRadjBMI on T2DM risk [1.993; 1.704–2.33], HbA1c [1.061; 1.042–1.08], fasting insulin [1.102; 1.068–1.136], and HOMA-IR [1.127; 1.088–1.167]. Both BMI (P = 0.546) and WHRadjBMI (P = 0.443) were unassociated with fasting glucose in the multivariable MR analysis. In conclusion,

overall and abdominal obesity have causal effects on T2DM risk and insulin resistance but no causal effect on fasting glucose. Individuals can substantially reduce their insulin resistance and T2DM risk through reduction of body fat mass and modification of body fat distribution.

Keywords: type 2 diabetes mellitus, insulin resistance, abdominal obesity, body fat mass, body fat mass distribution, Mendelian randomization

INTRODUCTION

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disease characterized by hyperglycemia secondary to insulin resistance and pancreatic β-cell failure (Alejandro et al., 2015). The findings of human epidemiologic studies indicate that the global prevalence of T2DM is increasing rapidly, and this increase parallels the increase in the prevalence of obesity (Sampath Kumar et al., 2019). The body mass index (BMI) is routinely used to quantify the overall obesity although body fat distribution of individuals can vary substantially. The waist-to-hip ratio (WHR) and WHR adjusted for BMI (WHRadjBMI) are frequently used surrogate measures of abdominal obesity. Many observational epidemiologic studies, including case–control and cohort studies, have demonstrated that higher WHR and BMI are two important risk factors for developing T2DM (Vazquez et al., 2007; Lv et al., 2017). Moreover, cohort and cross-sectional studies (Wang et al., 2018; Benites-Zapata et al., 2019) demonstrated that BMI and WHR were associated with glycemic traits, including fasting insulin, Hemoglobin A1c (HbA1c), and insulin resistance [measured by Homeostatic model assessment of insulin resistance (HOMA-IR)]. Longitudinal and cross-sectional studies have found an association between increased risk of T2DM and higher genetic predisposition to both BMI and WHRadjBMI in European and East Asian populations (Robiou-du-Pont et al., 2013; Zhu et al., 2014; Huang et al., 2015).

However, these observational studies cannot avoid the bias induced by the confounding effect and reverse causation and, therefore, are incapable of confirming whether these associations are causal (Smith and Ebrahim, 2003). Mendelian randomization (MR) is an approach that is used to unbiasedly test or estimate the causal relationship between an exposure and an associated outcome by using data on inherited genetic variants that influence exposure status in the presence of unmeasured confounding (Didelez and Sheehan, 2007; Lawlor et al., 2008). In the past few years, MR has been extensively used in epidemiology and other related areas of population science (Smit et al., 2019; Wainberg et al., 2019; Yang et al., 2019).

Previous MR analyses have demonstrated that per 1 standard deviation (SD) higher WHRadjBMI and BMI were causally associated with T2DM risk in European populations (Dale et al., 2017; Emdin et al., 2017). Wang et al. (2018) conducted an MR to further investigate the causal effect of both BMI and WHR on glycemic traits, and they found that BMI had a causal relevance for insulin secretion, whereas neither WHR and BMI was causally associated with HOMA-IR in a conventional MR in a Chinese Han population. However, there is a dearth of MR studies for testing and comparing the causal effect of both overall and abdominal obesity on glycemic in the European population. Epidemiologic studies have found differences in T2DM epidemiologic characteristics between the Asian and European population wherein, in comparison with South Asians, Europeans have a lower T2DM risk, typically develop T2DM 5–10 years later, and have a slower disease progression (Gujral et al., 2013; Admiraal et al., 2014; Gupta and Misra, 2016; Meeks et al., 2016; Banerjee and Shah, 2018). Moreover, Europeans developed many metabolic abnormalities, including hyperglycemia and elevated triacylglycerol and low high-density lipoprotein cholesterol (HDL-C), at a higher BMI and age (Raji et al., 2001; Razak et al., 2007). Therefore, the estimation and comparison of the causal effects of overall and abdominal obesity on glycemic traits could provide insights into the obesity-related mechanism of T2DM.

In this study, a two-sample Mendelian randomization (TSMR) with a large sample size was conducted to determine whether a genetic predisposition to increased BMI, WHR, and WHRadjBMI was causally associated with T2DM and glycemic traits, including HOMA-IR, fasting insulin, fasting glucose, and HbA1c. The causal effects were further compared to identify differences in the effect of overall and abdominal obesity on T2DM development and glycemic traits.

MATERIALS AND METHODS

Data Source

This study aimed to explore the causal effect of WHR, BMI, and WHRadjBMI on the risk of T2DM and glycemic traits (HOMA-IR, fasting insulin, fasting glucose, and HbA1c) in an European population, and used diastolic blood pressure (DBP), systolic blood pressure (SBP), HDL-C, and low-density lipoprotein cholesterol (LDL-C) as the covariates. The genome-wide association study summary statistics datasets used in this study were obtained from Zenodo[1] for WHR (Censin et al., 2019), BMI (Censin et al., 2019), and WHRadjBMI (Censin et al., 2019); the Program in Complex Trait Genomics[2] for T2DM (Xue et al., 2018); MAGIC Consortium[3] for HOMA-IR (Dupuis et al., 2010), fasting glucose (Manning et al., 2012), fasting insulin (Manning et al., 2012), and HbA1c (Wheeler et al., 2017); the

[1]https://zenodo.org

[2]https://cnsgenomics.com

[3]http://www.magicinvestigators.org/

MRBase platform[4] for HDL-C (Kettunen et al., 2016) and LDL-C (Kettunen et al., 2016); and the MRC-IEU Consortium[5] for SBP and DBP. Detailed information of the summary statistics datasets are displayed in **Table 1**. We obtained the β-coefficients and standard errors for the per allele association of each single-nucleotide polymorphism (SNP) as well as all exposures and outcomes from these data sources.

Selection of Genetic Instrumental Variables

In the TSMR analysis conducted in this study, the genetic variants for exposures (BMI, WHR, and WHRadjBMI) were used as instrumental variables (IVs) and were obtained by two steps. Firstly, SNPs that are strongly associated with exposures ($P < 5.0 \times 10^{-8}$) were extracted. Secondly, we pruned these extracted SNPs by linkage disequilibrium (LD; $r^2 = 0.001$, clumping distance = 10,000 kb) to ensure that each IV was independent of the others. To test the strength of the IVs, the *F*-statistics were calculated as previously described (Xu and Hao, 2017). *F*-statistics >10 are considered adequately strong to mitigate against any bias of the causal IV estimate.

Heterogeneity and Horizontal Pleiotropic Analysis

In MR, heterogeneity in the causal estimate may indicate that a variant has an effect on the outcome outside of its effect on the exposure (known as horizontal pleiotropy), and this can cause severe bias (Davey Smith and Hemani, 2014). Mendelian randomization–Egger (MR–Egger) regression was undertaken to assess the horizontal pleiotropy of the IVs, where a regression intercept that significantly differed from zero ($P < 0.05$) indicated the presence of horizontal pleiotropy exists or that the InSIDE (INstrument Strength Independent of Direct Effect) assumption was violated (Bowden et al., 2015). Heterogeneity between IVs in the conventional MR, with the inverse-variance weighted (IVW) method, was estimated by *Cochran's Q* statistic. The MR pleiotropy residual sum and outlier (MR–PRESSO) method can be used to test horizontal pleiotropic outliers and can obtain the corrected causal effect after removal of these outliers in MR (Verbanck et al., 2018). In the present study, both MR–Egger regression and MR–PRESSO tests were conducted using the TwoSampleMR and MRPRESSO R package in R (version 3.6.1), respectively.

Mendelian Randomization

Mendelian randomization can test and estimate the causal effect of an exposure on an outcome by using genetic variants as the IVs (Zheng et al., 2017). Firstly, Wald ratios were calculated for each IV by dividing the per-allele log-odds ratio (or beta) of that variant in the outcome data by the log-odds ratio (or beta) of the same variant in the exposure data. Then, the random-effects IVW method was applied to estimate the association between exposures and outcomes. In IVW, the Wald ratio for each SNP was weighted by its inverse variance, and the effect estimates were meta-analyzed using random effects.

Sensitivity Analysis

Sensitivity analysis was used to test the disproportionate effects of variants and the pleiotropy in the MR analysis (Mokry et al., 2016). A series of sensitivity analyses were conducted to test the robustness of the conventional MR results.

Multivariable IVW, which included the DBP, SBP, HDL-C, and LDL-C as covariates, was carried out in accordance with the method proposed by Rees et al. (2017) that was used to account for possible horizontal pleiotropy arising from the association of the instrument with these variables.

The MR–Egger regression and weighted median method are two pleiotropy-robust MR methods that are used to estimate consistent causal effects against unknown directional pleiotropy under the InSIDE assumptions (Bowden et al., 2015). In the MR–Egger regression method, the regression line fitted to the data is not constrained to pass through the origin, and the intercept represents the horizontal pleiotropic effect that may bias the IVW estimate, whereas the slope represents pleiotropy-corrected causal estimates. The weighted median method has considerable robustness to individual genetics with strongly outlying causal estimates and could provide a consistent causal estimate when the valid IVs exceed 50%.

The MR–PRESSO method was used to identify potential outliers in the conventional MR testing, and provided a robust estimate with outlier correction. Moreover, testing of significant distortion in the IVW causal estimate before and after MR–PRESSO correction, was undertaken and served as a sensitivity analysis.

The leave-one-out sensitivity analysis was conducted to ascertain whether the association was being disproportionately influenced by a single SNP. In this analysis, the random-effects IVW was repeated by leaving out each SNP in turn, and the overall analysis including all SNPs was used for the comparison. The variation of the results from before and after the removal of each SNP reflects the sensitivity of this SNP.

RESULTS

Genetic IVs

A total of 546, 356, and 330 IVs were identified for BMI, WHR, and WHRadjBMI, respectively. Some IVs were absent in the outcome data; however, the *F* statistics for BMI-IVs (86.250–89.078), WHR-IVs (67.502–67.991), and WHRadjBMI-IVs (90.758–96.860) that were used for MR were more than 10, which indicated that the weak instrument bias was negligible. Detailed information of IVs used in this study are shown in **Supplementary Table S1**.

Horizontal Pleiotropy and Heterogeneity Analysis

The MR–Egger regression intercepts obtained in this study (**Table 2**) showed that horizontal pleiotropy ($P = 0.029$) was

[4]http://www.mrbase.org

[5]http://www.bristol.ac.uk/integrative-epidemiology/

TABLE 1 | Summary statistics of data source.

Traits	Consortium	Data sources	No. of participants	No. of Variants	Population	Units in TSMR
WHR	ZENODO	Censin; PloS Genet; 2019	697,734	27,381,301	European	SD
BMI	ZENODO	Censin; PloS Genet; 2019	806,834	27,376,273	European	SD
WHRadjBMI	ZENODO	Censin; PloS Genet; 2019	694,649	27,375,636	European	SD
HDL-C	MRBase	Kettunen; Nat Commun; 2016	21,555	11,865,530	European	SD
LDL-C	MRBase	Kettunen; Nat Commun; 2016	21,559	11,871,461	European	SD
SBP	MRC-IEU	Ben Elsworth; 2018	436,419	9,851,867	European	SD
DBP	MRC-IEU	Ben Elsworth; 2018	436,424	9,851,867	European	SD
Fasting glucose	MAGIC	Manning, Nat Genet; 2012	58,074	2,628,880	European	mmol/L
Fasting insulin	MAGIC	Manning, Nat Genet; 2012	51,750	2,627,849	European	log pmol/L
HOMA-IR	MAGIC	Dupuis; Nat Genet; 2010	37,037	2,458,074	European	log HOMA
HbA1c	MAGIC	Wheeler, PloS Med; 2017	123,655	2,586,698	European	%
T2DM	Program in Complex Trait Genomics	Xue; Nat Commun; 2018	62,892/596,424	5,053,015	European	log odds

WHR, waist-to-hip ratio; BMI, body mass index; WHRadjBMI, waist-to-hip ratio adjusted for body mass index; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; T2DM, Type 2 diabetes; TSMR, two-sample Mendelian randomization; SD, standard deviation; HOMA-IR, Homeostasis model assessment of insulin resistance; HbA1c, Hemoglobin A1c.

only found in the MR with WHRadjBMI as exposure and fasting insulin as the outcome. Heterogeneity (**Table 2**) was observed between IVs; therefore, the random-effect IVW method was used in the subsequent stages of the research analysis. The MR–PRESSO test showed that horizontal pleiotropy was found in all IVW analyses in this study, and the horizontal pleiotropic outliers were identified and removed (**Supplementary Table S2**). After the removal of these outliers, the *F*-statistics of BMI-IVs (85.704–89.033), WHR-IVs (62.240–67.991), and WHRadjBMI-IVs (84.668–96.860) continued to remain well powered to estimate the causal effect of the exposure on the outcome.

Causal Effect of WHR, BMI, and WHRadjBMI on T2DM and Glycemic Traits

Table 3 and **Figure 1** show the causal effect estimates of WHR, BMI, and WHRadjBMI on T2DM and glycemic traits. The TSMR analysis by the IVW method showed a significant causal effect, wherein each SD of genetically higher BMI was associated with an increased T2DM risk [OR: 2.741; 95% confidence interval (CI): 2.421–3.104], higher fasting glucose [1.073; 1.048–1.099], higher fasting insulin [1.202; 1.173–1.231], higher HOMA-IR [1.221; 1.187–1.255], and higher HbA1c [1.054; 1.04–1.068]. Each SD of genetically higher WHR was associated with increased T2DM risk [3.12; 2.653–3.668], higher fasting glucose [1.087; 1.054–1.12], higher fasting insulin [1.193; 1.153–1.234], higher HOMA-IR [1.203; 1.155–1.252], and higher HbA1c [1.075; 1.056–1.095]. Each SD of genetically higher WHRadjBMI was associated with increased T2DM risk [1.993; 1.704–2.33], higher fasting glucose [1.039; 1.012–1.067], higher fasting insulin [1.102; 1.068–1.136], higher HOMA-IR [1.127; 1.088–1.167], and higher HbA1c [1.061; 1.042–1.08].

Sensitivity Analysis

In the leave-one-out sensitivity analysis, no single SNP strongly or reversely drove the overall effect of exposure on outcome in the IVW (**Supplementary Figure S1**). Consistent results were observed in the IVW after the MR–PRESSO correction, MR–Egger regression, and the weighted median method, with the exception of the causal estimates of WHR on HbA1c ($P = 0.058$) and fasting glucose ($P = 0.098$) in the MR–Egger regression. The MR–Egger regression could obtain pleiotropy-corrected causal estimates, although this method had less statistical power than an equivalent IVW method, and the CIs were wider and included the null value (Bowden, 2017; Weng et al., 2018). Because the intercept of the MR–Egger regression indicates that there was no horizontal pleiotropy in the MR–Egger regression between WHR and both HbA1c ($P = 0.695$) and fasting glucose ($P = 0.935$), the causal estimate was more convincing in the IVW. In the multivariable IVW (DBP, SBP, HDL-C, and LDL-C as covariates), BMI ($P = 0.546$) and WHRadjBMI ($P = 0.443$) were not causally associated with fasting glucose, whereas other multivariable IVW results persisted with that in the univariable IVW.

Taken together, the causal effect estimates of BMI and WHRadjBMI on fasting glucose in conventional MR might be biased by the horizontal pleiotropy of SBP, DBP, HDL-C, and LDL-C, while no significant bias was found in other causal effect estimates despite the existence of horizontal pleiotropy and heterogeneity.

DISCUSSION

Numerous observational studies indicated that obesity was strongly associated with T2DM risk and glycemic traits (Lv et al., 2017), however, a causal effect cannot be ascertained from these studies due to residual confounding or reverse causality. This present study utilized a TSMR design that was applied to the summary-level data from a large-scale genome-wide association study to address the potential causal role of overall obesity (measured by BMI) and abdominal obesity (measure by WHRadjBMI) on the risk of T2DM and glycemic traits. The well-powered conventional MR (random-effect IVW method)

TABLE 2 | Heterogeneity and horizontal pleiotropy analysis.

Exposure	Outcome	Heterogeneity			Horizontal pleiotropy			
					MR–PRESSO test		MR–Egger regression	
		Q	Q df	*P*	Global Test RSSobs	Global Test *P*	intercepts (95% CI)	*P*
BMI								
	T2DM	2867.057	445	<0.001	2885.029	<0.001	−0.001 (0.994,1.005)	0.837
	HOMA-IR	510.260	446	0.019	512.445	0.023	0 (0.998,1.001)	0.430
	HbA1c	654.863	445	<0.001	658.515	<0.001	0 (0.999,1)	0.714
	Fasting insulin	624.170	448	<0.001	626.889	<0.001	0 (0.999,1.001)	0.842
	Fasting glucose	619.187	448	<0.001	622.082	<0.001	0 (0.999,1.001)	0.838
WHR								
	T2DM	1718.377	275	<0.001	1739.466	<0.001	0.008 (1,1.017)	0.055
	HOMA-IR	368.360	277	<0.001	371.364	<0.001	0 (0.998,1.002)	0.729
	HbA1c	413.286	279	<0.001	416.611	<0.001	0 (0.999,1.001)	0.695
	Fasting insulin	431.511	280	<0.001	435.262	<0.001	0 (0.998,1.002)	0.853
	Fasting glucose	358.324	280	0.001	361.045	0.001	0 (0.998,1.002)	0.935
WHRadjBMI								
	T2DM	1590.053	236	<0.001	1609.124	<0.001	0.003 (0.995,1.011)	0.455
	HOMA-IR	290.496	237	0.010	293.251	0.011	−0.001 (0.997,1.001)	0.175
	HbA1c	383.587	238	<0.001	387.452	<0.001	0 (0.999,1)	0.355
	Fasting insulin	362.009	239	<0.001	365.491	<0.001	−0.002 (0.997,1)	0.029
	Fasting glucose	276.448	239	0.048	278.957	0.047	−0.001 (0.998,1.001)	0.319

WHR, waist-to-hip ratio; BMI, body mass index; WHRadjBMI, waist-to-hip ratio adjusted for body mass index; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; T2DM, Type 2 diabetes; SD, standard deviation; HOMA-IR, Homeostasis model assessment of insulin resistance; HbA1c, Hemoglobin A1c; Q, Cochran's Q test estimate; df, Cochran's Q test degrees of freedom; MR–egger, Mendelian randomization–Egger regression; MR–PRESSO, Mendelian randomization pleiotropy residual sum and outlier method; OR, odds ratio; CI, confidence interval.

confirmed that genetic predisposition to higher BMI, WHR, and WHRadjBMI are causally associated with higher fasting glucose, fasting insulin, HOMA-IR, HbA1c, and increased risk of T2DM in the European population.

However, heterogeneity and horizontal pleiotropy was found in the conventional MR analysis, a series of sensitivity analyses that included the multivariable MR (DBP, SBP, HDL-C, and LDL-C as covariates), MR–Egger regression, weighted median method, MR–PRESSO method, and leave-one-out analysis to test the robustness of the conventional MR results. The causal effect of BMI and WHRadjBMI on T2DM risk, HbA1c, fasting insulin, and HOMA-IR in the conventional MR were consistent with that in all the sensitivity analyses, which suggested that the causal estimate was robust and unbiased.

Each SD of genetically higher BMI [2.741; 2.421–3.104] and WHRadjBMI [1.993; 1.704–2.33] was associated with increased T2DM risk. Human epidemiologic studies have considered obesity to be a major risk factor of T2DM, and the substantial increase in the incidence of obesity contributes to the current T2DM epidemic (Sampath Kumar et al., 2019). Using the MR method in the European descendants, Emdin et al. confirmed that a 1 SD genetic increase in WHRadjBMI was associated with a higher risk of T2DM [1.77; 1.57–2.00] (Emdin et al., 2017), and Dale et al. revealed that each SD higher BMI was associated with increased T2DM risk [1.98; 1.41–2.78] (Dale et al., 2017). The results of this study are in agreement with those of previous observational studies (Lv et al., 2017; Sampath Kumar et al., 2019) and MR studies (Dale et al., 2017; Emdin et al., 2017) which suggested that both overall and abdominal obesity play a causal role on T2DM risk in the European population. In addition, our MR studies suggested that the causal effect of overall obesity on T2DM risk was greater than that of abdominal obesity. Moreover, both BMI [1.054; 1.04–1.068] and WHRadjBMI [1.061; 1.042–1.08] were found to have a causal effect on HbA1c, which suggested that overall and abdominal obesity have a similar but small causal effect on the HbA1c.

Insulin resistance refers to a decreased physiological response of peripheral tissues to insulin action, which implies an impaired effect of insulin in lowering the blood glucose (Gelaye et al., 2010). This serves as the key mechanism and a major global driver of the T2DM condition (Roglic, 2016; Czech, 2017). The accumulation of body fat and abdominal body fat are risk factors for increased insulin resistance (Kohrt et al., 1993; Gobato et al., 2014), and high BMI and WHR were found to be positively correlated with insulin resistance in observational epidemiological studies (Gobato et al., 2014; Benites-Zapata et al., 2019; Lin et al., 2019). Wang et al. reported that higher BMI was causally correlated with increased Stumvoll first- and second-phase insulin secretion and HOMA-IR, whereas no causal relationship between WHR and HOMA-IR was found in a conventional MR study in the Chinese Han population (Wang et al., 2018). In Europeans, a previous MR study found that higher WHRadjBMI was causally associated with higher

TABLE 3 | Mendelian randomization results.

		BMI			WHR			WHRadjBMI		
Outcome	**Method**	**nSNP**	**OR (95% CI)**	***P***	**nSNP**	**OR (95% CI)**	***P***	**nSNP**	**OR (95% CI)**	***P***
T2DM										
	IVW	446	2.741 (2.421, 3.104)	6.12E-57	276	3.12 (2.653, 3.668)	3.87E-43	237	1.993 (1.704, 2.33)	5.59E-18
	MR–PRESSO (Outlier-corrected)	433	3.064 (2.871, 3.271)	9.15E-123	254	3.543 (3.164, 3.968)	2.67E-60	211	2.117 (1.908, 2.35)	3.06E-32
	Multivariable MR	842	3.338 (2.625, 4.246)	1.47E-20	672	2.887 (2.54, 3.281)	3.66E-50	633	2.105 (1.755, 2.524)	8.57E-15
	Weighted median	446	2.835 (2.576, 3.12)	6.18E-101	276	2.769 (2.43, 3.156)	1.02E-52	237	1.983 (1.764, 2.229)	1.74E-30
	MR–Egger	446	2.829 (2.042, 3.918)	9.28E-10	276	1.894 (1.109, 3.234)	0.020	237	1.699 (1.087, 2.656)	0.0201
HOMA-IR										
	IVW	447	1.221 (1.187, 1.255)	6.69E-44	278	1.203 (1.155, 1.252)	4.03E-19	238	1.127 (1.088, 1.167)	3.69E-11
	MR–PRESSO (Outlier-corrected)	446	1.223 (1.19, 1.258)	5.15E-38	278	NA	NA	238	NA	NA
	Multivariable MR	723	1.294 (1.187, 1.411)	1.29E-08	554	1.192 (1.148, 1.237)	1.28E-18	529	1.197 (1.134, 1.262)	1.61E-10
	Weighted median	447	1.214 (1.161, 1.27)	2.79E-17	278	1.18 (1.114, 1.251)	2.25E-08	238	1.128 (1.07, 1.188)	6.90E-06
	MR–Egger	447	1.255 (1.165, 1.352)	4.41E-09	278	1.175 (1.024, 1.348)	0.022	238	1.199 (1.089, 1.32)	2.72E-3
HbA1c										
	IVW	446	1.054 (1.04, 1.068)	4.99E-14	280	1.075 (1.056, 1.095)	1.13E-14	239	1.061 (1.042, 1.08)	3.85E-11
	MR–PRESSO (Outlier-corrected)	441	1.05 (1.036, 1.064)	9.67E-13	278	1.075 (1.056, 1.094)	4.41E-14	235	1.054 (1.037, 1.072)	1.61E-09
	Multivariable MR	735	1.076 (1.031, 1.123)	8.11E-3	569	1.074 (1.055, 1.093)	5.04E-14	543	1.065 (1.038, 1.094)	3.79E-06
	Weighted median	446	1.06 (1.04, 1.081)	3.83E-09	280	1.07 (1.044, 1.096)	4.60E-08	239	1.059 (1.034, 1.085)	2.08E-06
	MR–Egger	446	1.06 (1.024, 1.099)	0.001	280	1.062 (0.998, 1.131)	0.058	239	1.084 (1.033, 1.137)	0.001
Fasting insulin										
	IVW	449	1.202 (1.173, 1.231)	4.73E-50	281	1.193 (1.153, 1.234)	2.28E-24	240	1.102 (1.068, 1.136)	6.37E-10
	MR–PRESSO (Outlier-corrected)	446	1.198 (1.17, 1.226)	4.55E-42	280	1.201 (1.162, 1.241)	3.45E-23	237	1.108 (1.077, 1.141)	2.80E-11
	Multivariable MR	742	1.298 (1.204, 1.4)	6.62E-11	574	1.184 (1.147, 1.223)	5.81E-23	548	1.179 (1.125, 1.235)	2.18E-11
	Weighted median	449	1.213 (1.168, 1.26)	1.54E-23	281	1.187 (1.136, 1.24)	1.41E-14	240	1.129 (1.083, 1.177)	1.03E-08
	MR–Egger	449	1.209 (1.134, 1.289)	1.06E-08	281	1.18 (1.052, 1.324)	0.005	240	1.2 (1.105, 1.304)	2.29E-05
Fasting glucose										
	IVW	449	1.073 (1.048, 1.099)	4.53E-09	281	1.087 (1.054, 1.12)	8.78E-08	240	1.039 (1.012, 1.067)	0.004
	MR–PRESSO (Outlier-corrected)	446	1.083 (1.059, 1.107)	7.61E-12	278	1.098 (1.067, 1.13)	7.19E-10	239	1.044 (1.017, 1.072)	0.001
	Multivariable MR	742	1.025 (0.947, 1.109)	0.546	574	1.067 (1.034, 1.102)	8.18E-05	548	1.019 (0.971, 1.069)	0.443
	Weighted median	449	1.077 (1.041, 1.114)	1.88E-05	281	1.092 (1.048, 1.138)	2.60E-05	240	1.058 (1.017, 1.101)	0.005
	MR–Egger	449	1.067 (1.002, 1.136)	0.043	281	1.091 (0.984, 1.209)	0.098	240	1.075 (1.001, 1.155)	0.049

WHR, waist-to-hip ratio; BMI, body mass index; WHRadjBMI, waist-to-hip ratio adjusted for body mass index; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; T2DM, type 2 diabetes; SD, standard deviation; HOMA-IR, homeostasis model assessment of insulin resistance; HbA1c, hemoglobin A1c; nSNP, numbers of the SNPs (instrumental variable) used in Mendelian randomization; IVW, inverse-variance weighted; MR–egger, Mendelian randomization–Egger regression; MR–PRESSO, Mendelian randomization pleiotropy residual sum and outlier method; OR, odds ratio; CI, confidence interval.

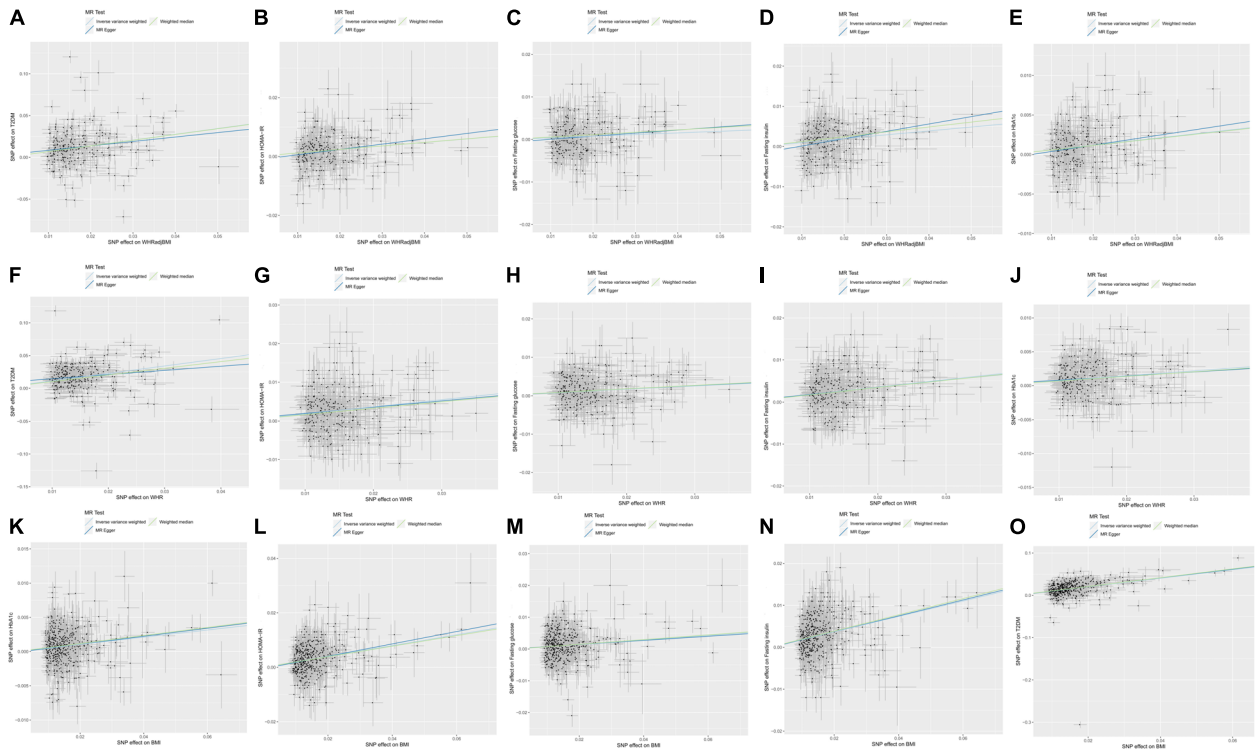

FIGURE 1 | Scatterplot and causal effect of body mass index (BMI), waist-to-hip ratio (WHR), and WHR adjusted for BMI (WHRadjBMI) on type 2 diabetes mellitus (T2DM) and glycemic traits including HOMA-IR, fasting insulin, fasting glucose and glycate hemoglobin (HbA1c). **(A)** Causal effect of WHRadjBMI on T2DM, **(B)** Causal effect of WHRadjBMI on HOMA-IR, **(C)** Causal effect of WHRadjBMI on fasting glucose, **(D)** Causal effect of WHRadjBMI on fasting insulin, **(E)** Causal effect of WHRadjBMI on HbA1c, **(F)** Causal effect of WHR on T2DM, **(G)** Causal effect of WHR on HOMA-IR, **(H)** Causal effect of WHR on fasting glucose, **(I)** Causal effect of WHR on fasting insulin, **(J)** Causal effect of WHR on HbA1c, **(K)** Causal effect of BMI on T2DM, **(L)** Causal effect of BMI on HOMA-IR, **(M)** Causal effect of BMI on fasting glucose, **(N)** Causal effect of BMI on fasting insulin, and **(O)** Causal effect of BMI on HbA1c. The *x*-axis presents the single nucleotide polymorphism (SNP) effect on exposure, and the *y*-axis presents the SNP effect on outcome. The light blue, dark blue and green regression line represents the inverse-variance weighted (IVW), Mendelian randomization (MR)–Egger, and weighted median estimate, respectively.

fasting insulin levels (Emdin et al., 2017). The present MR study provides a similar conclusion with regard to the European population, each SD of genetically higher WHRadjBMI [1.102; 1.068–1.136] and BMI [1.202; 1.173–1.231] was found to play a positive causal effect on higher fasting insulin. Furthermore, each SD of genetically higher BMI [1.221; 1.187–1.255] and WHRadjBMI [1.127; 1.088–1.167] was causally associated with the HOMA-IR. These results suggested that higher overall and abdominal obesity serve as causal risk factors of fasting insulin and insulin resistance in the European population. The findings of the present study are supported by experimental studies as well. Obesity could stimulate the formation of lipid metabolites, hormones, and cytokines, which involves changes in the insulin signaling pathway and the accelerated progression of insulin resistance (Patel and Abate, 2013; Balsan et al., 2015). Moreover, the causal effect of overall obesity on fasting insulin and insulin resistance is slightly greater than that of abdominal obesity. Thus, we highlighted that both mass and distribution of body fat play a causal role on insulin resistance and T2DM risk. This indicates that the development of therapies to modify the mass and distribution of body fat to reduce overall and abdominal obesity might contribute to the prevention and alleviation of T2DM and insulin resistance-related diseases.

Furthermore, although higher BMI and WHRadjBMI was found to be causally associated with higher fasting glucose in our conventional MR in the European population, no statistical significance was found between BMI and fasting glucose (P = 0.546) or with WHRadjBMI and fasting glucose (P = 0.443) in the multivariable MR (DBP, SBP, HDL-C, and LDL-C as covariates). The casual estimates of BMI and WHRadjBMI on fasting glucose in conventional MR might be biased by the horizontal pleiotropy of DBP, SBP, HDL-C, and LDL-C. These negative results warrant further investigation.

Through a comparison of the causal estimates of BMI and WHRadjBMI on glycemic traits (fasting glucose, fasting insulin, HOMA-IR, and HbA1c), this study further emphasizes that overall and abdominal obesity might increase the T2DM risk mainly via elevation of insulin resistance.

In conclusion, overall and abdominal obesity have a causal effect on the T2DM risk and insulin resistance, and overall obesity may have stronger effects, whereas they may have no causal effect on the fasting glucose. These results suggest that individuals can substantially reduce their insulin resistance and T2DM risk

through reduction of body fat mass and modification of body fat distribution.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The Zenodo (https://zenodo.org), the Program in Complex Trait Genomics (https://cnsgenomics.com), the MAGIC Consortium (http://www.magicinvestigators.org/), the MRBase platform (http://www.mrbase.org), and the MRC-IEU Consortium (http://www.bristol.ac.uk/integrative-epidemiology/).

AUTHOR CONTRIBUTIONS

HX and CJ collected, analyzed, and interpreted the data. All authors conceived and designed the project and wrote and approved the final version fo the manuscript.

FUNDING

This research was supported by Key Technology Research and Development Program of Shandong Province (2017CXGC1214).

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.com) for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00603/full#supplementary-material

FIGURE S1 | Leave-one-out sensitive analysis.

REFERENCES

Admiraal, W. M., Holleman, F., Snijder, M. B., Peters, R. J., Brewster, L. M., Hoekstra, J. B., et al. (2014). Ethnic disparities in the association of impaired fasting glucose with the 10-year cumulative incidence of type 2 diabetes. *Diabetes Res. Clin. Pract.* 103, 127–132. doi: 10.1016/j.diabres.2013.10.014

Alejandro, E. U., Gregg, B., Blandino-Rosano, M., Cras-Méneur, C., and Bernal-Mizrachi, E. (2015). Natural history of β-cell adaptation and failure in type 2 diabetes. *Mol. Aspects Med.* 42, 19–41. doi: 10.1016/j.mam.2014.12.002

Balsan, G. A., Vieira, J. L., Oliveira, A. M., and Portal, V. L. (2015). Relationship between adiponectin, obesity and insulin resistance. *Rev. Assoc. Med. Bras.* 61, 72–80. doi: 10.1590/1806-9282.61.01.072

Banerjee, A. T., and Shah, B. R. (2018). Differences in prevalence of diabetes among immigrants to Canada from South Asian countries. *Diabet. Med.* 35, 937–943. doi: 10.1111/dme.13647

Benites-Zapata, V. A., Toro-Huamanchumo, C. J., Urrunaga-Pastor, D., Guarnizo-Poma, M., Lazaro-Alcantara, H., Paico-Palacios, S., et al. (2019). High waist-to-hip ratio levels are associated with insulin resistance markers in normal-weight women. *Diabetes Metab. Syndr.* 13, 636–642. doi: 10.1016/j.dsx.2018.11.043

Bowden, J. (2017). Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. *Int. J. Epidemiol.* 46, 2097–2099. doi: 10.1093/ije/dyx192

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

Censin, J. C., Peters, S. A. E., Bovijn, J., Ferreira, T., Pulit, S. L., Magi, R., et al. (2019). Causal relationships between obesity and the leading causes of death in women and men. *PLoS Genet.* 15:e1008405. doi: 10.1371/journal.pgen.1008405

Czech, M. P. (2017). Insulin action and resistance in obesity and type 2 diabetes. *Nat. Med.* 23, 804–814. doi: 10.1038/nm.4350

Dale, C. E., Fatemifar, G., Palmer, T. M., White, J., Prieto-Merino, D., Zabaneh, D., et al. (2017). Causal associations of adiposity and body fat distribution with coronary heart disease, stroke subtypes, and Type 2 diabetes mellitus: a mendelian randomization analysis. *Circulation* 135, 2373–2388. doi: 10.1161/circulationaha.116.026560

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

Didelez, V., and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.* 16, 309–330. doi: 10.1177/0962280206077743

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* 42, 105–116. doi: 10.1038/ng.520

Emdin, C. A., Khera, A. V., Natarajan, P., Klarin, D., Zekavat, S. M., Hsiao, A. J., et al. (2017). Genetic association of waist-to-hip ratio with cardiometabolic traits. type 2 diabetes, and coronary heart disease. *JAMA* 317, 626–634. doi: 10.1001/jama.2016.21042

Gelaye, B., Revilla, L., Lopez, T., Suarez, L., Sanchez, S. E., Hevner, K., et al. (2010). Association between insulin resistance and c-reactive protein among Peruvian adults. *Diabetol. Metab. Syndrome* 2:30.

Gobato, A. O., Vasques, A. C., Zambon, M. P., Barros Filho Ade, A., and Hessel, G. (2014). Metabolic syndrome and insulin resistance in obese adolescents. *Rev. Paul. Pediatr.* 32, 55–62. doi: 10.1590/s0103-05822014000100010

Gujral, U. P., Pradeepa, R., Weber, M. B., Narayan, K. M., and Mohan, V. (2013). Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations. *Ann. N. Y. Acad. Sci.* 1281, 51–63. doi: 10.1111/j.1749-6632.2012.06838.x

Gupta, R., and Misra, A. (2016). Epidemiology of microvascular complications of diabetes in South Asians and comparison with other ethnicities. *J. Diabetes* 8, 470–482. doi: 10.1111/1753-0407.12378

Huang, T., Qi, Q., Zheng, Y., Ley, S. H., Manson, J. E., Hu, F. B., et al. (2015). Genetic predisposition to central obesity and risk of type 2 diabetes: two independent cohort studies. *Diabetes Care* 38, 1306–1311. doi: 10.2337/dc14-3084

Kettunen, J., Demirkan, A., Wurtz, P., Draisma, H. H., Haller, T., Rawal, R., et al. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* 7:11122. doi: 10.1038/ncomms11122

Kohrt, W. M., Kirwan, J. P., Staten, M. A., Bourey, R. E., King, D. S., and Holloszy, J. O. (1993). Insulin resistance in aging is related to abdominal obesity. *Diabetes Metab. Res. Rev.* 42, 273–281. doi: 10.2337/diab.42.2.273

Lawlor, D., Harbord, R., Jonathan, A. C. S., 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

Lin, C., Chen, K., Zhang, R., Fu, W., Yu, J., Gao, L., et al. (2019). The prevalence, risk factors, and clinical characteristics of insulin resistance in Chinese patients

with schizophrenia. *Compr. Psychiatry* 96:152145. doi: 10.1016/j.comppsych.2019.152145

Lv, J., Yu, C., Guo, Y., Bian, Z., Yang, L., Chen, Y., et al. (2017). Adherence to a healthy lifestyle and the risk of type 2 diabetes in Chinese adults. *Int. J. Epidemiol.* 46, 1410–1420. doi: 10.1093/ije/dyx074

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

Meeks, K. A., Freitas-Da-Silva, D., Adeyemo, A., Beune, E. J., Modesti, P. A., Stronks, K., et al. (2016). Disparities in type 2 diabetes prevalence among ethnic minority groups resident in Europe: a systematic review and meta-analysis. *Intern. Emerg. Med.* 11, 327–340. doi: 10.1007/s11739-015-1302-9

Mokry, L. E., Ross, S., Timpson, N. J., Sawcer, S., Davey Smith, G., and Richards, J. B. (2016). Obesity and multiple sclerosis: a mendelian randomization study. *PLoS Med.* 13:e1002053. doi: 10.1371/journal.pmed.1002053

Patel, P., and Abate, N. (2013). Body fat distribution and insulin resistance. *Nutrients* 5, 2019–2027. doi: 10.3390/nu5062019

Raji, A., Seely, E. W., Arky, R. A., and Simonson, D. C. (2001). Body fat distribution and insulin resistance in healthy Asian Indians and Caucasians. *J. Clin. Endocrinol. Metab.* 86, 5366–5371. doi: 10.1210/jcem.86.11.7992

Razak, F., Anand, S. S., Shannon, H., Vuksan, V., Davis, B., Jacobs, R., et al. (2007). Defining obesity cut points in a multiethnic population. *Circulation* 115, 2111–2118. doi: 10.1161/circulationaha.106.635011

Rees, J. M. B., Wood, A. M., and Burgess, S. (2017). Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat. Med.* 36, 4705–4718. doi: 10.1002/sim.7492

Robiou-du-Pont, S., Bonnefond, A., Yengo, L., Vaillant, E., Lobbens, S., Durand, E., et al. (2013). Contribution of 24 obesity-associated genetic variants to insulin resistance, pancreatic beta-cell function and type 2 diabetes risk in the French population. *Int. J. Obes.* 37, 980–985. doi: 10.1038/ijo.2012.175

Roglic, G. (2016). WHO Global report on diabetes: a summary. *Int. J. Noncommun. Dis.* 1:3. doi: 10.4103/2468-8827.184853

Sampath Kumar, A., Maiya, A. G., Shastry, B. A., Vaishali, K., Ravishankar, N., Hazari, A., et al. (2019). Exercise and insulin resistance in type 2 diabetes mellitus: A systematic review and meta-analysis. *Ann. Phys. Rehabil. Med.* 62, 98–103. doi: 10.1016/j.rehab.2018.11.001

Smit, R. A. J., Trompet, S., Leong, A., Goodarzi, M. O., Postmus, I., Warren, H., et al. (2019). Statin-induced LDL cholesterol response and type 2 diabetes: a bidirectional two-sample Mendelian randomization study. *Pharmacogenom. J.* 20, 462–470. doi: 10.1038/s41397-019-0125-x

Smith, G. D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22. doi: 10.1093/ije/dyg070

Vazquez, G., Duval, S., Jacobs, D. R. Jr., and Silventoinen, K. (2007). Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. *Epidemiol. Rev.* 29, 115–128. doi: 10.1093/epirev/mxm008

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

Wainberg, M., Mahajan, A., Kundaje, A., McCarthy, M. I., Ingelsson, E., Sinnott-Armstrong, N., et al. (2019). Homogeneity in the association of body mass index with type 2 diabetes across the UK Biobank: A Mendelian randomization study. *PLoS Med.* 16:e1002982. doi: 10.1371/journal.pmed.1002982

Wang, T., Zhang, R., Ma, X., Wang, S., He, Z., Huang, Y., et al. (2018). Causal Association of Overall Obesity and Abdominal Obesity with Type 2 Diabetes: A Mendelian Randomization Analysis. *Obesity* 26, 934–942. doi: 10.1002/oby.22167

Weng, L. C., Roetker, N. S., Lutsey, P. L., Alonso, A., Guan, W., Pankow, J. S., et al. (2018). Evaluation of the relationship between plasma lipids and abdominal aortic aneurysm: A Mendelian randomization study. *PLoS One* 13:e0195719. doi: 10.1371/journal.pone.0195719

Wheeler, E., Leong, A., Liu, C. T., Hivert, M. F., Strawbridge, R. J., Podmore, C., et al. (2017). Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS Med.* 14:e1002383. doi: 10.1371/journal.pmed.1002383

Xu, L., and Hao, Y. T. (2017). Effect of handgrip on coronary artery disease and myocardial infarction: a Mendelian randomization study. *Sci. Rep.* 7:954. doi: 10.1038/s41598-017-01073-z

Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., Zheng, Z., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* 9:2941. doi: 10.1038/s41467-018-04951-w

Yang, Q., Millard, L. A. C., and Davey Smith, G. (2019). Proxy gene-by-environment Mendelian randomization study confirms a causal effect of maternal smoking on offspring birthweight, but little evidence of long-term influences on offspring health. *Int. J. Epidemiol.* doi: 10.1093/ije/dyz250 [Epub ahead of print]

Zheng, J., Baird, D., Borges, M. C., Bowden, J., Hemani, G., Haycock, P., et al. (2017). Recent developments in mendelian randomization studies. *Curr. Epidemiol. Rep.* 4, 330–345. doi: 10.1007/s40471-017-0128-6

Zhu, J., Zong, G., Lu, L., Gan, W., Ji, L., Hu, R., et al. (2014). Association of genetic predisposition to obesity with type 2 diabetes risk in Han Chinese individuals. *Diabetologia* 57, 1830–1833. doi: 10.1007/s00125-014-3308-7

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 July 2020
doi: 10.3389/fgene.2020.00638

Characterization of Genetic Diversity and Genome-Wide Association Mapping of Three Agronomic Traits in Qingke Barley (*Hordeum Vulgare* L.) in the Qinghai-Tibet Plateau

Zhiyong Li[1†], Namgyal Lhundrup[2†], Ganggang Guo[1], Kar Dol[3], Panpan Chen[3], Liyun Gao[2], Wangmo Chemi[2], Jing Zhang[1], Jiankang Wang[1], Tashi Nyema[2], Dondrup Dawa[2] and Huihui Li[1,4*]*

[1] Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2] State Key Laboratory of Hulless Barley and Yak Germplasm Resources and Genetic Improvement, Tibet Academy of Agriculture and Animal Sciences, Lhasa, China, [3] Tibet Agricultural and Animal Husbandry College, Nyingchi, China, [4] International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico

OPEN ACCESS

Edited by:
Hailan Liu,
Sichuan Agricultural University, China

Reviewed by:
Haiming Xu,
Zhejiang University, China
Zaifeng Li,
Hebei Agricultural University, China

***Correspondence:**
Huihui Li
lihuihui@caas.cn;
h.li@cgiar.org
Dondrup Dawa
dwdunzhu@126.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: 09 March 2020
Accepted: 26 May 2020
Published: 03 July 2020

Citation:
Li Z, Lhundrup N, Guo G, Dol K, Chen P, Gao L, Chemi W, Zhang J, Wang J, Nyema T, Dawa D and Li H (2020) Characterization of Genetic Diversity and Genome-Wide Association Mapping of Three Agronomic Traits in Qingke Barley (Hordeum Vulgare L.) in the Qinghai-Tibet Plateau. Front. Genet. 11:638. doi: 10.3389/fgene.2020.00638

Barley (*Hordeum vulgare* L.) is one of the most important cereal crops worldwide. In the Qinghai-Tibet Plateau, six-rowed hulless (or naked) barley, called "qingke" in Chinese or "nas" in Tibetan, is produced mainly in Tibet. The complexity of the environment in the Qinghai-Tibet Plateau has provided unique opportunities for research on the breeding and adaptability of qingke barley. However, the genetic architecture of many important agronomic traits for qingke barley remains elusive. Heading date (HD), plant height (PH), and spike length (SL) are three prominent agronomic traits in barley. Here, we used genome-wide association (GWAS) mapping and GWAS with eigenvector decomposition (EigenGWAS) to detect quantitative trait loci (QTL) and selective signatures for HD, PH, and SL in a collection of 308 qingke barley accessions. The accessions were genotyped using a newly-developed, proprietary genotyping-by-sequencing (tGBS) technology, that yielded 14,970 high quality single nucleotide polymorphisms (SNPs). We found that the number of SNPs was higher in the varieties than in the landraces, which suggested that Tibetan varieties and varieties in the Tibetan area may have originated from different landraces in different areas. We have identified 62 QTLs associated with three important traits, and the observed phenotypic variation is well-explained by the identified QTLs. We mapped 114 known genes that include, but are not limited to, vernalization, and photoperiod genes. We found that 83.87% of the identified QTLs are located in the non-coding regulatory regions of annotated barley genes. Forty-eight of the QTLs are first reported here, 28 QTLs have pleotropic effects, and three QTL are located in the regions of the well-characterized genes *HvVRN1*, *HvVRN3*, and *PpD-H2*. EigenGWAS analysis revealed that multiple heading-date-related loci bear signatures of selection. Our results confirm that the barley panel used in this study is highly diverse, and showed a great promise for identifying the genetic basis of adaptive traits. This study should increase our understanding of complex traits in qingke barley, and should facilitate genome-assisted breeding for qingke barley improvement.

Keywords: qingke barley, genetic diversity, GWAS, EigenGWAS, adaptation

INTRODUCTION

Barley (*Hordeum vulgare* L.) was domesticated in Israel and Jordan in the southern part of the Fertile Crescent approximately 10,000 years ago (Badr et al., 2000). With an average world production of 120 Mt annually (Ullrich, 2010), barley ranks fourth among the most important cereal crops in the world (http://faostat.fao.org). Barley is mainly used for food, fodder, alcoholic beverage ingredient, and is generally considered to be a healthful food (Blake et al., 2010; Collins et al., 2010). In Qinghai-Tibet Plateau, six-rowed hulless (or naked) barley, called "qingke" in Chinese or "nas" in Tibetan, is mainly produced in Tibet, and Qinghai, Sichuan, and Yunnan provinces of China. In the Qinghai-Tibet plateau, Tibetans use qingke barley to make wine and for consumption (Tashi et al., 2013). As the main food of Tibetans, qingke barley has been grown on the Qinghai-Tibet Plateau for at least 3,500 years, most probably following its introduction via northern Pakistan, India and Nepal (Zeng et al., 2018). Tibetans have a rich spiritual and cultural connection to qingke barley on the Qinghai-Tibet Plateau due to its wide range of medicinal and nutritional uses. Therefore, analysis of the genetic diversity present in cultivated varieties of qingke barley is especially important.

The adaptation to diverse, high elevation environments makes qingke barley a unique resource for genetic study and barley breeding (Zeng et al., 2015). At present, the genetic architecture of grain starch quality (Li et al., 2014) and drought stress tolerance (Zeng et al., 2016) has been studied in qingke barley, and salt and aluminum tolerance have been studied in Tibetan wild barley (Qiu et al., 2011; Wu et al., 2011; Cai et al., 2013). In other studies, diverse barley lines from different regions, including the US (Zhou and Steffenson, 2013; Genievskaya et al., 2018), Europe (Xu et al., 2018), and India (Visioni et al., 2018), were used to identify the genetic architecture of complex traits (heading time, number of kernels per spike, grain yield) and disease resistance (durable spot, stripe rust) in barley. Although some studies used worldwide collections of barley germplasm, few have included barley varieties from Tibet (Pasam et al., 2012; Gyawali et al., 2017). Over the past decade, studies in barley (Cuesta-Marcos et al., 2008), wheat (Kiseleva et al., 2016), and rice (Yan et al., 2011) have shown that variations in heading date (HD), plant height (PH), and spike length (SL) contribute to environmental adaptation in cereal crops and also influence grain yield. In earlier studies, bi-parental mapping populations were used to reliably detect QTL for HD, PH, and spike morphological traits (Lin et al., 1998; Sameri et al., 2006; Zhang et al., 2009). With the emergence of more cost-effective, high-throughput genotyping technologies, single nucleotide polymorphisms (SNPs) related to HD have been identified by genome-wide association studies (GWAS) (Pasam et al., 2012; Visioni et al., 2013; Genievskaya et al., 2018), PH (Alqudah et al., 2016; Almerekova et al., 2019), leaf area (Alqudah et al., 2018), spike architecture (Comadran et al., 2011) and grain yield (Ingvordsen et al., 2015; Xu et al., 2018) in barley. However, the genetic study of complex agronomic traits in qingke barley is limited (Zhang et al., 2019).

For HD, important genes have been successfully isolated and characterized in barley. Exposure to low temperatures is known as vernalization, which is related to annual differences in seed production and flowering. This process protects the flowering meristem, which is sensitive to the cold, during winter (Yan et al., 2003; Trevaskis et al., 2006). Three genes control the vernalization parameters and growth conditions of barley: *HvVRN1*, *HvVRN2*, and *HvVRN3*. These are found on the respective chromosome arms 5HL, 4HL, and 7HS, all of which have been isolated (Laurie et al., 1995; Yan et al., 2003, 2004, 2006). A MADS-box transcription factor (TF) is encoded by *HvVRN1*, which shares homology with APETALA1, CAULIFLOWER, and FRUITFULL. These are transcription factors that promote flowering in the apical meristem of *Arabidopsis* (Trevaskis et al., 2003; Yan et al., 2003; Trevaskis, 2010). A transcription factor with a zinc finger-CCT domain is encoded by *HvVRN2*. While *Arabidopsis* has no homologous gene, its function is similar to *FLOWERING LOCUS C* (*FLC*), which inhibits flowering (Yan et al., 2004). *FLOWERING LOCUS T* (*FT*) in *Arabidopsis* is similar to *HvVRN3* in that it induces the expression of *HvVRN1* during periods of long daylight, promoting flowering (Yan et al., 2006; Distelfeld et al., 2009). In barley and wheat, *HvVRN3* integrates the photoperiod and vernalization pathways (Distelfeld et al., 2009). Another important pathway is that of the photoperiod, which regulates the date of flowering and heading and uses plant response daylight and optical cues from light receptors. It has been shown that *Ppd-H1* is the ortholog of the wheat *Ppd-D1* gene, a member of the pseudoresponse regulator (*PRR*) gene family via homology-based cloning (Beales et al., 2007). The major determinants of the long-day response in barley are the *Photoperiod-H1* (*Ppd-H1*) and *Photoperiod-H2* (*Ppd-H2*) genes on chromosomes 2H and 1H, respectively (Abdullaev et al., 2017). The results of the study of Turner et al. (2005) suggest that *Ppd-H1* might affect flowering by altering the expression of photoperiod pathway genes that are under circadian control. The dominant allele of *Ppd-H1* regulates response to increased photoperiod length and premature earing during long days. The recessive allele *ppd-H1* induces delays in heading during long days, while *Ppd-H2*, a dominant allele, quickens heading during short days. The recessive allele impedes it.

For PH, semi-dwarf genes include *uzu1*, *ari-e*, and *sdw1* genes are widely used in modern barley improvement (Kuczynska et al., 2013; Dockter and Hansson, 2015). The *ari-e* gene has served in European cultivars and been located on chromosome 5HL (Froster, 2001). The *uzu* gene, the primary dwarfing gene of East Asian barley strains, is located on chromosome 3HL (Zhang, 2000; Chono et al., 2003). Dwarfism regulated by *uzu* is induced by the mutation of one nucleotide interchange in the *HvBRI1* gene, which involves brassinolide in the response (Chono et al., 2003). The chromosome 3HL is also the site of the *sdw1* gene, which is an important dwarfing gene in Europe, North America, South America, and Australia breeding programs (Jia et al., 2009; Xu et al., 2017). The dwarfism controlled by *sdw1* caused by a deletion mutation in the gibberellin 20-oxidase gene (*HvGA20ox2*) (Xu et al., 2017). Previous studies have shown that the QTLs controlling PH and SL are distributed on multiple

TABLE 1 | Analysis of variance (ANOVA) of three traits across three locations.

Trait	Source	DF[e]	Sum of square	Mean square	*F*-Value	Pr > *F*
HD_LS[a]	Genotype	307	22587.70	73.58	9.44	0.00
	Replicate	2	5627.66	2813.83	361.03	0.00
PH_NM[b]	Genotype	307	142700.59	464.82	8.32	0.00
	Replicate	2	27.28	13.64	0.24	0.78
SL_NC[c]	Genotype	256	1016.82	3.97	5.04	0.00
	Replicate	2	6.21	3.11	3.95	0.02
SL_NM[d]	Genotype	307	1010.40	3.29	3.45	0.00
	Replicate	2	5.26	2.63	2.76	0.06

[a] *HD_LS, heading date in Lhasa.*
[b] *PH_NM, plant height in Namling.*
[c] *SL_NC, spike length in Nyingchi.*
[d] *SL_NM, spike length in Namling.*
[e] *DF, degree of freedom.*

chromosomes (Gyenis et al., 2007; Pasam et al., 2012; Fakheri et al., 2018), and that QTLs for PH and SL are identified on different chromosomes in different environments and treatments (Gyenis et al., 2007; Fakheri et al., 2018). In a wild x cultivated barley cross, Gyenis et al. (2007) identified QTLs for PH on chromosomes 1H, 2H, 3H, and 7H, and for SL on chromosomes 1H, 2H, 3H, and 6H. Another study identified QTLs for PH on chromosomes 2H, 3H, 4H, 5H, 6H, and 7H in a spring barley collection (Pasam et al., 2012). A recent study suggests that QTLs for PH are distributed on chromosomes 5H and 7H and for SL on chromosomes 1H, 2H, 5H, and 6H in Western European barley cultivars exposed to drought (Fakheri et al., 2018).

As the growth range of barley increased, it adapted to a wide spectrum of agricultural conditions. Studying selection signals in the barley genome is important to help us understand how this genome reacted to the various agricultural conditions experienced during domestication (Russell et al., 2016). Zeng et al. (2015) resequenced the genomes of 10 Tibetan wild barley accessions to uncover patterns of adaptation to the stressful environment of the Tibetan plateau. Further resequencing of 177 Tibetan barley genomes was performed to better understand the selection markers for the adaptation of local highland barley in the exome capture target range of the genome using the fixation index (F_{ST}) approach (Zeng et al., 2018). Eight regions as possible selective regions were identified, including the location near the *Naked caryopsis* (*nud*) on chromosome 7H. Recently, EigenGWAS, which combines the statistical framework of GWAS with eigenvector decomposition, is a novel approach for identifying regions of the genome under selection in any genetic data where the underlying population structure is unknown. EigenGWAS has been applied to studies in evolution, ecology, breeding, and human genetics (https://github.com/gc5k/GEAR/wiki/EigenGWAS).

In the present study, we collected old local qingke barley landraces in Tibet, modern qingke barley varieties, and representative qingke barley varieties from regions surrounding the Tibetan region. The genetic diversity was compared between the landraces and the two variety groups, and the trends in the changes in genetic structure from the landraces to the breeding varieties was considered. In this study, therefore, our objectives were to use the 14,970 high quality SNPs discovered using genotyping-by-sequencing (tGBS) in 308 qingke barley accessions to (1) understand the genetic diversity in the landraces and the modern varieties and the changes in population structure that occurred going from the landraces to the breeding varieties, (2) identify genetic loci associated with HD, PH, and SL by GWAS, and (3) identify loci that underwent selection for environmental adaptation using EigenGWAS. The findings of this study could facilitate a better understanding of the genetic mechanisms underlying the establishment of adaptive traits and genome-assisted selection in qingke barley breeding.

MATERIALS AND METHODS

Plant Materials

A total of 308 qingke barley accessions were used in this study; 206 qingke landraces, 72 qingke varieties, and 30 varieties (including 18, 5, 1, and 6 varieties from Qinghai, Gansu, Yunnan, and Sichuan provinces of China, respectively). All the 308 accessions were planted in Tibet at three locations; Lhasa (N29°36′, E91°06′) in April 2018, Namling (N29°18′, E88°46′) on May 2017, and Nyingchi (N29°39′, E94°21′) in October 2017 with three replicates each. We used a randomized design to construct the field experiment. At each location, 30 seeds of each accession were planted in a plot with two rows of 150 cm long and 30 cm between rows. HD was measured as the number of days when the head first emerged from the flag leaf sheath on the main shoot in a plot (Zadoks scale, $Z = 50$; Hemming et al., 2009). The PH was measured as the above-ground plant height without the awns. The SL was measured as the length from the base of main spike to the tip of main spike (excluding awns). All traits were measured as the average of five random plants.

Phenotypic Data Analysis

The Pearson's correlation coefficients between the traits and the broad sense heritability (H^2) of target traits were calculated by AOV functionality in QTL IciMapping v.4.1 (Meng et al., 2015). In the analysis of variance of the three traits, variance

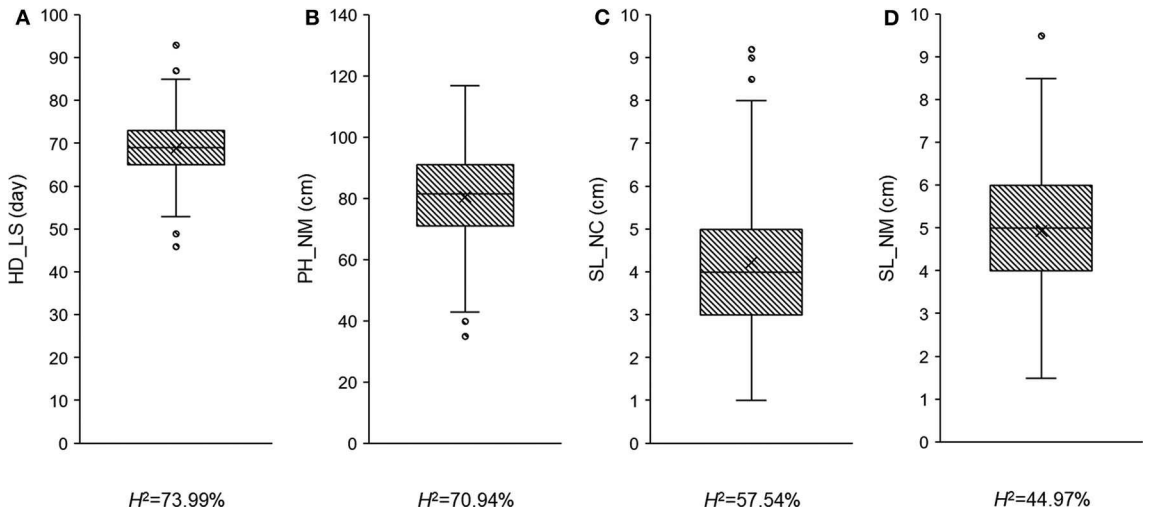

FIGURE 1 | The phenotypic distribution and heritability in broad sense for heading date in Lhasa (HD_LS) **(A)**, plant height in Namling (PH_NM) **(B)**, spike length in Nyingchi (SL_NC) **(C)**, and spike length in Namling (SL_NM) **(D)**.

components were estimated from a linear model; phenotype was partitioned into overall mean, genotypic effect, replication effect (i.e., location), and random error effect, all of which were treated as fixed effects. The H^2 on plot level was estimated from the following equation:

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_\varepsilon^2},$$

where σ_G^2 is the genetic variance and σ_ε^2 is the variance of the error. Although 308 accessions were planted at three locations, HD, PH, and SL were not all measured in three locations. Only SL has high-quality data in two locations (Nyingchi and Namling) in Tibet, abbreviated as SL_NC and SL_NM, respectively. For HD ad PH, phenotype from one location was used, HD in Lhasa (abbreviated as HD_LS) and PH in Namling (abbreviated as PH_NM), since from other two locations either the H^2 were lower than 30%, or only one measurement was available for each plant. Considering data with low heritability was not reliable to conduct GWAS, and data with no replication could not be used to estimate the H^2 and evaluate the data quality, we discarded the low-quality data. For clarity, HD_LS, PH_NM, SL_NC, and SL_NM were used in the following-up analysis.

SNP Genotyping and Genotypic Data Analyses

The 308 accessions were genotyped using a newly developed genotyping-by-sequencing technology (tGBS) that eases the process of sorting high-quality GBS sequencing libraries and results in more accurate SNP calling (Ott et al., 2017; Li et al., 2019). Sequence reads were aligned to the *Hordeum vulgare* Hv IBSC PGSB v2 reference genome (Mascher et al., 2017) after de-barcoding and trimming. SNP calling was conducted using only those reads that aligned to a single location in the reference genome. In total, 46,034 polymorphic sites for each accession were discovered, and the data was filtered as follows: missing values $\leq$0.4, heterozygosity rate (Het. Rate) $\leq$0.2, and minor allele frequency (MAF) $\geq$0.05 (**Supplementary Table 1**). After filtering, 14,970 high-quality SNPs were retained in the follow-up analysis. To assess population diversity, genome-wide pairwise linkage disequilibrium (LD) was calculated between SNP pairs to investigate the potential of the array to capture all significant regions associated with the observed phenotypes using the software package TASSEL v5.2 (Bradbury et al., 2007). LD was estimated by using the squared allele-frequency correlation (r^2; Weir and Cockerham, 1996) for pairs of loci, since r^2 is affected not only by recombination frequencies at the two sites, but also by the differences in allele frequencies between sites. Decay of LD was evaluated, as was the distance between sites in base pairs (bp) with non-linear regression as implemented in the R package (Remington et al., 2001). To avoid multiple significances within individual LD blocks, the support interval was determined when the decay distance of LD reached $r^2 = 0.5$. Nucleotide diversity (π) across the barley genome was calculated with TASSEL v5.2. The population structure of the 308 accessions was evaluated using principle component analysis (PCA) and a phylogenetic tree. Pairwise distances were estimated between genotyped individuals using an unbiased model of substitution frequencies. Distance estimates were then used to construct a phylogenetic tree using the Neighbor-Joining-like algorithm described by Saitou and Nei (1987) and implemented in the NJS module of the APE R package (Paradis et al., 2004). Unlike conventional neighbor-joining methods, the NJS algorithm is tolerant of missing data, enabling its use with GBS data. Relative branch lengths are proportional to the amount of divergence observed between individuals. The effective sample size was calculated according to the method in Powell et al. (2010) as implemented in the software GEAR.

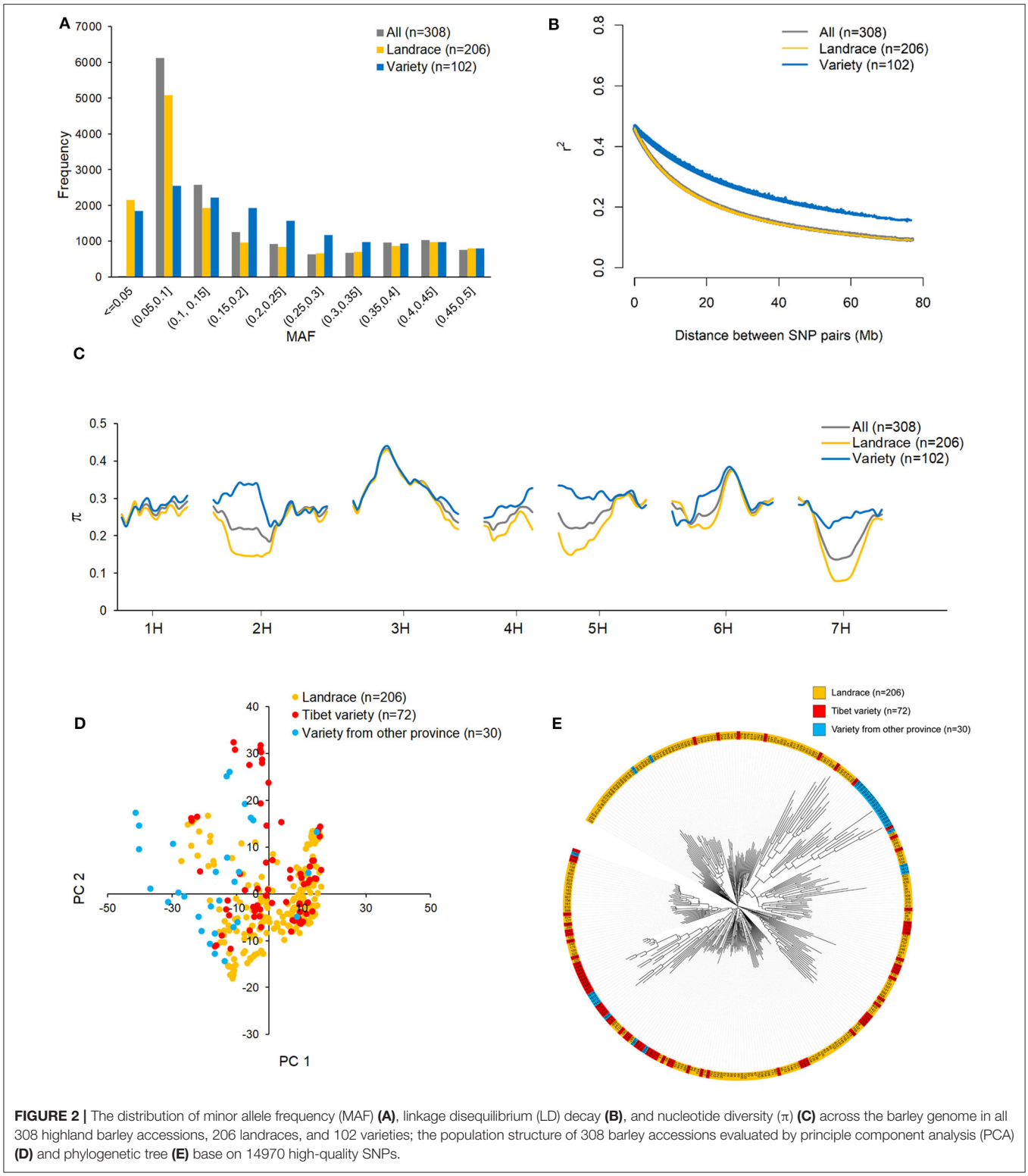

FIGURE 2 | The distribution of minor allele frequency (MAF) **(A)**, linkage disequilibrium (LD) decay **(B)**, and nucleotide diversity (π) **(C)** across the barley genome in all 308 highland barley accessions, 206 landraces, and 102 varieties; the population structure of 308 barley accessions evaluated by principle component analysis (PCA) **(D)** and phylogenetic tree **(E)** base on 14970 high-quality SNPs.

GWAS Analysis

A GWAS for the three agronomic traits was conducted with a general linear model (GLM) and a mixed linear model (MLM) as implemented in TASSEL v5.2 software (Bradbury et al., 2007). For both models, the first principal component of the PCA was fitted as the cofactor to exclude the effect of population structure. In MLM, a variance–covariance kinship matrix, as covariates to estimate the association between phenotypes and genotypes (Zhang et al., 2010), was also considered. To declare QTL from the GWAS results, the phenotypic observation of SL

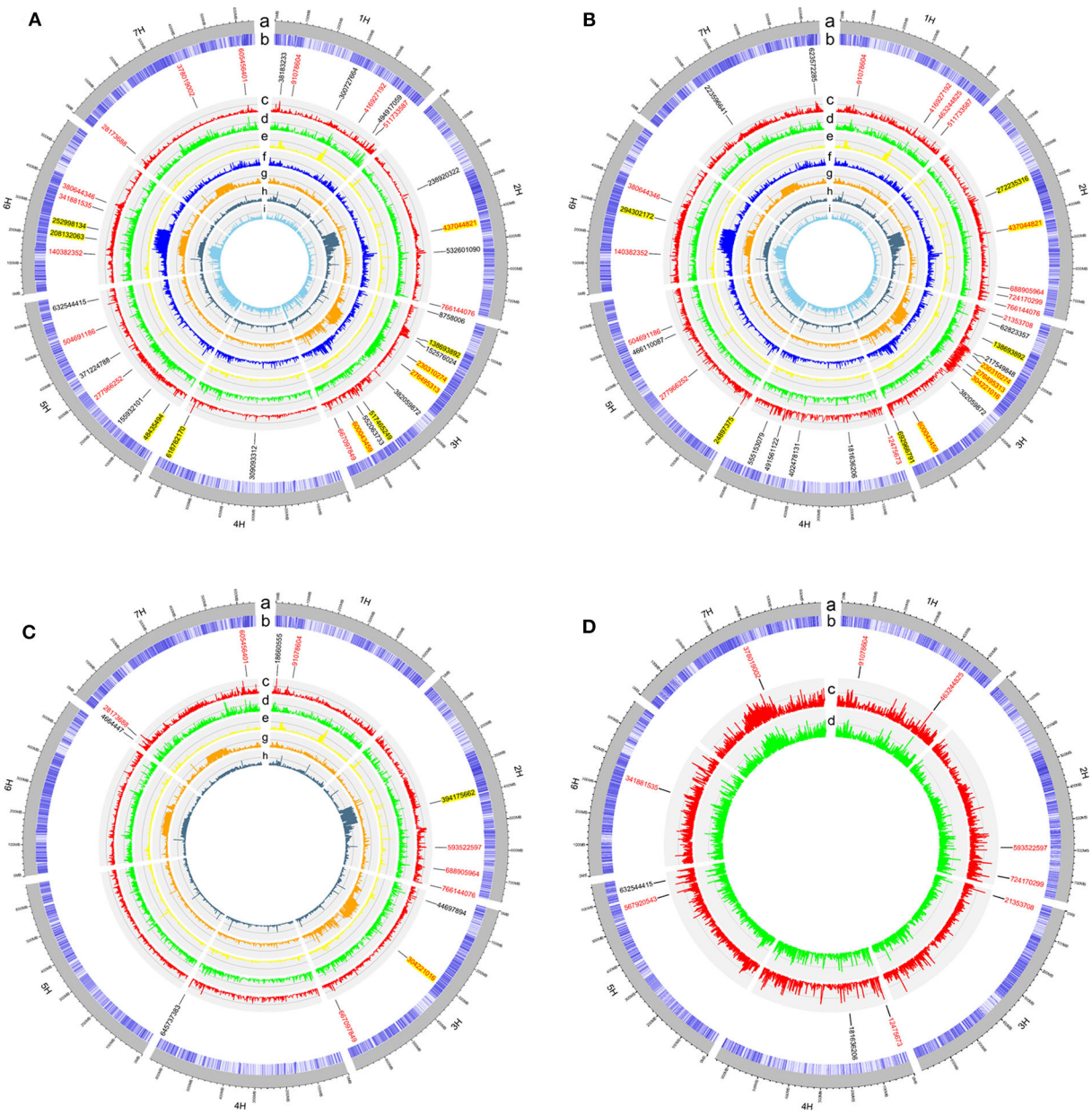

FIGURE 3 | The circular plots for heading date in Lhasa (HD_LS) **(A)**, plant height in Namling (PH_NM) **(B)**, spike length in Nyingchi (SL_NC) **(C)**, and spike length in Namling (SL_NM) **(D)**. From the outer circle to the inner circle, a is for the barley genome; b is for the SNP density; c is for the manhattan plot from generalized linear model (GLM); d is for the manhattan plot from mixed linear model (MLM); e is for the manhattan plot from EigenGWAS under the tenth eigenvector (EV10); f is for the manhattan plot from EigenGWAS under the seventh eigenvector (EV7); g is for the manhattan plot from EigenGWAS under the fifth eigenvector (EV5); h is for the manhattan plot from EigenGWAS under the third eigenvector (EV3); and i is for the manhattan plot from EigenGWAS under the second eigenvector (EV2). The SNP positions associated with the trait of interest were marked in black font; of which with pleotropic effects were highlighted in red font; and detected by EigenGWAS were highlighted in yellow background.

was reshuffled 1,000 times to analyze the null distribution. We calculated the 95th quantile of the 1,000 most significant p-values over 1,000 permutations to be 5.18 after log_{10} transformation. The Bonferroni correction, $-log_{10}(1/14{,}970) = 4.18$, was also calculated. To balance the false positives and false negatives, a $-log_{10}(P)$ threshold of 4.00 was used for the GLM and 3.00 was used for the MLM. To determine whether the uncovered genetic architecture was appropriate, the identified QTL was used to predict the performance of the corresponding trait. The most significant SNP in each QTL region was fitted in the linear model with the original trait performance as the dependent variable. The adjusted coefficient of determination (R^2) from the linear

model was then calculated. The performance of QTL in different locations were estimated by $a = \frac{1}{e}\sum_{i=1}^{e} a_i$ and $ae_i = a_i - a$, where a was the averaged effect of QTL across locations, a_i was the additive effect of QTL for each location, e is the number of locations, and ae_i was the additive by environment effect of QTL in each location (Li et al., 2015).

Analysis of Gene Annotation and Enrichment

We used SnpEff to conduct functional annotations and effect predictions of the target SNPs (Cingolani et al., 2012). The Barley Hv_IBSC_PGSB_v2 reference genome gene annotation was downloaded as a gff3 file from the Ensembl plants database (http://plants.ensembl.org/index.html). Gene annotation information was acquired by BARLEX: The Barley Genome Explorer (https://apex.ipk-gatersleben.de/apex/f?p=284:10:::::::; Colmsee et al., 2015). A Singular Enrichment Analysis (SEA) tool was used to perform a functional enrichment analysis of the annotated genes (Tian et al., 2017).

EigenGWAS Analysis

EigenGWAS is a regression approach based on principal component analysis (Chen et al., 2016; Li et al., 2019). It is similar to GWAS; however, the phenotype is replaced with an eigenvector (EV) to capture genetic variation in the studied population. In this study, EigenGWAS, implemented in the software GEAR (https://github.com/gc5k/GEAR), was used to separate loci under selection by treating top 10 eigenvectors (i.e., EV1-EV10) as phenotypes. We adjusted the p-value using a genomic control factor, denoted as P_{GC}, to exclude the effect of the genetic drift (Devlin and Roeder, 1999), and used the P_{GC} to identify loci under selection. We reshuffled the first eigenvector 1,000 times to identify the significance cutoff for the relevant loci, which helped us analyze the null distribution. We calculated the 95th quantile of the 1,000 most significant p-values over 1,000 permutations to be 5.75 after log_{10} transformation. Considering the Bonferroni correction 4.18 as mentioned above, a $-log_{10}(P)$ threshold of 4.00 was applied for EigenGWAS analyses in all 10 eigenvectors.

RESULTS

Phenotypic Variation and Correlation Analysis

To determine whether the observed traits exhibit wide variation, are highly heritable, and/or display a normal distribution, the recorded phenotypic data was analyzed using ANOVA (**Table 1** and **Supplementary Table 2**) and boxplots (**Figure 1**). Fifty-one plants had no measurement for SL_NC, so the degrees of freedom in this case was only 256 (**Table 1**). All the variance components were significant ($P < 0.05$) across trials, with the exception of the replicates in PH_NM and SL_NM (**Table 1**). Wide variations ranging from 46 to 93 days in HD_LS, from 35 to 117 cm in PH_NM, from 1 to 9.2 cm in SL_NC, and from 1.5 to 9.5 cm in SL_NM were observed in the collection of 308 qingke barley accessions (**Figure 1**). The SL distribution showed that the SL_NM mean was higher than it was for SL_NC (**Figures 1C,D**), and the correlation between SL_NM and SL_NC was 0.21 ($P < 0.01$; **Supplementary Figure 1**). The reason for this may be due to the big environmental difference between Namling (4,000 m above sea level) and Nyingchi (2,995 m above sea level) and the overcast and rainy weather in Nyingchi at flowering time, which was not conducive to pollination and thus decreased the effective seed-setting rate of the barley spikes. The broad-sense heritabilities for the three observed traits ranged from 44.97 to 73.99% (**Figure 1**). The highest correlation was between PH_NM and SL_NM (i.e., 0.48 with $P < 0.01$; **Supplementary Figure 1**), and there was a negative correlation between SL_NC and HD_LS. These observations are consistent with the general experience regarding the relationships between PH and SL (Wang et al., 2010), and between SL_NC and HD_LS (Wang et al., 2010; Al-Tabbal and Al-Fraihat, 2012).

Genetic Diversity and Population Structure in the 308 Qingke Barley Accessions

The MAF distributions of all 14,970 SNPs in the whole dataset and in the landrace and variety subpopulations are shown in **Figure 2A**. Because the 14,970 SNPs were filtered to remove those with MAF <0.05 in the 308 accessions, the minimum MAF here is 0.05, and the average MAF is 0.183. The MAF ranged from 0 to 0.5 in both subpopulations. SNPs with MAF <0.05 were considered to be rare SNPs. In this sense, more rare SNPs were observed in the landrace subpopulation (2,150) than in the variety subpopulation (1,841). The numbers of SNPs with MAFs ranging from 0.05 to 0.1 were 5,086 and 2,545 in the landrace and variety subpopulations, respectively. This suggests that more low MAF SNPs are present in the landrace subpopulation than in the variety subpopulation. Non-linear models of LD decay for the 206 landraces and 102 varieties are shown in **Figure 2B**. In general, LD in both datasets showed an intermediate rate of decline. The predicted value of r^2 declined to 0.5 within 1 Mb, which is considered to be the length of the support interval. As expected, LD decayed faster in the landrace subpopulation than in the variety subpopulation. The predicted value of r^2 declined to 0.2 within 23 Mb for the landraces and within 49 Mb for the varieties. It remained >0.1 for over 80 Mb in the varieties. Due to the different allele distribution of the SNPs in the two subpopulations, nucleotide diversity (π) in the variety subpopulation was higher than in the landrace subpopulation, particularly on chromosomes 2H, 4H, 5H, and 7H (**Figure 2C**).

To determine whether the population structure could be discerned from the whole-genome genotyping data, PCA (**Figure 2D**) and a phylogenetic analysis (**Figure 2E**) were conducted for the 308 accessions. Based on the PCA plot, the two subpopulations, landraces and varieties, could not be clearly separated. This is likely due to a large proportion of the varieties being derived from qingke barley landraces. However, from the phylogenetic tree, it cannot be ruled out that the 15 varieties from Gansu and Qinghai provinces were not derived from the Tibetan landraces (**Figure 2E**).

TABLE 2 | QTL identified by GWAS using generalized linear model (GLM) and mixed linear model (MLM) and EigenGWAS.

Chr.	Pos. (bp)	GLM		MLM		*F*st	$-log_{10}(P_{GC})$[a]	Annotation	References
		$-log_{10}(P)$	Trait	$-log_{10}(P)$	Trait				
1H	18,660,555	7.81	SL_NC					upstream_gene_variant	Mikołajczak et al., 2017; Hu et al., 2018
1H	38,183,233	12.52	HD_LS					intergenic_region	
1H	91,078,604	5.36	PH_NM, SL_NC, HD_LS	3.66	SL_NM, PH_NM			intergenic_region	
1H	300,727,664	6.41	HD_LS	4.68	HD_LS			intergenic_region	
1H	416,927,192	7.08	HD_LS, PH_NM	3.47	PH_NM			intergenic_region	Genievskaya et al., 2018; Hill et al., 2019
1H	463,244,825	5.18	SL_NM	4.67	SL_NM, PH_NM			intergenic_region	
1H	494,917,059	8.18	HD_LS					upstream_gene_variant	
1H	511,733,587	8.23	HD_LS, PH_NM					intergenic_region	Alqudah et al., 2014; Almerekova et al., 2019
2H	238,920,322	8.39	HD_LS					intergenic_region	
2H	272,235,316	5.16	PH_NM			0.34	6.06 (EV3)	intergenic_region	
2H	394,175,662	4.23	SL_NC			0.30	7.34 (EV3), 5.61 (EV10)	intergenic_region	
2H	437,044,821	6.30	PH_NM, HD_LS	3.37	PH_NM	0.13	4.97 (EV3)	intergenic_region	
2H	532,601,090	10.55	HD_LS					intergenic_region	Pasam et al., 2012; Pauli et al., 2014
2H	593,522,597	4.74	SL_NC, SL_NM	3.61	SL_NM			downstream_gene_variant	Wang et al., 2016
2H	688,905,964	4.70	PH_NM, SL_NC	3.21	SL_NC			intergenic_region	
2H	724,170,299	5.47	PH_NM, SL_NM	3.08	SL_NM, PH_NM			upstream_gene_variant	Comadran et al., 2011; Pasam et al., 2012; Hu et al., 2018
2H	766,144,076	6.92	HD_LS, SL_NC, PH_NM	4.82	HD_LS			intron_variant	Genievskaya et al., 2018
3H	8,758,006	7.46	HD_LS					intergenic_region	
3H	21,353,708	5.49	PH_NM, SL_NM	3.04	SL_NM			intergenic_region	
3H	44,697,894	4.23	SL_NC	3.90	SL_NC			intergenic_region	
3H	62,823,357	4.17	PH_NM					intergenic_region	
3H	138,693,892	8.62	HD_LS	3.14	PH_NM	0.10	6.17 (EV10)	intergenic_region	
3H	152,576,024	11.42	HD_LS					intergenic_region	
3H	217,549,848	7.23	PH_NM	3.99	PH_NM			intergenic_region	
3H	230,310,274	6.67	HD_LS, PH_NM			0.57	4.48 (EV5)	intergenic_region	
3H	276,495,313	6.28	HD_LS, PH_NM			0.72	4.47 (EV5)	intergenic_region	
3H	304,221,016	7.79	PH_NM	3.39	PH_NM, SL_NC	0.71	5.25 (EV5)	intergenic_region	
3H	382,059,872	10.31	HD_LS	3.36	PH_NM			intergenic_region	
3H	517,465,249	11.80	HD_LS			0.46	4.33 (EV7)	intergenic_region	
3H	552,063,733	10.20	HD_LS					intergenic_region	
3H	600,043,459	5.30	PH_NM, HD_LS			0.08	6.73 (EV10)	intergenic_region	Tondelli et al., 2013
3H	667,097,849	10.02	HD_LS, SL_NC					intergenic_region	
3H	692,966,791	8.20	PH_NM	3.78	PH_NM	0.18	6.56 (EV2)	intergenic_region	
4H	12,475,673	4.41	SL_NM, PH_NM	3.10	SL_NM, PH_NM			upstream_gene_variant	Pauli et al., 2014
4H	181,636,206	4.43	PH_NM	3.05	SL_NM			intergenic_region	
4H	309,093,312	7.06	HD_LS					intergenic_region	
4H	402,478,131	5.10	PH_NM	3.41	PH_NM			intergenic_region	
4H	491,561,122	6.24	PH_NM	3.37	PH_NM			intergenic_region	Tondelli et al., 2013
4H	555,153,079	5.60	PH_NM					intergenic_region	
4H	618,782,170	16.60	HD_LS			0.09	6.17 (EV10)	intergenic_region	Pauli et al., 2014; Almerekova et al., 2019
4H	645,737,383	4.84	SL_NC					intergenic_region	

(Continued)

TABLE 2 | Continued

Chr.	Pos. (bp)	GLM		MLM		*F*st	$-log_{10}(P_{GC})$[a]	Annotation	References
		$-log_{10}(P)$	Trait	$-log_{10}(P)$	Trait				
5H	24,897,375	8.10	PH_NM	3.16	PH_NM	0.17	4.44 (EV2)	downstream_gene_variant	
5H	48,435,494	9.26	HD_LS			0.47	4.91 (EV7)	intergenic_region	
5H	155,932,101	7.00	HD_LS	4.92	HD_LS			intergenic_region	
5H	277,966,252	6.31	HD_LS, PH_NM					intergenic_region	
5H	371,224,788	8.44	HD_LS					intergenic_region	
5H	466,110,087	5.15	PH_NM					intergenic_region	
5H	504,691,186			4.05	PH_NM, HD_LS			intergenic_region	
5H	567,920,543	4.13	SL_NM	3.41	SL_NM			intergenic_region	
5H	632,544,415	7.62	HD_LS	3.28	SL_NM			downstream_gene_variant	
6H	140,382,352	7.72	PH_NM, HD_LS					intergenic_region	
6H	208,132,063	4.43	HD_LS			0.27	4.82 (EV2), 5.21 (EV7)	intergenic_region	Genievskaya et al., 2018
6H	252,998,134	4.19	HD_LS			0.21	4.02 (EV2)	intergenic_region	
6H	294,302,172	4.97	PH_NM	4.07	PH_NM	0.33	4.88 (EV2), 4.09 (EV7)	intergenic_region	
6H	341,881,535	5.68	HD_LS	4.89	HD_LS, SL_NM			intergenic_region	
6H	380,644,346	6.30	HD_LS, PH_NM					intergenic_region	
7H	4,664,447	5.28	SL_NC					upstream_gene_variant	
7H	28,173,688	10.17	SL_NC, HD_LS					intergenic_region	
7H	223,596,641			4.86	PH_NM			intergenic_region	Pham et al., 2019
7H	378,019,002	4.09	SL_NM	4.17	HD_LS, SL_NM			intergenic_region	
7H	605,456,401	4.38	HD_LS	5.10	HD_LS, SL_NC			upstream_gene_variant	
7H	623,572,285	5.96	PH_NM	3.29	PH_NM			intergenic_region	Hu et al., 2018; Almerekova et al., 2019; Pham et al., 2019

[a]*Corrected p-value of EigenGWAS. Blank means the QTL was not identified by the corresponding method.*

GWAS and EigenGWAS

In the GWAS, a total of 62 QTLs distributed across the barley genome that control three agronomic traits were identified either by GLM or by MLM (**Figure 3** and **Table 2**). To evaluate if the first PC as cofactor was appropriate, no PC and PC number with 2–5 were also used to conduct GWAS (**Supplementary Figures 2–11**). Results showed that the parameter estimation would be inflated if no PC as cofactor in GWAS model. The parameter estimations from PC number 1–5 were fairly the same. Of the 62 QTLs, the largest number of QTLs, 16, was distributed on chromosome 3H, and the lowest number (6) was distributed on chromosomes 6H and 7H. There were 29 QTLs (46.7%) that were detected by both GLM and MLM; 16 QTLs were declared as selection loci by EigenGWAS under five eigenvectors (i.e., EV2, EV3, EV5, EV7, and EV10); six QTLs were reported by other studies, and six QTLs were consistently identified by GLM, MLM, and EigenGWAS (**Table 2**). In total, 28 QTLs had pleotropic effects (red text in **Figures 3**, **4A**). One QTL with pleotropic effects located at 91,078,604 bp on chromosome 1H was associated with all three traits in four trials. A QTL at 766,144,076 bp on chromosome 2H was related to the three traits HD_LS, PH_NM, and SL_NC, and was also associated with PH, days to seed maturation (SMT), peduncle length (PL), and HD, as reported by Genievskaya et al. (2018). Of 28 pleotropic-effect QTLs, six were detected by EigenGWAS as well, and these are shown in red text highlighted in yellow in **Figures 3A–C** and **Table 2**. These are the QTLs located at 437,044,821 bp on chromosome 2H by EV3, 138,693,892 bp on chromosome 3H by EV10, 230,310,274 bp on chromosome 3H by EV5, 276,495,313 bp on chromosome 3H by EV5, 304,221,016 bp on chromosome 3H by EV5, and 600,043,459 bp on chromosome 3H by EV10 (**Figure 3** and **Table 2**). Two QTLs at 91,078,604 bp on chromosome 1H and at 593,522,597 bp on chromosome 2H associated with SL were both detected in two locations (**Figures 3**, **4A** and **Table 2**).

In general, 36, 33, 12, and 11 QTLs were associated with HD_LS, PH_NM, SL_NC, and SL_NM, respectively (**Figure 4B** and **Table 2**), and are positively correlated with the broad sense heritabilities (**Figure 1**). In our study, we were able to investigate pleiotropy of QTLs on multiple traits. We observed that there were 16 QTLs for HD_LS and PH_NM in common. However, the correlation between HD_LS and PH_NM was not significant (**Supplementary Figure 1**), which may due to the repulsion linkage phase of the 16 QTLs (**Supplementary Table 4**). In contrast, six PH_NM QTLs were also significant for SL_NM, and the correlation between these two traits was 0.48, which was highly significant. The reason for this may be that the six QTLs are in coupling linkage

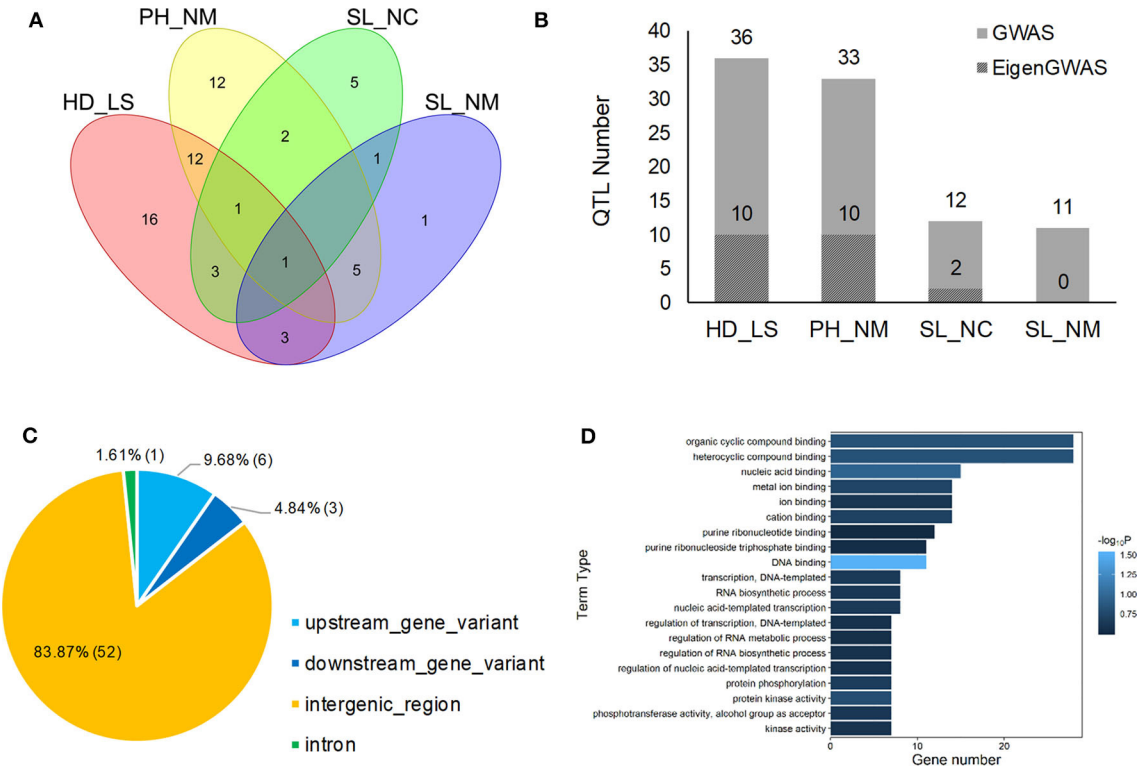

FIGURE 4 | Venn plot of QTL distribution for HD_LS, PH_NM, SL_NC, and SL_NM **(A)**, the distribution of QTL number for heading date in Lhasa (HD_LS), plant height in Namling (PH_NM), spike length in Nyingchi (SL_NC), and spike length in Namling (SL_NM) **(B)**; and gene annotation **(C)** and ontology **(D)** for the 62 QTL identified by GWAS.

phase (**Supplementary Table 4**). To evaluate the performance of QTL associated with SL in different locations, genotype by environment effects were estimated (**Supplementary Figure 12** and **Supplementary Table 5**). The most significant genotype-by-environment QTLs identified by both GLM and MLM were at 593,522,597 bp on chromosome 2H and 21,353,708 bp on chromosome 3H, since their additive effects were both significant in Nyingchi, but not in Namling. In addition, there were three genotype-by-environment QTLs identified by GLM on chromosomes 1H, 2H, and 7H, respectively.

Candidate Gene Annotation and Enrichment

The annotation conducted on the 62 significant QTLs identified by GWAS (**Table 2**) showed that 52 (83.87%) of QTL regions are intergenic, and that 10 (16.13%) are genic (**Figure 4C**). This is consistent with the *Hordeum vulgare* Hv IBSC PGSB v2 reference genome, where 19.2% of the barley genome is genic (Mascher et al., 2017) and a high ratio of loci (78.00%) related to phenotypic variation are identified in intergenic regions (Mei et al., 2017). Of the QTL, 9.68% and 4.84% were located in the upstream and downstream gene regions, and 1.61% of the QTL were in the intron regions (**Figure 4C**). In total, 114 known genes were mapped as significant QTLs in the GWAS, and most of them were assigned to the "molecular function" and "biological process" categories in gene ontology (GO) analysis (**Figure 4D**). One QTL, located at 605,456,401 bp on chromosome 7H, controls HD_LS and SL_NC, and is 2.1 Kb upstream of *HORVU7Hr1G100540*, a known gene that encodes an SBP (S-ribonuclease binding protein) family protein. A QTL at 24,897,375 on chromosome 5H significantly associated with PH_NM was found to be located 799 bp downstream of the gene *HORVU5Hr1G009980* that encodes a tetratricopeptide repeat (TPR)-like superfamily protein. For SL_NM and SL_NC, a stable QTL at 593,522,597 bp on chromosome 2H is located 554 bp downstream of *HORVU2Hr1G081800*, which encodes a WPP domain interacting protein 2 (**Supplementary Table 3**). In addition, screening of the associated mapping population identified variations in HD, and we found a significant SNP (5H: 599,361,872) near the vernalization gene *HvVRN1*, a significant SNP (7H: 38,508,938) near the vernalization gene *HvVRN3*, and a significant SNP (1H: 514,145,049) near the photoperiod gene *PpD-H2*.

Phenotype Prediction

To determine the accuracy of the QTL effect estimation, we used the significant QTL additive effect estimates to predict the phenotypic observations for the three traits, and were able to accurately predict HD_LS (R^2 = 69.20%), PH_NM (R^2 = 64.07%), and SL_NC (R^2 = 42.37%) (**Figure 5**). For SL_NM, the prediction was low, due in part to the low heritability of SL in

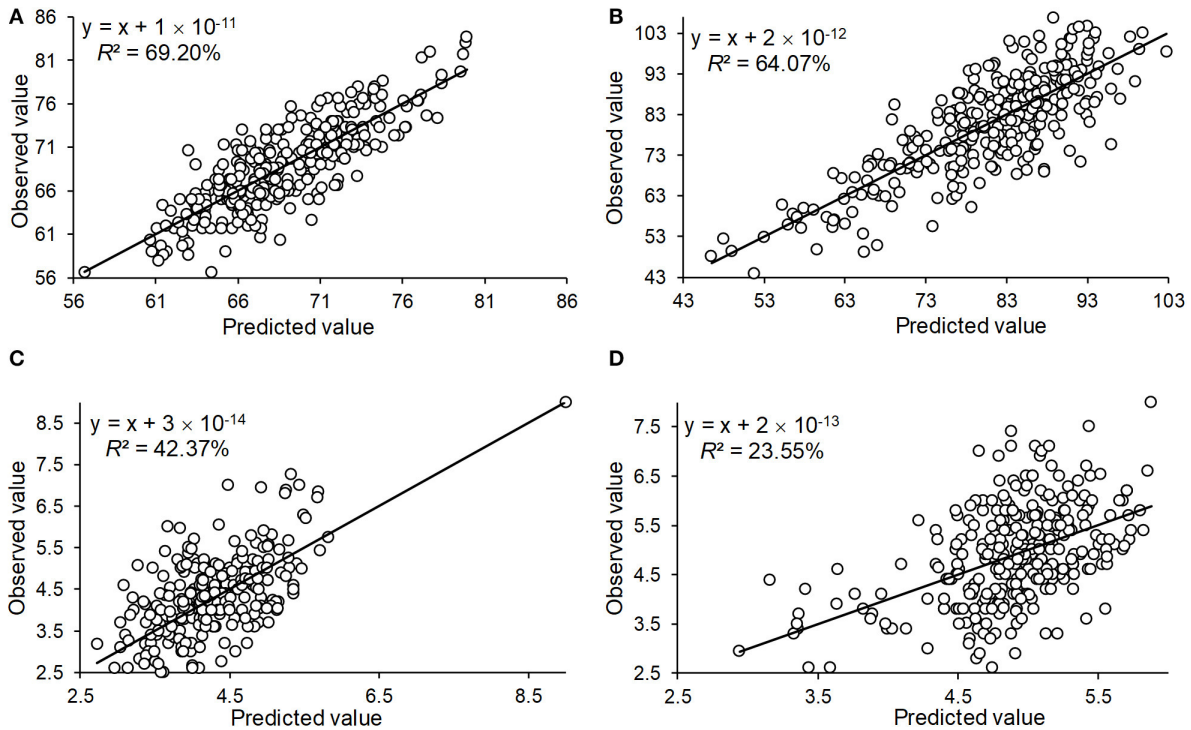

FIGURE 5 | The prediction of the observed phenotype by the QTL identified by GWAS for the traits of heading date in Lhasa (HD_LS) **(A)**, plant height in Namling (PH_NM) **(B)**, spike length in Nyingchi (SL_NC) **(C)**, and spike length in Namling (SL_NM) **(D)**.

NM (**Figure 1**). Looking at the broad sense heritabilities for the three traits (**Figure 1**) suggests that the QTL results presented in this study are reliable, and provide further evidence that a large proportion of the phenotypic variation can be explained by additive variance in this association panel.

DISCUSSION

To the best of our knowledge, few genetic studies have investigated the complex agronomic traits in Tibetan qingke barley (Zhang et al., 2019). Previous reports have included only a limited number of qingke barley accessions to identify potential signals of adaptation and domestication. For example, 95 wild barley accessions from Tibet and 28 six-rowed hulless barley varieties from Tibet and Xinjiang were used to show that the Tibetan Plateau and the surrounding areas are primary centers of barley cultivation (Dai et al., 2012); six wild-barley genotypes collected from the Tibetan Plateau were used in an RNA-seq analysis to reveal multiple origins of barley domestication (Dai et al., 2014); 10 Tibetan wild barley accessions were re-sequenced to uncover patterns of adaptation (Zeng et al., 2015); and 177 Tibetan barley accessions were re-sequenced to identify signals of selection in the genome (Zeng et al., 2018). In contrast, 308 qingke accessions from the Qinghai-Tibet Plateau, including 278 qinke barley accessions and 30 qingke varieties collected from five other Chinese provinces, were used for this study. The effective sample size is 272 in total, which is comprised of 182.63 in the landrace subpopulation and 89.37 in the variety subpopulation. Our results demonstrate that this panel has a large effective population size with high levels of intra-species genetic flow, making it a suitable candidate for the characterization of genetic structure and adaptation, and was appropriate for the genetic study of complex traits by GWAS.

Previous studies have shown that QTLs identified on all seven chromosomes are significantly associated with HD, except for photoperiod and vernalization loci (Pasam et al., 2012). QTLs located on chromosomes 1H, 2H, and 5H have been identified that are significantly associated with PH in a worldwide spring barley investigation (Alqudah et al., 2016). Another recent study shows that QTLs on chromosomes 5H and 7H have been identified to be significantly associated with PH, and QTLs on chromosomes 1H, 2H, 5H, and 6H were shown to be significantly associated with SL in spring barley exposed to drought (Fakheri et al., 2018). In our study, SNPs that are significantly associated with HD and PH were identified on almost all barley chromosomes, and the SNPs mainly identified on chromosomes 1H, 2H, 3H, 4H, and 7H had significant effects on SL. To validate the effect estimation of each QTL, QTL effect estimations were used to predict the observed phenotypic performance. The highest prediction accuracy was 69.2% for HD, and the lowest prediction accuracy was 23.55% for SL in Namling. In the present study, heritability for all traits ranged from 44.97 and 73.99%. For these traits, both the number of detected QTL, prediction accuracy, and broad-sense heritability showed the same trend in which higher heritability corresponded to high prediction accuracy and more detected QTLs.

In order to figure out the specific QTLs for qingke barley, QTLs reported in the reference and identified in this study were aligned to the barley Hordeum vulgare Hv IBSC PGSB v2 reference genome, and their physical positions of markers were queried in BARLEX database (https://apex.ipk-gatersleben.de/apex/f?p=284:48:::NO:RP:P48_MARKER_CHOICE:4). As a result, 48 of 62 identified QTLs were first reported in this study (**Table 2** and **Supplementary Figure 13**). For HD, 36 QTLs were identified, 7 of which were reported; for PH, 33 QTLs were identified, 9 of which were reported; and for SL, 22 QTLs were identified, 5 of which were reported. The possible reason for the high number of novel QTLs (viewed as qingke barley specific QTLs) identified in this study may be because (1) the genetics of qingke barley is lack of analysis; and (2) some reported QTLs based on SSR markers could not find the their physical positions, and some reported QTLs developed by in-house SNP chips could not match the chip version in the database. These qingke barley specific QTLs could be utilized for marker-assisted selection in qingke barley breeding programs focusing on adaption and high grain yield.

Among the five common vernalization and photoperiod loci (i.e., *HvVRN1*, *HvVRN2*, *HvVRN3*, *Ppd-H1*, and *Ppd-H2*), the SNPs near *HvVRN1*, *HvVRN3*, and *PpD-H2* were significant in this study. It suggests that *HvVRN1*, *HvVRN3*, and *Ppd-H2* play an important role in the qingke barley population, and should be prioritized when attempting to improve HD and plant growth in qingke barley cultivars in the qingke barley-growing regions of the Qinghai-Tibet Plateau. Previous studies have shown that QTLs for PH and SL are located on different chromosomes depending on the different environments and treatments (Gyenis et al., 2007; Fakheri et al., 2018). In the present study, we observed QTLs related to PH on all chromosomes, while QTLs associated with SL were detected on all chromosomes in Namling but on 1H, 2H, 3H, 4H, and 7H in Nyingchi. Due to the differences in locations of QTLs identified in different environments, the expression of genes controlling HD, PH, and SL are probably related to the environment and variety-specific adaptability. To validate this hypothesis, the study of selection signals in the qingke barley genome were conducted to help us understand how qingke barley how qingke barley responds to various historical environmental factors (Russell et al., 2016). Eight regions were identified as candidate selective regions, and these are distributed on all chromosomes except for chromosome 4H (Zeng et al., 2018). We used the first 10 eigenvectors for EigenGWAS and identified several selected loci in the qingke barley genome. We further compared the selected loci with the located QTLs, and found that some of these loci were located in the regions (1 Mb) of these QTLs (**Table 1**). Previous studies have shown that genes for HD and PH always influence barley maturity and adaptation (Barua et al., 1993; Laurie et al., 1994). In the present study, the results of EigenGWAS analysis indicated that the QTLs associated with HD and PH also bear signatures of genetic selection in this qingke barley population.

CONCLUSION

In this study, we identified several genetic loci associated with SL, PH, and HD in qingke barley from the Qinghai-Tibet Plateau using 14,970 SNPs in a tGBS genotyping assay. We found that more rare SNPs (2,150) were found in the landrace subpopulation than in the variety subpopulation. That is to say, the number of SNPs was higher in the varieties than in the landraces, indicating that the varieties grown in Tibet and the varieties from around the Tibetan area may be derived from the different landraces grown in the different regions. A GWAS identified 62 QTLs that are associated with HD, PH, and SL, and 114 known genes were mapped which include, but are not limited to, genes involved in vernalization and photoperiod. Of the 62 QTLs, 48 are first reported here as qingke specific QTLs, 52 (83.87%) were found to be in intergenic regions, 28 had pleotropic effects, and three QTL were in the regions of the well-characterized genes *HvVRN1*, *HvVRN3*, and *PpD-H2*. In addition, by comparing signatures of selection identified by EigenGWAS and novel QTLs, we found that six QTLs related to HD and PH in qingke barley cultivars from the Qinghai-Tibet Plateau were also under selection. The findings presented here could help increase our understanding of the genetic mechanisms underlying the establishment of adaptive traits, and also enable marker-assisted selection for important traits in qingke barley breeding.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://www.ncbi.nlm.nih.gov/sra/PRJNA606408.

AUTHOR CONTRIBUTIONS

This study was designed by HL and DD. The experiment was performed under the support of GG, JZ, and TN. The evaluation of traits was conducted by NL, KD, PC, LG, and WC. Data were analyzed by ZL, NL, and HL. The manuscript was drafted by ZL, NL, and HL, and revised by GG and JW. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (31660299), the Tibet Department of Key Projects (XZ201801NA01), and the National Key Research and Development Program of China (2015BAD02B01-2-2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00638/full#supplementary-material

REFERENCES

Abdullaev, R. A., Alpatieva, N. V., Zveinek, I. A., Batasheva, B. A., Anisimova, I. M., and Radchenko, E. E. (2017). Diversity of dagestan barleys for the duration of the period between shooting and earing stages and alleles in the Ppd-H1 and Ppd-H2 loci. *Russ. Agric. Sci.* 43, 99–103. doi: 10.3103/S1068367417020021

Almerekova, S., Sariev, B., Abugalieva, A., Chudinov, V., Sereda, G., Tokhetova, L., et al. (2019). Association mapping for agronomic traits in six-rowed spring barley from the USA harvested in Kazakhstan. *PLoS ONE* 14:e0221064. doi: 10.1371/journal.pone.0221064

Alqudah, A. M., Koppolu, R., Wolde, G. M., Graner, A., and Schnurbusch, T. (2016). The genetic architecture of barley plant stature. *Front. Genet.* 7:117. doi: 10.3389/fgene.2016.00117

Alqudah, A. M., Shama, R., Pasam, R. K., Graner, A., Kilian, B., et al. (2014). Genetic dissection of photoperiod response based on GWAS of pre-anthesis phase duration in spring barley. *PLoS ONE* 9:e113120. doi: 10.1371/journal.pone.0113120

Alqudah, A. M., Youssef, H. M., Graner, A., and Schnurbusch, T. (2018). Natural variation and genetic make-up of leaf blade area in spring barley. *Theor. Appl. Genet.* 131, 873–886. doi: 10.1007/s00122-018-3053-2

Al-Tabbal, J. A., and Al-Fraihat, A. H. (2012). Genetic variation, heritability, phenotypic and genotypic correlation studies for yield and yield components in promising barley genotypes. *J. Agric. Sci.* 4:193. doi: 10.5539/jas.v4n3p193

Badr, A., Muller, K., Sch, R., Rabey, H. E., Effgen, S., Ibrahim, H. H., et al. (2000). On the origin and domestication history of barley (*Hordeum vulgare*). *Mol. Biol. Evol.* 17, 499–510. doi: 10.1093/oxfordjournals.molbev.a026330

Barua, U. M., Chalmers, K. J., Thomas, W. T. B., Hackett, C. A., Lea, V., Jack, P., et al. (1993). Molecular mapping of genes determining height, time to heading, and growth habit in barley (*Hordeum vulgare*). *Genome* 36, 1080–1087. doi: 10.1139/g93-143

Beales, J., Turner, A., Griffiths, S., Snape, J. W., and Laurie, D. A. (2007). A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 115, 721–733. doi: 10.1007/s00122-007-0603-4

Blake, T., Blake, V. C., Bowman, J. G. P., and Abdel-Haleem, H. (2010). Barley feed uses and quality improvement. *Barley* 522–531. doi: 10.1002/9780470958636.ch16

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevents, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Cai, S., Wu, D., Jabeen, Z., Huang, Y. Q., Huang, Y. C., and Zhang, G. P. (2013). Genome-wide association analysis of aluminum tolerance in cultivated and Tibetan wild barley. *PLoS ONE* 8:e69776. doi: 10.1371/journal.pone.0069776

Chen, G. B., Lee, S. H., Zhu, Z. X., and Robinson, M. R. (2016). EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity* 117:51. doi: 10.1038/hdy.2016.25

Chono, M., Honda, I., Zeniya, H., Yoneyama, Z., Saisho, D., Takeda, K., et al. (2003). A semidwarf phenotype of barley uzu results from a nucleotide substitution in the gene encoding a putative brassinosteroid receptor. *Plant Physiol.* 133, 1209–12219. doi: 10.1104/pp.103.026195

Cingolani, P., Platts, A., Wang, L. 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* 6, 80–92. doi: 10.4161/fly.19695

Collins, H. M., Burton, R. A., Topping, D. L., Liao, M. L., Bacic, A., and Fincher, G. B. (2010). Variability in fine structures of noncellulosic cell wall polysaccharides from cereal grains: potential importance in human health and nutrition. *Cereal. Chem.* 87, 272–282. doi: 10.1094/CCHEM-87-4-0272

Colmsee, C., Beier, S., Himmelbach, A., Schmutzer, T., Stein, N., Scholz, U., et al. (2015). BARLEX–the barley draft genome explorer. *Mol. Plant.* 8, 964–966. doi: 10.1016/j.molp.2015.03.009

Comadran, J., Russell, J. R., Booth, A., Pswarayi, A., Ceccarelli, S., Grando, S., et al. (2011). Mixed model association scans of multi-environmental trial data reveal major loci controlling yield and yield related traits in *Hordeum vulgare* in Mediterranean environments. *Theor. Appl. Genet.* 122, 1363–1373. doi: 10.1007/s00122-011-1537-4

Cuesta-Marcos, A., Igartua, E., Codesal, P., Ciudad, F. J., Codesal, P., Russell, J. R., et al. (2008). Heading date QTL in a spring× winter barley cross evaluated in Mediterranean environments. *Mol. Breed.* 21, 455–471. doi: 10.1007/s11032-007-9145-3

Dai, F., Chen, Z. H., Wang, X., Li, Z., Jin, G., Wu, D., et al. (2014). Transcriptome profiling reveals mosaic genomic origins of modern cultivated barley. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13403–13408. doi: 10.1073/pnas.1414335111

Dai, F., Nevo, E., Wu, D., Comadran, J., Zhou, M., Qiu, L., et al. (2012). Tibet is one of the centers of domestication of cultivated barley. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16969–16973. doi: 10.1073/pnas.1215265109

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Distelfeld, A., Li, C., and Dubcovsky, J. (2009). Regulation of flowering in temperate cereals. *Curr. Opin. Plant Biol.* 12, 178–184. doi: 10.1016/j.pbi.2008.12.010

Dockter, C., and Hansson, M. (2015). Improving barley culm robustness for secured crop yield in a changing climate. *J. Exp. Bot.* 3499–3509. doi: 10.1093/jxb/eru521

Fakheri, B. A., Aghnoum, R., Nezhad, N. M., and Ataei, R. (2018). GWAS analysis in spring barley (*Hordeum vulgare* L.) for morphological traits exposed to drought. *PLoS ONE* 13:e0204952. doi: 10.1371/journal.pone.0204952

Froster, B. P. (2001). Mutation genetics of salt tolerance in barley: an assessment of Golden Promise and other semi-dwarf mutants. *Euphytica.* 120, 317–328. doi: 10.1023/A:1017592618298

Genievskaya, Y., Almerekova, S., Sariev, B., Chudinov, V., Tokhetova, L., Sereda, G., et al. (2018). Marker-trait associations in two-rowed spring barley accessions from Kazakhstan and the USA. *PLoS ONE* 13:e0205421. doi: 10.1371/journal.pone.0205421

Gyawali, S., Otte, M. L., Chao, S., Jilal, A., Jacob, D. L., Amezrou, R., et al. (2017). Genome wide association studies (GWAS) of element contents in grain with a special focus on zinc and iron in a world collection of barley (*Hordeum vulgare* L.). *J. Cereal Sci.* 77, 266–274. doi: 10.1016/j.jcs.2017.08.019

Gyenis, L., Yun, S. J., Smith, K. P., Steffenson, B. J., Bossolini, E., Sanguineti, M. C., et al. (2007). Genetic architecture of quantitative trait loci associated with morphological and agronomic trait differences in a wild by cultivated barley cross. *Genome* 50, 714–723. doi: 10.1139/G07-054

Hemming, M. N., Fieg, S., Peacock, W. J., and Dennis, E. S., Trevaskis, B. (2009). Regions associated with repression of the barley (*Hordeum vulgare*) VERNALIZATION1 gene are not required for cold induction. *Mol. Genet. Genomics* 282, 107–117. doi: 10.1007/s00438-009-0449-3

Hill, C. B., Angessa, T. T., McFawn, L. A., Wong, D., Tibbits, J., Zhang, X. Q., et al. (2019). Hybridisation-based target enrichment of phenology genes to dissect the genetic basis of yield and adaptation in barley. *Plant Biotechnol. J.* 17, 932–944. doi :10.1111/pbi.13029 doi: 10.1111/pbi.13029

Hu, X., Zuo, J., Wang, J., Liu, L., Sun, G., Li, C., et al. (2018). Multi-locus genome-wide association studies for 14 Main agronomic traits in barley. *Front. Plant Sci.* 9:1683. doi: 10.3389/fpls.2018.01683

Ingvordsen, C. H., Backes, G., Lyngkjær, M. F., Peltonen-Sainio, P., Jensen, J. D., Jalli, M., et al. (2015). GWAS of barley phenotypes established under future climate conditions of elevated temperature, CO_2, O_3 and elevated temperature and CO_2 combined. *Proc. Environ. Sci.* 29, 164–165. doi: 10.1016/j.proenv.2015.07.241

Jia, Q. J., Zhang, J. J., Westcott, S., Zhang, X. Q., Bellgard, M., Lance, R., et al. (2009). GA-20 oxidase as a candidate for the semidwarf gene sdw1/denso in barley. *Funct. Integr. Genomics.* 9, 255–262. doi: 10.1007/s10142-009-0120-4

Kiseleva, A. A., Shcherban, A. B., Leonova, I. N., Frenkel, Z., and Salina, E. A. (2016). Identification of new heading date determinants in wheat 5B chromosome. *BMC Plant Biol.* 16:8. doi: 10.1186/s12870-015-0688-x

Kuczynska, A., Surma, M., Adamski, T., Mikołajczak, K., Krystkowiak, K., and Ogrodowicz, P. (2013). Effects of the semi-dwarfing sdw1/denso gene in barley. *Plant Genet.* 54, 381–390. doi: 10.1007/s13353-013-0165-x

Laurie, D. A., Pratchett, N., Bezant, J. H., and Snape, J. W. (1994). Genetic analysis of a photoperiod response gene on the short arm of chromosome 2 (2H) of *Hordeum vulgare* (barley). *Heredity* 72:619. doi: 10.1038/hdy.1994.85

Laurie, D. A., Pratchett, N., Snape, J. W., and Bezant, J. H. (1995). RFLP mapping of five major genes and eight quantitative trait loci controlling flowering time in a winter× spring barley (*Hordeum vulgare* L.) cross. *Genome* 38, 575–585. doi: 10.1139/g95-074

Li, J., Chen, G. B., Rasheed, A., Li, D., Sonder, K., Zavala Espinosa, C., et al. (2019). Identifying loci with breeding potential across temperate and tropical adaptation via EigenGWAS and EnvGWAS. *Mol. Ecol.* 28, 3544–3560. doi: 10.1111/mec.15169

Li, Q., Pan, Z., Deng, G., Long, H., Li, Z., Deng, X., et al. (2014). Effect of wide variation of the waxy gene on starch properties in hull-less barley from Qinghai-Tibet Plateau in China. *J. Agric. Food Chem.* 62, 11369–11385. doi: 10.1021/jf5026746

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

Lin, S. Y., Sasaki, T., and Yano, M. (1998). Mapping quantitative trait loci controlling seed dormancy and heading date in rice, *Oryza sativa* L., using backcross inbred lines. *Theor. Appl. Genet.* 96, 997–1003. doi: 10.1007/s001220050831

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Mei, W., Stetter, M. G., Gates, D. J., Stitzer, M., and Ross-Ibarra, J. (2017). Adaptation in plant genomes: bigger isn't better, but it's probably different. *Am. J. Bot.* 105, 16–19. doi: 10.1101/196501

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, 269–283. doi: 10.1016/j.cj.2015.01.001

Mikołajczak, K., Kuczynska, A., Krajewski, P., Sawikowska, A., Surma, M., Piotr, O., et al. (2017). Quantitative trait loci for plant height in Maresi × CamB barley population and their associations with yield-related traits under different water regimes. *J. Appl. Genet.* 58, 23–35. doi: 10.1007/s13353-016-0358-1

Ott, A., Liu, S., Schnable, J. C., Yeh, C. T. E., Wang, K. S., and Schnable, P. S. (2017). tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. *Nucleic Acids Res.* 45:e178. doi: 10.1093/nar/gkx853

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 20, 289–290. doi: 10.1093/bioinformatics/btg412

Pasam, R. K., Sharma, R., Malosetti, M., Eeuwijk, F. A. V., Haseneyer, G., Kilian, B., et al. (2012). Genome-wide association studies for agronomical traits in a worldwide spring barley collection. *BMC Plant Biol.* 12:16. doi: 10.1186/1471-2229-12-16

Pauli, D., Muehlbauer, G. J., Smith, K. P., Cooper, B., Hole, D., Obert, D. E., et al. (2014). Association mapping of agronomic QTLs in US spring barley breeding germplasm. *Plant Genome* 7, 1–15. doi: 10.3835/plantgenome2013.11.0037

Pham, A. T., Maurer, A., Pillen, K., Brien, C., Dowling, K., Berger, B., et al. (2019). Genome-wide association of barley plant growth under drought stress using a nested association mapping population. *BMC Plant Biol.* 19:134. doi: 10.1186/s12870-019-1723-0

Powell, J. E., Visscher, P. M., and Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.* 11, 800–805. doi: 10.1038/nrg2865

Qiu, L., Wu, D., Ali, S., Cai, S., Dai, F., Jin, X., et al. (2011). Evaluation of salinity tolerance and analysis of allelic function of HvHKT1 and HvHKT2 in Tibetan wild barley. *Theor. Appl. Genet.* 122, 695–703. doi: 10.1007/s00122-010-1479-2

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., et al. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11479–11484. doi: 10.1073/pnas.201394398

Russell, J., Mascher, M., Dawson, I. K., Kyriakidis, S., Calixto, C., Freund, F., et al. (2016). Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. *Nat. Genet.* 48:1024. doi: 10.1038/ng.3612

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Sameri, M., Takeda, K., and Komatsuda, T. (2006). Quantitative trait loci controlling agronomic traits in recombinant inbred lines from a cross of oriental-and occidental-type barley cultivars. *Breed. Sci.* 56, 243–252. doi: 10.1270/jsbbs.56.243

Tashi, N., Yawei, T., and Xingquan, Z. (2013). "Food preparation from hulless barley in Tibet," in *Advance In Barley Sciences*, eds G. Zhang, C. Li, and X. Liu (Dordrecht: Springer), 151–158.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/nar/gkx382

Tondelli, A., Xu, X., Moragues, M., Sharma, R., Schnaithmann, F., Ingvardsen, C., et al. (2013). Structural and temporal variation in genetic diversity of european spring two-row barley cultivars and association mapping of quantitative traits. *Plant Genome* 6, 1–14. doi: 10.3835/plantgenome2013.03.0007

Trevaskis, B. (2010). The central role of the VERNALIZATION1 gene in the vernalization response of cereals. *Funct. Plant Biol.* 37, 479–487. doi: 10.1071/FP10056

Trevaskis, B., Bagnall, D. J., Ellis, M. H., Peacock, W. J., and Dennis, E. S. (2003). MADS box genes control vernalization-induced flowering in cereals. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13099–13104. doi: 10.1073/pnas.1635053100

Trevaskis, B., Hemming, M. N., Peacock, W. J., and Dennis, E. S. (2006). HvVRN2 responds to daylength, whereas HvVRN1 is regulated by vernalization and developmental status. *Plant Physiol.* 140, 1397–1405. doi: 10.1104/pp.105.073486

Turner, A., Beales, J., Faure, S., Faure, S., Dunford, R. P., and Laurie, D. A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science* 310, 1031–1034. doi: 10.1126/science.1117619

Ullrich, S. E. (2010). *Barley: Production, Improvement, and Uses,Vol. 12*. New Jersey, NJ: John Wiley & Sons.

Visioni, A., Tondelli, A., Francia, E., Pswarayi, A., Malosetti, M., Russell, J., et al. (2013). Genome-wide association mapping of frost tolerance in barley (*Hordeum vulgare* L.). *BMC Genomics* 14:424. doi: 10.1186/1471-2164-14-424

Visioni, A. V., Gyawali, S., Selvakumar, R., Gangwar, O. P., Shekhawat, P. S., and Bhardwaj, S. C. (2018). Genome wide association mapping of seedling and adult plant resistance to barley stripe rust (*Puccinia striiformis* f. sp. hordei) in India. *Front. Plant Sci.* 9:520. doi: 10.3389/fpls.2018.00520

Wang, J., Yang, J., McNeil, D. L., and Zhou, M. (2010). Identification and molecular mapping of a dwarfing gene in barley (*Hordeum vulgare* L.) and its correlation with other agronomic traits. *Euphytica* 175, 331–342. doi: 10.1007/s10681-010-0175-2

Wang, J. B., Sun, G., Ren, X. F., Li, C. D., Liu, L. P., Wang, Q. F., et al. (2016). QTL underlying some agronomic traits in barley detected by SNP markers. *BMC Genet.* 17:103. doi: 10.1186/s12863-016-0409-y

Weir, B. S., and Cockerham, C. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associates, Inc.

Wu, D., Qiu, L., Xu, L., Ye, L., Chen, M., Sun, D., et al. (2011). Genetic variation of HvCBF genes and their association with salinity tolerance in Tibetan annual wild barley. *PLoS ONE* 6:e22938. doi: 10.1371/journal.pone.0022938

Xu, X., Sharma, R., Tondelli, A., Russell, J., Comadran, J., Schnaithmann, F., et al. (2018). Genome-wide association analysis of grain yield-associated traits in a pan-European barley cultivar collection. *Plant Genome* 11:170073. doi: 10.3835/plantgenome2017.08.0073

Xu, Y. H., Jia, Q. J., Zhou, G. F., Zhang, X. Q., Angessa, T., Broughton, S., et al. (2017). Characterization of the sdw1 semi-dwarf gene in barley. *BMC Plant Biol.* 17:11. doi: 10.1186/s12870-016-0964-4

Yan, L., Fu, D., Li, C., Blechi, A., Tranquilli, G., Bonafede, M., et al. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19581–19586. doi: 10.1073/pnas.0607142103

Yan, L., Helguera, M., Kato, K., Fukuyama, S., Sherman, J., and Dubcovsky, J. (2004). Allelic variation at the VRN-1 promoter region in polyploid wheat. *Theor. Appl. Genet.* 109, 1677–1686. doi: 10.1007/s00122-004-1796-4

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., and Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6263–6268. doi: 10.1073/pnas.0937399100

Yan, W. H., Wang, P., Chen, H. X., Zhou, H. J., Li, Q. P., and Wang, C. R. (2011). A major QTL, Ghd8, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Mol. Plant*. 4, 319–330. doi: 10.1093/mp/ssq070

Zeng, X., Bai, L., Wei, Z., Yuan, H., Wang, Y., Xu, Q., et al. (2016). Transcriptome analysis revealed the drought-responsive genes in Tibetan hulless barley. *BMC Genomics* 17:386. doi: 10.1186/s12864-016-2685-3

Zeng, X., Guo, Y., Xu, Q., Mascher, M., Guo, G., Li, S., et al. (2018). Origin and evolution of qingke barley in Tibet. *Nat. Commun*. 9:5433. doi: 10.1038/s41467-018-07920-5

Zeng, X., Long, H., Wang, Z., Zhao, S., Tang, Y., Huang, Z., et al. (2015). The draft genome of Tibetan hulless barley reveals adaptive patterns to the high stressful Tibetan Plateau. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1095–1100. doi: 10.1073/pnas.1423628112

Zhang, J. (2000). Inheritance of agronomic traits from the Chinese barley dwarfing gene donors 'Xiaoshan Lixiahuang' and 'Cangzhou Luodama'. *Plant Breed.* 119, 523–524. doi: 10.1046/j.1439-0523.2000.00543.x

Zhang, K., Tian, J., Zhao, L., Liu, B., and Chen, G. (2009). Detection of quantitative trait loci for heading date based on the doubled haploid progeny of two elite Chinese wheat cultivars. *Genetica* 135, 257–265. doi: 10.1007/s10709-008-9274-6

Zhang, M., Fu, M. M., Qiu, C. W., Cao, F., Chen, Z. H., Zhang, G., et al. (2019). Response of tibetan wild barley genotypes to drought stress and identification of quantitative trait loci by genome-wide association analysis. *Int. J. Mol. Sci.* 20:791. doi: 10.3390/ijms20030791

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet*. 42:355. doi: 10.1038/ng.546

Zhou, H., and Steffenson, B. (2013). Genome-wide association mapping reveals genetic architecture of durable spot blotch resistance in US barley breeding germplasm. *Mol. Breed.* 32, 139–154. doi: 10.1007/s11032-013-9858-4

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

frontiers in Genetics

METHODS
published: 16 September 2020
doi: 10.3389/fgene.2020.540751

Prioritizing CircRNA–Disease Associations With Convolutional Neural Network Based on Multiple Similarity Feature Fusion

Chunyan Fan[1], Xiujuan Lei[1] and Yi Pan[2]**

[1] School of Computer Science, Shaanxi Normal University, Xi'an, China, [2] Department of Computer Science, Georgia State University, Atlanta, GA, United States

OPEN ACCESS

Edited by:
Hailan Liu,
Sichuan Agricultural University, China

Reviewed by:
Fangqing Zhao,
Beijing Institutes of Life Science (CAS), China
Francesco Manfrevola,
Second University of Naples, Italy
Quan Zou,
University of Electronic Science and Technology of China, China

***Correspondence:**
Xiujuan Lei
xjlei@snnu.edu.cn
Yi Pan
yipan@gsu.edu

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

Received: *25 March 2020*
Accepted: *12 August 2020*
Published: *16 September 2020*

Citation:
Fan C, Lei X and Pan Y (2020) Prioritizing CircRNA–Disease Associations With Convolutional Neural Network Based on Multiple Similarity Feature Fusion. Front. Genet. 11:540751. doi: 10.3389/fgene.2020.540751

Accumulating evidence shows that circular RNAs (circRNAs) have significant roles in human health and in the occurrence and development of diseases. Biological researchers have identified disease-related circRNAs that could be considered as potential biomarkers for clinical diagnosis, prognosis, and treatment. However, identification of circRNA–disease associations using traditional biological experiments is still expensive and time-consuming. In this study, we propose a novel method named MSFCNN for the task of circRNA–disease association prediction, involving two-layer convolutional neural networks on a feature matrix that fuses multiple similarity kernels and interaction features among circRNAs, miRNAs, and diseases. First, four circRNA similarity kernels and seven disease similarity kernels are constructed based on the biological or topological properties of circRNAs and diseases. Subsequently, the similarity kernel fusion method is used to integrate the similarity kernels into one circRNA similarity kernel and one disease similarity kernel, respectively. Then, a feature matrix for each circRNA–disease pair is constructed by integrating the fused circRNA similarity kernel and fused disease similarity kernel with interactions and features among circRNAs, miRNAs, and diseases. The features of circRNA–miRNA and disease–miRNA interactions are selected using principal component analysis. Finally, taking the constructed feature matrix as an input, we used two-layer convolutional neural networks to predict circRNA–disease association labels and mine potential novel associations. Five-fold cross validation shows that our proposed model outperforms conventional machine learning methods, including support vector machine, random forest, and multilayer perception approaches. Furthermore, case studies of predicted circRNAs for specific diseases and the top predicted circRNA–disease associations are analyzed. The results show that the MSFCNN model could be an effective tool for mining potential circRNA–disease associations.

Keywords: circRNA-disease associations, circRNA-miRNA interaction, similarity kernel fusion, feature matrix, convolutional neural network

INTRODUCTION

Circular RNAs (circRNAs) are a type of endogenous non-coding RNA with continuous covalently closed loop structures, which are produced by back-splicing or lariat events in genes (Barrett et al., 2015). Recently, with the development of high-throughput sequencing techniques and other technologies, a large number of circRNAs have been found in various organisms, including protists, plants, and metazoans (Danan et al., 2012; Memczak et al., 2013; Tang et al., 2018). The main functions of circRNAs include sequestration of microRNAs (miRNAs) and proteins (Salmena et al., 2011), regulation of transcription and splicing (Zhang et al., 2013; Conn et al., 2017), and even translation to produce polypeptides (Yang et al., 2017; Sun and Li, 2019). Accumulating evidence implicates mutation or alteration in expression of circRNAs in the initiation and progression of numerous diseases. For example, Chioccarelli et al. (2019) identified the differentially expressed circRNAs in human spermatozoa, and found that circRNAs are related to spermatozoa quality. By comparing the expression profiles of circRNAs in disease-specific tissues or cell lines with those in normal samples, significantly increased or decreased circRNAs can be identified. In addition, the intrinsic characteristics of circRNAs indicate they are stable both inside cells and in extracellular plasma (Bahn et al., 2015; Li et al., 2015; Memczak et al., 2015). Therefore, disease-associated circRNAs are considered to be promising novel biomarkers for diseases.

Recently, several studies have analyzed the roles of circRNAs in varies samples, and further explore their diversity, expression patterns, co-expression network, and so on. circAtlas integrates the most comprehensive circRNAs, their expression, and functional profiles in vertebrates (Wu et al., 2020). MiOncoCirc is a cancer-focused circRNA resource to be generated from an extensive array of tumor tissues (Vo et al., 2019). Ji et al. (2019) identifies full-length transcripts and evolutionarily conserved circRNAs, and infers circRNA functions on a global scale. Ruan et al. (2019) characterizes circRNAs expression profiles, and explores the potential mechanism of circRNA biogenesis as well as its therapeutic implications. exoRBase integrates and visualize the RNA expression profiles both normal individuals and patients with different diseases (Li et al., 2018). These studies will trigger functional implication for human diseases and benefit biomedical research community.

The de-regulated circRNAs in diseases can be identified for validation using low-throughput biological methods such as quantitative real-time PCR, northern blotting, and so on. However, these traditional experiments are costly and time-consuming. Therefore, computational approaches are important for exploring potential disease-causing circRNAs and understanding the associated mechanisms of pathogenicity. Several models have been proposed to forecast circRNA–disease associations; most of these approaches are based on the assumption that circRNAs with similar functions are likely to be associated with the same or similar diseases. Lei et al. (2018) developed a path-weighted model to predict circRNA–disease associations based on circRNA semantic similarity and disease functional similarity (Lei et al., 2018). KATZHCDA was used to calculate the number of walks between nodes and walk lengths for circRNA–disease associations, based on *a priori* knowledge of the circRNA expression similarity and disease phenotype similarity (Fan et al., 2018b). DWNN-RLS predicted circRNA–disease associations using regularized least squares of the Kronecker product kernel (Yan et al., 2018). Xiao et al. (2019) proposed a weighted dual-manifold regularized low-rank approximation model for disease-related circRNA prediction, called MRLDC (Xiao et al., 2019). Another model, iCircDA-MF, incorporated circRNA–gene, gene–disease, and circRNA–disease associations, together with disease semantic information, and used non-negative matrix factorization to predict circRNA–disease associations (Wei and Liu, 2019). Zhao et al. (2019) integrated the bipartite network projection algorithm and KATZ measure algorithm to explore novel circRNA–disease associations (Zhao et al., 2019). Deng et al. (2019) combined circRNAs, proteins, and diseases to predict circRNA–disease associations using the KATZ algorithm (Deng et al., 2019). Ge et al. (2019) developed the LLCDC model for prediction of human disease-associated circRNAs using locality-constrained linear coding and a label propagation algorithm (Ge et al., 2019). CD-LNLP calculated circRNA similarity and disease similarity using linear neighborhood similarity based on known associations, and then used the label propagation algorithm to mine circRNA–disease associations (Zhang et al., 2019). Wang Y. et al. (2019) used a graph-based recommendation algorithm, PersonalRank, to predict disease-related circRNAs based on circRNA expression profiles and functional similarities (Wang Y. et al., 2019). Lei and Fang (2019) used a gradient boosting decision tree with multiple biological data fusion for circRNA–disease prediction (Lei and Fang, 2019). Ding et al. (2020) developed the RWLR model based on the random walk and the logistic regression to predict circRNA-disease associations. iCDA-CGR quantified the sequence nonlinear relationship of circRNA by chaos game representation technology based on the biological sequence position information (Zheng et al., 2020). Lei and Bian (2020) integrated the random walk with restart and *k*-nearest neighbors to predict the associations between circRNAs and diseases. Although these computational models have achieved encouraging results, they represent the tip of the iceberg with respect to predicting circRNA–disease associations.

Several circRNAs can bind with the corresponding miRNAs and participate in multiple biological processes synchronously (Qu et al., 2018). Based on this theory, Fang and Lei (2019) used an improved random walk algorithm to predict circRNA–miRNA associations, named KRWRMC (Fang and Lei, 2019). As miRNAs have been implicated in various diseases, we consider that miRNA information should be included in the identification of circRNA–disease associations. However, there have been few studies of circRNA–miRNA interactions, and deep interaction patterns are rarely considered in prediction of circRNA–disease associations. In this work, we take circRNA–miRNA interactions and miRNA–disease associations into account, and capture the complex miRNA-based interaction features of circRNAs and diseases, respectively.

In recent years, deep learning architectures have attracted increasing attention in various fields, including image analysis (Yang and Xu, 2020), speech recognition (Graves et al., 2013), and bioinformatics (Min et al., 2017), etc. The convolutional neural network (CNN) is a well-known feed-forward artificial neural network inspired by biological processes that simulates the cognition function of human neural systems (LeCun et al., 2015). CNN architectures have the ability to automatically learn the meaning of combinations of features from the input data and simplify the process of manual feature selection (Liu et al., 2017). Recent applications of CNN-based methods indicate their effectiveness in computational biology (Liu et al., 2018), including in circRNA research. Wang and Wang (2019) developed the DeepCirCode model to discover the sequence code of back-splicing for circRNA formation, and sequence motifs were also extracted. The CSCRSites model was proposed to predict cancer-specific protein binding sites on circRNAs based on CNNs. The features learned by the CSCRSites model are converted to sequence motifs, some of which are involved in human diseases (Wang Z. et al., 2019). Inspired by the superior prediction performance of this approach, we used CNN architecture to detect combinations of features and predict potential circRNA–disease associations.

In this study, we present a novel computational model to predict potential associations between circRNAs and diseases, named MSFCNN. The main attributes of the MSFCNN model are as follows. (1) Four circRNA similarity kernels and seven disease similarity kernels are constructed using multiple biological and topological information, such as circRNA expression profiles, circRNA sequence information, disease-miRNA interactions, etc. (2) Whereas some existing methods simply use linear weighting to integrate the similarity kernels into one kernel, we considered that this may lead to information loss and noise. Hence, we used the similarity kernel fusion (SKF) method to fuse four circRNA similarity kernels and seven disease similarity kernels, thereby retaining the original information of each similarity kernel. A weight matrix is used to reduce the noise in the fused similarity kernel. (3) A feature matrix is constructed based on the fused circRNA similarity kernel, fused disease similarity kernel, and interactions and features among circRNAs, miRNAs, and diseases. Multiple biological premises are used to construct the feature matrix. On the one hand, two circRNAs (or diseases) are more similar could capture the relationships between the circRNA (or disease) similarities and circRNA–disease associations. On the other hand, circRNA–miRNA and miRNA–disease associations are also integrated, and the interaction features are captured using principal component analysis. (4) A two-layer CNN architecture is used to process the feature matrix and predict potential circRNA–disease associations. Five-fold cross-validation (CV) is used to assess the prediction performance of the MSFCNN model. The results indicate that the MSFCNN model outperforms several conventional machine learning classifiers. Furthermore, case studies of breast cancer, colorectal cancer, hepatocellular carcinoma, and acute myeloid leukemia indicate that MSFCNN could be an effective tool to infer potential circRNA–disease associations.

MATERIALS AND METHODS

A flow chart illustrating MSFCNN, our novel approach to predict potential circRNA–disease associations is shown in **Figure 1**. First, four circRNA similarity kernels and seven disease similarity kernels are computed based on their biological and topological properties. Then, these kernel similarities are combined into one circRNA similarity kernel and one disease similarity kernel by applying a similarity kernel fusion strategy. Subsequently, the feature matrix can be constructed based on the fused similarity kernels, and interactions and features among circRNAs, miRNAs, and diseases. Finally, we use a CNN to process the feature matrix and predict final scores for prediction of potential circRNA–disease associations.

Construction of the CircRNA–Disease, CircRNA–miRNA, and Disease–miRNA Networks

In this study, circRNA–disease associations, circRNA–miRNA associations, and disease–miRNA associations were used to predict circRNA–disease associations. Known circRNA–disease associations were downloaded from the CircR2Disease database (Fan et al., 2018a), which contained 739 entries including 725 experimentally validated circRNA–disease associations from four species. Only human circRNA–disease associations were used in this work. Interactions that did not correspond to circRNAs IDs in the circBase database and disease names were not recorded in the disease ontology database were removed (Glazar et al., 2014; Schriml et al., 2019). Thus, we retained 325 circRNAs, 53 diseases, and 371 circRNA–disease associations as the positive dataset. The circRNA–miRNA interactions were obtained from the CircBank database (Liu et al., 2019), and interactions overlapping with disease-related circRNAs were extracted. Thus, 24745 interactions between 322 circRNAs and 2545 miRNAs were obtained. In addition, the disease–miRNA associations that matched circRNA-related diseases were selected from the human microRNA disease database (Huang et al., 2019), and 4970 associations between 37 diseases and 873 miRNAs were obtained. Finally, all of these associations contained three types of nodes including 325 circRNAs, 53 diseases, and 3175 miRNAs.

Based on the circRNA–disease associations, an adjacency matrix $A(i,j)$ was constructed to represent associations between n_c circRNAs and n_d diseases; $A(i,j)$ was assigned a value of 1 if circRNA $c(i)$ was found to be related to disease $d(j)$, and 0 otherwise. Similarly, a circRNA–miRNA matrix $Y(i, j)$ was constructed to represent the associations between n_c circRNAs and n_m miRNAs, and the associations between n_d diseases and n_m miRNAs were represented by matrix $O(i, j)$. $Y(i, j)$ was set to 1 when there was an association between circRNA $c(i)$ and miRNA $m(j)$, and 0 otherwise. If disease $d(i)$ interacted with miRNA $m(j)$, $O(i, j)$ was set to 1, otherwise it was set to 0.

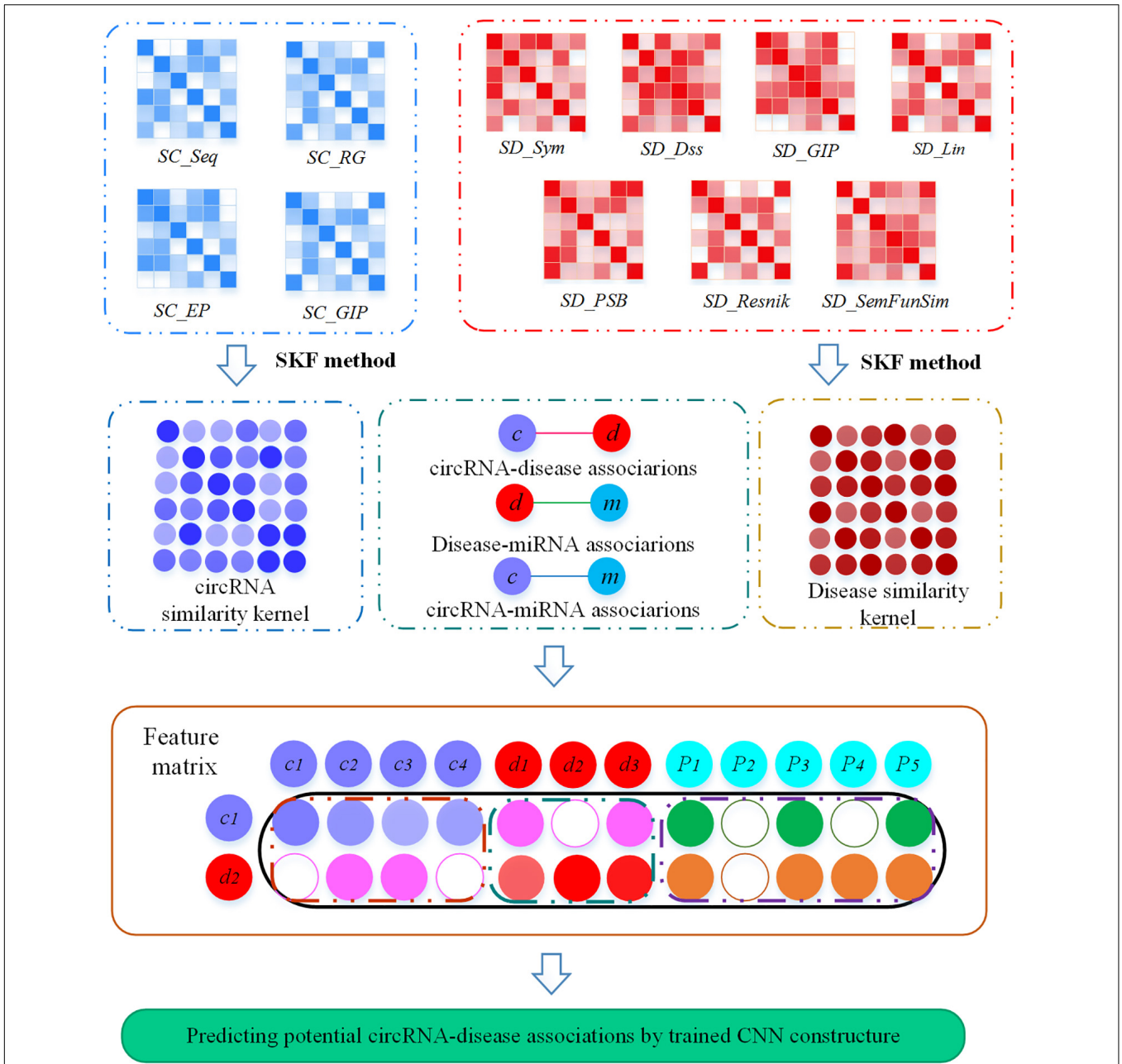

FIGURE 1 | Flow chart of the MSFCNN approach. Step 1: Four circRNA similarity kernels and seven disease similarity kernels are measured, respectively. Step 2: The similarity kernels for circRNAs (or diseases) is fused with SKF method. Step 3: The feature matrix for each circRNA–disease pair is constructed by integrating the fused similarity kernels, interactions and features among circRNAs, miRNAs, and diseases. Step 4: a CNN architecture is used to train MSFCNN approach and predict latent circRNA–disease associations.

Representation of CircRNA Similarity Kernels

CircRNA Sequence Similarity

The 325 circRNA sequences were obtained from the circBase database (Glazar et al., 2014), and the sequence similarity of each circRNA–circRNA pair was calculated using a modification of the Needleman–Wunsch algorithm with the Emboss-stretcher tool (Rice et al., 2000). Therefore, the circRNA sequence similarity score $SC\_Seq(c_i, c_j)$ could be obtained by setting the parameters as follows: Matrix = EDNAFULL, Gap open = 16, Gap extend = 4.

CircRNA Regulatory Similarity

Based on the assumption that circRNAs associated with the same miRNAs tend to have similar biological regulatory functions, we used the miRNA–circRNA interactions to measure the circRNA regulatory similarity (Huang et al., 2018). Given the two sets of

miRNAs, M_i and M_j, that had relationships with circRNAs c_i and c_j, respectively, the circRNA regulatory similarity kernel was calculated as follows:

$$SC\_RG(c_i, c_j) = \frac{card(M_i \bigcap M_j)}{\sqrt{card(M_i)} \cdot \sqrt{card(M_j)}} \quad (1)$$

CircRNA Expression Similarity

The circRNA expression profiles were derived from the exoRBase database (Li et al., 2018). Each circRNA record had 90 dimensions, representing the expression levels of a single type of circRNA. By extracting the common circRNAs between the CircR2Disease and exoRBase databases, circRNA expression profiles were obtained for calculation of the circRNA similarity kernel. We used the Pearson correlation coefficient to measure circRNA expression similarity, and let $SC\_EP(c_i, c_j)$ represent the expression similarity score between circRNAs c_i and c_i. The expression similarity kernel of the circRNAs was computed as follows:

$$SC\_EP(c_i, c_j) = \frac{\sum_{i=1}^{N}(xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum_{i=1}^{N}(xi - \bar{x})^2 \sum_{i=1}^{N}(yi - \bar{y})^2}} \quad (2)$$

where N represents the number of properties of the expression profiles, and x_i and y_i denote the expression values in different tissues. In general, a pair of circRNAs with a higher correlation score are considered to be more similarly expressed.

GIP Kernel Similarity for CircRNAs

The Gaussian interaction profile (GIP) kernel similarity was used to measure the similarity between circRNAs, based on the assumption that similar circRNAs are more likely exhibit a similar interaction or non-interaction pattern with miRNAs (Van Laarhoven et al., 2011). GIP kernel similarity for circRNAs was measured based on circRNA–miRNA interactions and defined as:

$$SC\_GIP(c_i, c_j) = exp(-\gamma_c \parallel c(i) - c(j) \parallel^2)$$
$$\gamma_c = \frac{1}{\frac{1}{n_c}\sum_{i=1}^{n_c} \parallel c(i) \parallel^2} \quad (3)$$

where the circRNA interaction profiles are represented by $c(i)$, a binary vector that encodes the interaction between circRNA i and all miRNAs, i.e., the i-th row of the circRNA–miRNA interaction matrix Y. The parameter γ_c controls the kernel bandwidth, and n_c is the number of circRNAs.

Representation of Disease Similarity Kernels

Disease Symptom Similarity

According to the co-occurrence of disease and symptom terms recorded in the PubMed bibliography, Zhou et al. (2014) considered that diseases are connected if they have a positive symptom similarity (Zhou et al., 2014). Thus, the disease similarity could be measured and a symptom-based human disease network was constructed. Here, the symptom-based disease similarity *SD_Sym* was obtained from the symptom profiles of diseases.

Disease Semantic Similarity

According to Medical Subject Headings descriptions, diseases can be described by a hierarchical directed acyclic graph (DAG). Here, disease semantic similarity is calculated using the method of Wang et al. (2007). $DAG_d = (d, T_d, E_d)$ represents the DAG of a disease, in which T_d denotes node d and its ancestor nodes, and E_d denotes the direct edges from a parent node to child nodes within T_d. Therefore, the semantic contribution of parent node t to d is defined as follows:

$$D_d(t) = \begin{cases} 1, & \text{if } t = d \\ \max\{\Delta * D_d(d') | d' \in \text{children of t}, & \text{if } t \neq d \end{cases} \quad (4)$$

where $\triangle$ represents the semantic contribution decay factor ($\triangle$ is set as 0.5). The semantic value of disease d can be calculated as follows:

$$DV(d) = \sum_{t \in T_d} D_d(t) \quad (5)$$

If two diseases share a larger part of DAGs, they tend to have higher similarity. The similarity score between d_i and d_j is defined as:

$$SD\_Dss(d_i, d_j) = \frac{\sum_{t \in Td_i \bigcap Td_j}(D_{d_i}(t) + D_{d_j}(t))}{DV(d_i) + DV(d_j)} \quad (6)$$

GIP Kernel Similarity for Diseases

Similar to the calculation of GIP kernel similarity for circRNAs, the disease GIP kernel similarity was measured based on disease–miRNA interaction profiles. It is defined as:

$$SD\_GIP(d(i), d(j)) = exp(-\gamma_d \parallel d(i) - d(j) \parallel^2)$$
$$\gamma_d = \frac{1}{\frac{1}{n_d}\sum_{i=1}^{n_d} \parallel d(i) \parallel^2} \quad (7)$$

where the disease interaction profiles are represented by $d(i)$, a binary vector that encodes the interaction between disease i and each miRNA, i.e., the i-th row of association matrix O. The parameter γ_d is also used to control the kernel bandwidth, and n_d is the number of diseases.

Other Disease Similarities

Besides disease symptom similarity, disease sematic similarity, and GIP kernel similarity, disease similarities can also be measured using the Lin (1998), PSB (Mathur and Dinakarpandian, 2012), Resnik (1995), and SemFunSim (Cheng et al., 2014) methods based on the DincRNA database (Cheng et al., 2018). Four disease similarity kernels were constructed using these methods and denoted *SD_Lin*, *SD_PSB*, *SD_Resnik*, and *SD_SemFunSim*, respectively.

Similarity Kernel Fusion

Next, we used the similarity kernel fusion method to integrate four circRNA similarity kernels and seven disease similarity kernels (Jiang et al., 2018). Let $S_{c,m}$ (m = 1,2,. . .4) represent the four circRNA similarity kernels and $S_{d,n}$ (n = 1,2,. . .7) the seven disease similarity kernels, respectively.

First, each original similarity kernel for circRNAs was normalized using Eq. (8):

$$NS_{c,m}(c_i, c_j) = \frac{S_{c,m}(c_i, c_j)}{\sum_{c_k \in C} S_{c,m}(c_k, c_j)} \tag{8}$$

where $NS_{c,m}$ denotes a normalized similarity kernel for circRNAs that satisfies $\sum_{c_k \in C} NS_{c,m}(c_k, c_j) = 1$.

Then, a sparse kernel for each circRNA similarity kernel was constructed using Eq. (9):

$$F_{c,m}(c_i, c_j) = \begin{cases} \dfrac{S_{c,m}(c_i, c_j)}{\sum_{c_k \in N_i} S_{c,m}(c_i, c_k)} & c_j \in N_i \\ 0 & c_j \notin N_i \end{cases} \tag{9}$$

where $F_{c,m}$ is a sparse kernel satisfying $\sum_{c_j \in C} F_{c,m}(c_k, c_j) = 1$, and N_i is a set of c_i's neighbors including c_i itself.

The four circRNA similarity kernels were computed using Eq. (10):

$$SC_{c,m}^{t+1} = \alpha\left(F_{c,m} \times \frac{\sum_{r \neq 1} SC_{c,r}^{t}}{2} \times F_{c,m}^{T}\right) + (1-\alpha)\left(\frac{\sum_{r \neq 1} SC_{c,r}^{0}}{2}\right) \quad \alpha \in (0, 1) \tag{10}$$

where $SC_{c,m}^{t+1}$ is the status matrix of m-th circRNA similarity kernel after $t+1$ iterations, and$SC_{c,r}^{0}$ denotes the initial status of $SC_{c,r}$.

After $t+1$ steps, the overall kernel for circRNAs was calculated using Eq. (11):

$$S_c = \frac{1}{4}\sum_{m=1}^{4} SC_{c,m}^{t+1} \tag{11}$$

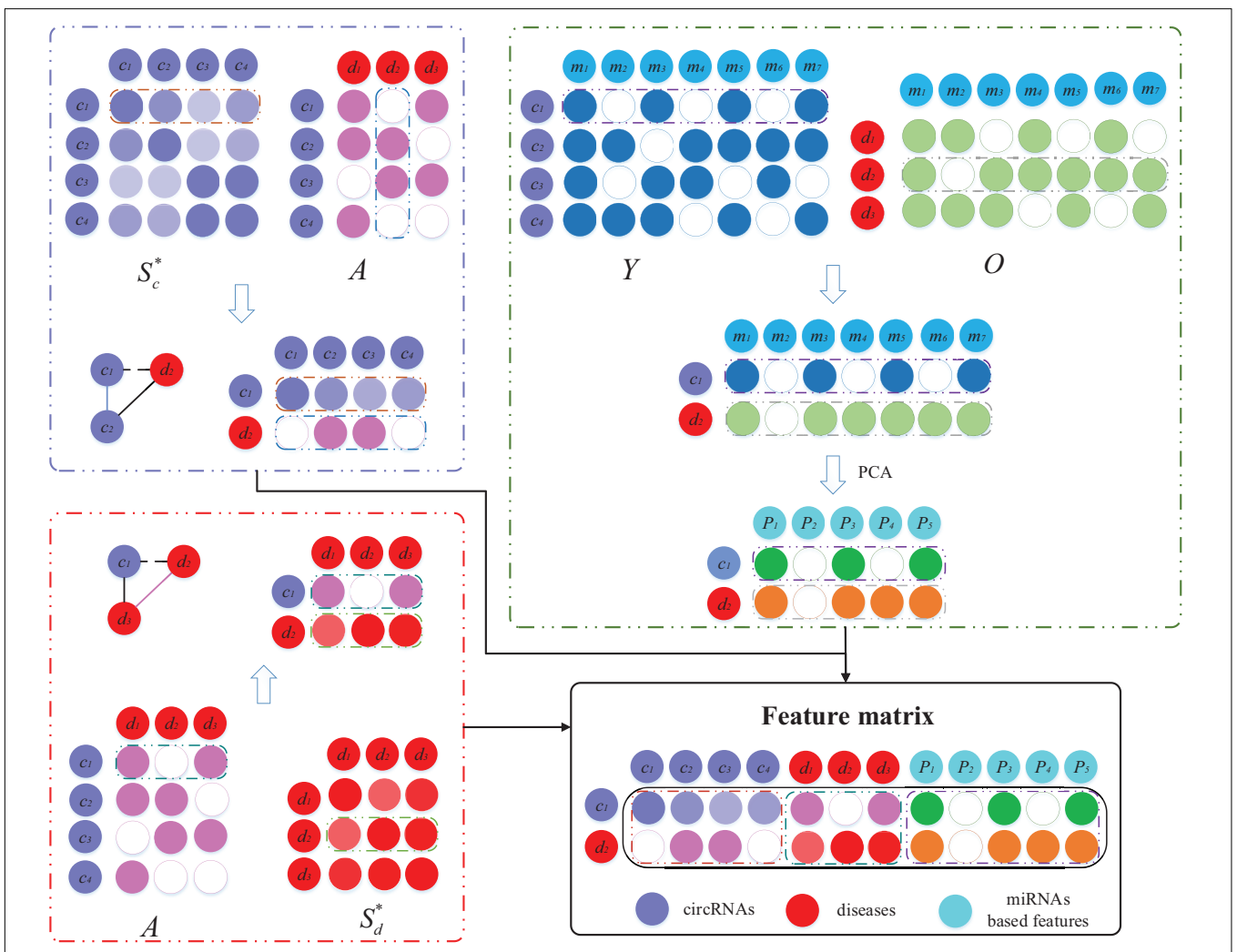

FIGURE 2 | Establishment of the feature matrix of circRNA c_1 and disease d_2. Based on three premises, c_1-d_2 feature matrix is constructed by combing fused similarities and associations among circRNAs, diseases and miRNAs.

Furthermore, a weight matrix w_c was used to eliminate the noise in matrix S_c, and the fused circRNA similarity kernel was computed using Eq. (12):

$$S_c^* = w_c \circ S_c \tag{12}$$

$$w_c(c_i, c_j) = \begin{cases} 1 & \text{if } c_i \in N_j \text{ and } c_j \in N_i \\ 0 & \text{if } c_i \notin N_j \text{ and } c_j \notin N_i \\ 0.5 & \text{otherwise} \end{cases} \tag{13}$$

Similarly, the seven disease similarity kernels were fused to form one disease similarity kernel, denoted by S_d^*.

Construction of the Feature Matrix

The feature matrix for each circRNA–disease pair was constructed by incorporating the fused circRNA similarity, fused disease similarity, circRNA–miRNA interactions, circRNA–disease associations, and disease–miRNA associations (**Figure 2**).

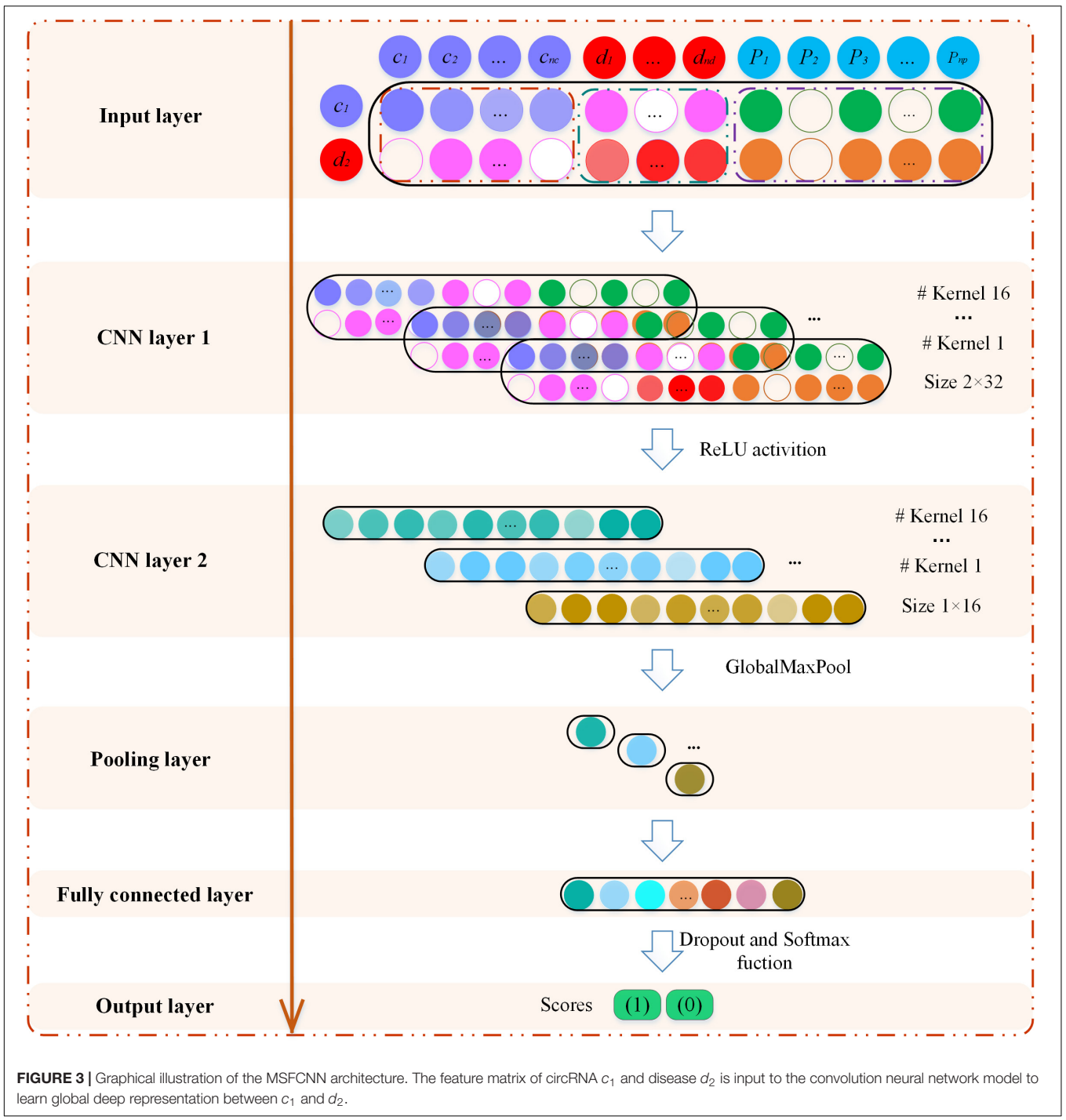

FIGURE 3 | Graphical illustration of the MSFCNN architecture. The feature matrix of circRNA c_1 and disease d_2 is input to the convolution neural network model to learn global deep representation between c_1 and d_2.

In the construction process of the feature matrix, three biological premises were used. Here, we take the construction of the c_1-d_2 feature matrix as an example. Based on the premise that the circRNAs should be more similar that have interaction with circRNA similarities and circRNA–disease associations, the first part of the feature matrix consists of the similarity between c_1 and all circRNAs, and the associations of d_2 with all circRNAs. If circRNA c_1 and c_2 or other circRNAs have similar functions, and at the same time d_2 has been shown to be associated with these circRNAs, c_1 has a large probability associated with d_2. The dimension of the first part of the feature matrix is $2 \times n_c$. Similarly, based on the premise that diseases should be more similar that have interaction with disease similarities and circRNA–disease associations, we integrate the associations between circRNA c_1 and all diseases, as well as the similarities between disease d_2 and all diseases. The second part of the feature matrix has dimension $2 \times n_d$. In addition, circRNA–miRNA and miRNA–disease is integrated to capture the relation features. When c_1 and d_2 have interactions with common miRNAs, they are more likely to be associated with each other. The interactions between c_1 and various miRNAs, as well as the associations between d_2 and miRNAs, are integrated to construct a matrix with dimension $2 \times n_m$. However, the matrix is very sparse, so we perform principal component analysis (PCA) to obtain miRNA-based features for the c_1-d_2 pair with dimension $2 \times n_p$ (n_p is set as 50). Finally, we concatenate these three matrices to form the feature matrix of circRNA c_1 and disease d_2 with dimension $2 \times (n_c+n_d+n_p)$.

Identification of CircRNA–Disease Associations Based on CNN

The MSFCNN architecture consists of an input layer, two convolutions, and an activation layer, polling layer, fully connected layer, and softmax layer (**Figure 3**). The feature matrix X of node pairs is used as an input to the CNN architecture to learn the representations of node-pair circRNAs and diseases. The MSFCNN can be summarized as:

$$Out = f^{Softmax} f^{Fully\_connected} f^{GlobalMaxPool} f^{Conv2D\_ReLU} f^{Conv2D\_ReLU}(X) \quad (14)$$

where X is the feature matrix that is fed to the two-dimensional convolution (Conv2D) layer. In the first convolutional layer, if the number of filters is n_{conv1}, the width of the kernel is n_w, and its length is set as n_l. The convolution filters are indicated as $W_{conv1} \in R^{nconv1 \times nw \times nl}$, and the feature maps are $Z_{conv1} \in R^{nconv1 \times (2-nw+1) \times (nc+nd+np-nl+1)}$. The convolution process can be described as follows:

$$X_{k,i,j} = X(i : i + n_w, j : j + n_l) \quad X_{k,i,j} \in R^{n_w \times n_l} \quad (15)$$

$$\begin{aligned} &Z_{conv1,k}(i,j) = g(W_{conv1}(k,:,:) * X_{conv1,i,j} + b_{conv1}(k)) \\ &k \in [1, n_{conv1}], i \in [1,2], j \in [1, n_c + n_d + n_p - n_l + 1] \end{aligned}, \quad (16)$$

where $X(i,j)$ is the element of matrix X in the i-th row and j-th column, and $X_{k,i,j}$ represents the region in the filter where the k-th filter slides to the position $X(i,j)$. g is a rectified linear units ($relu$) function (Nair and Hinton, 2010), b_{conv1} is the bias vector, $*$ represents the convolution operation, and $Z_{conv1,k}(i,j)$ represents the convolution result of the k-th filter sliding to the j-th column of the i-th row.

Similarly, the second Conv2D layer is also used to learn the higher-level features. To compress data and reduce over-fitting, the polling layer is used to obtain robust features. Here, the max-pooling operation is employed for each feature map (Collobert et al., 2011). Then, the outputs of the pooling layer are concatenated together from all kernels into one feature vector and input into the fully connected layer. The nonlinear softmax activation function is used to perform the task of classification.

To avoid over-fitting, a dropout layer is implemented before the output, in which the output of every neuron is set to zero with a probability of 0.5. The dropped-out neurons are not included in the forward pass or the back-propagation (Hinton et al., 2012).

Prediction of Novel CircRNA–Disease Associations

Next, we used all the positive and negative circRNA–disease association samples to train the MSFCNN architecture. Then, MSFCNN was used to score the unlabeled associations between circRNAs and diseases. Owing to the different negative samples used to train the model in each iteration of the five-fold cross validation (five-fold CV), we scored the candidate associations 10 times. Finally, we calculated the average scores for the candidate associations, and the candidate circRNAs related to specific diseases were analyzed using case studies.

RESULTS

Performance Evaluation

The performance of MSFCNN and other conventional machine learning-based methods for predicting circRNA–disease associations was evaluated using five-fold CV. If the circRNA $c(i)$ was found to be related to disease $d(j)$, the node pair c_i-d_j was considered as a positive example. Hence, the validated circRNA–disease associations were regarded as the positive set. However, because of the unavailability of a dataset for negative samples, we randomly selected a negative set from unobserved associations that was the same size as the positive set. All the positive samples were divided into five subsets of equal size, and each subset was tested once. For each CV, we took four positive subsets and the same number of negative subsets from five subsets to train the models; the remaining one positive subset and one negative subset were used for testing to evaluate the prediction performance. To lessen the bias resulting from sample division, we performed 10 repetitions of five-fold CV and obtained the average values of five experiments.

Receiver operating characteristic (ROC) curves were plotted to show the prediction performance by calculating the true positive rate and false positive rate. The area under the curve (AUC) was calculated to evaluate the overall performance. In addition, five metrics, precision (Pre), sensitivity (Sen), accuracy (Acc), $F1$-$score$, and Matthews's correlation coefficient (MCC) were used

to evaluate the capability of the MSFCNN model. The detailed calculation of these metrics was as follows:

$$Pre = \frac{TP}{TP + FP} \tag{17}$$

$$Sen = \frac{TP}{TP + FN} \tag{18}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$F1 - score = \frac{2 \times Sen \times Pre}{Sen + Pre} \tag{20}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FN) * (TN + FP)}} \tag{21}$$

where TP and TN represent the number of true positives and true negatives, respectively, and FP and FN represent the number of positives and negatives, respectively, that were wrongly predicted.

Parameter Setting

Convergence and parameter selection are important factors in the SKF method, that is, the number of iterations and two parameters, α and the size of neighbors. Following a previous study (Jiang et al., 2018), we set these two parameters to 0.1 and 36, respectively. As the number of iterations is important for the convergence of the SKF method, we also analyzed whether the number of iterations was sufficient for convergence in the four circRNA similarity kernels and seven disease similarity kernels. The relative error of the process of iteration was denoted EC_t and ED_t for circRNA similarity fusion and disease similarity fusion, respectively. The number of iterations ranged from 1 to 25 with steps of 1, and EC_t and ED_t were computed after every iteration. The convergence processes of the four circRNA similarity kernels and seven disease similarity kernels are shown in **Figure 4**. The results indicate that the convergence process was fast, and the EC_t and ED_t values reached 10^{-10} after 10 iterations. Therefore, we set the number of iterations to 10 for both circRNA similarity fusion and disease similarity fusion.

$$EC_t = \frac{\| SC_{c,m}^{t+1} - SC_{c,m}^{t} \|}{\| SC_{c,m}^{t} \|} \tag{22}$$

$$ED_t = \frac{\| SD_{d,n}^{t+1} - SD_{d,n}^{t} \|}{\| SD_{d,n}^{t} \|} \tag{23}$$

In the convolution operation of the MSFCNN model, the number of filters was set to 8. The kernel size was set to 2×32 in the first convolutional layer and 1×16 in the second convolutional layer. We implemented the MSFCNN model using the Keras 2.2.4 library in Python 3.7.3.

Evaluation of Prediction Performance

To assess the performance of the MSFCNN model for prediction of circRNA–disease associations, we used five-fold CV with 10 experiments (see **Table 1** and **Figure 5** for details). MSFCNN achieved average precision, sensitivity, *F1-score*, *Acc*, *MCC*, and AUC values of 0.9030, 0.9464, 0.9240, 0.9220, 0.8452, and 0.9525, with standard deviations of 0.0360, 0.0256, 0.0292, 0.0305, 0.0605, and 0.0202, respectively. Furthermore, the ROC curves for the MSFCNN model were at the upper left of the picture. These results indicate that our proposed model performs well in prediction of circRNA–disease associations.

Comparison With Average Kernel Fusion Strategy

In the MSFCNN model, the SKF method is used to fuse the four circRNA similarity kernels and seven disease similarity kernels into one circRNA similarity kernel and one disease similarity kernel, respectively. We compared the performance of the SKF method when integrating several similarity kernels with that of an average kernel fusion strategy. The average fusion strategy calculated the average similarity scores for four circRNA similarity matrix or seven disease similarity matrices, respectively. Five-fold CV was performed 10 times for predicting

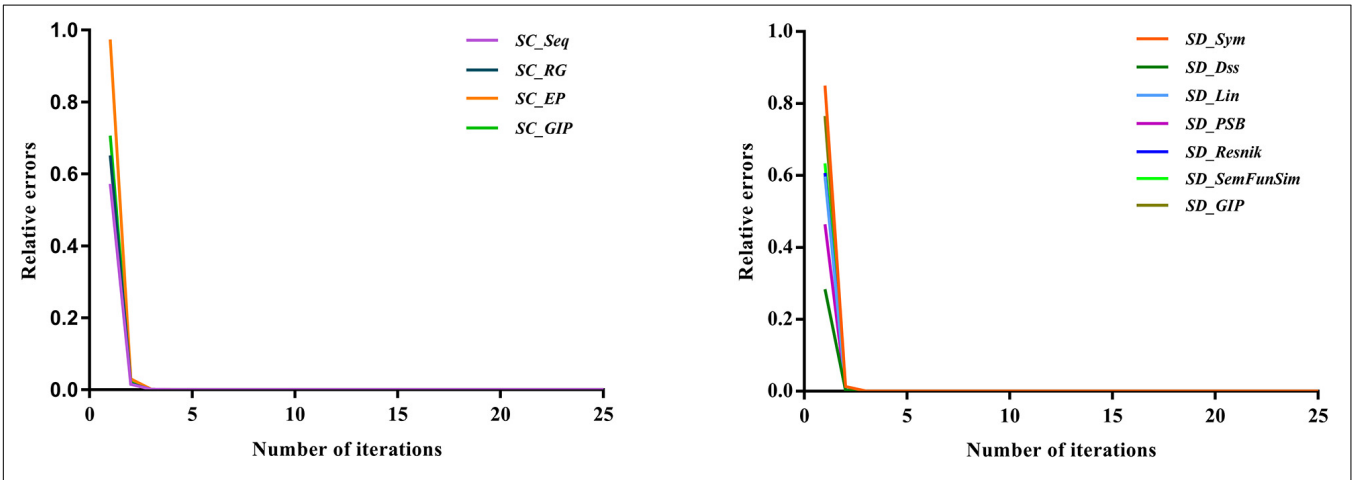

FIGURE 4 | Relative errors of the SKF method with various numbers of iteration for the four circRNA similarity matrices and seven disease similarity matrices.

TABLE 1 | Evaluation metrics for performance of the MSFCNN approach.

Times	*Pre*	*Sen*	*F1-score*	*Acc*	*MCC*
1	0.9573	0.9677	0.9625	0.9623	0.9246
2	0.8488	0.9380	0.8912	0.8854	0.7752
3	0.9251	0.9326	0.9289	0.9286	0.8572
4	0.8660	0.9057	0.8854	0.8827	0.7663
5	0.9203	0.9650	0.9421	0.9407	0.8824
6	0.9010	0.9568	0.9281	0.9259	0.8534
7	0.9258	0.9757	0.9501	0.9488	0.8989
8	0.8641	0.9084	0.8857	0.8827	0.7665
9	0.8835	0.9407	0.9112	0.9084	0.8184
10	0.9377	0.9730	0.9550	0.9542	0.9090
Average	0.9030+/−0.0360	0.9464+/−0.0256	0.9240+/−0.0292	0.9220+/−0.0305	0.8452+/−0.0605

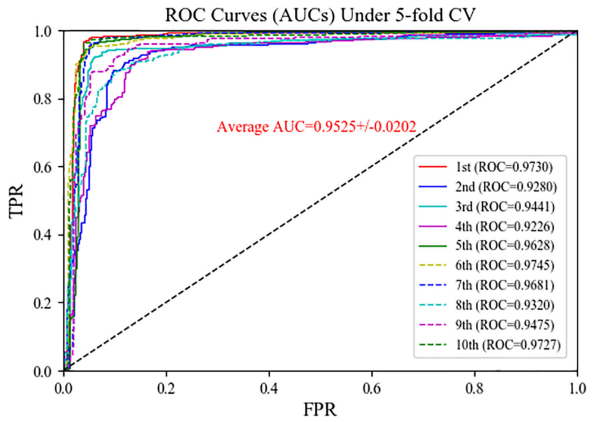

FIGURE 5 | ROC curves of MSFCNN model for the task of circRNA–disease association prediction.

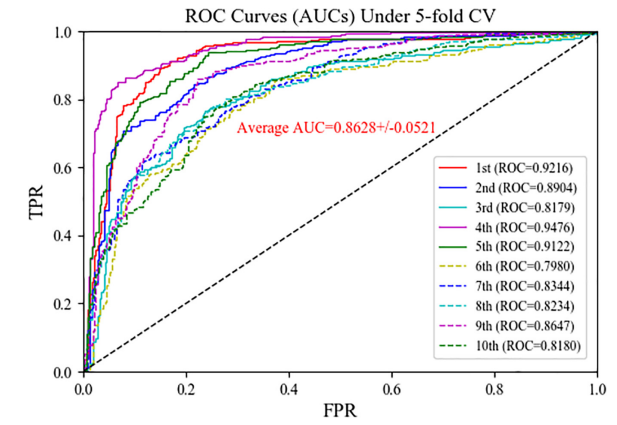

FIGURE 6 | ROC curves of the MSFCNN model with average kernel fusion strategy.

TABLE 2 | Evaluation metrics for performance of the MSFCNN model with average kernel fusion strategy.

Times	*Pre*	*Sen*	*F1-score*	*Acc*	*MCC*
1	0.8448	0.8948	0.8691	0.8653	0.7317
2	0.7889	0.8464	0.8166	0.8100	0.6216
3	0.7834	0.7116	0.7458	0.7574	0.5170
4	0.8832	0.8760	0.8796	0.8801	0.7601
5	0.8342	0.8410	0.8376	0.8369	0.6738
6	0.7186	0.7709	0.7438	0.7345	0.4702
7	0.7171	0.7925	0.7529	0.7398	0.4825
8	0.7649	0.7278	0.7459	0.7520	0.5046
9	0.7778	0.8679	0.8204	0.8100	0.6242
10	0.7357	0.7951	0.7642	0.7547	0.5111
Average	0.7848+/−0.0553	0.8123+/−0.0629	0.7976+/−0.0534	0.7941+/−0.0537	0.5897+/−0.1070

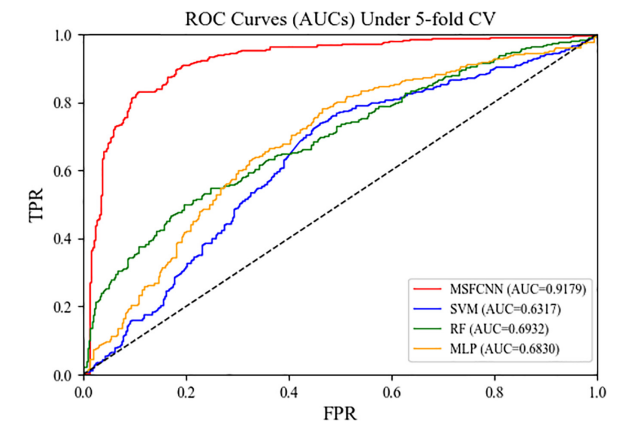

FIGURE 7 | ROC curves of the MSFCNN model and other machine learning methods.

circRNA–disease associations. The average kernel fusion-based MSFCNN model had an average AUC of 0.8628 (**Figure 6**); by comparison, the SKF-based MSFCNN model had an AUC of 0.9525 (an improvement of 0.0897). Other evaluation metrics also indicated that the SKF method performs better than the average kernel fusion strategy in MSFCNN (**Table 2**). Hence, the SKF method is an effective fusion strategy for prediction of circRNA–disease associations.

Comparison With Conventional Machine Learning Approaches

To demonstrate the reliability and robustness of the MSFCNN method, we made comparisons with state-of-the-art machine learning approaches: support vector machine (SVM), random forest (RF), and multilayer perception (MLP). For each of these machine learning approach, the feature matrix fed into the model was consistent with that used for MSFCNN to ensure the fairness of the experiments. As shown in **Figure 7**, the average AUC of the MSFCNN model in the five-fold CV was 0.9179 higher than those of the SVM, RF, and MLP methods. In addition, MSFCNN achieved higher precision, sensitivity, *F1-score*, *Acc*, and *MCC* values than the other machine learning approaches (**Table 3**). Therefore, the proposed method is more suitable than these conventional approaches for the task of circRNA–disease association prediction.

TABLE 3 | Evaluation metrics for performance of the MSFCNN and other tmachine learning methods.

Methods	*Pre*	*Sen*	*F1-score*	*Acc*	*MCC*
MSFCNN	0.8468	0.8491	0.8479	0.8477	0.6954
SVM	0.6166	0.6415	0.6288	0.6213	0.2428
RF	0.6851	0.5337	0.6000	0.6442	0.2957
MLP	0.6455	0.6577	0.6515	0.6482	0.2965

TABLE 4 | Candidate circRNAs predicted by the MSFCNN model for four diseases.

Diseases	circRNAs	Rank	Evidence
Acute myeloid leukemia	hsa_circ_0000677	3	Circ2Traits
	hsa_circ_0000175	6	Circ2Traits
Breast cancer	hsa_circ_0000677	8	Circ2Traits
	hsa_circ_0000175	11	Circ2Traits
	hsa_circ_0001417	25	Circ2Traits
Colorectal cancer	hsa_circ_0001417	16	Circ2Traits
	hsa_circ_0000175	19	Circ2Traits
	hsa_circ_0001283	40	Circ2Traits
	hsa_circ_0000615	56	Circ2Traits
Hepatocellular	hsa_circ_0000677	10	Circ2Traits
	hsa_circ_0001417	24	Circ2Traits
	hsa_circ_0001283	48	Circ2Traits

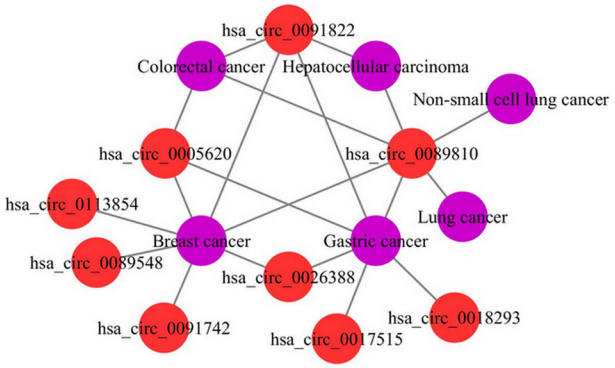

FIGURE 8 | Top 20 predicted circRNA–disease associations.

Case Study

To further demonstrate the ability of the MSFCNN model to discover potential circRNA–disease associations, we scored unlabeled associations between circRNAs and diseases using the trained model. Average scores were obtained from 10 applications of the MSFCNN model, and candidate circRNA–disease associations were identified based on their ranked scores. Case studies were performed for breast cancer, colorectal cancer, hepatocellular carcinoma, and acute myeloid leukemia. Some of the predicted specific disease-related circRNAs were found in the Circ2Traits database (Ghosal et al., 2013), which collects circRNAs and miRNAs related to diseases and traits (**Table 4**). In addition, we plotted the top 20 predicted circRNA–disease associations; the results show that these circRNAs may be related to the same diseases, and the diseases may also be associated with the same circRNAs (**Figure 8**). Hence, these results show that the MSFCNN model could be an effective tool for the prediction of circRNA–disease associations.

CONCLUSION

Prioritizing potential disease-related circRNAs based on various types of prior information is beneficial to understanding disease mechanisms, diagnosis, and treatment. In this study, we developed a novel computational method named MSFCNN to predict potential circRNA–disease associations, using a two-layer two-dimensional CNN and integrating multiple biological data. First, one of the crucial technical points for predicting circRNA–disease associations is the similarity calculation for circRNA–circRNA and disease–disease pairs. Therefore, we calculated four circRNA similarity kernels and seven disease similarity kernels based on multiple biological and topological information. In addition, similarity kernel fusion was used to integrate various similarity kernels into one circRNA similarity kernel and one disease similarity kernel. Based on these fused similarity kernels and interactions/features among circRNAs, miRNA, and diseases, a feature matrix was constructed for each circRNA–disease pair. Finally, a two-layer CNN architecture was used to predict circRNA–disease associations. The MSFCNN approach showed good performance based on the five-fold CV, outperforming the SVM, RF, and MLP classifiers. Furthermore, case studies of breast cancer, colorectal cancer, hepatocellular carcinoma, and acute myeloid leukemia demonstrated that the MSFCNN framework could be an effective tool for successfully inferring potential circRNA–disease associations and providing a basis for biological validation.

The good performance of MSFCNN method mainly conclude following aspects. Firstly, multiple similarity kernels for circRNAs and diseases are effectively introduced to measure the biological and topological features of circRNAs and diseases. Secondly, the relationships of circRNA–miRNA and disease–miRNA are also used to construct the feature matrix for each circRNA–disease pair. Furthermore, the application of CNN architecture guarantees the effectiveness of learning the meaning of combinations of features from the feature matrix. Hence, MSFCNN method is an effective biomedical resource to predict the circRNA–disease associations.

Despite its promising prediction performance, the MSFCNN approach has some limitations. First, incomplete and noisy circRNA–disease associations were used as positive samples, and negative samples are randomly selected, limiting the prediction performance. This could be improved as more associations are discovered. Furthermore, more reliable biological information should be considered, such as circRNA coding potential and circRNA functional information, as well as disease phenotypes and functional information, etc. In addition, optional similarity measurements would be integrated based on comparing the prediction results of different similarity measures. Therefore,

more data sources should be collected, and a more effective model needs to be developed to address the above limitations.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://bioinfo.snnu.edu.cn/CircR2Disease/, http://www.circbank.cn/, https://disease-ontology.org/, http://bioannotation.cn:18080/DincRNAClient/#/Home, https://www.nlm.nih.gov/mesh/, http://www.cuilab.cn/hmdd, and http://www.exorbase.org.

AUTHOR CONTRIBUTIONS

XL and YP conceptualized the study. CF and XL performed the data collection, designed the method, and drafted the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61972451, 61672334, and 61902230) and the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (GK201901010).

REFERENCES

Bahn, J. H., Zhang, Q., Li, F., Chan, T.-M., Lin, X., Kim, Y., et al. (2015). The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clin. Chem.* 61, 221–230. doi: 10.1373/clinchem.2014.230433

Barrett, S. P., Wang, P. L., and Salzman, J. (2015). Circular RNA biogenesis can proceed through an exon-containing lariat precursor. *eLife* 4:e07540. doi: 10.7554/eLife.07540

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Li, J., Ju, P., Peng, J., and Wang, Y. (2014). SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One* 9:e99415. doi: 10.1371/journal.pone.0099415

Chioccarelli, T., Manfrevola, F., Ferraro, B., Sellitto, C., Cobellis, G., Migliaccio, M., et al. (2019). Expression patterns of circular RNAs in high quality and poor quality human spermatozoa. *Front. Endocrinol.* 10:435. doi: 10.3389/fendo.2019.00435

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* 12, 2493–2537.

Conn, V. M., Hugouvieux, V., Nayak, A., Conos, S. A., Capovilla, G., Cildir, G., et al. (2017). A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nat. Plants* 3:17053. doi: 10.1038/nplants.2017.53

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* 40, 3131–3142. doi: 10.1093/nar/gkr1009

Deng, L., Zhang, W., Shi, Y., and Tang, Y. (2019). Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci. Rep.* 9:9605. doi: 10.1038/s41598-019-45954-x

Ding, Y., Chen, B., Lei, X., Liao, B., and Wu, F. X. (2020). Predicting novel CircRNA-disease associations based on random walk and logistic regression model. *Comput. Biol. Chem.* 87:107287. doi: 10.1016/j.compbiolchem.2020.107287

Fan, C., Lei, X., Fang, Z., Jiang, Q., and Wu, F.-X. (2018a). CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. *Database* 2018:bay044. doi: 10.1093/database/bay044

Fan, C., Lei, X., and Wu, F. X. (2018b). Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. *Int. J. Biol. Sci.* 14, 1950–1959. doi: 10.7150/ijbs.28260

Fang, Z., and Lei, X. (2019). Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous network. *Big Data Min. Anal.* 2, 248–272.

Ge, E., Yang, Y., Gang, M., Fan, C., and Zhao, Q. (2019). Predicting human disease-associated circRNAs based on locality-constrained linear coding. *Genomics* 112, 1335–1342. doi: 10.1016/j.ygeno.2019.08.001

Ghosal, S., Das, S., Sen, R., Basak, P., and Chakrabarti, J. (2013). Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* 4:283. doi: 10.3389/fgene.2013.00283

Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *RNA* 20, 1666–1670. doi: 10.1261/rna.043687.113

Graves, A., Mohamed, A.-R., and Hinton, G. (2013). "Speech recognition with deep recurrent neural networks," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, 6645–6649.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv [Preprint]* Available online at: arXiv.org > cs > arXiv:1207.0580 (accessed July 3, 2012).

Huang, Y.-A., Chan, K. C. C., and You, Z.-H. (2018). Constructing prediction models from expression profiles for large scale lncRNA-miRNA interaction profiling. *Bioinformatics* 34, 812–819. doi: 10.1093/bioinformatics/btx672

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47, D1013–D1017. doi: 10.1093/nar/gky1010

Ji, P., Wu, W., Chen, S., Zheng, Y., Zhou, L., Zhang, J., et al. (2019). Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell Rep.* 26, 3444–3460.e5. doi: 10.1016/j.celrep.2019.02.078

Jiang, L., Ding, Y., Tang, J., and Guo, F. (2018). MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association. *Front. Genet.* 9:618. doi: 10.3389/fgene.2018.00618

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Lei, X., and Bian, C. (2020). Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association. *Sci. Rep.* 10:1943.

Lei, X., and Fang, Z. (2019). GBDTCDA: predicting circRNA-disease associations based on gradient boosting decision tree with multiple biological data fusion. *Int. J. Biol. Sci.* 15, 2911–2924. doi: 10.7150/ijbs.33806

Lei, X., Fang, Z., Chen, L., and Wu, F. X. (2018). PWCDA: path weighted method for predicting circRNA-disease associations. *Int. J. Mol. Sci.* 19:3410. doi: 10.3390/ijms19113410

Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., et al. (2018). exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res.* 46, D106–D112. doi: 10.1093/nar/gkx891

Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015). Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.* 25, 981–984. doi: 10.1038/cr.2015.82

Lin, D. (1998). "An information-theoretic definition of similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, Manitoba, 296–304.

Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., et al. (2018). Applications of deep learning to MRI images: a survey. *Big Data Min. Anal.* 1, 1–18. doi: 10.26599/bdma.2018.9020001

Liu, M., Wang, Q., Shen, J., Yang, B. B., and Ding, X. (2019). Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol.* 16, 899–905. doi: 10.1080/15476286.2019.1600395

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing* 234, 11–26. doi: 10.1016/j.neucom.2016.12.038

Mathur, S., and Dinakarpandian, D. (2012). Finding disease similarity based on implicit semantic similarity. *J. Biomed. Inform.* 45, 363–371. doi: 10.1016/j.jbi.2011.11.017

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Memczak, S., Papavasileiou, P., Peters, O., and Rajewsky, N. (2015). Identification and characterization of circular RNAs as a new class of putative biomarkers in human blood. *PLoS One* 10:e0141214. doi: 10.1371/journal.pone.0141214

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Brief. Bioinform.* 18, 851–869. doi: 10.1093/bib/bbw068

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, 807–814.

Qu, S., Liu, Z., Yang, X., Zhou, J., Yu, H., Zhang, R., et al. (2018). The emerging functions and roles of circular RNAs in cancer. *Cancer Lett.* 414, 301–309. doi: 10.1016/j.canlet.2017.11.022

Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Adelaide, 448–453.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/s0168-9525(00)02024-2

Ruan, H., Xiang, Y., Ko, J., Li, S., Jing, Y., Zhu, X., et al. (2019). Comprehensive characterization of circular RNAs in ∼ 1000 human cancer cell lines. *Genome Med.* 11:55.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., et al. (2019). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* 47, D955–D962. doi: 10.1093/nar/gky1032

Sun, P., and Li, G. (2019). CircCode: a powerful tool for identifying circRNA coding ability. *Front. Genet.* 10:981. doi: 10.3389/fgene.2019.00981

Tang, B., Hao, Z., Zhu, Y., Zhang, H., and Li, G. (2018). Genome-wide identification and functional analysis of circRNAs in *Zea mays*. *PLoS One* 13:e0202375. doi: 10.1371/journal.pone.0202375

Van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Vo, J. N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., et al. (2019). The landscape of circular RNA in cancer. *Cell* 176, 869–881.e13. doi: 10.1016/j.cell.2018.12.021

Wang, J., and Wang, L. (2019). Deep learning of the back-splicing code for circular RNA formation. *Bioinformatics* 35, 5235–5242. doi: 10.1093/bioinformatics/btz382

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, Y., Nie, C., Zang, T., and Wang, Y. (2019). Predicting circRNA-disease associations based on circRNA expression similarity and functional similarity. *Front. Genet.* 10:832. doi: 10.3389/fgene.2019.00832

Wang, Z., Lei, X., and Wu, F.-X. (2019). Identifying cancer-specific circRNA-RBP binding sites based on deep learning. *Molecules* 24:4035. doi: 10.3390/molecules24224035

Wei, H., and Liu, B. (2019). iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief. Bioinform.* 21, 1356–1367. doi: 10.1093/bib/bbz057

Wu, W., Ji, P., and Zhao, F. (2020). CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* 21:101. doi: 10.1186/s13059-020-02018-y

Xiao, Q., Luo, J., and Dai, J. (2019). Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework. *IEEE J. Biomed. Health Inform.* 23, 2661–2669. doi: 10.1109/jbhi.2019.2891779

Yan, C., Wang, J., and Wu, F. X. (2018). DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC Bioinformatics* 19(Suppl. 19):520. doi: 10.1186/s12859-018-2522-6

Yang, M., and Xu, S. (2020). A novel patch-based nonlinear matrix completion algorithm for image analysis through convolutional neural network. *Neurocomputing* 389, 56–82. doi: 10.1016/j.neucom.2020.01.037

Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., et al. (2017). Extensive translation of circular RNAs driven by N(6)-methyladenosine. *Cell Res.* 27, 626–641. doi: 10.1038/cr.2017.31

Zhang, W., Yu, C., Wang, X., and Liu, F. (2019). Predicting CircRNA-disease associations through linear neighborhood label propagation method. *IEEE Access* 7, 83474–83483. doi: 10.1109/access.2019.2920942

Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., et al. (2013). Circular intronic long noncoding RNAs. *Mol. Cell* 51, 792–806. doi: 10.1016/j.molcel.2013.08.017

Zhao, Q., Yang, Y., Ren, G., Ge, E., and Fan, C. (2019). Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans. Nanobiosci.* 18, 578–584. doi: 10.1109/tnb.2019.2922214

Zheng, K., You, Z. H., Li, J. Q., Wang, L., Guo, Z. H., and Huang, Y. A. (2020). iCDA-CGR: identification of circRNA-disease associations based on chaos game representation. *PLoS Comput. Biol.* 16:e1007872. doi: 10.1371/journal.pcbi.1007872

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms-disease network. *Nat. Commun.* 5:4212. doi: 10.1038/ncomms5212

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 November 2020
doi: 10.3389/fgene.2020.581594

Exploring the Link Between Additive Heritability and Prediction Accuracy From a Ridge Regression Perspective

Arthur Frouin[1*], *Claire Dandine-Roulland*[1], *Morgane Pierre-Jean*[1], *Jean-François Deleuze*[1,2], *Christophe Ambroise*[3*†] and *Edith Le Floch*[1*†]

[1] CNRGH, Institut Jacob, CEA - Université Paris-Saclay, Évry, France, [2] Centre d'Etude du Polymorphisme Humain, Fondation Jean Dausset, Paris, France, [3] LaMME, Université Paris-Saclay, CNRS, Université d'Évry val d'Essonne, Évry, France

Genome-Wide Association Studies (GWAS) explain only a small fraction of heritability for most complex human phenotypes. Genomic heritability estimates the variance explained by the SNPs on the whole genome using mixed models and accounts for the many small contributions of SNPs in the explanation of a phenotype. This paper approaches heritability from a machine learning perspective, and examines the close link between mixed models and ridge regression. Our contribution is two-fold. First, we propose estimating genomic heritability using a predictive approach via ridge regression and Generalized Cross Validation (GCV). We show that this is consistent with classical mixed model based estimation. Second, we derive simple formulae that express prediction accuracy as a function of the ratio $\frac{n}{p}$, where n is the population size and p the total number of SNPs. These formulae clearly show that a high heritability does not imply an accurate prediction when $p > n$. Both the estimation of heritability via GCV and the prediction accuracy formulae are validated using simulated data and real data from UK Biobank.

Keywords: heritability, prediction accuracy, ridge regression, mixed model, generalized cross validation, best linear unbiased predictor

OPEN ACCESS

Edited by:
Lide Han,
Vanderbilt University Medical Center, United States

Reviewed by:
Kui Zhang,
Michigan Technological University, United States
Xiaowei Wu,
Virginia Tech, United States
Guo-Bo Chen,
Zhejiang Provincial People's Hospital, China

***Correspondence:**
Christophe Ambroise
christophe.ambroise@univ-evry.fr
Arthur Frouin
arthurs.frouin@gmail.com
Edith Le Floch
edith.lefloch@cnrgh.fr

†These authors share last authorship

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

Received: *09 July 2020*
Accepted: *29 September 2020*
Published: *04 November 2020*

Citation:
Frouin A, Dandine-Roulland C, Pierre-Jean M, Deleuze J-F, Ambroise C and Le Floch E (2020) Exploring the Link Between Additive Heritability and Prediction Accuracy From a Ridge Regression Perspective. Front. Genet. 11:581594. doi: 10.3389/fgene.2020.581594

1. INTRODUCTION

The old nature vs. nurture debate is about whether a complex human trait is determined by a person's genes or by the environment. It is a longstanding philosophical question that has been reinvestigated in the light of statistical genetics (Feldman and Lewontin, 1975). The concept of heritability was introduced by Fisher (1918) and Wright (1920, 1921) in the context of pedigree data. It has proved highly useful in animal (Meuwissen et al., 2001) and plant genetics (Xu, 2003) for selection purposes because of its association with accurate prediction of a trait from genetic data. In the last decades, Genome-Wide Association Studies (GWAS) have become highly popular for identifying variants associated with complex human traits (Hirschhorn and Daly, 2005). They have recently been used for heritability estimations (Yang et al., 2010). A shortcut is often made between the heritability of a trait and the prediction of this trait. However, heritable complex human traits are often caused by a large number of genetic variants that individually make small contributions to the trait variation, which is often referred to as polygeny. In this context, the relation between heritability and prediction accuracy may not hold (de Vlaming and Groenen, 2015).

The goal of this paper is to establish a clear relation between prediction accuracy and heritability, especially when the number of genetic markers is much higher than the population size, which is typically the case in GWAS. Based on the linear model, statistical analyses of SNP data address very different and sometimes unrelated questions. The most commonly performed analyses tend to be association studies, where multiple hypothesis testing makes it possible to test the link between any SNP and a phenotype of interest. In genomic selection, markers are selected to predict a phenotype with a view to selecting an individual in a breeding population. Association studies and genomic selection may identify different sets of markers, since even weak associations might be of interest for prediction purposes, while not all strongly associated markers are necessarily useful, because of redundancy through linkage disequilibrium. Genomic heritability allows quantifying the amount of genomic information relative to a given phenotype via mixed model parameter estimation. The prediction of the phenotype using all genomic information via the mixed model is a closely related but different problem.

We approach the problem of heritability estimation from a machine learning perspective. This is not a classical approach in genetics, where inferential statistics is the usual preferred tool. In this context, heritability is considered as a parameter to be inferred from a small sample of the population. The machine learning approach places the emphasis on prediction accuracy. It makes a clear distinction between performance on training samples and performance on testing samples, whereas inferential statistics focuses on parameter estimation on a single dataset.

1.1. Classical Approach via Mixed Models

Heritability is defined as the proportion of phenotypic variance due to genetic factors. A quantitative definition of heritability requires a statistical model. The model commonly adopted is a simple three-term model without gene-environment interaction (Henderson, 1975):

$$\mathbf{y} = \mathbf{g} + \mathbf{f} + \mathbf{e},$$

where $\mathbf{y} \in \mathbb{R}^n$ is a quantitative phenotype vector describing n individuals, $\mathbf{f} \in \mathbb{R}^n$ is a non-genetic covariate term, $\mathbf{g} \in \mathbb{R}^n$ is a genetic term and $\mathbf{e} \in \mathbb{R}^n$ an environmental residual term. The term $\mathbf{g}$ will depend on the diploid genotype matrix $\mathbf{M} \in \mathcal{M}_{n,p}(\mathbb{R})$ of the p causal variants.

There are two definitions of heritability in common use: first, there is H^2, heritability in the broad sense, measuring the overall contribution of the genome; and second, there is h^2 , heritability in the narrow sense (also known as additive heritability), defined as the proportion of phenotypic variance explained by the additive effects of variants.

The quantification of narrow-sense heritability goes back to family studies by Fisher (1918), who introduced the above model with the additional hypothesis that $\mathbf{g}$ is the sum of independent genetic terms, and with $\mathbf{e}$ assumed to be normal. This heritability in the narrow sense is a function of the correlation between the phenotypes of relatives.

Although Fisher's original model makes use of pedigrees for parameter estimation, some geneticists have proposed using the same model with genetic data from unrelated individuals (Yang et al., 2011a).

1.1.1. Polygenic Model

In this paper, we focus on the version of the additive polygenic model with a Gaussian noise where $\mathbf{g} = \mathbf{Z}u$, $\mathbf{f} = \mathbf{X}\beta$, with $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$ a standardized (by columns) version of $\mathbf{M}$, $u \in \mathbb{R}^p$ a vector of genetic effects, $\mathbf{X} \in \mathcal{M}_{n,r}(\mathbb{R})$ a matrix of covariates, $\beta \in \mathbb{R}^r$ a vector of covariate effects, μ an intercept and $\mathbf{e} \sim \mathcal{N}\left(\mathbf{0}_n, \sigma^2\mathbf{I}_n\right)$ a vector of environmental effects.

The model thus becomes

$$\mathbf{y} = \mu\mathbb{1}_n + \mathbf{Z}u + \mathbf{X}\beta + \mathbf{e}, \tag{1}$$

where $\mathbb{1}_n \in \mathbb{R}^n$ a vector of ones.

1.1.2. Estimation of Heritability From GWAS Results

To estimate heritability in a GWAS context, a first intuitive approach would be to estimate u with a least squares regression to solve problem (1). Unfortunately, this is complicated in practice for three reasons: the causal variants are not usually available among genotyped variants; genotyped variants are in linkage disequilibrium (LD); and the least squares estimate is only defined when $n > p$, which is not often the case in a GWAS (Yang et al., 2010).

One technique for obtaining a solvable problem is to use the classical GWAS approach to determine a subset of variants significantly associated with the phenotype. The additive heritability can then be estimated by summing their effects estimated by simple linear regressions. In practice this estimation tends to greatly underestimate h^2 (Manolio et al., 2009). It only takes into account variants that have passed the significance threshold after correction for multiple comparisons (strong effects) and does not capture the variants that are weakly associated with the phenotype (weak effects).

1.1.3. Estimating Heritability via the Variance of the Effects

Yang et al. (2010) suggest that most of the missing heritability comes from variants with small effects. In order to be able to estimate the information carried by weak effects they assume a linear mixed model where the vector of random genetic effects follows a normal homoscedastic distribution $u \sim \mathcal{N}\left(0_p, \tau\mathbf{I}_p\right)$. They propose estimating the variance components τ and σ^2, and defining genomic heritability as $h^2_G = \frac{p\tau}{p\tau+\sigma^2}$. An example of an algorithm for estimating variance components is the Average Information—Restricted Maximum Likelihood (AI-REML) algorithm, implemented in software such as Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011a) or `gaston` (Perdry and Dandine-Roulland, 2018). More recent methods that are much faster than REML have also been proposed, such as the modified Haseman-Elston regression (Chen, 2014) or methods based on summary statistics such as the LD-score regression (Bulik-Sullivan et al., 2015) or the MQS (MinQue for Summary statistics) approach (Zhou, 2017).

1.2. A Statistical Learning Approach via Ridge Regression

The linear model is used in statistical genetics for exploring and summarizing the relation between a phenotype and one or more genetic variants, and it is also used in predictive medicine and genomic selection for prediction purposes. When used for prediction, the criterion for assessing performance is the prediction accuracy.

Although least squares linear regression is the baseline method for quantitative phenotype prediction, it has some limitations. As mentioned earlier, the estimator is not defined when the number of descriptive variables p is greater than the number of individuals n. Even when $n > p$, the estimator may be highly variable when the descriptive variables are correlated, which is clearly the case in genetics.

Ridge regression is a penalized version of least squares that can overcome these limitations (Hoerl and Kennard, 1970). Ridge regression is strongly related to the mixed model and is prediction-oriented.

1.2.1. Ridge Regression

The ridge criterion builds on the least squares criterion, adding an extra penalization term. The penalization term is proportional to the ℓ_2 norm of the parameter vector. The proportionality coefficient λ is also called the penalization parameter. The penalty tends to shrink the coefficients of the least squares estimator, but never cancels them out. The degree of shrinkage is controlled by λ: the higher the value of λ, the greater the shrinkage:

$$\hat{u}_R = \arg\min_u \left\|\mathbf{y} - \mathbf{Z}u\right\|_2^2 + \lambda \left\|u\right\|_2^2, \tag{2}$$

$$= \left(\mathbf{Z}^T\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1} \mathbf{Z}^T\mathbf{y}, \tag{3}$$

$$= \mathbf{Z}^T \left(\mathbf{Z}\mathbf{Z}^T + \lambda \mathbf{I}_n\right)^{-1} \mathbf{y}. \tag{4}$$

Ridge regression can be seen as a *Bayesian Maximum a Posteriori* estimation of the linear regression parameters considering a Gaussian prior with hyperparameter λ.

The estimator depends on a λ that needs to be chosen. In a machine learning framework, a classical procedure is to choose the λ that minimizes the squared loss over new observations.

The practical effect of the penalty term is to add a constant to the diagonal of the covariance matrix, which makes the matrix non-singular, even in the case where $p > n$. When the descriptive variables are highly correlated, this improves the conditioning of the $\mathbf{Z}^T\mathbf{Z}$ matrix, while reducing the variance of the estimator.

The existence theorem states that there always exists a value of $\lambda > 0$ such that the Mean Square Error (MSE) of the ridge regression estimator (variance plus the squared bias) is smaller than the MSE of the Maximum Likelihood estimator (Hoerl and Kennard, 1970). This is because there is always an advantageous bias-variance compromise that reduces the variance without greatly increasing the bias.

Ridge regression also allows us to simultaneously estimate all the additive effects of the genetic variants without discarding any, which reflects the idea that all the variants make a small contribution.

1.2.2. Link Between Mixed Model and Ridge Regression

This paper builds on the parallel between BLUPs (Best Linear Unbiased Predictions) derived from the mixed model and ridge regression (Meuwissen et al., 2001). The use of ridge regression in quantitative genetics has already been discussed (De los Campos et al., 2013; de Vlaming and Groenen, 2015) We look at a machine-learning oriented paradigm for estimating the ridge penalty parameter, which provides us with a direct link to heritability. There is an equivalence between maximizing the posterior $p\left(u|\mathbf{y}\right)$ and minimizing a ridge criterion (Bishop, 2006) under the assumptions that $u \sim \mathcal{N}\left(0_p, \tau\mathbf{I}_p\right)$ and $\mathbf{e} \sim \mathcal{N}\left(0_n, \sigma^2\mathbf{I}_n\right)$ (see section 6 in **Supplementary Material** for details). The optimal penalty hyperparameter of the ridge criterion λ can be used to estimate the heritability. It is indeed defined as the ratio of the variance parameters of the mixed model:

$$\arg\max_u p\left(u|\mathbf{y}\right) = \arg\min_u \left\|\mathbf{y} - \mathbf{Z}u\right\|_2^2 + \lambda \left\|u\right\|_2^2 \text{ with } \lambda = \frac{\sigma^2}{\tau}. \tag{5}$$

The relation between λ and h_G^2 (de Vlaming and Groenen, 2015) is thus:

$$h_G^2 = \frac{p}{p+\lambda}; \lambda = p\frac{1-h_G^2}{h_G^2}. \tag{6}$$

Consequently, the BLUP has a similar formulation to the ridge estimator. Indeed, as shown in the section 6.2 of the article by Robinson et al. (1991), its general definition is:

$$\hat{u}_{BLUP} = \widehat{\mathbb{E}(u|\mathbf{y})} = \hat{\mathbf{W}}\mathbf{Z}^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}), \tag{7}$$

where $u \sim \mathcal{N}\left(0_p, \mathbf{W}\right)$, $\mathbf{e} \sim \mathcal{N}\left(0_n, \mathbf{E}\right)$ and $\mathbf{\Sigma} = \mathbf{Z}\mathbf{W}\mathbf{Z}^T + \mathbf{E}$.

When we further assume $\beta = 0_r$, $\mathbf{W} = \tau\mathbf{I}_p$ and $\mathbf{E} = \sigma^2\mathbf{I}_n$, it becomes:

$$\hat{u}_{BLUP} = \tau\mathbf{Z}^T(\tau\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} = \mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \frac{\sigma^2}{\tau}\mathbf{I}_n)^{-1}\mathbf{y}, \tag{8}$$

which is exactly the ridge estimator.

1.2.3. Over-Fitting

Interestingly, ridge regression and the mixed model can be seen as two similar ways to deal with the classical over-fitting issue in machine learning, which is where a learner becomes overspecialized in the dataset used for the estimation of its parameters and is unable to generalize (Bishop, 2006). When $n > p$, estimating the parameters of a fixed-effect linear model via maximum likelihood estimation may lead to over-fitting, when too many variables are considered. A classical way of reducing over-fitting is regularization, and in order to set the value of the regularization parameter there are two commonly adopted approaches: first, the Bayesian approach, and second, the use of additional data.

Mixed Model parameter estimation via maximum likelihood can be seen as a type of self-regularizing approach (see Equation 5). Estimating the variance components of the mixed model may

be interpreted as a kind of empirical Bayes approach, where the ratio of the variances is the regularization parameter that is usually estimated using a single dataset. In contrast to this, in order to properly estimate the ridge regression regularization hyperparameter that gives the best prediction, two datasets are required. If a single dataset were to be used, this would result in an insufficiently regularized (i.e., excessively complex) model offering too high prediction performances on the present dataset but unable to predict new samples well. This over-fitting phenomenon is particularly evident when dimensionality is high.

The fact that the complexity of the ridge model is controlled by its hyperparameter can be intuitively understood when considering extreme situations. When λ tends to infinity, the estimated effect vector (i.e., $\hat{u}_R$) tends to the null vector. Conversely, when λ tends to zero, the model approaches maximum complexity. One solution for choosing the right complexity is therefore to use both a training set to estimate the effect vector for different values of the hyperparameter and a validation set to choose the hyperparameter value with the best prediction capacity on this independent sample. An alternative solution, when data is sparse, is to use a cross-validation approach to mimic a two-set situation. Finally, it should be noted that the estimation of prediction performance on a validation dataset is still overoptimistic, and consequently a third dataset, known as a test set, is required to assess the real performance of the model.

1.2.4. Prediction Accuracy in Genetics

In genomic selection and in genomic medicine, several authors have been interested in predicting complex traits that show a relatively high heritability using mixed model BLUPs (Speed and Balding, 2014). The literature defined the prediction accuracy as the correlation between the trait and its prediction, which is unusual in machine learning where the expected loss is often preferred. Several approximations of this correlation have been proposed in the literature (Brard and Ricard, 2015), either in a low-dimensional context (where the number of variants is lower than the number of individuals) or in a high-dimensional context.

Daetwyler et al. (2008) derived equations for predicting the accuracy of a genome-wide approach based on simple least-squares regressions for continuous and dichotomous traits. They consider one univariate linear regression per variant (with a fixed effect) and combine them afterwards, which is equivalent to a Polygenic Risk Score (PRS) (Pharoah et al., 2002; Purcell et al., 2009). Goddard (2009) extended this prediction to Genomic BLUP (GBLUP), which used the concept of an effective number of loci. Rabier et al. (2016) proposed an alternative correlation formula conditionally on a given training set. Their formula refines the formula proposed by Daetwyler et al. (2008). Elsen (2017) used a Taylor development to derive the same formula in small dimension.

Using intensive simulation studies, de Vlaming and Groenen (2015) showed a strong link between PRS and ridge regression in terms of prediction accuracy, when the population size is limited. However, with ridge regression, predictive accuracy improves substantially as the sample size increases.

It is important to note a difference in the prediction accuracy of GBLUP when dealing with human populations as opposed to breeding populations (De los Campos et al., 2013). De los Campos et al. (2013) show that the squared correlation between GBLUP and the phenotype reaches the trait heritability, asymptotically when considering unrelated human subjects. Dandine-Roulland and Perdry (2015) also proposed a theoretical formula of the performance of BLUPs for prediction in the context of human genetics, which is proportional to the number of individuals, to the squared heritability and to the variance of the off-diagonal terms of the Genetic Relatedness Matrix.

Zhao and Zhu (2019) studied cross trait prediction in high dimension. They derive generic formulae for in and out-of sample squared correlation. They link the marginal estimator to the ridge estimator and to GBLUP. Their results are very generic and generalize formulae proposed by Daetwyler et al. (2008).

1.2.5. Outline of the Paper

While some authors have proposed making use of the equivalence between ridge regression and the mixed model for setting the hyperparameter of ridge regression according to the heritability estimated by the mixed model, we propose on the contrary to estimate the optimal ridge hyperparameter using a predictive approach via Generalized Cross Validation. We derive approximations of the squared correlation and of the expected loss, both in high and low dimensions.

Using synthetic data and real data from UK Biobank, we show that our results are consistent with classical mixed model based estimation and that our approximations are valid.

Finally, with reference to the ridge regression estimation of heritability, we discuss how heritability is linked to prediction accuracy in highly polygenic contexts.

2. MATERIALS AND METHODS

2.1. Generalized Cross Validation for Speeding Up Heritability Estimation via Ridge Regression

2.1.1. Generalized Cross Validation

A classical strategy for choosing the ridge regression hyperparameter uses a grid search and k -fold cross validation. Each grid value of the hyperparameter is evaluated by the cross validated error. This approach is time-consuming in high dimension, since each grid value requires k estimations. In the machine learning context, we propose using Generalized Cross Validation (GCV) to speed up the estimation of the hyperparameter λ and thus to estimate the additive heritability h^2_G using the link described in Equation (6).

The GCV error in Equation (9) (Golub et al., 1978) is an approximation of the Leave-One-Out error (LOO) (see section 2 in **Supplementary Material**). Unlike the classical LOO, GCV does not require n ridge regression estimations (where n is the number of observations) at each grid value, but involves a single run. It thus provides a much faster and convenient alternative for choosing the hyperparameter. We have

$$\mathrm{err}^{GCV} = \frac{\left\|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\right\|_2^2}{\left[\frac{1}{n}\mathrm{tr}\left(\mathbf{I}_n - \mathbf{H}_\lambda\right)\right]^2}, \tag{9}$$

where $\hat{\mathbf{y}}(\lambda) = \mathbf{Z}\hat{u}_R(\lambda) = \mathbf{H}_\lambda \mathbf{y}$ is the prediction of the training set phenotypes using the same training set for the estimation of $\hat{u}_R$ and where $\mathbf{H}_\lambda$ is defined as (see section 1 in **Supplementary Material** for details):

$$\begin{aligned}\mathbf{H}_\lambda &= \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1}\mathbf{Z}^T \\ &= \mathbf{Z}\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T + \lambda \mathbf{I}_n\right)^{-1}.\end{aligned}$$

A Singular Value Decomposition (SVD) of the $\mathbf{H}_\lambda$ can be used advantageously to speed up GCV computation (see section 3 in **Supplementary Material**).

2.1.2. Empirical Centering Can Lead to Issues in the Choice of Penalization Parameter in a High-Dimensional Setting

In high dimensional settings ($p > n$), the use of GCV after empirical centering of the data can lead to a strong bias in the choice of λ and thus in heritability estimation. Let us illustrate the problem with a simple simulation. We simulate a phenotype from synthetic genotype data with a known heritability of $h^2 = 0.25$, $n = 1,000$ individuals, $p = 10,000$ variants and 100% causal variants. The simulation follows the additive polygenic model without intercept or covariates, as described in section 2.3. Before applying GCV, genotypes are standardized in the most naive way: the genotype matrix $\mathbf{M}$ is empirically centered and scaled column-wise, resulting in the matrix $\mathbf{Z}$. Since we want to mimic an analysis on real data, let us assume that there is a potential intercept in our model (in practice the empirical mean of our simulated phenotype is likely to be non-null):

$$\mathbf{y} = \mu \mathbb{1}_{\boldsymbol{n}} + \mathbf{Z}u + \mathbf{e}. \tag{10}$$

GCV expects all the variables to be penalized, but penalizing the intercept is not relevant. We therefore consider a natural two-step procedure: first the model's intercept is estimated via the empirical mean of the phenotype $\hat{\mu} = \frac{1}{n}\sum_i y_i$, and, second, GCV is applied on the empirically centered phenotype $\mathbf{y} = \mathbf{y} - \hat{\mu}\mathbb{1}_{\boldsymbol{n}}$.

Figure 1 shows the GCV error (dotted line). Heritability is strongly overestimated. The GCV error appears to tend toward its minimum as λ approaches 0 (i.e., when h^2 tends to 1).

This is a direct consequence of the empirical standardization of $\mathbf{M}$ and of the phenotype. By centering the columns of $\mathbf{M}$ with the empirical means of those columns, a dependency is introduced, and each line of the resulting standardized genotype matrix $\mathbf{Z}$ becomes a linear combination of all the others. The same phenomenon of dependency can be observed with the phenotype when using empirical standardization. Given the nature of the LOO in general (where each individual is considered successively as a validation set), this kind of standardization introduces a link between the validation set and the training set at each step: the "validation set individual" can be written as a linear combination of the individuals in the training set. In high dimension, this dependency leads to $\text{err}^{LOO} \xrightarrow[\lambda \to 0]{} 0$ (see section 4 in **Supplementary Material**), due to over-fitting occurring in the training set.

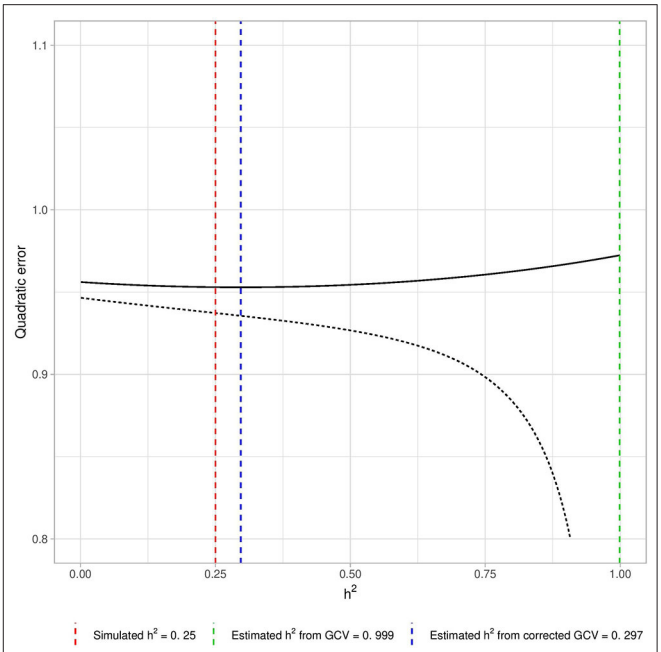

FIGURE 1 | Example of biased estimation by GCV if $p > n$. We computed the GCV error curve with $n = 1,000$ individuals, $p = 10,000$ causal variants and simulated heritability $h^2_{sim} = 0.25$. We used a grid of λ corresponding to the grid of heritability $\{0.01, 0.02, ..., 0.99\}$ using the link described in Equation (6) and computed the GCV error for those λ after empirical standardization of the data (dotted line). The λ that minimizes the GCV error corresponds to the heritability estimation. Here the GCV error tends to its minimum as h^2 tends to 1, and heritability is thus largely over-estimated. The plain line corresponds to the GCV error obtained after correction of this bias by the projection approach (see section 2.1.3), which provides a satisfactory estimation of h^2. The three vertical lines correspond respectively to the simulated heritability (red line), the heritability estimated using uncorrected GCV (green line) and the heritability estimated using corrected GCV (blue line).

From a GCV perspective, a related consequence of the empirical centering of the genotype data is that the matrix $\mathbf{Z}\mathbf{Z}^T$ has at least one null eigenvalue and an associated constant eigenvector in a high dimensional setting (see section 4 in **Supplementary Material**). This has a direct impact on GCV: using the singular value decomposition of the empirically standardized matrix $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{U} \in \mathcal{O}_n(\mathbb{R})$, $\mathbf{V} \in \mathcal{O}_p(\mathbb{R})$ two orthogonal squared matrices spanning, respectively, the lines and columns spaces of $\mathbf{Z}$ while $\mathbf{D} \in \mathcal{M}_{n,p}(\mathbb{R})$ is a rectangular matrix with singular values $\{d_1, ..., d_n\}$ on the diagonal. In a high dimensional context: $\text{err}^{GCV}(\mathbf{y}, \mathbf{Z}, \lambda) \xrightarrow[\lambda \to 0]{d_n^2 = 0} (\mathbb{1}_{\boldsymbol{n}}{}^T\mathbf{y})^2$. Performing the "naive" empirical centering of the phenotype results in

$$\text{err}^{GCV}(\mathbf{y} - \hat{\mu}\mathbb{1}_{\boldsymbol{n}}, \mathbf{Z}, \lambda) \xrightarrow[\lambda \to 0]{d_n^2 = 0} (\mathbb{1}_{\boldsymbol{n}}{}^T\mathbf{y} - \mathbb{1}_{\boldsymbol{n}}{}^T\hat{\mu}\mathbb{1}_{\boldsymbol{n}})^2 = 0.$$

The very same problem is observed for a more general model with covariates (see section 4 in **Supplementary Material**).

2.1.3. A First Solution Using Projection

A better solution for dealing with the intercept [and a matrix of covariates $\mathbf{X} \in \mathcal{M}_{n,r}(\mathbb{R})$] in ridge regression is to use a projection matrix as a contrast and to work on the orthogonal of the space spanned by the intercept (and the covariates).

Contrast matrices are a commonly used approach in the field of mixed models for REstricted Maximum Likelihood computations (REML) (Patterson and Thompson, 1971). REML provides maximum likelihood estimation once fixed effects are taken into account. Contrast matrices are used to "remove" fixed effects from the likelihood formula. If we are only interested in the estimation of the component of variance, we do not even need to make this contrast matrix explicit: any semi-orthogonal matrix $\mathbf{C} \in \mathcal{M}_{n-r-1,n}(\mathbb{R})$ such that $\mathbf{C}\mathbf{C}^T = \mathbf{I}_{n-r-1}$ and $\mathbf{C} \times (\mu \mathbb{1}_n + \mathbf{X}\beta) = 0_{n-r-1}$ provides a solution (see section 6 in **Supplementary Material** for details). In a ridge regression context, an explicit expression of $\hat{u}$ is needed for choosing the optimal complexity. An explicit form for $\mathbf{C}$ is therefore necessary.

In the presence of covariates, a QR decomposition can be used to obtain an explicit form for $\mathbf{C}$ (see section 5 in **Supplementary Material** for details). In the special case of an intercept without covariates, there is a convenient choice of $\mathbf{C}$. Since the eigenvector of $\mathbf{Z}\mathbf{Z}^T$ associated with the final null eigenvalue is constant, $\mathbf{C} = [\mathbf{U}_1, ..., \mathbf{U}_{n-1}]^T \in \mathcal{M}_{n-1,n}(\mathbb{R})$ is a contrast matrix adapted for our problem. Additionally, by considering $\mathbf{C}\mathbf{Z}$ instead of $\mathbf{Z}$, we have $\mathbf{C}\mathbf{Z} = \mathbf{D}_{-n}\mathbf{V}^T \rightarrow \mathbf{C}\mathbf{Z}\mathbf{Z}^T\mathbf{C}^T = \mathbf{D}_{-n}\mathbf{D}_{-n}^T$ with $\mathbf{D}_{-n}$ the matrix $\mathbf{D}$ deprived of row n. This choice of contrast matrix thus simplifies the GCV formula and allows extremely fast computation.

2.1.4. A Second Solution Using 2 Data Sets

Dependency between individuals can be a problem when we use the same data for the standardization (including the estimation of potential covariate effects) and for the estimation of the genetic effects. This can be overcome by partitioning our data. Splitting our data into a standardization set and a training set, we will first use the standardization set to estimate the mean and the standard deviation of each variant, the intercept, and the potential covariate effects. Those estimators will then be used to standardize the training set on which GCV can then be applied.

This method has two main drawbacks. The first is that the estimation of the non-penalized effects is done independently of the estimation of the genetic effects, even though in practice we do not expect covariates to be highly correlated with variants. The other drawback is that it reduces the number of individuals for the heritability estimation (which is very sensitive to the number of individuals). This approach therefore requires a larger sample than when using projection.

2.2. Prediction vs. Heritability in the Context of Small Additive Effects

Ridge regression helps to highlight the link between heritability and prediction accuracy. What is the relation between the two concepts? Is prediction accuracy an increasing function of heritability?

In a machine learning setting, we have training and testing sets. The classical bias-variance trade-off formulation considers the expectation of the loss over both the training set and the test individual phenotype. It breaks down the prediction error into three terms commonly called variance, bias, and irreducible error. In this paper we do consider the genotypes of the training set as fixed and the genotype of a test individual as random, and somewhat abusively continue to employ the terms variance, bias, and irreducible error:

$$\begin{aligned}\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] &= \mathbb{E}_{z_{te}}\left[\mathbb{E}_{\mathbf{y}_{tr},y_{te}|z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right]\right] \\ &= \mathbb{E}_{z_{te}}\left[\mathrm{var}(y_{te}|z_{te}) + \mathrm{var}(\hat{y}_{te}|z_{te})\right] + \\ &\quad \mathbb{E}_{z_{te}}\left[\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right] - \mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right].\end{aligned}$$

where the index $_{tr}$ refers to the training set, while $_{te}$ refers to the test set.

Assuming a training set genotype matrix $\mathbf{Z} \in \mathcal{M}_{n,p}(\mathbb{R})$ (without index $_{tr}$ to lighten notations) whose columns have zero mean and unit variances, we denote $\mathbf{K}_\lambda = \left(\mathbf{Z}^T\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1}\mathbf{Z}^T$. Assuming the independence of the variants $\mathbb{E}_{z_{te}}[z_{te}] = 0_p$ and $\mathrm{var}(z_{te}) = \mathbf{I}_p$, irreducible error, variance, and bias become:

$$\begin{aligned}\mathbb{E}_{z_{te}}\left[\mathrm{var}(y_{te}|z_{te})\right] &= \sigma^2 \\ \mathbb{E}_{z_{te}}\left[\mathrm{var}(\hat{y}_{te}|z_{te})\right] &= \sigma^2 \mathrm{tr}\left(\mathbf{K}_\lambda \mathbf{K}_\lambda^T\right) \\ \mathbb{E}_{z_{te}}\left[\left(\mathbb{E}_{\mathbf{y}_{tr}|z_{te}}\left[\hat{y}_{te}\right] - \mathbb{E}_{y_{te}|z_{te}}[y_{te}]\right)^2\right] &= u^T\left(\mathbf{K}_\lambda \mathbf{Z} - \mathbf{I}_p\right)^2 u.\end{aligned}$$

where u is the vector of the ridge parameters.

Since individuals are assumed to be unrelated, the covariance matrix of the individuals is diagonal. The covariance matrix of the variants is also diagonal, since variants are assumed independent. Assuming scaled data, $\mathbf{Z}\mathbf{Z}^T$ and $\mathbf{Z}^T\mathbf{Z}$ are the empirical estimations of covariance matrices of respectively the individuals and the variants (up to a p or n scaling factor). Two separate situations can be distinguished according to the n/p ratio. In the high-dimensional case where $p > n$, the matrix $\mathbf{Z}\mathbf{Z}^T$ estimates well the individuals' covariance matrix up to a factor p. Where $n > p$, on the other hand, $\mathbf{Z}^T\mathbf{Z}$ estimates well the covariance matrix of variants up to a factor n. Eventually, $\mathbf{Z}\mathbf{Z}^T \simeq p\mathbf{I}_n$ when $n < p$ and $\mathbf{Z}^T\mathbf{Z} \simeq n\mathbf{I}_p$ when $n > p$.

Assuming further that

- $\forall i \in [\![1, n]\!]\ \mathrm{var}(y_i) = 1$, we then have $\sigma^2 = 1 - h^2$,
- heritability is equally distributed among normalized variants i.e., $\forall j \in [\![1, p]\!]\ \mathrm{var}(u_j) = \frac{h^2}{p}$ (which is indeed the mixed model hypothesis),
- $u^T u \simeq p \times \frac{h^2}{p}$ and $(\mathbf{Z}u)^T(\mathbf{Z}u) \simeq nh^2$,

the expected prediction error can be stated more simply, according to the $\frac{n}{p}$ ratio (see section 8 in **Supplementary Material** for details):

$$\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] \simeq \begin{cases} 1 - \frac{n}{p}(h^2)^2, \text{if } p \geq n \\ (1-h^2)\frac{1+\frac{n}{p}h^2}{1+h^2(\frac{n}{p}-1)}, \text{otherwise.} \end{cases} \tag{11}$$

When considering the theoretical quadratic error with respect to the log ratio of the number of individuals over the number

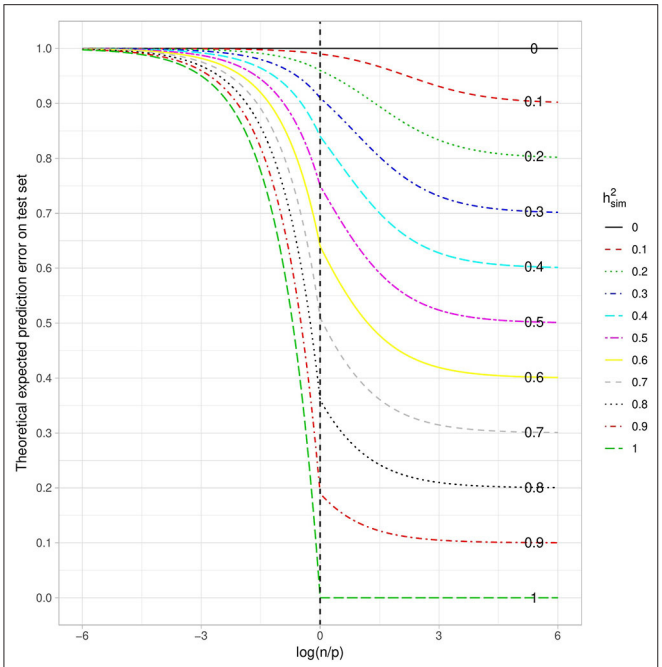

FIGURE 2 | Theoretical Quadratic Error with respect to the log ratio of the number of individuals over the number of variants in the training set. Each curve corresponds to a given heritability (in the narrow sense). Note that the total variance is assumed to be 1.

of variants in the training set (**Figure 2**), as expected we have a decreasing function. This means that the larger the number of individuals in the training sample, the smaller the error. We also observe that the higher the heritability, the smaller the error. Both of these things are intuitive, and as a consequence the error tends toward the irreducible error when n becomes much larger than p. What is more surprising is that the prediction error is close to the maximum, whatever the heritability, when n is much smaller than p. Paradoxically, even with the highest possible heritability, if the number of variants is too large in relation to the number of individuals, no prediction is possible. This can be explained by the fact that the penalization plays a very important role in that case and thus strongly increases the bias, while reducing the variance. The squared bias and the variance with respect to the log ratio of the number of individuals over the number of variants in the training set are shown in **Supplementary Figures 3**, **4**. The irreducible error is only a function of heritability and is not affected by the dimension of the training set.

Similarly, the prediction error can be computed on the training set instead of on the test set. Using the same assumptions as before, the expected prediction error on the training set can be approximated by:

$$\mathbb{E}_{\mathbf{y}_{tr}}\left[\frac{1}{n}(\mathbf{y}_{tr}-\hat{\mathbf{y}}_{tr})^T(\mathbf{y}_{tr}-\hat{\mathbf{y}}_{tr})\right] \simeq \begin{cases} (1-h^2)^2 \text{ if } p > n, \\ 1-2\frac{n}{n+\lambda}\left(\frac{p}{n}(1-h^2)+h^2\right)+\left(\frac{n}{n+\lambda}\right)^2\left(\frac{p}{n}(1-h^2)+h^2\right) \text{ otherwise.} \end{cases}$$

A graph similar to **Figure 2** for this expected error can be found in **Supplementary Figure 5**. Interestingly, when $p > n$, the error on the training set does not depend on the n/p ratio. When n becomes greater than p, it increases and tends toward the irreducible error $1 - h^2$ when $n \gg p$. As shown in **Figure 2**, the error on the test set is always higher than the irreducible error and thus higher than the error on the training set, which is a sign of over-fitting. However, the difference between the error on the test set and the error on the training set is a decreasing function of the n/p ratio, which is linear when $p > n$ and tends toward zero when $n \gg p$.

Another popular way of looking at the predictive accuracy is to consider the squared correlation between y_{te} and $\hat{y}_{te}$ (Goddard, 2009; Daetwyler et al., 2010):

$$\text{corr}^2(y_{te}, \hat{y}_{te}) = \frac{\text{cov}^2(y_{te}, \hat{y}_{te})}{\text{var}\left[y_{te}\right]\text{var}\left[\hat{y}_{te}\right]}.$$

Although correlation and prediction error both provide information about the prediction accuracy, correlation may have an interpretation that is intuitive, but it does not take the scale of the prediction into account. From a predictive point of view, this is clearly a disadvantage. Considering y_{te}, z_{te}, and y_{tr} to be random, and using the same assumptions that were made in relation to prediction error, the three terms of the squared correlation become:

$$\begin{aligned} \text{cov}^2(y_{te}, \hat{y}_{te}) &= (u^T\mathbf{K}_\lambda\mathbf{Z}_{tr}u)^2, \\ \text{var}\left[\hat{y}_{te}\right] &= \text{tr}(\mathbf{K}_\lambda^T\mathbf{K}_\lambda \times \sigma^2\mathbf{I}_n) + (\mathbf{Z}_{tr}u)^T\mathbf{K}_\lambda^T\mathbf{K}_\lambda(\mathbf{Z}_{tr}u), \\ \text{var}\left[y_{te}\right] &= 1. \end{aligned}$$

Like in the case of prediction error, replacing $\mathbf{ZZ}^T$ or $\mathbf{Z}^T\mathbf{Z}$ by their expectations, the squared correlation simplifies to:

$$\text{corr}^2(y_{te}, \hat{y}_{te}) \simeq \begin{cases} \frac{n}{p}(h^2)^2 \text{ if } n < p, \\ \frac{(h^2)^2}{\frac{p}{n}(1-h^2)+h^2} \text{ otherwise.} \end{cases} \quad (12)$$

When considering this theoretical squared correlation with respect to the log ratio of the number of individuals over the number of variants in the training set (**Figure 3**), we have, as expected, an increasing function. Similarly, the higher the heritability, the higher the squared correlation. We also observe that when $n \gg p$, the squared correlation tends toward the simulated heritability. Conversely, when $p \gg n$, it is close to zero whatever the heritability.

2.3. Simulations and Real Data

Since narrow-sense heritability is a quantity that relates to a model, we will first illustrate our contributions via simulations where the true model is known. We perform two different types of simulation: fully synthetic simulations where both genotypes and phenotypes are drawn from statistical distributions, and semi-synthetic simulations where UK Biobank genotypes are used to simulate phenotypes. We also illustrate our contributions using height and body mass index (BMI) from the UK Biobank dataset.

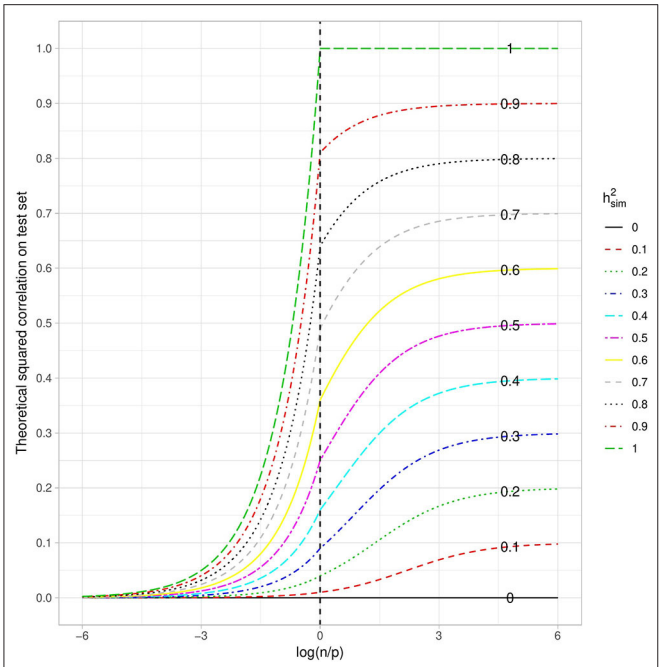

FIGURE 3 | Theoretical squared correlation between phenotype and its prediction with respect to the log ratio of the number of individuals over the number of variants in the training set. Each curve corresponds to a given heritability (in the narrow sense).

We first assess the performance of GCV for heritability estimation and then look at the accuracy of the prediction when the ratio of the number of individuals to the number of variants varies in the training set.

2.3.1. UK Biobank Dataset

The present analyses were conducted under UK Biobank data application number 45,408. The UK Biobank dataset consists of $\simeq$ 784 K autosomal SNPs describing $\simeq$ 488 K individuals. We applied relatively stringent quality control and minor allele frequency filters to the dataset (callrate for individuals and variants $>$ 0.99; p-values of Hardy-Weinberg equilibrium test $>$ 1e-7; Minor Allele Frequency $>$ 0.01), leading to 473,054 and 417,106 remaining individuals and SNPs, respectively.

Two phenotypes were considered in our analyses: height (standing) and BMI. In order to have a homogeneous population for the analysis of these real phenotypes, we retained only those individuals who had reported their ethnicity as white British and whose Principal Component Analysis (PCA) results obtained by UK Biobank were consistent with their self-declared ethnicity. In addition, each time we subsampled individuals we removed related individuals [one individual in all pairs with a Genetic Relatedness Matrix (GRM) coefficient $>$0.025 was removed], as in Yang et al. (2011b) in order to avoid confusion between shared genetic factors and shared environmental factors. Several covariates were also considered in the analysis of these phenotypes: the sex, the year of birth, the recruitment center, the genotyping array, and the first 10 principal components computed by UK Biobank.

TABLE 1 | Table of the parameters sets of the simulations.

Parameters	Levels
n/p	Simulation: 1,000/10,000; 5,000/10,0000; 10,000/500,000
	Data-based: 1,000/10,000; 5,000/10,0000; 10,000/417,106
f_c	0.1; 0.5; 1
h^2_{sim}	{0.1, ..., 0.9}

n/p: the ratio of the dimensions of the genotype matrix. f_c: proportion of causal variants. h^2_{sim}: simulated heritability.

2.3.2. Synthetic Genotype Data

The synthetic genotype matrices are simulated as in Golan et al. (2014) and de Vlaming and Groenen (2015). This corresponds to a scenario with independent loci or perfect linkage equilibrium. To simulate synthetic genotypes for p variants, we first set a vector of variant frequencies $f \in \mathbb{R}^p$, with these frequencies independently following a uniform distribution $\mathcal{U}([0.05, 0.5])$. Individual genotypes are then drawn from binomial distributions with proportions f, to form the genotype matrix $\mathbf{M}$. A matrix of standardized genotypes $\mathbf{Z}^*$ can be obtained by standardizing $\mathbf{M}$ with the true variant frequencies f.

2.3.3. Simulations to Assess Heritability Estimation Using GCV

We consider both synthetic and real genetic data, and simulate associated phenotypes.

In the two simulation scenarios we investigate the influence on heritability estimation of the following three parameters: the shape of the genotype matrix in the training set (the ratio between n the number of individuals and p the number of variants), the fraction of variants with causal effects f_c, and the true heritability h^2_{sim}. The tested levels of these quantities are shown in **Table 1**.

For each simulation scenario and for a given a set of parameters (n, p, f_c, h^2_{sim}), the simulation of the phenotype starts with a matrix of standardized genotypes (either a synthetic genotype matrix $\mathbf{Z}^*$ standardized with the true allele frequencies, as described in section 2.3.2, or a matrix of empirically standardized genotypes $\mathbf{Z}$ obtained from UK Biobank data). To create the vector of genotype effects u, $p \times f_c$ causal SNPs are randomly sampled and their effects are sampled from a multivariate normal distribution with zero mean and a covariance matrix $\tau \mathbf{I}_{p \times f_c}$ (where $\tau = \frac{h^2_{sim}}{p \times f_c}$), while the remaining $p \times (1 - f_c)$ effects are set to 0. The vector of environmental effects $\mathbf{e}$ is sampled from a multivariate normal distribution with zero mean and a covariance matrix $\sigma^2 \mathbf{I}_n$, where $\sigma^2 = 1 - h^2_{sim}$. The phenotypes are then generated as $\mathbf{y} = \mathbf{Z}^* u + \mathbf{e}$ and $\mathbf{y} = \mathbf{Z} u + \mathbf{e}$, for the fully synthetic scenario and the semi-synthetic scenario, respectively. A standardization set of 1,000 individuals (that will be used for the GCV approach based on two datasets) is also generated for each scenario in the same way.

Applying GCV to large-scale matrices can be extremely time-consuming, since it requires the computation of the GRM associated with $\mathbf{Z}^*$ or $\mathbf{Z}$ and the eigen decomposition of

the GRM. For this reason we employed the same strategy as de Vlaming and Groenen (2015) in order to speed up both simulations and analyses by making it possible to test more than one combination of simulation parameters. We simulated an ($n_{max} = 10,000 \times p_{max} = 500,000$) genotype matrix for the training set in the fully synthetic scenario and used this simulated matrix for all the $9 \times 3 \times 3 = 81$ ($h^2_{sim} \times f_c \times n/p$) parameter combinations. Similarly, we sampled $n_{max} = 10,000$ individuals from the UK Biobank dataset to obtain an ($n_{max} = 10,000 \times p_{max} = 417,106$) genotype matrix for the training set in the semi-synthetic scenario. Smaller matrices were then created from a subset of these two large matrices (note that for subsets of the real genotype matrix we took variants in the original order to keep the linkage disequilibrium structure). Consequently, computation of the GRM and its eigen decomposition needed to be performed only once for each n/p ratio considered. The fully synthetic and the semi-synthetic scenarios were each replicated 30 times.

2.3.4. Simulations to Assess Prediction Accuracy

We performed fully synthetic simulations for different ratios $\frac{n}{p}$ in order to study the behavior of the mean prediction error and the correlation between the phenotype and its prediction. We considered a training set of size $n = 1,000$, and a test set of size $n_{te} = 5,000$. The maximum number of variants was set to $p_{max} = 50,000$ and the heritability to $h^2 = 0.6$. We first simulated a global allelic frequency vector $f \sim \mathcal{U}_{p_{max}}(0.05, 0.5)$ and a global vector of genetic effects $u \sim \mathcal{N}\left(p_{max}, \frac{h^2}{p_{max}} \mathbf{I}_{p_{max}}\right)$.

For each subset of variants of size $p < p_{max}$, we selected a vector of genetic effects composed of the p first components of u multiplied by a $\sqrt{\frac{p_{max}}{p}}$ factor assuring a total variance of 1 and $\text{var}(u^p) = \frac{h^2}{p}\mathbf{I}_p$: $u^p = (u_1, ..., u_p) \times \sqrt{\frac{p_{max}}{p}}$. The genotype matrix $\mathbf{M}_{te}$ was then simulated and its normalized version $\mathbf{Z}^*_{te}$ computed as described in section 2.3.2. The normalization used the first p components of f. The noise vector $\mathbf{e}_{te} \sim \mathcal{N}(\mathbf{0}_{n_{te}}, (1 - h^2)\mathbf{I}_{n_{te}})$ and a vector of phenotypes $\mathbf{y}_{te} = \mathbf{Z}^*_{te}u^p + \mathbf{e}_{te}$ were eventually simulated.

We generated 300 training sets by simulating the normalized genotype matrix, noise, and phenotype using the same process as for the test set. Here, the training set index is denoted as k. A prediction $\hat{\mathbf{y}}_{te,k}$ for the test set was made with each training set using the ridge estimator of $\boldsymbol{u}^p$ obtained with $\lambda = p\frac{1-h^2}{h^2}$, and the following empirical quantities were estimated: $\text{err}_p = \frac{1}{300}\sum_k \frac{1}{n_{te}} \left\| \mathbf{y}_{te,k} - \hat{\mathbf{g}}_p \right\|^2_2$, $\text{bias}^2_p = \frac{1}{n_{te}} \sum_{i \in [\![1, n_{te}]\!]} \left(\left[\mathbf{Z}_{te}u^p - \hat{\mathbf{g}}_p \right]_i \right)^2$ and $\text{var}_p = \frac{1}{300}\sum_k \frac{1}{n_{te}} \left\| \hat{\mathbf{y}}_{te,k} - \hat{\mathbf{g}}_p \right\|^2_2$, where $\hat{\mathbf{g}}_p = \left(\frac{1}{300} \sum_{k \in [\![1,300]\!]} \left[\hat{\mathbf{y}}_{te,k} \right]_i \right)_{i \in [\![1, n_{te}]\!]}$. The squared correlation between $\hat{\mathbf{y}}_{te,k}$ and $\mathbf{y}_{te,k}$ was also estimated.

We considered the following numbers of variants:

$$p \in \{50,000;\ 25,000;\ 16,667;\ 12,500;\ 10,000;\ 5,000;\ 3,333;\ 2,500;\ 2,000;\ 1,667;\ 1,429;\ 1,250;\ 1,111;\ 1,000;\ 500;\ 136;\ 79;\ 56;\ 43;\ 35;\ 29;\ 25;\ 22;\ 20\}.$$

TABLE 2 | Size (number of individuals) of training, standardization, and test sets for assessing predictive power on real data.

Set	Size
Training	{1,000; 2,000; 5,000; 10,000; 20,000}
Standardization	1,000
Test	1,000

TABLE 3 | Number of repetitions for the evaluation of the predictive power on real data.

Size of the training set	1,000	2,000	5,000	10,000	20,000
Number of repetitions	100	70	50	20	10

2.4. Prediction of Height and BMI Using UK Biobank Data

To experiment on UK Biobank for assessing the prediction accuracy, for each phenotype we considered three sets of data: a training set for the purpose of learning genetic effects, a standardization set for learning non-penalized effects (covariates and intercept), and a test set for assessing predictive power. Pre-treatment filters (as described in section 2.3.1) were systematically applied on the training set. We computed the estimation of genetic effects using the projection-based approach to take into account non-penalized effects, where the penalty parameter was obtained by GCV with the same projection approach:

$$\hat{u}_R = \mathbf{Z}^T_{tr}\mathbf{C}^T_{tr}\left(\mathbf{C}_{tr}\mathbf{Z}_{tr}\mathbf{Z}^T_{tr}\mathbf{C}^T_{tr} + \hat{\lambda}_{GCV}\mathbf{I}_{n-r}\right)^{-1}\mathbf{C}_{tr}\mathbf{y}_{tr}.$$

We then estimated non-penalized effects (here $\mathbf{X}$ contains the intercept):

$$\hat{\beta} = \left(\mathbf{X}^T_{std}\mathbf{X}_{std}\right)^{-1}\mathbf{X}^T_{std}\left(\mathbf{y}_{std}\right). \tag{13}$$

Finally, we applied these estimations on the test set:

$$\hat{\mathbf{g}}_{te} = \mathbf{Z}_{te}\hat{u}_R,$$
$$\hat{\mathbf{f}}_{te} = \mathbf{X}_{te}\hat{\beta},$$
$$\tilde{\mathbf{y}}_{te} = \mathbf{y}_{te} - \hat{\mathbf{f}}_{te},$$

in order to compute the Mean Square Error = $\frac{1}{n_{te}}(\tilde{\mathbf{y}}_{te} - \hat{\mathbf{g}}_{te})^T(\tilde{\mathbf{y}}_{te} - \hat{\mathbf{g}}_{te})$ between the phenotype residuals $\tilde{\mathbf{y}}_{te}$ after removal of non-penalized effects and $\hat{\mathbf{g}}_{te}$.

This procedure was performed for different ratios $\frac{n}{p}$ using different sized subsets of individuals for the training set, while keeping all the variants that passed pre-treatment filters (see **Table 2**).

For each number n of individuals considered in the training set, the sampling of these individuals was repeated several times, as seen in **Table 3**, in order to account for the variance of the estimated genetic effects due to sampling.

3. RESULTS

3.1. Generalized Cross Validation for Heritability Estimation

3.1.1. Simulation Results

For the two simulation scenarios we look at the difference between the estimation of h_g^2 by GCV and the simulated heritability h_{sim}^2 in different configurations of study size n/p, h_{sim}^2 and the fraction of causal variants f_c. Similarly, we look at the difference between the estimation by the classical mixed model approach and the simulated heritability. In our simulations f_c was seen to have no influence, and so only the influence of the remaining parameters is shown in **Figure 4** and f_c is fixed at 10%. For full results, see **Supplementary Figures 1**, **2**.

For the fully-simulated scenario, the two GCV approaches give very similar results and appear to provide an unbiased estimator of h^2. They compare very well with the estimation of heritability by ridge regression with a 10-fold CV. Moreover, the variance of the GCV estimators does not appear higher than the variance of 10-fold CV.

In the case of the semi-synthetic simulations, here too both GCV approaches and the 10-fold CV provide a satisfactory heritability estimation. Our choice of using GCV in place of a classical CV approach for estimating heritability by ridge regression is therefore validated.

For both simulation scenarios we also note that the classical mixed model approach (using the AI-REML method in the `gaston` R package) gives heritability estimations that are very similar to those obtained using the GCV approaches. The value of simulated heritability does not appear to have a strong effect on the quality of the heritability estimation. On the other hand, the ratio n/p seems to have a real impact on estimation variance, with lower ratios leading to lower variances, which initially might appear surprising. One possible explanation for this is that in our simulations n increases as the ratio n/p decreases. Visscher and Goddard (2015) showed that the variance of the heritability is a decreasing function of n, which could explain the observed behavior.

3.1.2. Illustration on UK Biobank

We now compare heritability estimations between the two GCV approaches and the classical mixed model approach for height and BMI, on a training set of 10,000 randomly sampled individuals (the training set being of the same size as for the simulated data). All three approaches take account of covariates and the intercept. The AI-REML approach also uses a projection matrix to deal with covariates. For the GCV approach based on two datasets, a standardization set of 1,000 individuals is also sampled, and for comparison purposes we have chosen to apply this two-set strategy to the classical mixed model approach as well.

Since the true heritability is of course unknown with real data, the sampling of the training and standardization sets is repeated 10 times in order to account for heritability estimation variability. Note that the SNP quality control and MAF filters were repeated at each training set sampling and applied to the standardization set.

Figure 5 shows that for each phenotype the two GCV approaches and the classical mixed model approach (AI-REML) give similar estimations. There is relatively little estimation variability, and any variability observed seems depend more on the individuals sampled for the training set than on the approach used.

3.2. Prediction vs. Heritability in the Context of Small Additive Effects

3.2.1. Prediction From Synthetic Data

As expected, the mean of the test set error follows closely the theoretical curve when the $\log \frac{n}{p}$ varies (**Figure 6**). When $n > p$, the mean of the test set is close to the minimum possible error, which means that the ridge regression provides a reliable prediction on average.

Interestingly, if the mean error behaves as expected by our approximation, the standard deviation of the error may be very large. **Figures 6A,B** show the same mean error with different error bars. **Figure 6A** plots the error bars corresponding to the training set variation: the mean test set error is computed for each training set and the error bars show one standard deviation across the 300 training sets. **Figure 6B** plots the error bars corresponding to the variation of the errors across the test set.

The error bars in **Figure 6B** are much larger than those in **Figure 6A**, which shows that the variation in the prediction error is mostly due to the test individual whose phenotype we wish to predict, and depends little on the training set. This may be explained by the fact that the environmental residual term can be very large for some individuals. For these individuals the phenotype will be predicted with a very large error even when $n \gg p$, that is to say when the genetic term is correctly estimated, irrespective of the training set.

The squared correlation between the phenotype and its prediction, as a function of $\log \frac{n}{p}$, is also in line with our approximation (**Figure 7**). As expected, when $n \gg p$, the squared correlation tends toward the simulated heritability. We compared our approximation with the approximation obtained by Daetwyler et al. (2008) and observed that although Daetwyler's approximation is very similar to ours when $p \gg n$, our simulation results make Daetwyler's approximation appear under-optimistic when $n \gg p$. Finally, we also compared our approximation with that obtained by Rabier et al. (2016), which is the same as ours when $n > p$. However, when $p > n$, Rabier's approximation appears over-optimistic.

3.2.2. Prediction From UK Biobank Data

Let us consider the proposed theoretical approximation of the predictive power of ridge regression with respect to the n/p ratio applied to the UK Biobank data, for height and BMI residuals (after removal of covariate effects and intercept).

The two phenotypes differ considerably as regards heritability: we estimate by the projection-based GCV approach that 73.33% of height is "heritable" whereas only 33.91% of BMI is (on average over the 10 training samples of 20,000 individuals).

These estimated values are close to those currently found in the literature (Ge et al., 2017). It is important to note that

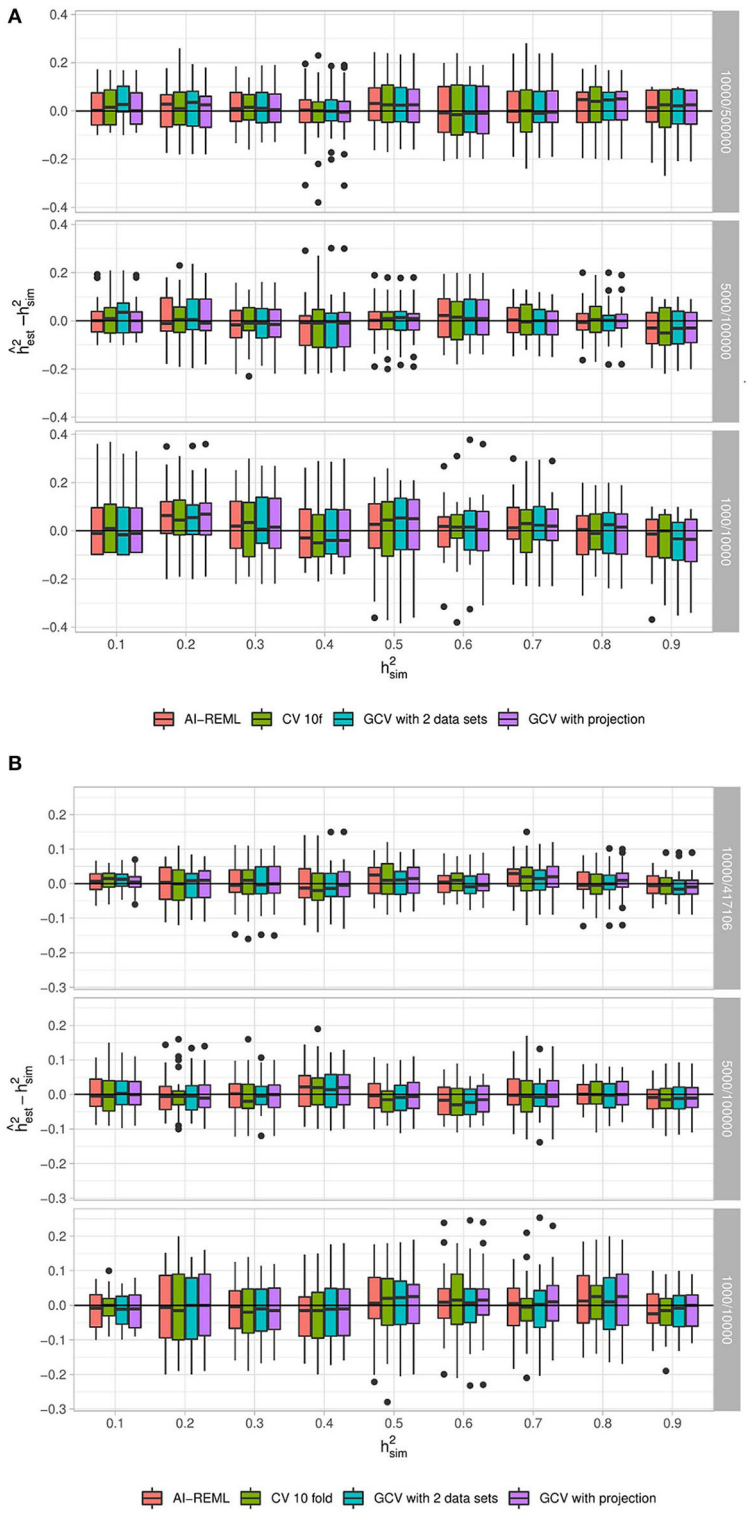

FIGURE 4 | Distribution of $(h^2_{est} - h^2_{sim})$ for different parameter combinations with 30 replications. **(A)** Corresponds to data simulated under the "fully synthetic" procedure, while **(B)** corresponds to the "semi-synthetic simulation" procedure. Each sub-panel corresponds to a different value of n/p. In both scenarios 10% of the variants have causal effects (i.e., $f_c = 0.1$). For each panel, the horizontal axis corresponds to the simulated heritability $h^2_{sim} \in \{0.1, ..., 0.9\}$ and the vertical axis corresponds to $(h^2_{est} - h^2_{sim})$. Heritability estimations are done with the random effects model using AI-REML and with ridge regression using three approaches for the choice of λ: GCV with a projection correction, GCV with a two-dataset correction and a 10-fold cross-validation (CV 10f).

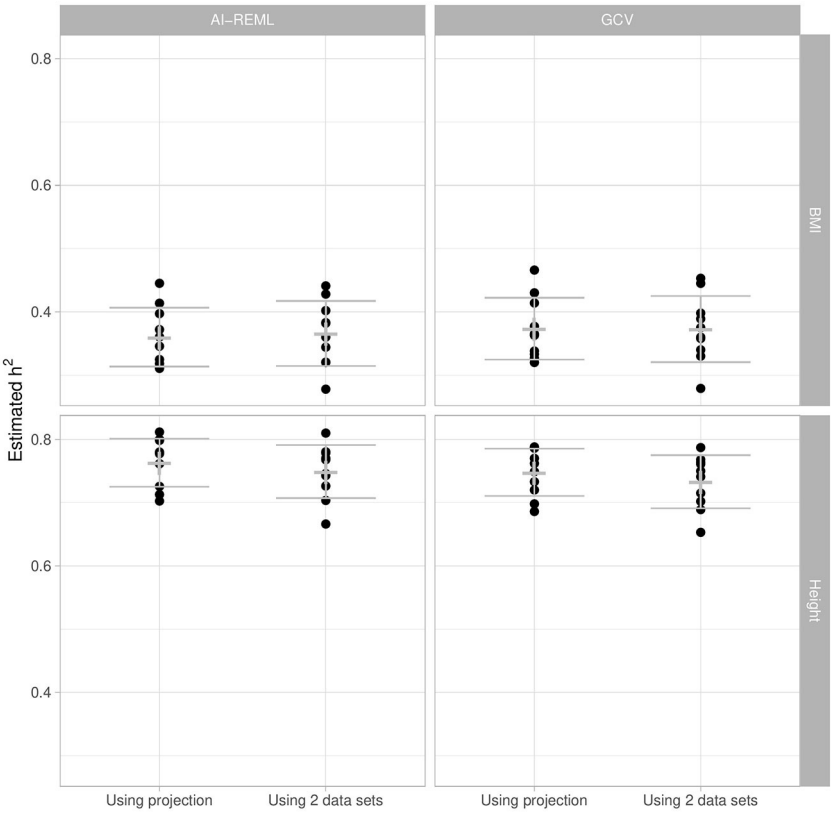

FIGURE 5 | Heritability estimation of BMI and height using AI-REML and GCV, with the projection-based approach and with the two-set approach. We sub-sampled the original UK Biobank dataset 10 times for replication. The cross corresponds to the mean and the error bar to the mean ± one standard deviation.

the heritability estimation is strongly dependent on the filters. Variations of up to 20% were observed in the estimations when the filtering procedure setup was slightly modified.

A major difference between UK Biobank data and our simulations designed to check the proposed approximation lies in the strong linkage disequilibrium present in the human genome. Several papers have proposed using the effective number of independent markers to make adjustments in the multiple testing framework (Li et al., 2012), and we likewise propose adjusting our prediction model by taking into account an effective number of SNPs (p_e). We estimate the effective $\frac{n}{p_e}$ ratio for each training set and for each considered n value using the observed mean square errors, the estimated heritability, and the theoretical relation in the case of independent variants $\mathbb{E}_{\mathbf{y}_{tr},y_{te},z_{te}}\left[(y_{te}-\hat{y}_{te})^2\right] = 1-\frac{n}{p}(h^2)^2$ when $p > n$. We then use a simple linear regression to find the coefficient between these estimated $\frac{n}{p_e}$ ratios and the corresponding real $\frac{n}{p}$ ratios.

Table 4 shows different but close effective numbers of SNPs for the two phenotypes.

We also consider normalizing the test set errors using the mean square error of phenotype residuals (after removing non-penalized effects). Using this error normalization and adjusting the theoretical curve for an effective number of SNPs, we observe a close fit between the estimated errors on the test set and their theoretical values (**Figure 8**).

4. DISCUSSION

In this work we investigated an alternative computation of genomic heritability based on ridge regression. We proposed a fast, reliable way to estimate the optimal penalization parameter of the ridge via Generalized Cross Validation adapted for high dimension. The genomic heritability estimated from the GCV gives results comparable to mixed model AIREML estimates. It clearly demonstrates that a predictive criterion allows a reliable choice of the penalization parameter and associated heritability, even when the prediction accuracy of ridge regression is low. Moreover, even though our approach does not formally consider Linkage Disequilibrium, simulations showed that it still provides reliable genomic heritability estimates in presence of realistic Linkage Disequilibrium.

We also provided theoretical approximations of the ridge regression prediction accuracy, in terms of both error and correlation between the phenotype and its prediction on new samples. These approximations perform well on synthetic data, in both high and low dimensions. They rely on the assumption that individuals and markers are independent in approximating

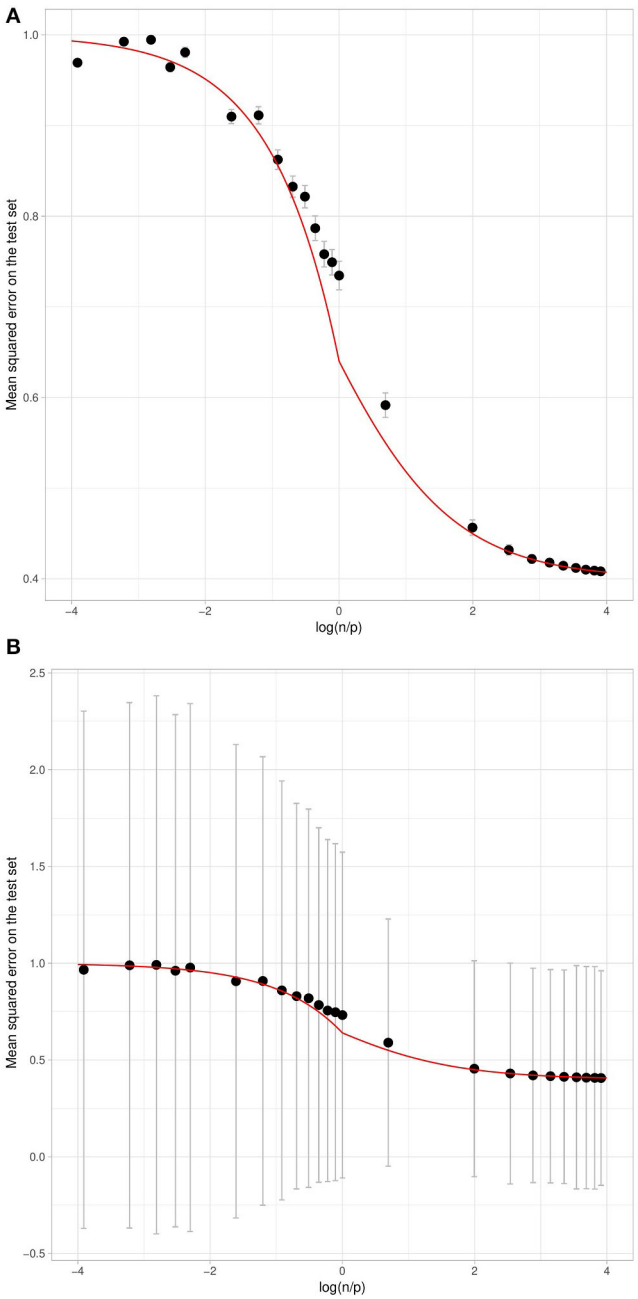

FIGURE 6 | Mean Squared Error of the prediction on the test set with respect to $\log(\frac{n}{p})$, using simulated data with $h^2 = 0.6$. The curves correspond to the theoretical link for $h^2 = 0.6$. The black points correspond to the mean expectation for each value of $\log(\frac{n}{p})$ over 300 repetitions. The error bars in **(A,B)** correspond respectively to one standard deviation of the test set error obtained using two different aggregation strategies. On the left **(A)**, we consider an aggregation strategy where each of the 300 training sets results in a mean test set error, whereas on the right **(B)** each individual in the test set results in an error averaged over all training sets.

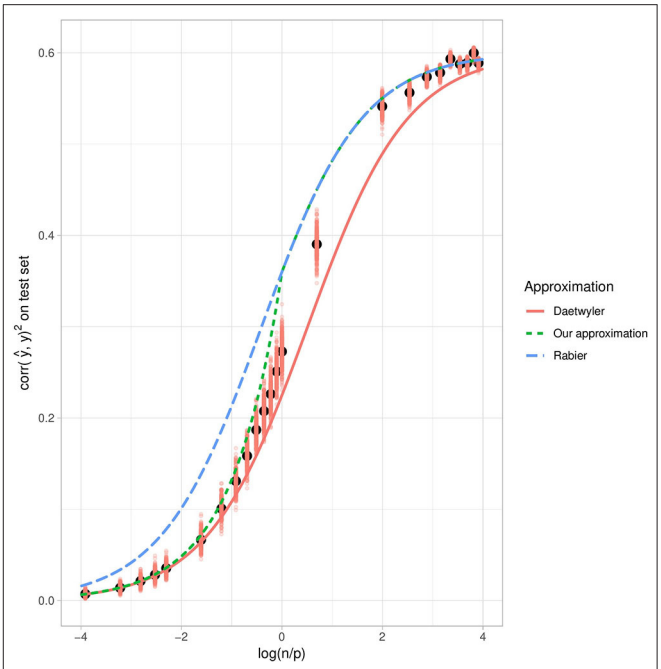

FIGURE 7 | Mean squared correlation between the phenotype and its prediction on the test set with respect to $\log(\frac{n}{p})$, using simulated data with $h^2 = 0.6$. The salmon points correspond to the evaluation of the squared correlation and the black points correspond to the mean expectation for each value of $\log(\frac{n}{p})$ over 300 repetitions. Red dots correspond to training set replications. The red plain curve corresponds to Daetwyler's approximation for $h^2 = 0.6$, while the blue dashed curve corresponds to Rabier's approximation and the green dotted curve corresponds to ours.

TABLE 4 | Effective number of SNPs.

Phenotype	p/p_e
Height	5.01
BMI	3.48

the empirical covariance matrices. Our approximation of the prediction accuracy in terms of correlation proposes a good compromise between existing approximations. In particular, it exhibits similar performances to Daetwyler et al. (2008) when $p > n$ and to Rabier et al. (2016) when $p < n$.

Our theoretical approximation of the prediction error is also consistent with the error observed on real genetic data when $p > n$, after adjusting for the effective number of independent markers. Unfortunately, due to computational issues, we were unable to perform the analysis in the $n \simeq p$ case with real data. However, we observed that the prediction accuracy already reaches almost 15% of the heritability of height when $n/p \simeq 5\%$, while De los Campos et al. (2013) suggested that its asymptotic upper bound is of the order of 20% of the heritability because of incomplete LD between causal loci and genotyped markers. Interestingly, ridge regression is not affected by correlated predictors, and consequently it is not affected by high LD between markers. When LD is high, this has the effect of reducing the degrees of freedom of the model (Dijkstra, 2014), which results in an improved prediction accuracy in comparison with a problem having the same number of independent predictors and the same heritability.

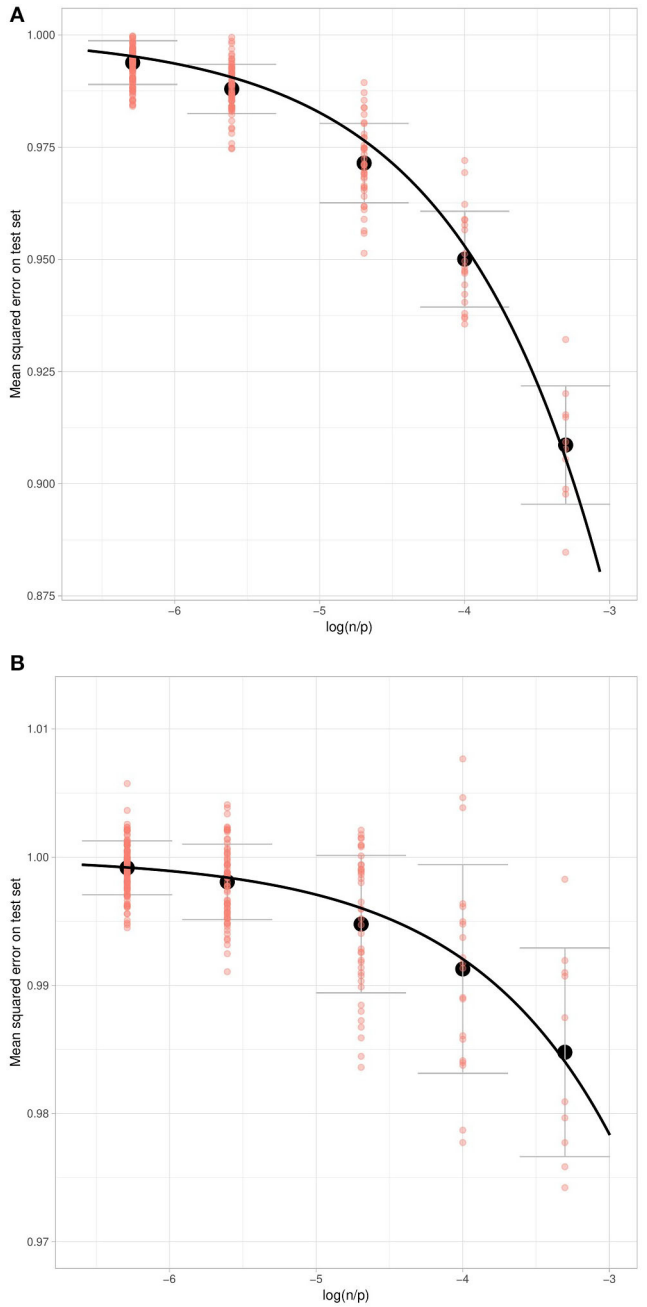

FIGURE 8 | Normalized Mean Squared Error on the test set for the prediction of Height **(A)** and BMI **(B)** with respect to the log ratio of the number of individuals over the number of markers in the training set. Salmon dots correspond to training set replications, black dot to the mean of replications for different ratio and we show one standard deviation (over the training sets) of the mean test set error. The theoretical curves are fitted using the estimated heritability and an effective number of markers.

Although our approximations and simulation results tend to show that the prediction accuracy can reach the heritability value when $n \gg p$, as already suggested by previous works (Daetwyler et al., 2008; de Vlaming and Groenen, 2015; Rabier et al., 2016), the large standard deviation of the prediction error that we observed between simulated individuals suggests that disease risk prediction from genetic data alone is not accurate at the individual level, even for a relatively high heritability value in the context of a small additive effect hypothesis.

In direct continuity of this work, it would be interesting to investigate the behavior of prediction accuracy on real human data where $n \simeq p$. This would enable us to determine whether our approximations still hold in that case, and even in the case where $n > p$ (where we approximate the empirical covariance matrix of the markers to be diagonal). It would show whether it is possible for the prediction accuracy to exceed the upper bound proposed by De los Campos et al. (2013). A further prospect would be to consider a nonlinear model extension via kernel ridge regression, which may improve the prediction (Morota and Gianola, 2014).

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: There is a charge for access to the UK Biobank dataset. Requests to access these datasets should be directed to: https://www.ukbiobank.ac.uk/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by NHS National Research Ethics Service North West (11/NW/0382). The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

EL and CA designed and directed the study. AF, EL, and CA wrote the manuscript. AF created the synthetic datasets. AF performed all the analyses with substantial input from EL, CA, CD-R, and MP-J. All authors discussed the results and commented on the manuscript.

FUNDING

Financial support was obtained by the Laboratory of Excellence GENMED (Medical Genomics) grant no. ANR-10-LABX-0013 managed by the French National Research Agency (ANR) part of the Investment for the Future program.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource under application number 45408.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.581594/full#supplementary-material

REFERENCES

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information science and statistics. New York, NY: Springer.

Brard, S., and Ricard, A. (2015). Is the use of formulae a reliable way to predict the accuracy of genomic selection? *J. Anim. Breed. Genet*. 132, 207–217. doi: 10.1111/jbg.12123

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

Chen, G.-B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front. Genet*. 5:107. doi: 10.3389/fgene.2014.00107

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., and Woolliams, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185, 1021–1031. doi: 10.1534/genetics.110.116855

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

Dandine-Roulland, C., and Perdry, H. (2015). The use of the linear mixed model in human genetics. *Hum. Hered*. 80, 196–206. doi: 10.1159/000447634

De los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet*. 9:e1003608. doi: 10.1371/journal.pgen.1003608

de Vlaming, R., and Groenen, P. J. F. (2015). The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Res. Int*. 2015:143712. doi: 10.1155/2015/143712

Dijkstra, T. K. (2014). Ridge regression and its degrees of freedom. *Qual. Quant*. 48, 3185–3193. doi: 10.1007/s11135-013-9949-7

Elsen, J.-M. (2017). An analytical framework to derive the expected precision of genomic selection. *Genet. Sel. Evol*. 49:95. doi: 10.1186/s12711-017-0366-6

Feldman, M. W., and Lewontin, R. C. (1975). The heritability hang-up. *Science* 190, 1163–1168. doi: 10.1126/science.1198102

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* 52, 399–433. doi: 10.1017/S0080456800012163

Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R., and Smoller, J. W. (2017). Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet*. 13:e1006711. doi: 10.1371/journal.pgen.1006711

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–257. doi: 10.1007/s10709-008-9308-0

Golan, D., Lander, E. S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. U.S.A*. 111, E5272–E5281. doi: 10.1073/pnas.1419064111

Golub, G. H., Heath, M., and Wahba, G. (1978). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21, 215–233. doi: 10.1080/00401706.1979.10489751

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi: 10.2307/2529430

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet*. 6, 95–108. doi: 10.1038/nrg1521

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Li, M.-X., Yeung, J. M., Cherny, S. S., and Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet*. 131, 747–756. doi: 10.1007/s00439-011-1118-2

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

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet*. 5:363. doi: 10.3389/fgene.2014.00363

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554. doi: 10.1093/biomet/58.3.545

Perdry, H., and Dandine-Roulland, C. (2018). *gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models. R package version 1.5.4*, Villejuif.

Pharoah, P. D., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F., and Ponder, B. A. (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet*. 31, 33–36. doi: 10.1038/ng853

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

Rabier, C.-E., Barre, P., Asp, T., Charmet, G., and Mangin, B. (2016). On the accuracy of genomic selection. *PLoS ONE* 11:e0156086. doi: 10.1371/journal.pone.0156086

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Stat. Sci*. 6, 15–32. doi: 10.1214/ss/1177011926

Speed, D., and Balding, D. J. (2014). Multiblup: improved SNP-based prediction for complex traits. *Genome Res*. 24, 1550–1557. doi: 10.1101/gr.169375.113

Visscher, P. M., and Goddard, M. E. (2015). A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. *Genetics* 199, 223–232. doi: 10.1534/genetics.114.171017

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proc. Natl. Acad. Sci. U. S. A*. 6:320. doi: 10.1073/pnas.6.6.320

Wright, S. (1921). Correlation and causation. *J. Agric. Res*. 7, 557–585.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* 163, 789–801.

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

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011a). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*. 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., et al. (2011b). Genome partitioning of genetic variation for complex traits using common SNPS. *Nat. Genet*. 43:519. doi: 10.1038/ng.823

Zhao, B., and Zhu, H. (2019). Cross-trait prediction accuracy of high-dimensional ridge-type estimators in genome-wide association studies. *arXiv:1911.10142 [stat]*.

Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann. Appl. Stat*. 11:2027. doi: 10.1214/17-AOAS1052

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 November 2020
doi: 10.3389/fgene.2020.589047

Modification of Experimental Design and Statistical Method for Mapping Imprinted QTLs Based on Immortalized F_2 Population

Kehui Zheng[1,2†], Jiqiang Yan[2†], Jiacong Deng[3], Weiren Wu[4,5†] and Yongxian Wen[2*†]*

[1] College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [2] College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, [3] School of Ocean and Biochemical Engineering, Fuqing Branch of Fujian Normal University, Fuzhou, China, [4] Fujian Provincial Key Laboratory of Crop Breeding by Design, Fujian Agriculture and Forestry University, Fuzhou, China, [5] Key Laboratory of Genetics, Breeding and Multiple Utilization of Crops, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou, China

Genomic imprinting is an epigenetic phenomenon, which plays important roles in the growth and development of animals and plants. Immortalized F_2 (imF_2) populations generated by random cross between recombinant inbred (RI) or doubled haploid (DH) lines have been proved to have significant advantages for mapping imprinted quantitative trait loci (iQTLs), and statistical methods for this purpose have been proposed. In this paper, we propose a special type of imF_2 population (R-imF_2) for iQTL mapping, which is developed by random reciprocal cross between RI/DH lines. We also propose two modified iQTL mapping methods: two-step point mapping (PM-2) and two-step composite point mapping (CPM-2). Simulation studies indicated that: (i) R-imF_2 cannot improve the results of iQTL mapping, but the experimental design can probably reduce the workload of population construction; (ii) PM-2 can increase the precision of estimating the position and effects of a single iQTL; and (iii) CPM-2 can precisely map not only iQTLs, but also non-imprinted QTLs. The modified experimental design and statistical methods will facilitate and promote the study of iQTL mapping.

Keywords: genomic imprinting, imprinted quantitative trait loci, point mapping, composite point mapping, immortalized F_2 population

OPEN ACCESS

Edited by:
Lide Han,
Vanderbilt University Medical Center, United States

Reviewed by:
Fa Cui,
University of Chinese Academy of Sciences, China
Huihui Li,
Chinese Academy of Agricultural Sciences, China
Haiming Xu,
Zhejiang University, China

****Correspondence:***
Weiren Wu
wuwr@fafu.edu.cn
Yongxian Wen
wenyx9681@fafu.edu.cn

† These authors have contributed equally to this work

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

Received: *30 July 2020*
Accepted: *29 October 2020*
Published: *20 November 2020*

Citation:
Zheng K, Yan J, Deng J, Wu W and Wen Y (2020) Modification of Experimental Design and Statistical Method for Mapping Imprinted QTLs Based on Immortalized F_2 Population. Front. Genet. 11:589047. doi: 10.3389/fgene.2020.589047

INTRODUCTION

Genomic imprinting is a phenomenon found in animal and plant, in which two alleles of a gene show unequal expression depending on their parental origins. Genes involved in such phenomenon are called imprinted genes. Many imprinted genes have been found in animal and human (Nolan et al., 2001; Morison et al., 2005; Long and Cai, 2007; Babak et al., 2008; Hagan et al., 2009; Girardot et al., 2013; Pembrey et al., 2014; Hur et al., 2016; Jiang et al., 2017; Mackay and Temple, 2017). In plant, the first imprinted gene, which is involved in the coloration of maize kernel endosperm, was

discovered as early as about half centuries ago (Kermicle, 1970). Compared with those in animal, however, the imprinted genes identified in plant so far are still very limited, of which most are found from *Arabidopsis*, rice and maize (Danilevskaya et al., 2003; Haun et al., 2007; Luo et al., 2009; Bauer and Fischer, 2011; Raissig et al., 2011; Zhang et al., 2011; Ikeda, 2012; Pei et al., 2019).

It has been found that many quantitative traits are affected by genomic imprinting (Spencer, 2002; Croteau and Croteau, 2004; Sandovici et al., 2005; Santure and Spencer, 2011; Wang et al., 2012). The quantitative trait loci (QTLs) showing imprinting effect are called imprinted QTL (iQTL). A number of different experimental designs and corresponding statistical methods have been proposed for mapping iQTLs (Knott et al., 1998; de Koning et al., 2000; Pratt et al., 2000; Strauch et al., 2000; Hanson et al., 2001; Haghighi and Hodge, 2002; Shete and Amos, 2002; Shete et al., 2003; Knapp and Strauch, 2004; Mantey et al., 2005; Cui et al., 2006, 2007, 2008, Cui, 2007; Liu et al., 2007; Li et al., 2008, 2012a; Hager et al., 2008; Yang et al., 2010; Zhou et al., 2012; Karami et al., 2019). F_2 (outbred or inbred) and BC_1 populations are usually used for iQTL mapping (Haley et al., 1994; de Koning et al., 2000; Cui et al., 2006; Li et al., 2012a), but they have obvious shortcomings, such as relatively low power in iQTL detection, low accuracy in estimating the positions and effects of iQTLs, inability of permanent preservation of the population, and unrepeatability. Besides, determination of the parental origins of alleles is also difficult or problematic under the F_2 and BC_1 designs (Wu et al., 2002; Cui et al., 2006; Wolf et al., 2008; Lawson et al., 2013). In addition, the imprinting effect cannot be separated from the maternal effect in the BC_1 design.

Wen and Wu (2014) proposed statistical methods for iQTL mapping using an immortalized F_2 (abbreviated as imF_2) population generated from random crosses between recombinant inbred (RI) lines or doubled haploid (DH) lines. Compared with the previous designs, the imF_2 design has significant advantages for iQTL mapping. First, the parental origins of marker alleles in each imF_2 line can be exactly known from the cross. Second, analysis based on imF_2 lines can reduce environmental error so as to increase the statistical power of iQTL mapping. Third, a very large imF_2 population can be produced without increasing the cost of molecular marker assay. However, there are also shortcomings in the experimental design and mapping methods proposed by Wen and Wu (2014). In the experimental design, the work of constructing an imF_2 population is laborious. In the statistical methods, iQTL mapping is performed only by testing the imprinting effect without making use of the information of additive effect and dominance effect. This may reduce the precision of iQTL mapping.

In this paper, to overcome the above shortcomings, we propose a modified imF_2 design and modified statistical methods for iQTL mapping based on the work of Wen and Wu (2014). We demonstrate by simulation studies that the modified methods can map both iQTLs and non-imprinted QTLs (niQTLs) simultaneously as well as improve the accuracies of estimation of the positions and effects of iQTLs. In addition, the modified design can potentially reduce the workload in the construction of the imF_2 population.

THEORY

Modification of Experimental Design

Suppose there is a DH or RI population derived from a cross between two pure lines, P_1 and P_2. The experimental design proposed by Wen and Wu (2014) for iQTL mapping is to develop an imF_2 population by randomly crossing DH or RI lines, namely, Line i × Line j (i, j = 1, 2, 3, …; $i \neq j$). Consider a QTL with two alleles, Q_1 and Q_2. The two alleles can form four genotypes: Q_1Q_1, Q_1Q_2, Q_2Q_1, and Q_2Q_2, with one allele (say, Q_1) from the male gamete and the other (Q_2) from the female gamete in each genotype. Let a, d and i be the additive effect, dominance effect and imprinting effect of the QTL, respectively. Thus, in an imF_2 population, the single-QTL model would be (Wen and Wu, 2014):

$$y_j = \mu + ax_j + dz_j + it_j + \varepsilon_j \quad (1)$$

where y_j is the trait value of the j^{th} combination (or hybrid line); μ is the population mean; x_j, z_j and t_j are dummy variables taking values depending on the QTL genotype in the j^{th} combination (**Table 1**); and ε_j is residual error following a normal distribution $N(0, \sigma^2)$.

In the above design, the cross in each combination is "unidirectional," namely, one line is used as female parent and the other as male parent. However, there can be an alternative genetic mating design, in which reciprocal crosses are performed for every combination, namely, Line i × Line j (positive cross, PC) and Line j × Line i (negative cross, NC; i, j = 1, 2, 3, …; $i < j$). This modified experimental design generates a special imF_2 population. We call it reciprocal-cross imF_2 (R-imF_2) population. For distinction, we shall call the usual imF_2 population as unidirectional-cross imF_2 (U-imF_2) population. Genetically, Eq. (1) is still applicable to R-imF_2. Therefore, the iQTL mapping methods for U-imF_2 are also applicable to R-imF_2.

Modification of Point Mapping Method

Suppose the parental DH or RI population has been genotyped and therefore a genetic map has been constructed. Thus, the genotypes of imF_2 lines can be deduced from the parental DH or RI lines and the genetic map can be used for iQTL mapping. Suppose the genetic map is of ultrahigh density so that the markers can well represent the whole genome. Thus, iQTLs can be mapped by testing every marker throughout the genome. We call this approach point mapping (PM; Wen and Wu, 2014).

Suppose the size (total number of hybrid lines) of the imF_2 population is $2n$ (for R-imF_2 population, there are n PC and n NC hybrid lines, respectively). Let RSS_0, RSS_1 and RSS_2 be the

TABLE 1 | Values of dummy variables indicating the QTL genotype in Eq. (1).

QTL genotype (♀/♂)	x_j	z_j	t_j
Q_1/Q_1	1	0	0
Q_1/Q_2	0	1	1
Q_2/Q_1	0	1	−1
Q_2/Q_2	−1	0	0

minimum residual sum of squares calculated based on Eq. (1) under the hypotheses H_0: $a = d = i = 0$, H_1: $i = 0$ but not $a = d = 0$, and H_2: not $a = d = i = 0$, respectively. Thus, two approximate log-likelihood ratio tests can be performed as below:

$$\mathrm{LOD}_1 = n\left[log_{10}\left(RSS_1\right) - log_{10}\left(RSS_2\right)\right] \quad (2)$$

and

$$\mathrm{LOD}_2 = n\left[log_{10}\left(RSS_0\right) - log_{10}\left(RSS_2\right)\right] \quad (3)$$

The PM method proposed by Wen and Wu (2014) maps iQTLs by checking the imprinting effect of every marker in the genome using Eq. (2). The LOD_1 significance threshold is estimated by permutation tests (Churchill and Doerge, 1994). A genomic region covered by a LOD_1 peak exceeding the threshold is thought to harbor an iQTL and the highest point of the peak is the most probable position of the iQTL. Obviously, this is a one-step method (denoted as PM-1), in which an iQTL is mapped based on its imprinting effect only.

The modified PM method proposed here is a two-step method (denoted as PM-2). The first step is QTL mapping, namely, to map QTLs (including imprinted and non-imprinted) by testing every marker in the genome using Eq. (3). Similarly, the LOD_2 significance threshold used in this step can be estimated by permutation tests. The second step is iQTL identification, namely, to identify iQTLs among the mapped QTLs by checking the imprinting effect of each QTL using Eq. (2). A QTL is taken as an iQTL if its imprinting effect is significant. Otherwise, it is taken as a usual niQTL. The LOD_1 significance threshold used in the second step can also be estimated by permutation tests, but the tests are performed only on the mapped QTLs rather than on every marker in the genome.

Modification of Composite Point Mapping Method

The PM method can be extended to composite point mapping (CPM) by adding some markers as cofactors into Eq. (1), namely (Wen and Wu, 2014):

$$y_j = \mu + ax_j + dz_j + it_j + \sum_{k_1=1}^{m_1} a_{k_1}^* x_{k_1 j}^* + \sum_{k_2=1}^{m_2} d_{k_2}^* z_{k_2 j}^* + \sum_{k_3=1}^{m_3} i_{k_3}^* t_{k_3 j}^* + \varepsilon_j \quad (4)$$

where m_1, m_2, and m_3, $a_{k_1}^*$, $d_{k_2}^*$, and $i_{k_3}^*$, and $x_{k_1 j}^*$, $z_{k_2 j}^*$, and $t_{k_3 j}^*$ are the numbers, effects and corresponding dummy variables of additive, dominance and imprinting cofactors, respectively; other symbols are the same as those in Eq. (1). Cofactors can be selected by stepwise regression. Note that the three effects of a marker are orthogonal or independent to each other, among which only the significant ones are selected by the stepwise regression. Therefore, the markers selected as cofactors based on different effects can be different (Zeng, 1994). The CPM method proposed by Wen and Wu (2014) is a one-step method, corresponding to PM-1. Similarly, there can be an alternative two-step CPM method (CPM-2). CPM-1 and CPM-2 have a similar procedure to that of PM-1 and PM-2, respectively. The only difference of CPM from PM is that the RSS_0, RSS_1 and RSS_2 in Eqs. (2 and 3) are calculated based on Eq. (4) rather than on Eq. (1) under the corresponding hypotheses (H_0, H_1, and H_2).

Simulation Studies

To examine the feasibility and efficiency of the modified imF_2 design ($\mathrm{R\text{-}imF}_2$) and the modified statistical methods (PM-2 and CPM-2) for iQTL mapping in comparison with the previous design ($\mathrm{U\text{-}imF}_2$) and methods (PM-1 and CPM-1), two simulation studies were conducted. The first study was to compare the performances of $\mathrm{R\text{-}imF}_2$ and $\mathrm{U\text{-}imF}_2$, and of PM-1 and PM-2 in the mapping of a single iQTL; the second study was to compare the performances of CPM-1 and CPM-2 as well as PM-1 and PM-2 in the mapping of multiple iQTLs. The programs for PM and CPM were written in Visual Basic 6.0[1].

Simulation Study I

In this simulation study, we assumed that an iQTL was located at the middle of a chromosome, which was 100 cM in length and had one marker every cM. The iQTL segregated in a DH population of 100 lines, from which a $\mathrm{U\text{-}imF}_2$ or $\mathrm{R\text{-}imF}_2$ population comprising 800 hybrid lines was generated. The imprinting effect of the iQTL explained 2% of the phenotypic variance in the $\mathrm{U\text{-}imF}_2$ or $\mathrm{R\text{-}imF}_2$ population. Six different types of iQTL in terms of the effects (a, d, and i) were investigated, including the full-effect type, in which all sorts of effect exist, and five partial-effect types, in which either additive effect or dominance effect, or both do not exist (**Table 2**; Cheverud et al., 2008). The simulated data were analyzed with PM-1 and PM-2, respectively. Each case was simulated for 500 times. LOD thresholds for PM-1 and PM-2 at the overall significance level of 0.05 were estimated by simulation under the null hypothesis with 5,000 replicates. The results (**Table 2**) showed:

(i) When other conditions (iQTL type and mapping method) were fixed, the results (means and standard deviations of estimates and statistical powers) obtained under the two designs were all very similar, suggesting that the two designs are basically equivalent for iQTL mapping.

(ii) Except in the case of type DIB, which had no additive and dominance effects, the power of QTL detection in PM-2 (step 1) was always higher than the power of iQTL detection in PM-1, and the difference was especially large in the cases of type FULL, PEP and PEM. However, the power of iQTL detection in PM-2 (step 2) was always lower than that in PM-1, and the difference was especially large in the case of type DIB.

(iii) The mean estimates of iQTL position obtained by PM-1 and PM-2 were very close to the real value in all the cases, suggesting that a single iQTL can be unbiasedly mapped by both methods. However, the standard deviation of iQTL position obtained by PM-2 was always significantly smaller than that obtained by PM-1 except in the case of type DIB. Even for type DIB, the former was still a little

[1]The software package of our methods is available by contacting us through email.

TABLE 2 | Simulation results of point mapping of a single iQTL.

Type[a]	Design	Method	Pos. (cM)[b]	*a*	*d*	*i*	Power (%)[c]
FULL	U-imF_2	PM-1	49.65 ± 9.01	1.80 ± 0.60	1.60 ± 0.86	2.22 ± 0.52	92.2
(0.05)		PM-2	49.91 ± 2.86	2.04 ± 0.51	2.10 ± 0.74	2.15 ± 0.41	86.6 (99.6)
	R-imF_2	PM-1	50.27 ± 7.71	1.80 ± 0.57	1.66 ± 0.81	2.15 ± 0.48	92.8
		PM-2	49.96 ± 3.25	1.99 ± 0.51	2.00 ± 0.76	2.11 ± 0.40	88.2 (99.8)
		Real value	50	2	2	2	
DIPOD	U-imF_2	PM-1	49.98 ± 7.32	−0.03 ± 0.53	1.57 ± 0.79	2.21 ± 0.41	90.8
(0.03)		PM-2	49.90 ± 3.49	−0.03 ± 0.57	2.01 ± 0.71	2.17 ± 0.42	82.6 (95.0)
	R-imF_2	PM-1	49.88 ± 7.03	0.01 ± 0.53	1.70 ± 0.81	2.16 ± 0.43	91.8
		PM-2	49.76 ± 4.40	−0.01 ± 0.56	2.09 ± 0.76	2.14 ± 0.40	85.8 (92.2)
		Real value	50	0	2	2	
DIPUD	U-imF_2	PM-1	49.86 ± 6.42	−0.04 ± 0.51	1.68 ± 0.81	−2.21 ± 0.50	89.8
(0.03)		PM-2	49.58 ± 3.79	−0.04 ± 0.56	2.04 ± 0.75	−2.21 ± 0.43	82.4 (92.8)
	R-imF_2	PM-1	50.02 ± 7.10	0.03 ± 0.48	1.66 ± 0.82	−2.19 ± 0.43	91.6
		PM-2	50.44 ± 5.15	0.04 ± 0.53	1.99 ± 0.79	−2.17 ± 0.40	87.4 (93.0)
		Real value	50	0	2	−2	
PEP	U-imF_2	PM-1	50.43 ± 9.80	1.77 ± 0.57	−0.02 ± 0.72	2.22 ± 0.46	91.6
(0.04)		PM-2	50.02 ± 4.46	2.03 ± 0.53	−0.01 ± 0.81	2.17 ± 0.43	86.4 (99.0)
	R-imF_2	PM-1	50.57 ± 7.43	1.83 ± 0.58	0.01 ± 0.72	2.21 ± 0.41	92.4
		PM-2	50.01 ± 4.43	2.02 ± 0.54	0.02 ± 0.82	2.16 ± 0.43	88.8 (97.8)
		Real value	50	2	0	2	
PEM	U-imF_2	PM-1	49.69 ± 8.61	1.79 ± 0.54	−0.06 ± 0.74	−2.23 ± 0.40	90.2
(0.04)		PM-2	49.90 ± 5.17	2.00 ± 0.53	−0.05 ± 0.82	−2.16 ± 0.43	87.8 (99.2)
	R-imF_2	PM-1	49.69 ± 7.45	1.79 ± 0.56	−0.03 ± 0.71	−2.18 ± 0.44	91.2
		PM-2	49.94 ± 4.76	2.02 ± 0.54	−0.03 ± 0.81	−2.13 ± 0.42	89.0 (98.6)
		Real value	50	2	0	−2	
DIB	U-imF_2	PM-1	50.07 ± 7.87	0.01 ± 0.50	0.07 ± 0.70	2.23 ± 0.45	89.0
(0.02)		PM-2	50.05 ± 7.32	0.01 ± 0.59	0.05 ± 0.87	2.33 ± 0.38	71.0 (76.4)
	R-imF_2	PM-1	49.61 ± 8.22	−0.02 ± 0.51	0.02 ± 0.70	2.17 ± 0.45	89.2
		PM-2	49.70 ± 7.42	−0.02 ± 0.59	0.00 ± 0.91	2.27 ± 0.43	69.6 (71.0)
		Real value	50	0	0	2	

[a]FULL, full-effect type ($a = d = i$); DIPOD, dominance imprinting, polar, over dominance ($a = 0 \cap d = i$); DIPUD, dominance imprinting, polar, under dominance ($a = 0 \cap d = -i$); PEP, parental expression, paternal ($a = i \cap d = 0$); PEM, parental expression, maternal ($a = -i \cap d = 0$); DIB, dominance imprinting, bipolar ($a = 0 \cap d = 0$). The data in parenthesis are the total proportion of phenotypic variance explained by the iQTL (the proportion of phenotypic variance explained by the imprinting effect of the iQTL is 0.02 in each type). [b]The estimates of position and effects are shown as "mean ± standard deviation." [c]The powers were estimated based on 500 times of simulation. The data in parenthesis are the power of QTL detection in PM-2 (step 1). The LOD thresholds for PM-1, step 1 of PM-2, and step 2 of PM-2 were 1.97, 3.25, and 1.79 in U-imF_2 and 1.97, 3.36, and 1.80 in R-imF_2, respectively.

smaller than the latter. This suggests that PM-2 is more precise than PM-1 for iQTL mapping in general.

(iv) The estimation results of imprinting effect obtained by PM-1 and PM-2 were similar in all the cases. Noticeably, the means were always a little larger than the real value, suggesting that both methods may slightly overestimate the imprinting effect. For the additive and dominance effects, the means obtained by PM-2 were very close to the real values, suggesting that the estimation is unbiased; but the means obtained by PM-1 were always obviously smaller than the real values, suggesting that PM-1 may underestimate the additive and dominance effects. These results suggest that PM-2 is better than PM-1 for estimating the additive and dominance effects of iQTL.

Simulation Study II

In this simulation study, we assumed that a species had three pairs of chromosomes, each of which was 150 cM in length and had one marker every cM. There were seven QTLs in the genome, including three iQTLs on chromosome 1, one iQTLs and one niQTL on chromosome 2, and two iQTLs on chromosome 3 (**Table 3**). An R-imF_2 population comprising 800 hybrid lines was generated from a DH population of 100 lines. Each QTL accounted for ∼7% of the phenotypic variance in the R-imF_2 population. The simulated data were analyzed with PM-1, PM-2, CPM-1, and CPM-2, respectively. Cofactors for CPM-1 and CPM-2 were selected by stepwise regression at the significance level of 0.05. LOD thresholds at the overall significance level of 0.05 were estimated by permutation test with 1,000 replicates. The results (**Figure 1** and **Table 3**) showed:

(i) All of the iQTLs were precisely mapped by both CPM-1 and CPM-2, and the estimates of iQTL positions obtained by the two methods were almost completely the same (with only a slight difference at Q1-3). PM-1 and PM-2 also precisely mapped some of the iQTLs. These two

TABLE 3 | Simulation results of mapping multiple iQTLs using PM-1, PM-2, CPM-1, and CPM-2.

Chr.	QTL[a]	Type	Method	Pos. (cM)[b]	a^b	d^b	i^b
1	Q1-1	PEP	PM-1	17	1.23	0.28	0.93
	(7.06, 3.53)		PM-2	17	1.23	0.28	0.93
			CPM-1	17	0.76	0.13	0.90
			CPM-2	17	0.76	0.13	0.90
			Real value	17	0.84	0	0.84
	Q1-2	DIPUD	PM-1	(62)	(0.24)	(1.07)	(−0.85)
	(7.06, 4.70)		PM-2	(62)	(0.24)	(1.07)	(−0.85)
			CPM-1	62	−0.09	086	−1.01
			CPM-2	62	−0.09	0.86	−1.01
			Real value	62	0	0.97	−0.97
	Q1-3	PEM	PM-1	102	0.86	0.08	−1.29
	(7.06, 3.53)		PM-2	102	0.86	0.08	−1.29
			CPM-1	102	0.77	0.02	−0.89
			CPM-2	103	0.79	0.03	−0.86
			Real value	103	0.84	0	−0.84
2	Q2-1	Non-imprinted	PM-1	ns	ns	ns	ns
	(7.06, 0)		PM-2	25	1.42	1.30	0.37
			CPM-1	ns	ns	ns	ns
			CPM-2	25	1.03	1.17	0.03
			Real value	25	0.97	0.97	0
	Q2-2	DIPOD	PM-1	70	0.38	0.90	0.96
	(7.06, 4.70)		PM-2	70	0.38	0.90	0.96
			CPM-1	70	−0.01	0.89	0.94
			CPM-2	70	−0.01	0.89	0.94
			Real value	70	0	0.97	0.97
3	Q3-1	DIB	PM-1	45	0.46	0.30	1.56
	(6.96, 6.96)		PM-2	(45)	(0.46)	(0.30)	(1.56)
			CPM-1	45	−0.08	0.08	1.29
			CPM-2	45	−0.08	0.08	1.29
			Real value	45	0	0	1.18
	Q3-2	FULL	PM-1	90	1.15	0.97	1.48
	(7.03, 2.81)		PM-2	90	1.15	0.97	1.48
			CPM-1	90	0.83	0.79	0.81
			CPM-2	90	0.83	0.79	0.81
			Real value	90	0.75	0.75	0.75

[a]The data in parenthesis are the percentages of phenotypic variance explained by the QTL and its imprinting effect, respectively. [b]The estimates in parenthesis were obtained based on unremarkable or unclear LOD peaks. ns, not significant. The LOD thresholds for PM-1, step 1 of PM-2, step 2 of PM-2, CPM-1, step 1 of CPM-2, and step 2 of CPM-2 were 2.42, 3.82, 2.37, 2.85, 4.15, and 2.79, respectively.

methods obtained exactly the same estimates of iQTL positions. However, the LOD peaks of PM-1 and PM-2 were broad. In addition, there were many small peaks, which may make it difficult to identify the peaks of true iQTLs (such as the peaks of Q1-2 in PM-1 and PM-2, and the peak of Q3-1 in PM-2) and result in ghost or false iQTLs (such as the peak on the left of Q2-1 and that between Q2-1 and Q2-2 in PM-1, and the peaks between Q1-1 and Q1-2, between Q2-1 and Q2-2, and between Q3-1 and Q3-1 in PM-2).

(ii) Corresponding to the estimation of iQTL positions, the estimates of iQTL effects were also completely the same between PM-1 and PM-2 and almost completely the same between CPM-1 and CPM-2, respectively. In most of the cases, the estimates of effects obtained by CPM-1 and CPM-2 were closer to the real values than those obtained by PM-1 and PM-2.

(iii) The LOD peaks of PM-2 were always higher than those of PM-1. This is consistent with the results of simulation study I. For the same reason, the LOD peaks of CPM-2 were higher than those of CPM-1 except for the DIB-type iQTL (Q3-1). In addition, as expected, the niQTL (Q2-1) was mapped only by PM-2 and CPM-2, respectively.

DISCUSSION

The advantages of iQTL mapping based on imF_2 populations have been demonstrated before (Wen and Wu, 2014). R-imF_2 is a special type of imF_2 population. Although the simulation study

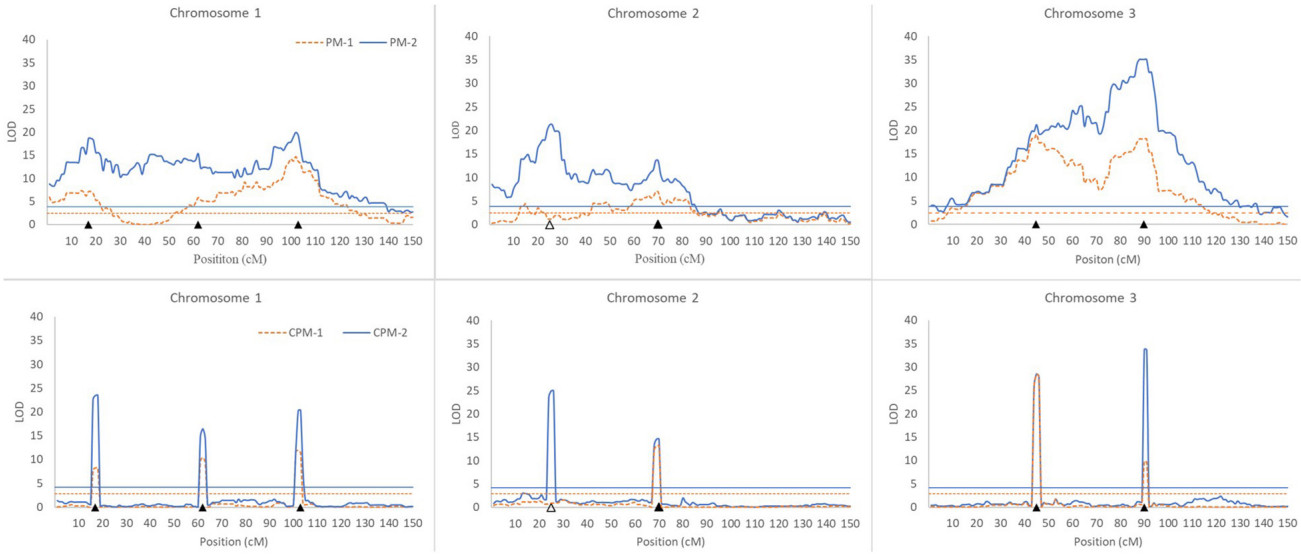

FIGURE 1 | Results of QTL scanning by PM methods **(upper)** and CPM methods **(lower)** in an assumed genome consisting of three chromosomes. The horizontal lines indicate LOD thresholds at the overall significance level of 0.05. The filled and blank triangles indicate the positions of iQTL and niQTL, respectively.

results suggest that R-imF_2 does not apparently improve the result of iQTL mapping in comparison with U-imF_2 (**Table 2**), it is expected to be advantageous for experimental operation. For each cross combination, only one hybrid line is produced in U-imF_2, while two hybrid lines are produced in R-imF_2. Therefore, R-imF_2 only needs half of the number of combinations used in U-imF_2. This makes the cross work more convenient and may, to some extent, alleviate the workload of developing the imF_2 population.

According to the simulation study results, PM-2 can estimate the position and the additive and dominance effects of an iQTL more precisely than PM-1 (**Table 2**). This is understandable. PM-1 estimates the position of an iQTL based on its imprinting effect only, while PM-2 estimates the position of an iQTL based on not only its imprinting effect, but also its additive and dominance effects. Obviously, the latter would have a higher statistical power as long as the additive and dominance effects exist (**Table 2**). This would surely increase the estimation precision of iQTL position and therefore increase the estimation precision of iQTL effects.

Although PM-2 can noticeably improve the estimation of iQTL position and effects, the power of detecting iQTL in PM-2 is always lower than that in PM-1 (**Table 2**). This means that there is a cost of losing power for gaining precision in PM-2. The reason may be that an iQTL detected by PM-1 is at the position, where the imprinting effect has the highest significance, while the iQTL position estimated by PM-2 may not be at the point, where the imprinting effect is the most significant. Nevertheless, the power difference of iQTL detection between PM-1 and PM-2 is not large except in the case of type DIB (**Table 2**). Therefore, the improvement of estimation precision achieved by PM-2 is cost worthy.

Although the PM methods behave well in the mapping of a single iQTL, they are not ideal for multiple iQTL mapping (**Figure 1**). In practice, therefore, it is more appropriate to use the CPM methods. PM-2 demonstrates the merit of two-step analysis. CPM-2 also exhibits the merit of higher LOD score in the identification of QTL position (**Figure 1**). However, the LOD peaks obtained by CPM-1 and CPM-2 usually have the same width for the same iQTL (**Figure 1**). This suggests that the two methods have similar resolution in iQTL mapping. Therefore, the higher LOD score of CPM-2 might have little help for increasing the precision of iQTL mapping, probably due to the role of cofactors. Nevertheless, CPM-2 still has an advantage over CPM-1, namely, it can map both iQTLs and niQTLs.

Considering that the basic principle and conclusions of iQTL mapping may not change with the density of markers (Wen and Wu, 2014), we did not consider in this paper the situation of iQTL mapping based on conventional low-density maps. The methods described above can be directly applied to the conventional map as long as the values of the dummy variables at the position to be tested in Eqs. (1 and 4) are replaced with the expected values calculated according to the flanking markers (Wen and Wu, 2014).

Power is the most frequently used index for evaluating the efficiency of a QTL mapping method, which can reflect the probability of type II error. Besides, false discovery rate (FDR) may be also an important index for the evaluation because it can reflect the probability of type I error (Li et al., 2012b). A good QTL mapping method should have higher power but lower FDR. Similar to the power, the FDR in QTL mapping can also be estimated by computer simulation (Li et al., 2012b). In the simulation study I of this study, one QTL was set at the center of a chromosome of 50 cM in length in each case. Since the QTL really existed,

a single QTL detected on the chromosome could be always regarded as true, although the estimated QTL position was very imprecise (far from the real position) sometimes. Certainly, if there were two or more QTLs detected on the chromosome simultaneously, the additional QTL should be false. However, such situations did not occur in the simulations. Therefore, the FDR was always zero in our simulation study. In other words, the conditions assumed in our simulation study avoided the occurrence of false discovery. This was beneficial to the comparative study based on the power.

In this study, we only consider the iQTL mapping based on one-environment experiment. However, the genetic model can be easily extended to adapt the data of multi-environment experiment, from which the QTL-by-environment interaction can be analyzed using suitable statistical methods such as the mixed linear model approach, which has been used for mapping QTLs with the digenic epistasis and QTL-by-environment interaction as well as additive and dominance effects based on imF_2 population (Gao and Zhu, 2007).

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 32071892 and 31571558) and the Sci-Tech Innovation Special Fund of Fujian Agriculture and Forestry University (Grant Nos. CXZX2017248 and CXZX2019127G).

REFERENCES

Babak, T., Deveale, B., Armour, C., Raymond, C., Cleary, M. A., van der Kooy, D., et al. (2008). Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.* 18, 1735–1741. doi: 10.1016/j.cub.2008.09.044

Bauer, M. J., and Fischer, R. L. (2011). Genome demethylation and imprinting in the endosperm. *Curr. Opin. Plant Biol.* 14, 162–167. doi: 10.1016/j.pbi.2011.02.006

Cheverud, J. M., Hager, R., Roseman, C., Fawcett, G., Wang, B., and Wolf, J. B. (2008). Genomic imprinting effects on adult body composition in mice. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4253–4258. doi: 10.1073/pnas.0706562105

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Croteau, A. K., and Croteau, N. (2004). Mechanisms of epigenetic variation: polymorphic imprinting. *Curr. Genomics* 5, 417–429.

Cui, Y. (2007). A statistical framework for genome-wide scanning and testing of imprinted quantitative trait loci. *J. Theor. Biol.* 244, 115–126. doi: 10.1016/j.jtbi.2006.07.009

Cui, Y., Cheverud, J. M., and Wu, R. (2007). A statistical model for dissecting genomic imprinting through genetic mapping. *Genetica* 130, 227–239. doi: 10.1007/s10709-006-9101-x

Cui, Y., Li, S., and Li, G. (2008). Functional mapping imprinted quantitative trait loci underlying developmental characteristics. *Theor. Biol. Med. Modelling* 5:6. doi: 10.1186/1742-4682-5-6

Cui, Y., Lu, Q., Cheverud, J. M., Littell, R. C., and Wu, R. (2006). Model for mapping imprinted quantitative trait loci in an inbred F2 design. *Genomics* 87, 543–551. doi: 10.1016/j.ygeno.2005.11.021

Danilevskaya, O. N., Hermon, P., Hantke, S., Muszynski, M. G., Kollipara, K., and Ananiev, E. V. (2003). Duplicated fie genes in maize: expression pattern and imprinting suggest distinct functions. *Plant Cell* 15, 425–438. doi: 10.1105/tpc.006759

de Koning, D. J., Rattink, A. P., Harlizius, B., van Arendonk, J. A., Brascamp, E. W., and Groenen, M. A. (2000). Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proc. Natl. Acad. Sci. U.S.A.* 97, 7947–7950. doi: 10.1073/pnas.140216397

Gao, Y. M., and Zhu, J. (2007). Mapping QTLs with digenic epistasis under multiple environments and predicting heterosis based on QTL effects. *Theor. Appl. Genet.* 115, 325–333.

Girardot, M., Feil, R., and Llères, D. (2013). Epigenetic deregulation of genomic imprinting in humans: causal mechanisms and clinical implications. *Epigenomics* 5, 715–728. doi: 10.2217/epi.13.66

Hagan, J. P., O'Neill, B. L., Stewart, C. L., Kozlov, S. V., and Croce, C. M. (2009). At least ten genes define the imprinted Dlk1-Dio3 cluster on mouse chromosome 12qF1. *PLoS One* 4:e4352. doi: 10.1371/journal.pone.0004352

Hager, R., Cheverud, J. M., and Wolf, J. B. (2008). Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics* 178, 1755–1762. doi: 10.1534/genetics.107.080697

Haghighi, F., and Hodge, S. E. (2002). Likelihood formulation of parent-of-origin effects on segregation analysis, including ascertainment. *Am. J. Hum. Genet.* 70, 142–156. doi: 10.1086/324709

Haley, C. S., Knott, S. A., and Elsen, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* 136, 1195–1207.

Hanson, R. L., Kobes, S., Lindsay, R. S., and Knowler, W. C. (2001). Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *Am. J. Hum. Genet.* 68, 951–962. doi: 10.1086/319508

Haun, W. J., Laoueillé-Duprat, S., O'connell, M. J., Spillane, C., Grossniklaus, U., Phillips, A. R., et al. (2007). Genomic imprinting, methylation and molecular evolution of maize Enhancer of zeste (Mez) homologs. *Plant J.* 49, 325–337.

Hur, S. K., Freschi, A., Ideraabdullah, F., Thorvaldsen, J. L., Luense, L. J., Weller, A. H., et al. (2016). Humanized H19/Igf2 locus reveals diverged imprinting mechanism between mouse and human and reflects silver-russell syndrome phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 113, 10938–10943. doi: 10.1073/pnas.1603066113

Ikeda, Y. (2012). Plant imprinted genes identified by genome-wide approaches and their regulatory mechanisms. *Plant Cell Physiol.* 53, 809–816. doi: 10.1093/pcp/pcs049

Jiang, J., Shen, B., O'Connell, J. R., VanRaden, P. M., Cole, J. B., and Ma, L. (2017). Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. *BMC Genomics* 18:425. doi: 10.1186/s12864-017-3821-4

Karami, K., Zerehdaran, S., Javadmanesh, A., Shariati, M. M., and Fallahi, H. (2019). Characterization of bovine (*Bos taurus*) imprinted genes from genomic to amino acid attributes by data mining approaches. *PLoS One* 14:e0217813. doi: 10.1371/journal.pone.0217813

Kermicle, J. L. (1970). Dependence of the R-mottled aleurone phenotype in maize on mode of sexual transmission. *Genetics* 66, 69–85.

Knapp, M., and Strauch, K. (2004). Affected-sib-pair test for linkage based on constraints for identical-by-descent distributions corresponding to disease

models with imprinting. *Genet. Epidemiol.* 26, 273–285. doi: 10.1002/gepi.10320

Knott, S. A., Marklund, L., Haley, C. S., Andersson, K., Davies, W., Ellegren, H., et al. (1998). Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* 149, 1069–1080.

Lawson, H. A., Cheverud, J. M., and Wolf, J. B. (2013). Genomic imprinting and parent-of-origin effects on complex traits. *Nat. Rev. Genet.* 14, 609–617. doi: 10.1038/nrg3543

Li, H. H., Zhang, L. Y., and Wang, J. K. (2012b). Estimation of statistical power and false discovery rate of QTL mapping methods through computer simulation. *Chin. Sci. Bull.* 57:27012710. doi: 10.1007/s11434-012-5239-3

Li, S., Wang, X., Li, J., Yang, T., Min, L., Liu, Y., et al. (2012a). Bayesian mapping of genome-wide epistatic imprinted loci for quantitative traits. *Theor. Appl. Genet.* 124, 1561–1571. doi: 10.1007/s00122-012-1810-1

Li, Y., Coelho, C. M., Liu, T., Wu, S., Wu, J., Zeng, Y., et al. (2008). A statistical model for estimating maternal-zygotic interactions and parent-of-origin effects of QTLs for seed development. *PLoS One* 3:e3131. doi: 10.1371/journal.pone.0003131

Liu, T., Todhunter, R. J., Wu, S., Hou, W., Mateescu, R., Zhang, Z., et al. (2007). A random model for mapping imprinted quantitative trait loci in a structured pedigree: an implication for mapping canine hip dysplasia. *Genomics* 90, 276–284. doi: 10.1016/j.ygeno.2007.04.004

Long, J. E., and Cai, X. (2007). Igf-2r expression regulated by epigenetic modification and the locus of gene imprinting disrupted in cloned cattle. *Gene* 388, 125–134. doi: 10.1016/j.gene.2006.10.014

Luo, M., Platten, D., Chaudhury, A., Peacock, W. J., and Dennis, E. S. (2009). Expression, imprinting, and evolution of rice homologs of the polycomb group genes. *Mol. Plant* 2, 711–723. doi: 10.1093/mp/ssp036

Mackay, D., and Temple, I. K. (2017). Human imprinting disorders: principles, practice, problems and progress. *Eur. J. Med. Genet.* 60, 618–626. doi: 10.1016/j.ejmg.2017.08.014

Mantey, C., Brockmann, G. A., Kalm, E., and Reinsch, N. (2005). Mapping and exclusion mapping of genomic imprinting effects in mouse F2 families. *J. Hered.* 96, 329–338. doi: 10.1093/jhered/esi044

Morison, I. M., Ramsay, J. P., and Spencer, H. G. (2005). A census of mammalian imprinting. *Trends Genet.* 21, 457–465. doi: 10.1016/j.tig.2005.06.008

Nolan, C. M., Killian, J. K., Petitte, J. N., and Jirtle, R. L. (2001). Imprint status of M6P/IGF2R and IGF2 in chickens. *Dev. Genes Evol.* 211, 179–183. doi: 10.1007/s004270000132

Pei, L., Zhang, L., Li, J., Shen, C., Qiu, P., Tu, L., et al. (2019). Tracing the origin and evolution history of methylation-related genes in plants. *BMC Plant Biol.* 19:307. doi: 10.1186/s12870-019-1923-7

Pembrey, M., Saffery, R., Bygren, L. O., Network in epigenetic epidemiology, and Network in epigenetic epidemiology (2014). Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research. *J. Med. Genet.* 51, 563–572. doi: 10.1136/jmedgenet-2014-102577

Pratt, S. C., Daly, M. J., and Kruglyak, L. (2000). Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *Am. J. Hum. Genet.* 66, 1153–1157. doi: 10.1086/302830

Raissig, M. T., Baroux, C., and Grossniklaus, U. (2011). Regulation and flexibility of genomic imprinting during seed development. *Plant Cell* 23, 16–26. doi: 10.1105/tpc.110.081018

Sandovici, I., Kassovska-Bratinova, S., Loredo-Osti, J. C., Leppert, M., Suarez, A., Stewart, R., et al. (2005). Interindividual variability and parent of origin DNA methylation differences at specific human Alu elements. *Hum. Mol. Genet.* 14, 2135–2143. doi: 10.1093/hmg/ddi218

Santure, A. W., and Spencer, H. G. (2011). Quantitative genetics of genomic imprinting: a comparison of simple variance derivations, the effects of inbreeding, and response to selection. *G3* 1, 131–142. doi: 10.1534/g3.111.000042

Shete, S., and Amos, C. I. (2002). Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *Am. J. Hum. Genet.* 70, 751–757. doi: 10.1086/338931

Shete, S., Zhou, X., and Amos, C. I. (2003). Genomic imprinting and linkage test for quantitative-trait Loci in extended pedigrees. *Am. J. Hum. Genet.* 73, 933–938. doi: 10.1086/378592

Spencer, H. G. (2002). The correlation between relatives on the supposition of genomic imprinting. *Genetics* 161, 411–417.

Strauch, K., Fimmers, R., Kurz, T., Deichmann, K. A., Wienker, T. F., and Baur, M. P. (2000). Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *Am. J. Hum. Genet.* 66, 1945–1957. doi: 10.1086/302911

Wang, C., Wang, Z., Prows, D. R., and Wu, R. (2012). A computational framework for the inheritance pattern of genomic imprinting for complex traits. *Brief. Bioinform.* 13, 34–45. doi: 10.1093/bib/bbr023

Wen, Y., and Wu, W. (2014). Mapping of imprinted quantitative trait loci using immortalized F2 populations. *PLoS One* 9:e92989. doi: 10.1371/journal.pone.0092989

Wolf, J. B., Hager, R., and Cheverud, J. M. (2008). Genomic imprinting effects on complex traits: a phenotype-based perspective. *Epigenetics* 3, 295–299. doi: 10.4161/epi.3.6.7257

Wu, R. L., Ma, C. X., Wu, S. S., and Zeng, Z. B. (2002). Linkage mapping of sex-specific differences. *Genet Res.* 79, 85–96.

Yang, R., Wang, X., Wu, Z., Prows, D. R., and Lin, M. (2010). Bayesian model selection for characterizing genomic imprinting effects and patterns. *Bioinformatics* 26, 235–241. doi: 10.1093/bioinformatics/btp620

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136, 1457–1468.

Zhang, M., Zhao, H., Xie, S., Chen, J., Xu, Y., Wang, K., et al. (2011). Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20042–20047. doi: 10.1073/pnas.1112186108

Zhou, X., Fang, M., Li, J., Prows, D. R., and Yang, R. (2012). Characterization of genomic imprinting effects and patterns with parametric accelerated failure time model. *Mol. Genet. Genomics* 287, 67–75. doi: 10.1007/s00438-011-0661-9

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 January 2021
doi: 10.3389/fgene.2020.602526

How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy?

Samuel B. Fernandes, Kevin S. Zhang, Tiffany M. Jamann and Alexander E. Lipka**

Department of Crop Science, University of Illinois at Urbana-Champaign, Urbana, IL, United States

Quantification of the simultaneous contributions of loci to multiple traits, a phenomenon called pleiotropy, is facilitated by the increased availability of high-throughput genotypic and phenotypic data. To understand the prevalence and nature of pleiotropy, the ability of multivariate and univariate genome-wide association study (GWAS) models to distinguish between pleiotropic and non-pleiotropic loci in linkage disequilibrium (LD) first needs to be evaluated. Therefore, we used publicly available maize and soybean genotypic data to simulate multiple pairs of traits that were either (i) controlled by quantitative trait nucleotides (QTNs) on separate chromosomes, (ii) controlled by QTNs in various degrees of LD with each other, or (iii) controlled by a single pleiotropic QTN. We showed that multivariate GWAS could not distinguish between QTNs in LD and a single pleiotropic QTN. In contrast, a unique QTN detection rate pattern was observed for univariate GWAS whenever the simulated QTNs were in high LD or pleiotropic. Collectively, these results suggest that multivariate and univariate GWAS should both be used to infer whether or not causal mutations underlying peak GWAS associations are pleiotropic. Therefore, we recommend that future studies use a combination of multivariate and univariate GWAS models, as both models could be useful for identifying and narrowing down candidate loci with potential pleiotropic effects for downstream biological experiments.

Keywords: Simulation, multi-trait, Unified Mixed-Model, QTN, maize, soybean, LD

OPEN ACCESS

Edited by:
Riyan Cheng,
University of California, San Diego,
United States

Reviewed by:
Alexander V. Favorov,
Johns Hopkins University,
United States
Kui Zhang,
Michigan Technological University,
United States

****Correspondence:***
Samuel B. Fernandes
samuelf@illinois.edu
Alexander E. Lipka
alipka@illinois.edu

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

Received: *03 September 2020*
Accepted: *11 December 2020*
Published: *08 January 2021*

Citation:
Fernandes SB, Zhang KS, Jamann TM and Lipka AE (2021) How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy? Front. Genet. 11:602526. doi: 10.3389/fgene.2020.602526

1. INTRODUCTION

The number of traits available from state-of-the-art phenotyping techniques typically exceeds the number of genes in many species' genomes. For instance, the human genome contains over 20,000 genes (Wagner and Zhang, 2011), but the Human Metabolome Database (Wishart et al., 2007) alone has collected more than 114,000 metabolite traits. A direct consequence is that many genes likely control more than one of these traits, a phenomenon known as pleiotropy (Visscher and Yang, 2016). The identification and characterization of this phenomenon has been the subject of extensive research in the 100+ years following the first attributed use of the term "pleiotropy" in Platt (1910) Stearns (2010). Examples of important genes with pleiotropic effects in plant science include *Lg1* and its contribution to inflorescence and leaf traits in maize (Foster et al., 2004; Lewis et al., 2014) and multiple disease resistance attributed to *GH3-2* in rice (Fu et al., 2011) and *Lr67* in wheat (Moore et al., 2015). With the recent acquisition of high-throughput phenotype and genotype data, it is now possible to directly identify pleiotropic causal mutations (Wagner and Zhang, 2011).

The abundance of such high-throughput data in conjunction with a plethora of tools available for quantifying genotype-to-phenotype associations (Marchini et al., 2007; Purcell et al., 2007; Lipka et al., 2012; Zhou and Stephens, 2014) is providing increasing evidence for pleiotropic genes involved in evolution (Smith, 2016; Auge et al., 2019), disease resistance (Wisser et al., 2011; Lopez-Zuniga et al., 2019; Qiu et al., 2020), yield (Ward et al., 2019), and many other traits (Jiang et al., 2019; Rice et al., 2020). These analyses have also led to opposing views for (Boyle et al., 2017) and against (Wray et al., 2018) the ubiquitousness of pleiotropy in complex trait variation, particularly in the form of the omnigenic model. This model assumes that the same set of small-effect regulatory genes explain the vast majority of complex disease resistance traits expressed in a disease-relevant cell (Boyle et al., 2017).

One of the most commonly used approaches for quantifying genotype-to-phenotype relationships is the genome-wide association study (GWAS), which has been used to investigate pleiotropy (Wisser et al., 2011; Schaid et al., 2016; Rice et al., 2020). However, a significant drawback of a GWAS is that most of the markers available in typical high-throughput genotypic data are not causal. Instead, they are in imperfect linkage disequilibrium (LD) with the causal mutations of a given trait. This LD obfuscates the ability to distinguish a single pleiotropic causal mutation underlying multiple traits from multiple non-pleiotropic causal mutations in LD with each other (Gianola et al., 2015). Furthermore, it would only be possible to differentiate between a set of multiple non-pleiotropic causal mutations and one pleiotropic causal mutation if the former were in imperfect LD (Kemper et al., 2018). The scenario of tightly-linked non-pleiotropic causal mutations being mistaken for one pleiotropic causal mutation is known as spurious pleiotropy (Solovieff et al., 2013; van Rheenen et al., 2019). In addition to hindering the characterization of biological processes underlying trait variability, the presence of spurious pleiotropy in GWAS results could have serious negative downstream breeding ramifications (Chen and Lübberstedt, 2010). For instance, if two separate causal mutations in LD with antagonistic effects each control one of two correlated traits, breeders could allocate resources toward increasing population size to find individuals with recombination between these causal mutations (Schulthess et al., 2017). However, if a set of GWAS results are misinterpreted as suggesting that one pleiotropic causal mutation is present (i.e., the scenario of spurious pleiotropy is realized), then such efforts to increase the population size may never be undertaken.

Many studies use the term cross-phenotype to refer to markers with strong statistical associations with multiple traits (Tyler et al., 2016). Several univariate and multivariate GWAS approaches have been implemented to detect cross-phenotype associations (Zhou and Stephens, 2014; Cichonska et al., 2016; Joo et al., 2016), with multi-trait models shown to be optimum under many circumstances (Yang and Wang, 2012; Porter and O'Reilly, 2017; Melo et al., 2019; Pitchers et al., 2019; Rice et al., 2020). Although there is great value in detecting cross-phenotype associations, there is still a critical need to distinguish whether the underlying causal mutation(s) are pleiotropic or are non-pleiotropic but in strong LD. We hypothesized that one of the major reasons underlying the difficulty in distinguishing between these two scenarios is that the most widely-used univariate and multivariate GWAS models are insufficient for making such a distinction. Therefore, we used publicly available maize and soybean genotypic data to simulate pairs of correlated traits that were either (i) controlled by non-pleiotropic quantitative trait nucleotides (QTNs) on separate chromosomes, (ii) controlled by non-pleiotropic QTNs in various degrees of LD with each other, or (iii) controlled by a single pleiotropic QTN. We then assessed the ability of state-of-the-art univariate and multivariate GWAS models to identify these QTNs. We predicted that as the amount of LD between the non-pleiotropic QTNs increased, the multivariate GWAS results would more closely resemble those from traits controlled by a single pleiotropic QTN.

2. MATERIALS AND METHODS

2.1. Maize and Soybean Data

In this study, we used publicly available molecular marker data from two crop species, specifically maize (*Zea mays* L.) and soybean (*Glycine max* L.). These two species were selected because of their contrasting rates of LD decay; while soybean tends to have long-range LD (Hyten et al., 2007; Zhang et al., 2015), more rapid LD decay is typically observed in maize (Gore et al., 2009; Romay et al., 2013). The maize data were comprised of 2,815 accessions from the North Central Regional Plant Introduction Station (NCRPIS) panel (Romay et al., 2013), while the soybean data consisted of a random sample of 2,815 accessions in maturity groups III and IV from SoyBase (Song et al., 2015). To investigate the impact of sample size on the results, for each data set, we considered the full set of $S_1 = 2,815$ accessions, a subsample of $S_2 = 1,000$ accessions, and a subsample of $S_3 = 500$ individuals. The accessions of S_3 were randomly sampled from S_2, whereas the accessions of S_2 were randomly sampled from S_1, i.e., $S_3 \subset S_2 \subset S_1$. All subsamples were obtained using the (*vcftools --max-indv*) command in vcftools (Danecek et al., 2011). Details on how to access the datasets are provided in the **Supplementary Material**.

The maize data included 681,257 single-nucleotide polymorphisms (SNPs) obtained through genotyping-by-sequencing (Romay et al., 2013), available at http://cbsusrv04.tc.cornell.edu/users/panzea/download.aspx?filegroupid=6. The soybean data were downloaded from SoyBase (Song et al., 2015) at http://soybase.org/snps/download.php, and consisted of 42,291 SNPs obtained with the SoySNP50K (Song et al., 2013). The same filters were applied to both datasets using vcftools. These filters included removing all SNPs with more than 5% missing data. Additionally, Plink was used to conduct LD pruning, where the LD parameter was set to $r^2 = 0.9$ (*--indep-pairwise 100 10 0.9*) (Purcell et al., 2007). Thus, only markers that were in an LD of $r^2 \geq 0.9$ were filtered out. Only SNPs from chromosomal DNA that passed the minor allele count threshold of 5 in S_3 were included in this simulation study. Consequently, the final data sets used for simulation were 44,930 SNPs for maize, and 18,364 for soybean.

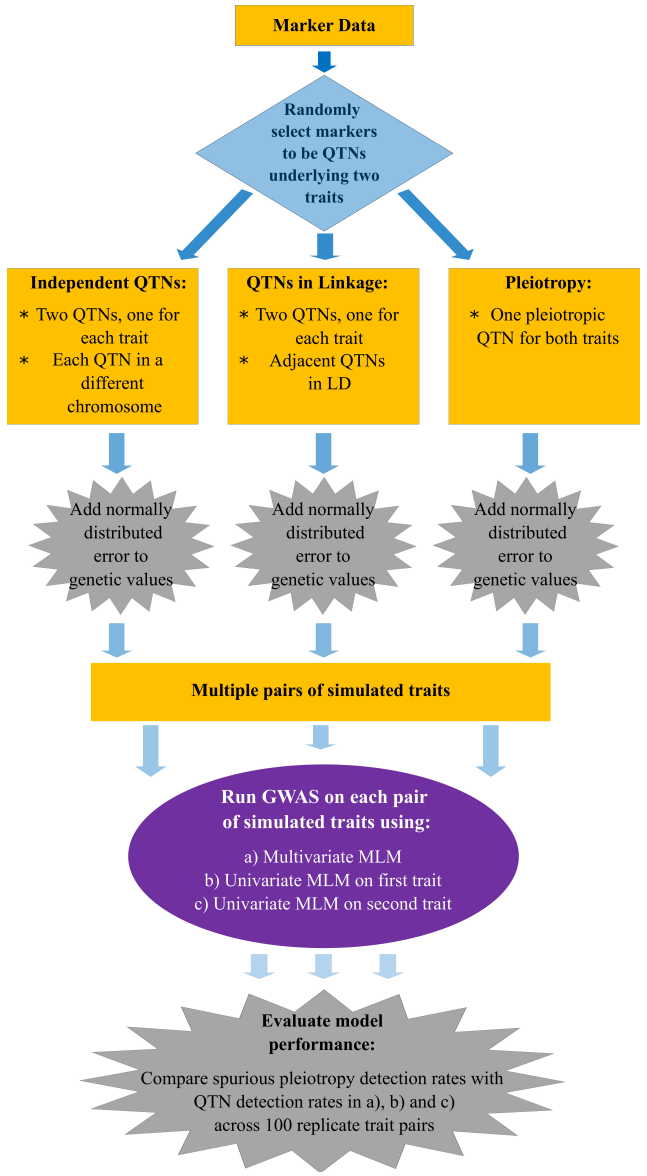

FIGURE 1 | Flow chart with the methods used to simulate, conduct a genome-wide association study (GWAS), and detect quantitative trait nucleotides (QTNs) in multiple traits.

2.2. Trait Simulation

The flow chart presented in **Figure 1** summarizes the main aspects of the simulation study we conducted. In brief, we simulated pairs of traits controlled by either pleiotropic or non-pleiotropic QTNs. Each pair of traits was simulated with the simplePHENOTYPES (Fernandes and Lipka, 2020) package in the R software (R Core Team, 2020). We were specifically interested in comparing and contrasting the behavior of single peak-associated SNPs from GWAS, similar in magnitude to those reported in Rice et al. (2020), over multiple simulation replicates. Thus, all individual traits were controlled by exactly one additive QTN selected from either the maize or soybean marker data. For each pair of replicate traits, a maximum of two QTNs were selected. To investigate the impact of LD between two non-pleiotropic QTNs on the GWAS results, we sampled QTNs in three different scenarios. First, the QTNs were sampled from different chromosomes (called "Independent QTNs" in **Figure 1**). Such a configuration of QTNs was achieved by simulating trait pairs independently in simplePHENOTYPES. For a given set of input parameter values (**Table 1**), this process was repeated until 100 replicate trait pairs, with each trait in a pair controlled by a QTN on a different chromosome, were obtained. Next, we simulated trait pairs where the maximum amount of LD between the two linked non-pleiotropic QTNs controlling each trait was specified (called "QTNs in linkage" in **Figure 1**). We simulated this configuration in simplePHENOTYPES by specifying *architecture* = "*LD*," and indicated the amount of maximum desired LD between pairs of selected SNPs through the *ld* input parameter. We controlled this amount of LD both directly (i.e., the LD between the two QTNs) and indirectly (the LD between each QTN and a marker located in between). Finally, we simulated pairs of correlated traits that were controlled by a single pleiotropic QTN (called "Pleiotropy" in **Figure 1**). This configuration was specified in simplePHENOTYPES by *architecture* = "*pleiotropic*."

Table 1 provides a summary of the input parameters considered in the simulation study. Briefly, three configurations of narrow-sense heritability (h^2) were simulated: two with the same h^2 for both traits and one with a different h^2 for each trait. The latter configuration is a common situation in breeding programs, where a trait of interest with small heritability is correlated to a trait of less interest but with a higher h^2 (Fernandes et al., 2018). Because a single QTN controlled each trait, and the non-genetic variance was a function of the inputted heritability, the additive effect size of every QTN in this study was set to the same value, namely 0.10. To evaluate the impact of rare vs. common variants on the results, we also considered the minor allele frequencies (MAFs) of the selected markers as an input parameter (see **Table 2** for details). Altogether, we simulated 216 scenarios, where each scenario consisted of a unique combination of input parameters. Each scenario was replicated 100 times using the option *vary_QTN* = *TRUE* in simplePHENOTYPES, meaning that a different pair of QTNs were selected for each replicate.

We used simplePHENOTYPES' option "*remove_QTN* = *TRUE*" to simulate the frequently occurring scenario of the causal mutations not being included in the marker sets. Thus, for each of the 100 replicate trait pairs evaluated at a given scenario, the marker data were saved without the SNPs used as QTNs. Accordingly, for all traits, we conducted GWAS on all markers except the one selected to be the QTN.

2.3. Genome-Wide Association Studies

Multivariate and univariate GWAS was conducted on all simulated traits. For each replicate trait pair, we used the multivariate version of the unified mixed linear model (MLM) (Yu et al., 2006) implemented in GEMMA (Zhou and Stephens, 2014) to conduct the multivariate GWAS. In this analysis, a given replicate trait pair was included in this model as the multivariate

TABLE 1 | Description of the input parameter values considered to simulate each pair of traits in the simulation study.

QTN selection[a]	Type of LD[b]	MAF[c]	h^{2d}		Sample size	Species
			Trait 1	Trait 2		
QTNs independently selected	Direct	0.05	0.30	0.30	500	Maize
LD controlled at < 0.01	Indirect	0.40	0.30	0.80	1,000	Soybean
LD controlled at < 0.98			0.80	0.80	2,815	
One pleiotropic QTN						

Each combination of input parameter values resulted in 216 simulation scenarios.
[a]*QTN, quantitative trait nucleotide.*
[b]*LD, linkage disequilibirum (r^2).*
[c]*MAF, minor allele frequency.*
[d]*h^2, narrow-sense heritability.*

TABLE 2 | Description of how minor allele frequency (MAF) was controlled in the simulation study.

QTN[a] configuration	MAF control
QTNs independently selected	Both QTNs selected based on MAF
LD[b] between QTNs directly controlled	QTN for first trait selected based on MAF
LD between QTNs indirectly controlled	Common marker located between QTNs selected based on MAF
One pleiotropic QTN	Pleiotropic QTN selected based on MAF

[a]*QTN, quantitative trait nucleotide.*
[b]*LD, linkage disequilibirum.*

response variable. The multivariate MLM was fitted in GEMMA using the commands (*"gemma --bfile bed_file -lmm 2 -miss 0.001 -maf 0.001 -r2 0.999999 -n 1 2 -k kinship.txt -o output"*), with the kinship matrix (VanRaden, 2008) calculated with the AGHmatrix R package (Amadeu et al., 2016). Similarly, for each of the two simulated traits contributing to a replicate trait pair, an analogous univariate unified MLM was fitted in the GEMMA software using all of the same commands except for *-n 1*. No fixed-effect covariates accounting for subpopulation structure were included in any GWAS model because (i) subpopulation structure did not explicitly contribute to the variability of the simulated traits, and (ii) all QTNs were randomly sampled irrespective of the degree to which their alleles segregated by subpopulations.

2.4. QTN Detection Rate for Univariate and Multivariate GWAS

For each simulation scenario, we compared the proportion of 100 replicate trait pairs in which the multivariate MLM identified a signal in the vicinity of the QTN(s) and the proportion in which the univariate MLM identified a signal in the vicinity of the QTN controlling the tested trait. We applied the Benjamini and Hochberg (1995) procedure to control the genome-wide false discovery rates (FDR) at 10%, and 5% for each model ran on each replicate trait pair. A SNP-trait association passing this threshold was deemed to be in the vicinity of a given QTN if it was within 10 kb (in maize) or 1 Mb (in soybean) of the QTN. These physical window sizes roughly correspond to a pairwise LD decay of $r^2 = 0.10$ in both species (**Supplementary Figures 1**, **2**). To compare the influence of window sizes on the results, we also considered window sizes of 1 kb in maize and 10 kb in soybean; these results are presented in **Supplementary Figures 20**–**29**, **40**–**49**.

For a given replicate trait pair, the multivariate MLM (which tested H_0: No association between the tested SNP and any trait in the multivariate model) was said to have identified a QTN if at least one SNP with an FDR-adjusted P-value <0.10 (or 0.05 when the FDR was controlled at 5%) was located within the surrounding physical window. Similarly, for a given trait in a replicate trait pair, the univariate MLM (which tested H_0: No association between the tested SNP and the trait in the univariate model) was said to have correctly identified the QTN underlying that trait if at least one SNP with an FDR-adjusted P-value <0.10 (or 0.05) was located within the physical window of that QTN. Thus, across the 100 replicate trait pairs simulated at each setting, we recorded the following percentages:

1. The percentage of replicate trait pairs where a given GWAS model identified the QTN underlying the first trait.
2. The percentage of replicate trait pairs where a given GWAS model identified the QTN underlying the second trait.
3. The percentage of replicate trait pairs where a given GWAS model identified both QTNs underlying both traits.
4. The percentage of replicate trait pairs where a given GWAS model identified at least one statistically significantly associated marker outside of both windows for both traits.

When these percentages 1–3 were calculated for the multivariate GWAS under the "Independent QTNs" and "QTNs in linkage" scenario, they were referred to as the spurious pleiotropy detection rate. Otherwise, these proportions were called QTN detection rates. For both multivariate and univariate GWAS, the percentages calculated in 4 were called the error rate. Finally, as a measure of regional LD, i.e., LD in the region surrounding the selected QTN, we calculated the LD (r^2) between the selected QTN and the 20 SNPs upstream and the 20 SNPs downstream.

3. RESULTS

In general, the results were similar across sample sizes and heritabilities. Unless noted otherwise, we highlight below the findings at the relatively moderate sample size of 1,000

individuals, heritability of trait pairs set to $h^2 = (0.30, 0.80)$, 10% FDR and window size of 10 kb for maize and 1 Mb for soybean. We chose to present these particular heritabilities because of the aforementioned interest in correlated traits with contrasting heritabilities among breeders (Fernandes et al., 2018). For completeness, results for the remaining sample sizes and heritabilities are included in the **Supplementary Material**.

3.1. Observed MAFs Were Similar to User-Inputted Values, but Observed LD Was Lower

The various user-inputted parameters in simplePHENOYPTES enabled control of the MAFs of QTNs, as well as the LD between non-pleiotropic QTNs, to a certain extent. For QTNs where we specified the MAFs as an input parameter (indicated by a darker color in **Figure 2**; **Supplementary Figures 6–9**), the observed MAF distributions were similar to the user-inputted values. For QTNs where the MAFs were not directly controlled as an input parameter (indicated by a lighter color in **Figure 2**), most observed MAFs tended to be lower in maize than in soybean.

As expected, the observed LD between non-pleiotropic QTN pairs tended to be higher in soybean than in maize, although outlying instances of similar levels of high LD were observed in maize (**Figure 2**; **Supplementary Figures 6–9**). Surprisingly, the distribution of LD between non-pleiotropic QTN under the independent QTNs scenario yielded outlying LD values greater than what was observed under the direct control of LD at $r^2 = 0.01$. Because each pair of independent QTNs were simulated on separate chromosomes, we attribute these outlying values to interchromosomal LD. Thus, these simulated traits yielded pairs of non-pleiotropic QTNs with contrasting levels of LD between each other, enabling a thorough evaluation of the performance of univariate and multivariate GWAS models.

3.2. QTN and Spurious Pleiotropy Detection Rates Varied Across Sample Sizes, Heritabilities and QTN MAFs

The QTN and spurious pleiotropy detection rates generally increased as the sample size increased (**Supplementary Figures 10–12**, **20–22**, **30–32**, **40–42**). Similarly, these rates increased monotonically as the heritabilities increased (**Figure 3**, **Supplementary Figures 10–12**, **20–22**, **30–32**, **40–42**). The overall high QTN and spurious pleiotropy detection rates in soybean precluded the discernment of any notable trends in the GWAS approaches' performance across the observed MAFs (**Supplementary Figures 10–12**, **20–22**, **30–32**, **40–42**). However, in maize, we noted that for most settings, higher QTN and spurious pleiotropy detection rates tended to be observed for QTNs where the MAFs were specified to be around 0.40 instead of 0.05 (**Figure 4**). In general, all the conclusions were similar when varying the FDR and window size. The largest difference in this regard was noted in soybean, specifically in that a considerably higher QTN and spurious pleiotropy detection rate was noted whenever the multiple testing was adjusted at 10% FDR and the window size was 1 Mb (**Figure 5**; **Supplementary Figures 16–19**).

3.3. Observed Multivariate GWAS Performance for Non-pleiotropic QTNs in Linkage and a Single Pleiotropic QTN

The multivariate GWAS results are presented in their entirety in **Figure 5** and **Supplementary Figures 16–19**, **26–29**, **36–39**, **46–49**). In general, high spurious pleiotropy detection rates were observed under the "QTNs in Linkage" scenario. Specifically, for QTNs that were in high LD, we observed that the multivariate GWAS spurious pleiotropy detection rate of both QTNs (depicted as the light green bar in **Figure 5**) tended to be similar to or greater than the multivariate GWAS detection rate of the pleiotropic QTNs (yellow bar in **Figure 5**). Interestingly, we also noted a trend in the ability of multivariate GWAS to identify each individual non-pleiotropic QTN in LD (depicted as the purple and blue-green bars in **Figure 5**). That is, with the exception of indirect LD of 0.01, we observed that all individual multivariate GWAS spurious pleiotropy detection rates were higher than the corresponding multivariate GWAS detection rates on the pleiotropy scenario. The pattern of error rate, i.e., significant markers detected outside the predefined window size, was similar to the QTN and spurious pleiotropy detection rate (**Supplementary Figures 50–89**). The only notably different result when considering the error rate was observed in the independent QTNs scenario, where it resulted in a reduced error rate compared to the other genetic architectures.

3.4. Univariate GWAS Displayed Distinct Detection Patterns for Non-pleiotropic QTNs in High LD and Single Pleiotropic QTNs

Univariate GWAS tended to yield distinct patterns of QTN detection under both (i) high LD between non-pleiotropic QTNs and (ii) pleiotropy (depicted as the two rightmost columns of **Figure 5**; **Supplementary Figures 16–19**, **26–29**, **36–39**, **46–49**). Specifically, the simultaneous detection rate of the QTNs for both traits (depicted as the green bars in **Figure 5**) tended to be relatively similar to the individual QTN detection rates for each trait (depicted as the purple and blue-green bars **Figure 5**). For the remaining scenarios where non-pleiotropic QTN were simulated (presented in the four leftmost columns of **Figure 5**), we contrastingly observed that the simultaneous detection rate of each pair of non-pleiotropic QTNs tended to be less similar to the individual QTN detection rates. These results suggest that univariate GWAS could be extremely useful for distinguishing between a single pleiotropic QTN and two or more non-pleiotropic QTNs in linkage. In the scenarios of high LD, the SNPs selected to be QTNs were located in regions of slightly higher LD (**Supplementary Figures 3–5**). Consequently, the univariate QTN detection rate was slightly higher in the instances where QTNs in LD were simulated. In most cases, the error rate was similar across different settings.

4. DISCUSSION

The full potential of GWAS to contribute to the identification of pleiotropy will not be realized until its ability to distinguish

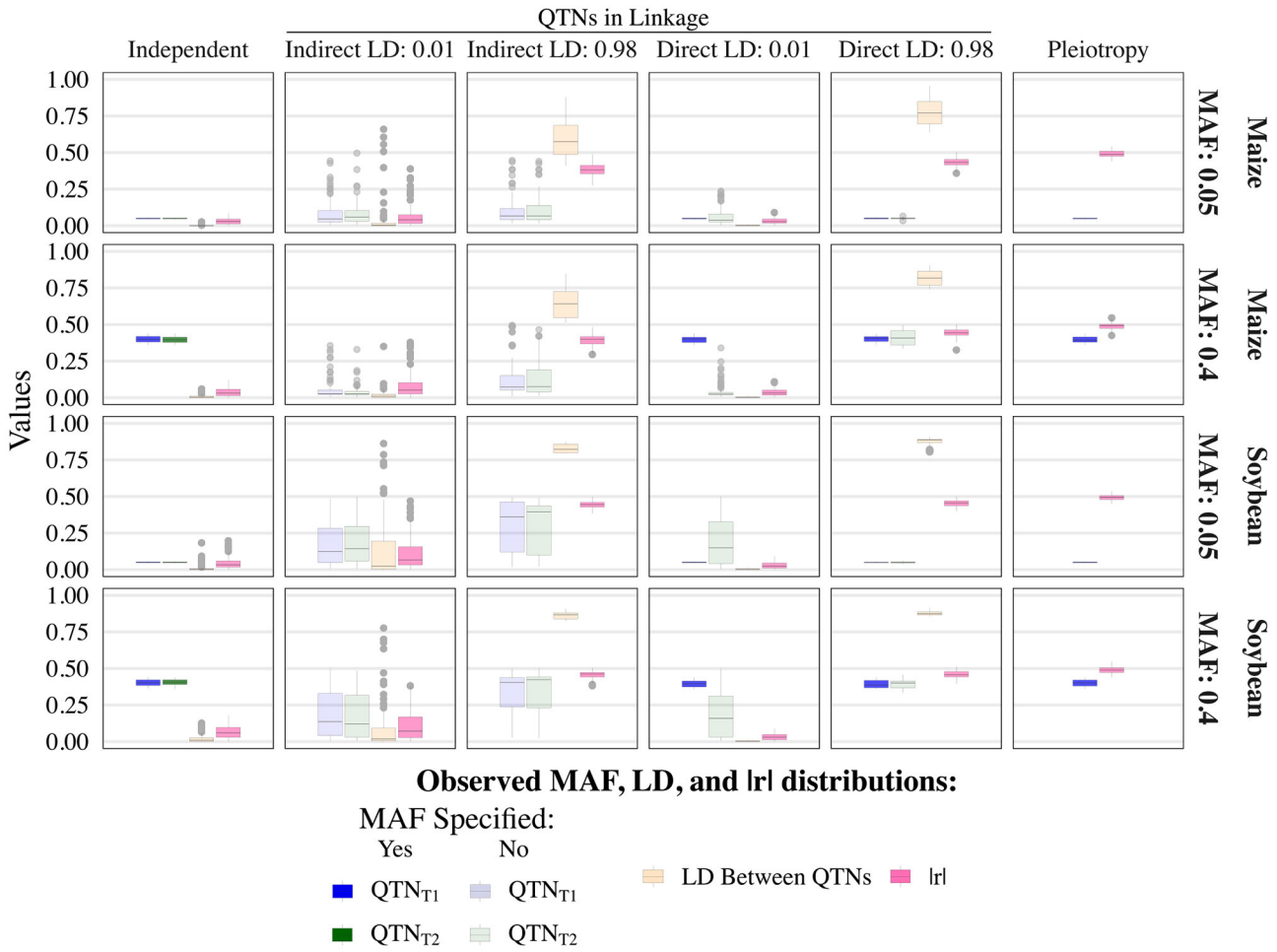

FIGURE 2 | Observed minor allele frequencies (MAF) for quantitative trait nucleotides (QTN) controlling trait 1 (QTN_{T1}) and trait 2 (QTN_{T2}), and the observed linkage disequilibrium (LD) between them, measured as r^2, for the sample size of 1,000 and narrow-sense heritability of 0.3 and 0.8, for traits 1 and 2, respectively. Darker colors indicate QTNs that had MAF directly controlled by an input parameter of the simulation, whereas lighter colors indicate QTNs where MAF was not controlled. The simulated genetic architecture is listed in the horizontal and vertical titles.

between a single pleiotropic causal mutation and multiple non-pleiotropic causal mutations in LD is scrutinized in real genomic data. Therefore, we used publicly available maize and soybean marker data to conduct a simulation study that quantified the QTN and spurious pleiotropy detection rates of both pleiotropic and non-pleiotropic QTNs for two widely-used statistical models in plant GWAS. We specifically used the univariate and multivariate MLM and controlled for multiple testing at 10% FDR. Our results showed that even at surprisingly small LD between non-pleiotropic QTNs, the multivariate GWAS model tended to yield high spurious pleiotropy detection rates. Because of the high spurious pleiotropy detection rates we inferred that multivariate GWAS was unable to distinguish between a single pleiotropic QTN and two non-pleiotropic QTNs in LD. We also observed that for pleiotropic QTNs, the univariate GWAS model's simultaneous QTN detection rates for both traits were similar to the QTN detection rates for the individual traits; such a degree of similarity was observed only at non-pleiotropic QTNs pairs in the highest amount of pairwise LD that we specified in our simulation parameters. Collectively, these results suggest that the univariate GWAS model might be useful in conjunction with multivariate GWAS model for distinguishing between true and spurious pleiotropy.

4.1. High Spurious Pleiotropy Detection Rates From Multivariate GWAS Were Observed Under LD

The potential of multivariate GWAS models has been demonstrated in many studies (Galesloot et al., 2014; Zhou and Stephens, 2014; Pitchers et al., 2019; Rice et al., 2020). Our results agree with this previous work, as the observed ability of multivariate GWAS to identify QTNs was generally high for all scenarios particularly in soybean. The fact that the multivariate GWAS was able to detect non-pleiotropic QTNs is not surprising because the null

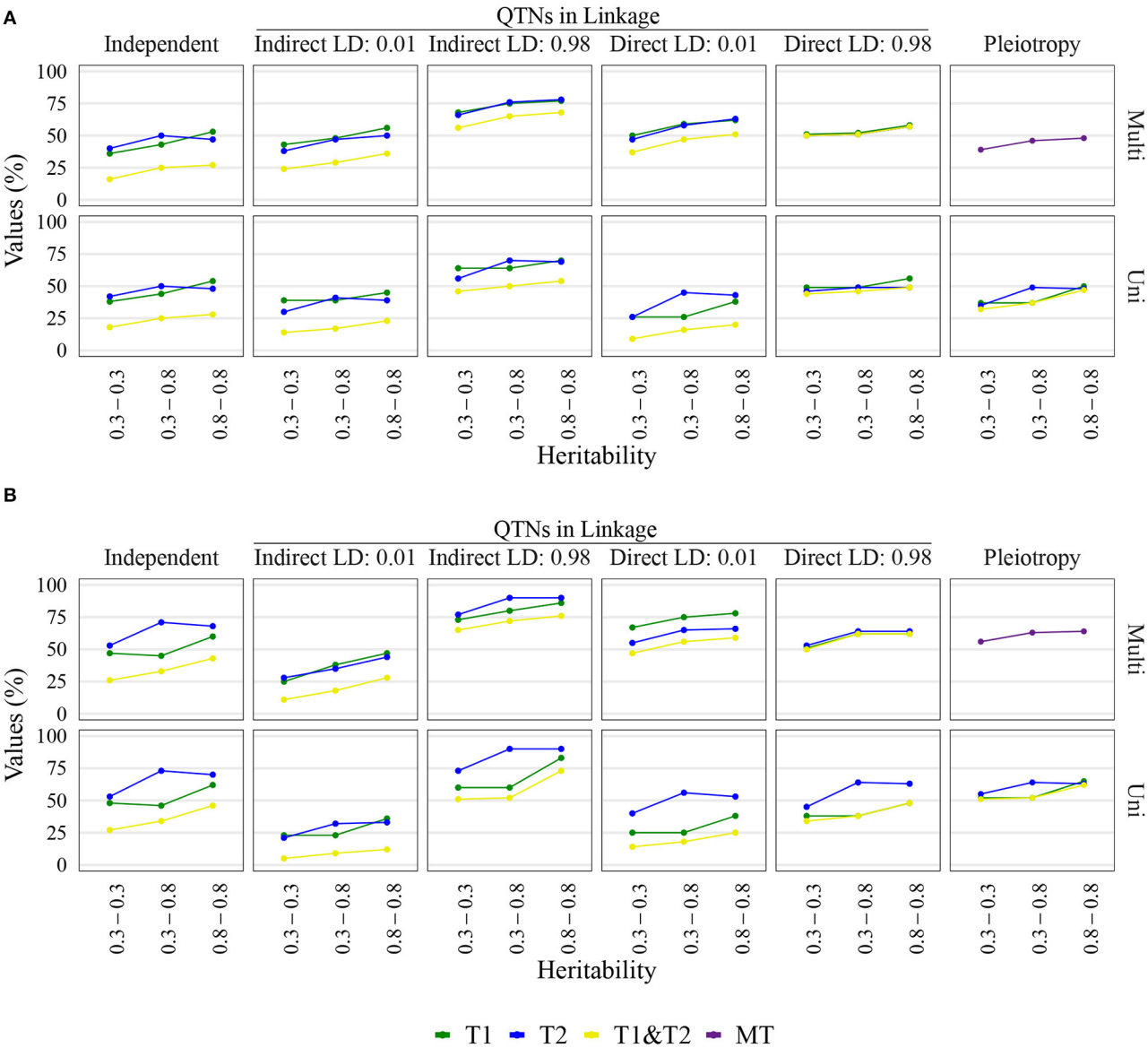

FIGURE 3 | Quantitative trait nucleotide (QTN) and spurious pleiotropy detection rate (Y-axis) achieved by multivariate (Multi) and univariate (Uni) GWAS, relative to the QTN controlling trait 1 (T1), trait 2 (T2), and both QTN simultaneously (T1&T2) or, in the pleiotropic scenario, relative to the pleiotropic QTN (MT). These values were obtained for maize with a sample size of 1,000. The X-axis displays the narrow-sense heritability for Trait 1 (bottom value) and Trait 2 (top value). **(A)** Inputted minor allele frequency (MAF) of 0.05; **(B)** MAF of 0.4.

hypothesis for most multivariate tests of association, including those used for the multivariate MLM, is H_0: No association between the tested SNP and any trait (Schaid et al., 2016; Salinas et al., 2018). Thus, the multivariate MLM's detection of non-pleiotropic QTN, and more specifically spurious pleiotropy under the "QTNs in linkage" scenario, should not be regarded as false positives because these events technically occur in the alternative hypothesis. Nevertheless, the outcome of spurious pleiotropy underscores an intrinsic lack of resolution to distinguish between pleiotropic and non-pleiotropic QTNs.

The observed performance of multivariate GWAS at the various levels of LD between non-pleiotropic QTNs on the same chromosome was insightful. Although a previous study showed that multivariate GWAS could not distinguish between a single pleiotropic QTN and multiple non-pleiotropic QTNs in LD (Chebib and Guillaume, 2019), we expected that at low levels of LD between non-pleiotropic QTNs, the spurious pleiotropy detection rates would be similar to QTN detection rates under the scenario where non-pleiotropic QTNs were simulated on separate chromosomes. Furthermore, we predicted that as the amount of LD between the non-pleiotropic QTNs

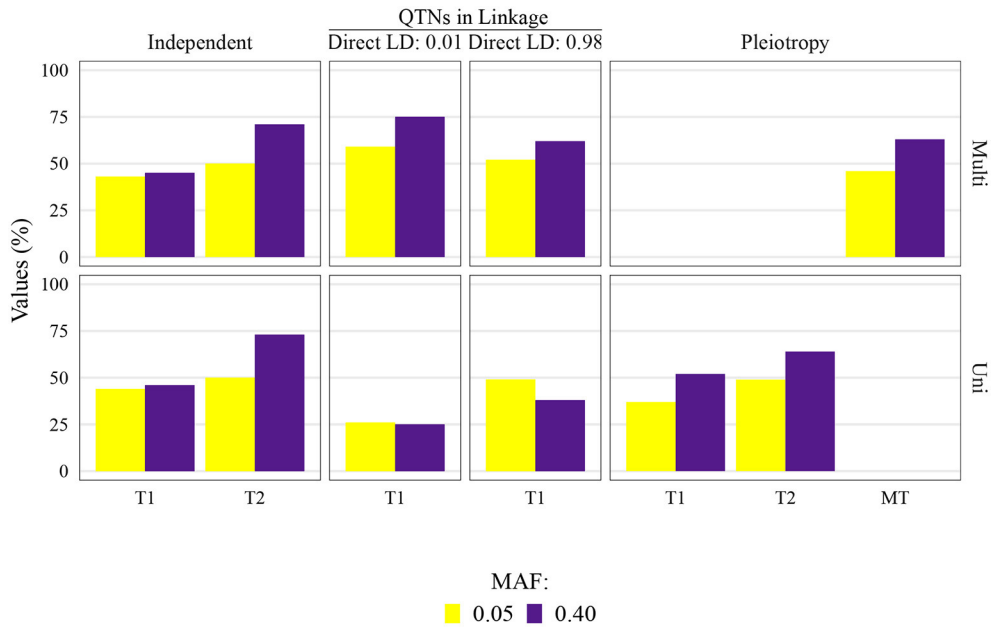

FIGURE 4 | Quantitative trait nucleotide (QTN) and spurious pleiotropy detection rate (Y-axis) in scenarios for which minor allele frequency (MAF) was directly controlled by a simulation input parameter. These values were obtained by multivariate (Multi) and univariate (Uni) GWAS, relative to the QTN controlling trait 1 (T1), and trait 2 (T2), or in the pleiotropic scenario, relative to the pleiotropic QTN (MT). This figure shows results for maize with a sample size of 1,000, and a narrow-sense heritability of 0.3 and 0.8, for Trait 1 and Trait 2, respectively.

increased, the spurious pleiotropy detection rate of multivariate GWAS would become similar to the observed multivariate GWAS detection rate of a single pleiotropic QTN. Instead, we observed that even at LD levels of $r^2 < 0.01$ between non-pleiotropic QTNs, the multivariate GWAS model yielded high spurious pleiotropy detection rates, a trend that was analogous to the QTN detection rates observed for traits controlled by one pleiotropic QTN. Interestingly, for the most stringent control of LD between non-pleiotropic QTNs on the same chromosome (i.e., $r^2 < 0.01$), the maximum amount of observed LD was less than some outlying values of interchromosomal LD between non-pleiotropic QTNs simulated on separate chromosomes (**Figure 2**). These results were contrary to our prior expectations, and we consequently made two main conclusions. First, we confirmed that multivariate GWAS is a potentially useful tool for identifying causal mutations. Second, multivariate GWAS, particularly the multivariate unified MLM, alone is insufficient for distinguishing between multiple QTNs in LD and a single pleiotropic QTN, irrespective of the amount of LD between the QTNs.

4.2. Univariate GWAS Is Potentially Useful for Identifying Pleiotropy

One of the most useful findings from this study was the subtle differences in univariate GWAS QTN detection rates for both non-pleiotropic QTNs in high LD and pleiotropic QTNs. We hypothesize that if incorporated into standard GWAS analyses, these subtle differences could play a crucial role in inferring whether or not a certain set of GWAS results suggest pleiotropy. Although there is a critical need for future studies to investigate the most appropriate use of univariate GWAS in such a role, our results suggest two steps for using univariate GWAS for this purpose. First, a univariate GWAS could be conducted on each trait separately. Second, an *a posteriori* analysis could then be used to determine how frequently each univariate GWAS detects a signal. If a signal is consistently detected across several univariate analyses of individual traits, this could provide evidence that a pleiotropic causal mutation is underlying the signals detected from GWAS.

4.3. Considerations for Further Studies on the Ability of GWAS to Identify Loci Controlling Multiple Traits

Our findings build upon other studies (e.g., Chebib and Guillaume, 2019), indicating that caution should be used when interpreting multivariate GWAS results. Moreover, it highlights the usefulness of univariate GWAS in making conclusions regarding trait genetic architecture. However, some potential weaknesses of our study should be considered when designing future research. In particular, the inconsistent amount of local LD levels surrounding QTNs selected from different genetic architectures is a potential source of bias. We opted to consider a fixed window size when detecting the QTNs; this favors a comparison across different sample sizes, but because the LD will vary, so will the chance of detecting a QTN in that specific window.

In particular, when we simulated traits with QTNs in high LD "QTNs in Linkage" scenario, we observed that they were typically

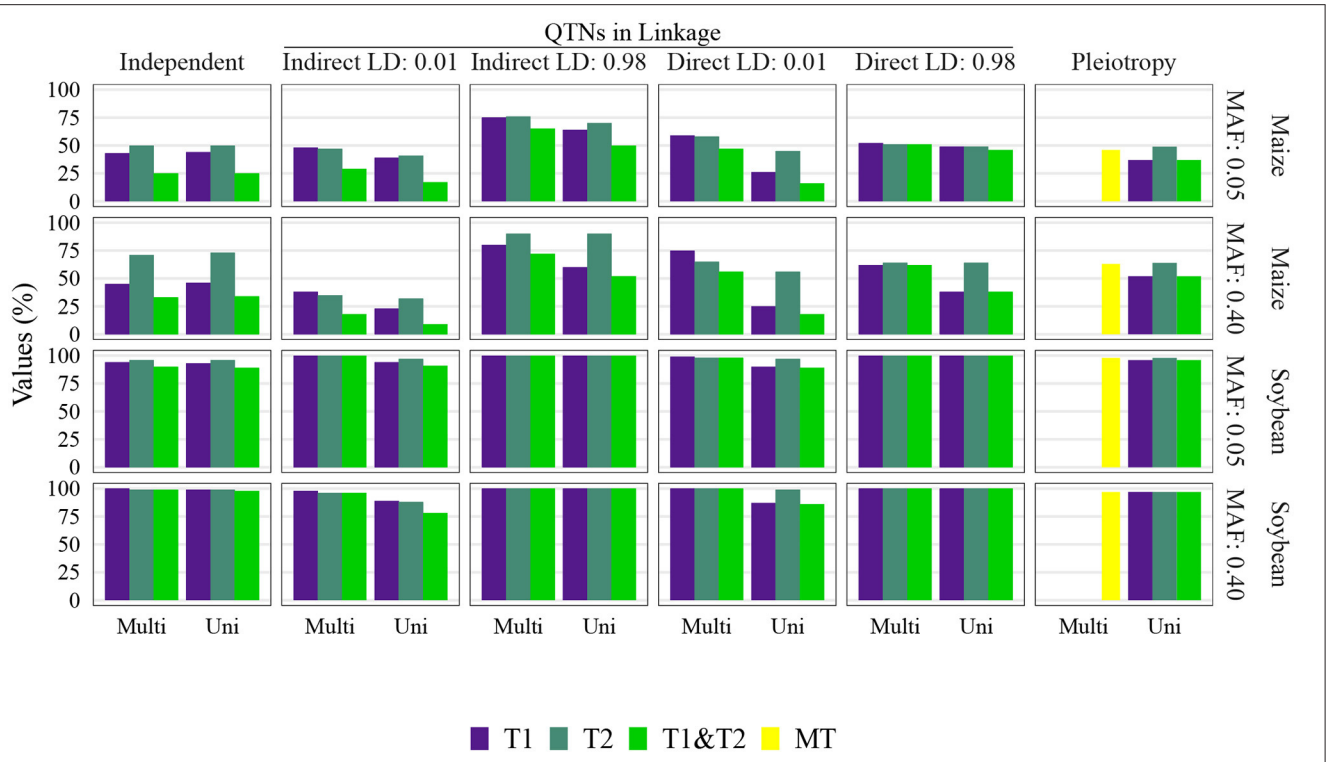

FIGURE 5 | Quantitative trait nucleotide (QTN) and spurious pleiotropy detection rate (Y-axis) achieved by multivariate (Multi) and univariate (Uni) GWAS (X-axis), relative to the QTN controlling trait 1 (T1), trait 2 (T2), and both QTN simultaneously (T1&T2) or, in the pleiotropic scenario, relative to the pleiotropic QTN (MT). The simulated genetic architecture is listed in the horizontal and vertical titles. These values were obtained with a sample size of 1,000, and a narrow-sense heritability of 0.3 and 0.8, for traits 1 and 2, respectively. MAF, minor allele frequencies.

selected from genomic regions that contained at least one pair of SNPs in high LD. Thus, the local LD in these regions tended to be biased upwards. For the remaining simulation scenarios, the amount of local LD was not biased upwards, as can be seen in **Supplementary Figures 3–5**. We infer that these differences in local LD might have influenced the observed QTN and spurious pleiotropy detection rates in this study. A potential solution for this issue would be to simulate pleiotropy and linked QTNs based on marker data with SNPs evenly spaced.

One final suggestion for future research is to investigate the impact of (i) the residual correlation between traits and (ii) the sign of QTN effect sizes on the performance of univariate and multivariate GWAS. As described in Jiang and Zeng (1995), the power of multivariate approaches should be less than those of the univariate ones whenever the direction of residual correlation (i.e., whether the sign of the residual correlation is positive or negative) is the same as those of the product of QTN effect sizes, regardless of whether these QTNs are in linkage or are pleiotropic. Thus, it is critical to determine if the overall patterns of QTN and spurious pleiotropy detection observed in this study are similar under genetic architectures where multivariate GWAS is theoretically expected to yield lower power than univariate GWAS.

5. CONCLUSION

The main conclusion from this study is that the use of either univariate or multivariate GWAS alone is insufficient for rigorously dissecting the genetic architecture of multiple traits. Association studies should instead use both univariate and multivariate models, as we demonstrated that both of these models are useful. Although our results suggest that multivariate GWAS cannot distinguish between a single pleiotropic QTN and multiple non-pleiotropic QTNs in LD, we confirmed that multivariate models are potentially useful for analyzing traits that are controlled by causal mutations that are either pleiotropic or in LD with each other. Once the genomic regions most likely to contain relevant causal mutations are identified through multivariate GWAS, univariate analyses could then be applied, potentially through the *a posteriori* analysis proposed in the Discussion, to shed light on whether or not the underlying causal mutations are pleiotropic. Such use of univariate and multivariate analyses in a concerted manner could maximize the amount of information ascertained from the GWAS of multiple traits, and potentially provide biological researchers with a smaller list of candidate loci that are likely to contribute to their variability.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SF, AL, and TJ designed the experiments. SF and KZ conducted the experiments. SF, AL, KZ, and TJ wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by National Science Foundation Plant Genome Research Project, Grant Number 1733606 and USDA-NIFA Grant Number 2019-67032-31623. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

ACKNOWLEDGMENTS

The authors would like to acknowledge the availability of advanced computational resources at the University of Illinois, which made it possible to conduct this simulation study in a reasonable amount of time.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.602526/full#supplementary-material

R scripts can also be found at: https://github.com/samuelbfernandes/fernandes_et_al_2020.

REFERENCES

Amadeu, R. R., Cellon, C., Olmstead, J. W., Garcia, A. A. F., Resende, M. F. R, Muñoz P. R., et al. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a blueberry example. *Plant Genome* 9:plantgenome2016.01.0009. doi: 10.3835/plantgenome2016.01.0009

Auge, G. A., Penfield, S., and Donohue, K. (2019). Pleiotropy in developmental regulation by flowering-pathway genes: is it an evolutionary constraint? *New Phytol.* 224, 55–70. doi: 10.1111/nph.15901

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

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Chebib, J., and Guillaume, F. (2019). Pleiotropy or linkage? Their relative contributions to the genetic correlation of quantitative traits and detection by multi-trait GWA studies. *bioRxiv* 656413. doi: 10.1101/656413

Chen, Y., and Lübberstedt, T. (2010). Molecular basis of trait correlations. *Trends Plant Sci.* 15, 454–461. doi: 10.1016/j.tplants.2010.05.004

Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimäki, T., et al. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* 32, 1981–1989. doi: 10.1093/bioinformatics/btw052

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Fernandes, S. B., Dias, K. O., Ferreira, D. F., and Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor. Appl. Genet.* 131, 747–755. doi: 10.1007/s00122-017-3033-y

Fernandes, S. B., and Lipka, A. E. (2020). simplePHENOTYPES: SIMulation of pleiotropic, linked and epistatic phenotypes. *BMC Bioinformatics* 21:491. doi: 10.1186/s12859-020-03804-y

Foster, T., Hay, A., Johnston, R., and Hake, S. (2004). The establishment of axial patterning in the maize leaf. *Development* 131, 3921–3929. doi: 10.1242/dev.01262

Fu, J., Liu, H., Li, Y., Yu, H., Li, X., Xiao, J., et al. (2011). Manipulating broad-spectrum disease resistance by suppressing pathogen-induced auxin accumulation in rice. *Plant Physiol.* 155, 589–602. doi: 10.1104/pp.110.163774

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:e95923. doi: 10.1371/journal.pone.0095923

Gianola, D., de los Campos, G., Toro, M. A., Naya, H., Schön, C. C., and Sorensen, D. (2015). Do molecular markers inform about pleiotropy? *Genetics* 201, 23–29. doi: 10.1534/genetics.115.179978

Gore, M. A., Chia, J.-M., Elshire, R. J., Sun, Q., Ersoz, E. S., Hurwitz, B. L., et al. (2009). A first-generation haplotype map of maize. *Science* 326, 1115–1117. doi: 10.1126/science.1177837

Hyten, D. L., Choi, I.-Y., Song, Q., Shoemaker, R. C., Nelson, R. L., Costa, J. M., et al. (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175, 1937–1944. doi: 10.1534/genetics.106.069740

Jiang, C., and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140, 1111–1127.

Jiang, J., Ma, L., Prakapenka, D., VanRaden, P. M., Cole, J. B., and Da, Y. (2019). A large-scale genome-wide association study in U.S. Holstein cattle. *Front. Genet.* 10:412. doi: 10.3389/fgene.2019.00412

Joo, J. W. J., Kang, E. Y., Org, E., Furlotte, N., Parks, B., Hormozdiari, F., et al. (2016). Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure. *Genetics* 204, 1379–1390. doi: 10.1534/genetics.116.189712

Kemper, K. E., Bowman, P. J., Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genet. Select. Evol.* 50, 1–13. doi: 10.1186/s12711-018-0377-y

Lewis, M. W., Bolduc, N., Hake, K., Htike, Y., Hay, A., Candela, H., et al. (2014). Gene regulatory interactions at lateral organ boundaries in maize. *Development* 141, 4590–4597. doi: 10.1242/dev.111955

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Lopez-Zuniga, L. O., Wolters, P., Davis, S., Weldekidan, T., Kolkman, J. M., Nelson, R., et al. (2019). Using maize chromosome segment substitution line populations for the identification of loci associated with multiple disease resistance. *G3 Genes Genomes Genet.* 9, 189–201. doi: 10.1534/g3.118.200866

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

Melo, D., Marroig, G., and Wolf, J. B. (2019). Genomic perspective on multivariate variation, pleiotropy, and evolution. *J. Hered.* 110, 479–493. doi: 10.1093/jhered/esz011

Moore, J. W., Herrera-Foessel, S., Lan, C., Schnippenkoetter, W., Ayliffe, M., Huerta-Espino, J., et al. (2015). A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. *Nat. Genet*. 47, 1494–1498. doi: 10.1038/ng.3439

Pitchers, W., Nye, J., Márquez, E. J., Kowalski, A., Dworkin, I., and Houle, D. (2019). A multivariate genome-wide association study of wing shape in drosophila melanogaster. *Genetics* 211, 1429–1447. doi: 10.1534/genetics.118.301342

Porter, H. F., and O'Reilly, P. F. (2017). Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci. Rep*. 7:38837. doi: 10.1038/srep38837

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81, 559–575. doi: 10.1086/519795

Qiu, Y., Cooper, J., Kaiser, C., Wisser, R., Mideros, S. X., and Jamann, T. M. (2020). Identification of loci that confer resistance to bacterial and fungal diseases of maize. *G3 (Bethesda)* 10, 2819–2828. doi: 10.1534/g3.120.401104

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing

Rice, B. R., Fernandes, S. B., and Lipka, A. E. (2020). Multi-trait genome-wide association studies reveal loci associated with maize inflorescence and leaf architecture. *Plant Cell Physiol*. 61, 1427–1437. doi: 10.1093/pcp/pcaa039

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol*. 14:R55. doi: 10.1186/gb-2013-14-6-r55

Salinas, Y. D., Wang, Z., and Dewan, A. T. (2018). Statistical analysis of multiple phenotypes in genetic epidemiologic studies: from cross-phenotype associations to pleiotropy. *Am. J. Epidemiol*. 187, 855–863. doi: 10.1093/aje/kwx296

Schaid, D. J., Tong, X., Larrabee, B., Kennedy, R. B., Poland, G. A., and Sinnwell, J. P. (2016). Statistical methods for testing genetic pleiotropy. *Genetics* 204, 483–497. doi: 10.1534/genetics.116.189308

Schulthess, A. W., Reif, J. C., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., et al. (2017). The roles of pleiotropy and close linkage as revealed by association mapping of yield and correlated traits of wheat (*Triticum aestivum* L.). *J. Exp. Bot*. 68, 4089–4101. doi: 10.1093/jxb/erx214

Smith, S. D. (2016). Pleiotropy and the evolution of floral integration. *New Phytol*. 209, 80–85. doi: 10.1111/nph.13583

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

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* 8:e54985. doi: 10.1371/journal.pone.0054985

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2015). Fingerprinting soybean germplasm and its utility in genomic research. *G3 Genes Genomes Genet*. 5, 1999–2006. doi: 10.1534/g3.115.019000

Stearns, F. W. (2010). One hundred years of pleiotropy: a retrospective. *Genetics* 186, 767–773. doi: 10.1534/genetics.110.122549

Tyler, A. L., Crawford, D. C., and Pendergrass, S. A. (2016). The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief. Bioinformatics* 17, 13–22. doi: 10.1093/bib/bbv050

van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet*. 20, 567–581. doi: 10.1038/s41576-019-0137-z

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91, 4414–4423. doi: 10.3168/jds.2007-0980

Visscher, P. M., and Yang, J. (2016). A plethora of pleiotropy across complex traits. *Nat. Genet*. 48, 707–708. doi: 10.1038/ng.3604

Wagner, G. P., and Zhang, J. (2011). The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet*. 12, 204–213. doi: 10.1038/nrg2949

Ward, B. P., Brown-Guedira, G., Kolb, F. L., Van Sanford, D. A., Tyagi, P., Sneller, C. H., et al. (2019). Genome-wide association studies for yield-related traits in soft red winter wheat grown in Virginia. *PLoS ONE* 14:e0208217. doi: 10.1371/journal.pone.0208217

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., et al. (2007). HMDB: the human metabolome database. *Nucleic Acids Res*. 35, D521–D526. doi: 10.1093/nar/gkl923

Wisser, R. J., Kolkman, J. M., Patzoldt, M. E., Holland, J. B., Yu, J., Krakowsky, M., et al. (2011). Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. *Proc. Natl. Acad. Sci. U.S.A*. 108, 7339–7344. doi: 10.1073/pnas.1017391108

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J., and Visscher, P. M. (2018). Common disease is more complex than implied by the core gene omnigenic model. *Cell* 173, 1573–1580. doi: 10.1016/j.cell.2018.05.051

Yang, Q., and Wang, Y. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *J. Probab. Stat*. 2012:652569. doi: 10.1155/2012/652569

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet*. 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (glycine max) germplasm. *BMC Genomics* 16:217. doi: 10.1186/s12864-015-1441-4

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* 11, 407–409. doi: 10.1038/nmeth.2848

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

frontiers in Genetics

METHODS
published: 18 January 2021
doi: 10.3389/fgene.2020.590012

A Multi-Locus Association Model Framework for Nested Association Mapping With Discriminating QTL Effects in Various Subpopulations

Suhong Bu[1,2], Weiren Wu[2] and Yuan-Ming Zhang[3]**

[1] College of Agriculture, South China Agricultural University, Guangzhou, China, [2] Key Laboratory of Genetics, Breeding and Multiple Utilization of Crops, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou, China, [3] Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China

Nested association mapping (NAM) has been an invaluable approach for plant genetics community and can dissect the genetic architecture of complex traits. As the most popular NAM analysis strategy, joint multifamily mapping can combine all information from diverse genetic backgrounds and increase population size. However, it is influenced by the genetic heterogeneity of quantitative trait locus (QTL) across various subpopulations. Multi-locus association mapping has been proven to be powerful in many cases of QTL mapping and genome-wide association studies. Therefore, we developed a multi-locus association model of multiple families in the NAM population, which could discriminate the effects of QTLs in all subpopulations. A series of simulations with a real maize NAM genomic data were implemented. The results demonstrated that the new method improves the statistical power in QTL detection and the accuracy in QTL effect estimation. The new approach, along with single-family linkage mapping, was used to identify QTLs for three flowering time traits in the maize NAM population. As a result, most QTLs detected in single family linkage mapping were identified by the new method. In addition, the new method also mapped some new QTLs with small effects, although their functions need to be identified in the future.

Keywords: nested association mapping (NAM), multi-locus association model, joint-family, subpopulation, maize

OPEN ACCESS

Edited by:
Hailan Liu,
Sichuan Agricultural University, China

Reviewed by:
Quan Long,
University of Calgary, Canada
Tomas Drgon,
United States Food and Drug Administration, United States

****Correspondence:***
Weiren Wu
wuwr@fafu.edu.cn
Yuan-Ming Zhang
soyzhang@mail.hzau.edu.cn

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

Received: *31 July 2020*
Accepted: *04 December 2020*
Published: *18 January 2021*

Citation:
Bu S, Wu W and Zhang Y-M (2021) A Multi-Locus Association Model Framework for Nested Association Mapping With Discriminating QTL Effects in Various Subpopulations. Front. Genet. 11:590012. doi: 10.3389/fgene.2020.590012

INTRODUCTION

Association mapping of large genetically diverse population has advantages over quantitative trait locus (QTL) mapping of biparental segregation population, such as the ability to access multiple gene alleles and higher mapping resolution (Zhang et al., 2005; Korte and Farlow, 2013). This is because the former carries more recombination breakpoints in history. However, the genetic structure of genome-wide association study (GWAS) population leads to high false positive rates (FPRs; Yu and Buckler, 2006). Moreover, low allele frequencies confer low statistical power (Rafalski, 2010). To address these issues, multiparental population or next-generation mapping populations, such as nested association mapping (NAM) and multiparent advanced generation intercross (MAGIC), were proposed (Cavanagh et al., 2008; Yu et al., 2008; Morrell et al., 2012). It was proved to have sufficient power and resolution to detect genomic associations for plant complex traits.

The NAM population was a special kind of multiparental panel, which was first proposed in maize (Yu et al., 2008). They crossed 25 representative lines with homozygous B73 line to generate 25 populations that consisted of 5,000 recombinant inbred lines (RILs; McMullen et al., 2009) and demonstrated that the NAM population method was powerful in dissecting the genetic architecture of complex traits, including flowering time, leaf architecture, stalk strength, and plant height (Buckler et al., 2009; Tian et al., 2011; Peiffer et al., 2013, 2014; Li et al., 2016). This initial success prompted the development of the NAM population in other crops, such as rice, wheat, barley, soybean, and sorghum (Maurer et al., 2015; Schmutzer et al., 2015; Bajgain et al., 2016; Bouchet et al., 2017; Fragoso et al., 2017; Song et al., 2017). Taking a wide view of all NAM methods applied in previous studies, they were prone to joint linkage across all subpopulations over single population mapping, as single population analysis has far less power and accuracy than joint mapping, although it will not position QTL inaccurately (Buckler et al., 2009). However, these approaches did not take into account the potential difference of QTL effects across families.

Genetic heterogeneity from different parents is likely to contribute to potential diversity of genetic architecture across subpopulations. Buckler et al. (2009) investigated the difference of allelic effects across different founder lines and demonstrated that the difference of QTL effects across subpopulations is related to latitudinal variation. Given that this diversity exists, the above methods, considering all QTL with same effects across all subpopulations, are not appropriate. To address this issue, we conducted a series of composite interval mapping (CIM; Zeng, 1994) for each RIL population in the maize NAM population. The results showed that QTLs detected in different subpopulations did not share either the same position or effect (**Supplementary Table 1**). For instance, different RIL populations might detect different QTLs; even if QTLs were detected across more than one population, these QTLs could rarely share the same effect. **Figure 1** shows an example of overlapped QTL. Within a distance of 10 cM, there were three QTLs identified in various subpopulations and having quite different effects. Because their peaks were very close, these QTLs were treated as an overlapped QTL. The results confirmed our suspicion. In association mapping in multiparental population, therefore, it is necessary to discriminate QTL effects in various subpopulations.

In this study, we proposed a speculation that QTL shared across multiple subpopulations of NAM has different effects in genetic mapping model. It was a specialty for the NAM design and also other similar multiple populations from multiple parents. A multi-locus association model was introduced to dissect the genetic basis of complex traits. In this kind of statistical model, variables involved are extremely colossal when single-nucleotide polymorphism (SNP) makers are numerous. Thus, we suggested a new matrix transform approach to address the problem of super-high dimensions. A series of Monte Carlo simulation experiments based on NAM marker data were performed to demonstrate the performance of this new method. Additionally, the validated approach was applied in genetic analysis for three flowering time traits in maize.

MATERIALS AND METHODS

NAM Population

We used the maize NAM population data (Buckler et al., 2009) from the Panzea website.[1] The NAM population consists of 4,699 RILs derived from the crosses between 25 diverse lines and the common parent B73. All the RILs from each cross were considered as a subpopulation. A total of 1,106 SNP markers were genotyped for each RIL, covering a genetic map of 1,400 cM and one marker every 1.3 cM on average. The best linear unbiased predictions (BLUPs) of three flowering time traits, including days to anthesis (DA, male flowering), days to silking (DS, female flowering), and anthesis-silking interval (ASI), were used as the phenotypic data in following analysis.

Genetic Model

Suppose that a general NAM design is as follows: k selected founder lines are crossed to a common parent, followed by selfing to generate k segregation F_2 populations, and each F_2 population are used to generate a half-sib subpopulation composed of n RILs by selfing for multiple generations. The phenotypic value of a quantitative trait may be described by the following model:

$$Y = \lambda\mu + \sum_{i=1}^{q}\sum_{j=1}^{k} X_{ij}\beta_{ij} + \varepsilon \quad (1)$$

where $Y = (y_1, y_2, \cdots y_{kn})'$; μ is a 25×1 matrix of covariant components; each element represents one subpopulation phenotype mean; λ is the $kn \times 25$ indicator matrix relating to each subpopulation; q is the number of QTL associated with interested trait; k is the number of sub-populations; and ε is the vector of residual error with a $N(0, \sigma^2)$ distribution. β_{ij} represents the additive effect of the ith QTL in the jth subpopulation. Namely, we gave k effects for one QTL across the k subpopulations. $X_{ij}$is a $kn \times 1$incidence vector of the ith QTL in the jth subpopulation. In this incidence vector, the n elements corresponding to the jth subpopulation are coded $(-1, 1)$, representing the genotype of SNP (AA and aa), and the other $(k-1)n$ elements are assigned 0, suggesting the absence of this QTL effect in other subpopulations. In multi-locus model, all available SNPs are considered as candidate QTL to be incorporated in the genetic model. Thus, the numerous variables in the model from huge number of SNPs and many subpopulations make a big burden for computing.

In order to relieve the computing burden, the dimensions of incidence matrix need to be reduced. Thus, we proposed a strategy to achieve dimension reduction and also make sure that the incidence matrix still involves different subpopulation information. In this method, k column original incidence vectors, corresponding to one QTL in all subpopulations, are emerged into one column vector. Here is the process: as for an SNP, we first calculate the main effect of each genotype in all subpopulations, respectively, $\omega_{ij} = \bar{y}_{ij} - \bar{y}$, where $i = 1, \cdots, k$, $j = 1, 2$ respects AA and aa. Thus, a vector ω, consisted of $2k$ indicators, is

[1] http://www.panzea.org

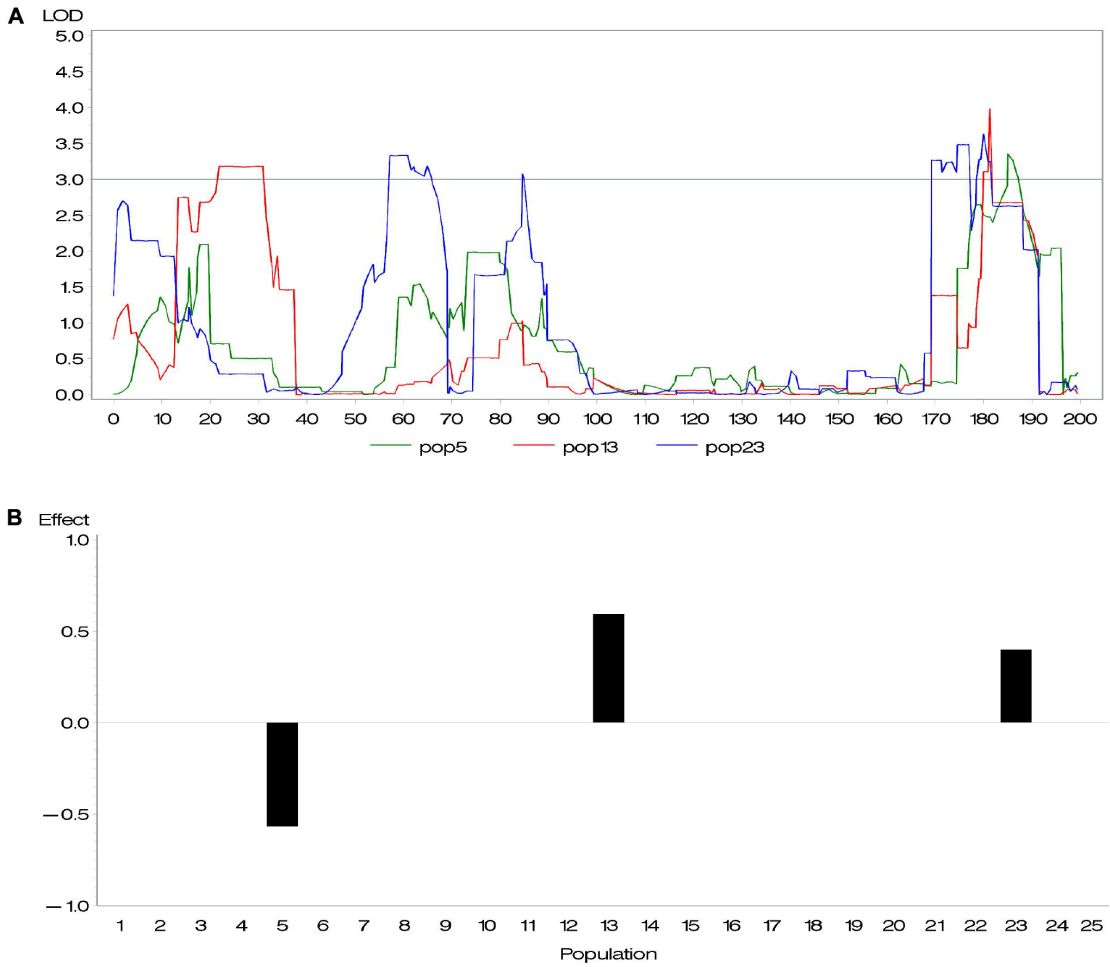

FIGURE 1 | (A) An overlapped quantitative trait locus (QTL) identified in subpopulations 5, 13, and 23 and **(B)** its effects.

obtained and then sorted. Next, we recode the genotypes across all subpopulations according to their effects' order and obtain a transformed incidence vector Z_i for a given QTL (SNP) (Lü et al., 2011). In addition, Z_i could be also transformed according to segmented ω. The genetic model is transformed as:

$$Y = \lambda\mu + \sum_{i=1}^{q} Z_i\gamma_i + \varepsilon \qquad (2)$$

where Z_i is a kn× 1 incidence vector of the ith QTL in all subpopulations, and γ_i is the corresponding effect of the ith putative QTL.

Multi-Locus Association Analysis

To select, estimate, and validate loci associated with interested trait, we proposed a multi-locus association of two-stage processes. Based on the genetic model (2), genome SNPs scanning needs to further select, estimate, and validate SNPs associated with given trait. We proposed the following two-stage selection process to screen. In the first stage, shrinkage estimate algorithm was used to estimate the additive effect of SNPs, and all SNPs with $t_i = |\hat{\gamma}_j/\hat{\sigma}_j| > 10^{-4}$ are picked up. Considering stability, effectiveness, and computing time, we adopted the empirical Bayes (E-Bayes) method (Xu, 2010). Compared with other shrinkage estimation (Zhang and Xu, 2005; Yi and Banerjee, 2009; Feng et al., 2013), E-Bayes provides a more robust shrinkage that the large effect subjects are shrunk to virtually no shrinkage while small effects to zero, so that nonsignificant SNP is estimated toward zero. Simulation studies showed that the E-Bayes is predominant compared with other shrinkage estimation methods in terms of small mean squares error (Xu, 2010). For the technical details of the method, refer to the original study by Xu (2010). The method is briefly described here.

The parameters β and σ^2 are always included in the model; the uniform prior is assigned to the two parameters: $P(\beta) \propto 1$ and $P(\sigma^2) \propto 1$ (Zhang and Xu, 2005). We adopt the normal prior for each of the genetic effects (γ_k) in model (2): $P(\gamma_k) \propto N(0, \sigma_k^2)$. The scaled inverse χ^2 prior distribution is further assigned to σ_k^2: $P(\sigma_k^2) = \text{Inv}{-}\chi^2(\sigma_k^2|\tau, \omega) \propto \left(\sigma_k^2\right)^{-\frac{\tau+2}{2}} \exp\left(-\omega/2\sigma_k^2\right)$ (Xu, 2010). Clearly, Y in model (2) follows a multivariate

normal distribution with mean $\mu = X\beta$ and variance–covariance $V = \sum_k Z_k Z_k^T \sigma_k^2 + I\sigma^2$. Let $\theta = (\beta, \gamma, \sigma^2)$. Therefore, the main steps for parameter estimation are described as below.

Step (0): Let $\xi = (\tau, \omega) = (0, 0)$, $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$, $\hat{\sigma}^2 = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right)/n$, and γ_k and σ_k^2 were initialized $(k = 1, 2, \cdots, 2m^2)$;

Steps (1): Using $E(\gamma_k) = \sigma_k^2 Z_k^T V^{-1}(y - X\beta)$ and $var(\gamma_k) = I\sigma_k^2 - \sigma_k^2 Z_k^T V^{-1} Z_k \sigma_k^2$, $E\left(\gamma_k^T \gamma_k\right)$ was estimated by $E\left(\gamma_k^T\right) E(\gamma_k) + tr[var(\gamma_k)]$. This is the E-step;

Step (2): update β, σ^2 and σ_k^2: $\sigma_k^2 = \left[E\left(\gamma_k^T \gamma_k\right) + \omega\right] / (\tau + 2 + 1)$,$\beta = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y$, and $\sigma^2 = \left(y - X\beta\right)^T \left[y - X\beta - \sum_{k=1}^{m} Z_k E(\gamma_k)\right] \Big/ n$. This is the M-step;

Step (3): Repeat the E-step and the M-step until convergence is reached.

After the reduction in dimension in the first stage, maximum likelihood method could be used to reanalyze the reduced model and perform the likelihood ratio test (LRT) in the second stage. LRT was aimed to decide the inclusion and retention of a SNP in the model based on LR score:

$$LR_j = -2\ln[L(\theta_{-j}) / L(\theta)]$$

where θ is the parameter vector in the reduced genetic model; θ_{-j} is the parameter vector in θ excluding the currently tested genetic effect $\hat{\gamma}$. $L(\theta)$ and $L(\theta_{-j})$ are the maximum likelihood function for θ and θ_{-j}, respectively. If LR_jexceeds one given threshold, then it indicates that this SNP could significantly improve model fit. For simplicity, we suggested an alternative statistical parameter $LOD = LR_j / 4.61$ and 3.0 as the critical value in our association mapping process.

Monte Carlo Simulation Design

For ease of computation, only few subpopulation data from the maize NAM population including 100 SNP markers from chromosome 1 were used to perform the simulation experiments. The length of chromosome segment was 153.4 cM. We investigated four simulation scenarios, and each simulation had 10 assumed QTLs locating at the given chromosome segment evenly. All the QTL were overlapped with the markers and listed in **Supplementary Table 2**.

In the first scenario, the effect of QTL heritability on the new method was assessed in five populations with 964 RILs. We assumed 10 QTLs in each of three simulations. The size (or heritability, h_i^2) of each QTL, the proportion of total phenotypic variance explained by the QTL, was all set to 0.03 in the first simulation, 0.05 in the second simulation, and 0.08 in the third simulation. We supposed that each of 10 QTLs had different fixed effects α_i (i = 1, 2, …, 5) among the five populations, and $\sum_i \alpha_i = 0$. The breeding value of each RIL i from population k was calculated as$a_{ki} = \sum_j X_j \alpha_{kij}$ (j = 1, 2, …, 10), and the phenotypic value $y_{ki} = a_{ki} + e_{ki}$,$e_{ki}$was a residual effect sampled from a normal distribution with mean 0 and variance $\sigma_e^2 = 1$. The additive genetic variance of the ith QTL,σ_{ai}^2, was calculated from $\sigma_{ai}^2 = h_i^2 \sigma_e^2 / (1 - \sum h_i^2)$. Then, the QTL effects within a given populations were calculated by relating σ_{ai}^2 to the allelic frequencies and effects.

In the second scenario, we evaluated the effect of sample size on the new method by setting the sample size as 400 (four subpopulations each with 100 RILs), 600 (six subpopulations each with 100 RILs), and 800 (eight subpopulations each with 100 RILs). Each QTL size was set as 0.07. Other parameters were the same as those in the first scenario.

In the third scenario, we explored the feasibility of a new method on random-effect QTLs. Ten assumed QTLs have the same positions with those in the former two scenarios. The first five QTLs shared a fixed effect (1.5) across all subpopulations. For the jth of the latter five QTLs, five effects α_{ij}(i = 1, 2, . . ., 5) were randomly sampled from multivariate normal distribution with mean 1.5 and variance–covariance structure

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} .\rho$$

was the correlation of QTL effects between two nearest populations and set with two levels ($\rho = 0.2$ and $\rho = 0.8$). The proportion of nongenetic variance σ_{ek}^2 to total additive genetic variance σ_{ak}^2 in population k was related to a magnitude of heritability for a trait. In this scenario, the total heritability was set to 0.6.

RESULTS

Mapping QTLs for DA, DS, and ASI in Single Maize NAM Subpopulation via the CIM Method

We performed CIM mapping, implemented by Windows QTL Cartographer V2.5,[2] for DA, DS, and ASI in each NAM subpopulation. For DA in maize, approximately five to six QTLs were detected in each subpopulation. A total of 137 QTLs were identified, with a LOD threshold of 3. Among the 137 QTLs, 10 QTLs clusters (defined with more than five QTLs within a 30 cM interval) were dispersed across 10 chromosomes (**Supplementary Table 1** and **Supplementary Figure 1**). We found 28 overlapped loci (where more than two QTLs from various subpopulations totally or partially overlapped), whereas no same QTL was found across all 25 subpopulations. For most of those overlapped loci, one QTL contributed different effects in different subpopulations. **Figure 1** gave an example of an overlapped QTL. Three subpopulations (5, 13, and 23) detected one QTL in a

[2]https://brcwebportal.cos.ncsu.edu/qtlcart/WQTLCart.htm

small 177–189 cM interval on chromosome 1 (**Figure 1A**), where their effects in the three subpopulations are -0.57, 0.40, and 0.60, respectively (**Figure 1B**). Yet, there were few overlapped QTLs with similar effects across various subpopulations (**Supplementary Table 1**). In addition, we found a relatively large proportion of total phenotypic variance explained by all the QTLs, such as 66.4% for DA, 74.6% for DS, and 94.4% for ASI.

Simulation Results

Effect of QTL Size on Mapping QTL

In the first simulation experiment, the effect of QTL size on mapping QTL in the maize NAM population was evaluated. QTL size was set as 3, 5, and 8%. Ten assumed QTLs were uniformly distributed across the genome in the three cases. Each sample was analyzed by the new method, and the results are shown in **Figure 2** and **Supplementary Table 2.1**. The average power for 10 assumed QTLs in each case was 59.2, 81, and 91.1% for the QTL sizes of 3, 5, and 8%, respectively, indicating the increase in average power of all the 10 assumed QTLs with the increase of QTL size (**Figure 2A**). The FPR was less in both 5 and 8% cases than in 3% case (**Figure 2B**). The bias of QTL position estimate was relatively low, and it had a negative correlation with QTL size (**Figure 2C**). Besides, **Figure 2D** shows a relatively small bias (−0.068 to 0.040) between estimated and assumed effects for each QTL in three simulation cases.

Effect of Sample Size on Mapping QTL

In the second simulation, we investigated the effect of sample size on mapping QTL. The sample sizes were set as 400, 600, and 800 (k subpopulations each with 100 RILs), all the QTL sizes were set as 0.07, and other parameters were the same as those in the first simulation. The results are shown in **Figure 3** and **Supplementary Table 2.2**. The results indicated the increase in statistical power in QTL detection and accuracy in QTL position estimation with the increase in sample size. FPR still stays on a low level (<2.1%). The effect estimates in this simulation showed more bias than those in the first simulation. This is possibly caused by smaller sample (400, 600, and 800) than those in the first simulation (964).

Random Effect Simulation

We also conducted a simulation experiment to investigate how fixed or random effect of QTL would influence our association mapping. A fixed effect was assigned to the first five QTLs, and there were no differences for these QTLs across various subpopulations, while random effects were assigned to the last five QTLs, and there were various values of ρ across various subpopulations. As a result, no significant difference between fixed and random effects in fixed ρ value was observed (**Figure 4** and **Supplementary Table 2.3**), although their powers were more than 80%. Meanwhile, no significant difference among various ρ values in the same setup (fixed or random) of QTL effect was observed. FPR and the bias of QTL position and effect stayed a quite low level.

Mapping QTLs for DA, DS, and ASI in Joint Maize NAM Subpopulations

The new method was used to identify QTLs for three flowering time traits in the joint maize NAM subpopulations. As a result, 77, 79, and 75 QTLs were identified, and these QTLs accounted for 90.11, 89.44, and 82.50% of the total phenotypic variances for the above three traits, respectively. Most QTLs detected by the CIM method in the single-maize NAM subpopulation were also identified by the new method in the joint maize NAM subpopulations. As for DA, the new results covered 127 of 137 QTLs from the CIM method (**Supplementary Figure 1**), including all the seven extremely large QTLs ($r^2 > 15\%$, light blue), 24 of 25 large QTLs ($10\% < r^2 < 15\%$, deep green), 56 of 59 relative large QTLs ($5\% < r^2 < 10\%$, deep blue), and 40 of 46 small QTLs ($2\% < r^2 < 5\%$, pink) (**Supplementary Figure 1**). As for DS, 132 of 138 QTLs from the CIM method were covered by our new method, including all the five extremely large QTLs, 21 of 23 large QTLs, 71 of 75 relative large QTLs, and all the 35 small QTLs. As for ASI, 81 of 89 QTLs from the CIM method were found by the new method, including 21 of 22 large QTLs, 55 of 62 relative large QTLs, and all the 5 small QTLs. Clearly, the above results validated our new method.

As compared with the CIM results, we detected 25 additional QTLs for DA, 29 for DS, and 32 for ASI (**Supplementary Table 3**). The genetic variances of all additional QTLs were quite small. Most QTL for DA and DS accounted for < 1% phenotypic variance by a single QTL, although 8 of 32 QTLs for ASI accounted for more than 1% by a single QTL, and 5 QTLs accounted for more than 3% by a single QTL. This indicated that our new method had a high power for detecting minor alleles.

To validate these additional QTLs, we mined candidate genes around the above additional QTLs via phytozome v9.1.[3] All the additional QTLs were found to be very close to their candidate genes, and these candidate genes were listed in **Supplementary Table 3**. For example, 19 of 25 candidate genes for DA, as well as 21 of 29 candidate genes for DS, were found to be within the distance of 1 kb from their associated SNPs. Among candidate genes for ASI, 23 of 31 genes were within 1 kb, and only two genes were found to be within >5 kb. The close distance indicated a strong linkage between associated SNPs and their candidate genes. Some evidence for candidate genes were described as below (**Supplementary Table 3**). GRMZM2G154896, near the SNP PZA00368.1 associated with DA, is a pollen tube developmental gene; GRMZM2G177151, near the SNP associated with DS, is C2H2-type zinc finger protein gene, which may play an important role in spike development; and GRMZM2G061900, near the SNP PZA00276.18 associated with ASI, is Ras protein gene that affects cell growth, differentiation, cytoskeleton, protein transport, and secretion.

DISCUSSION

Compared with QTL mapping in biparental segregation population, multiparental population could provide high power

[3]http://www.phytozome.net/

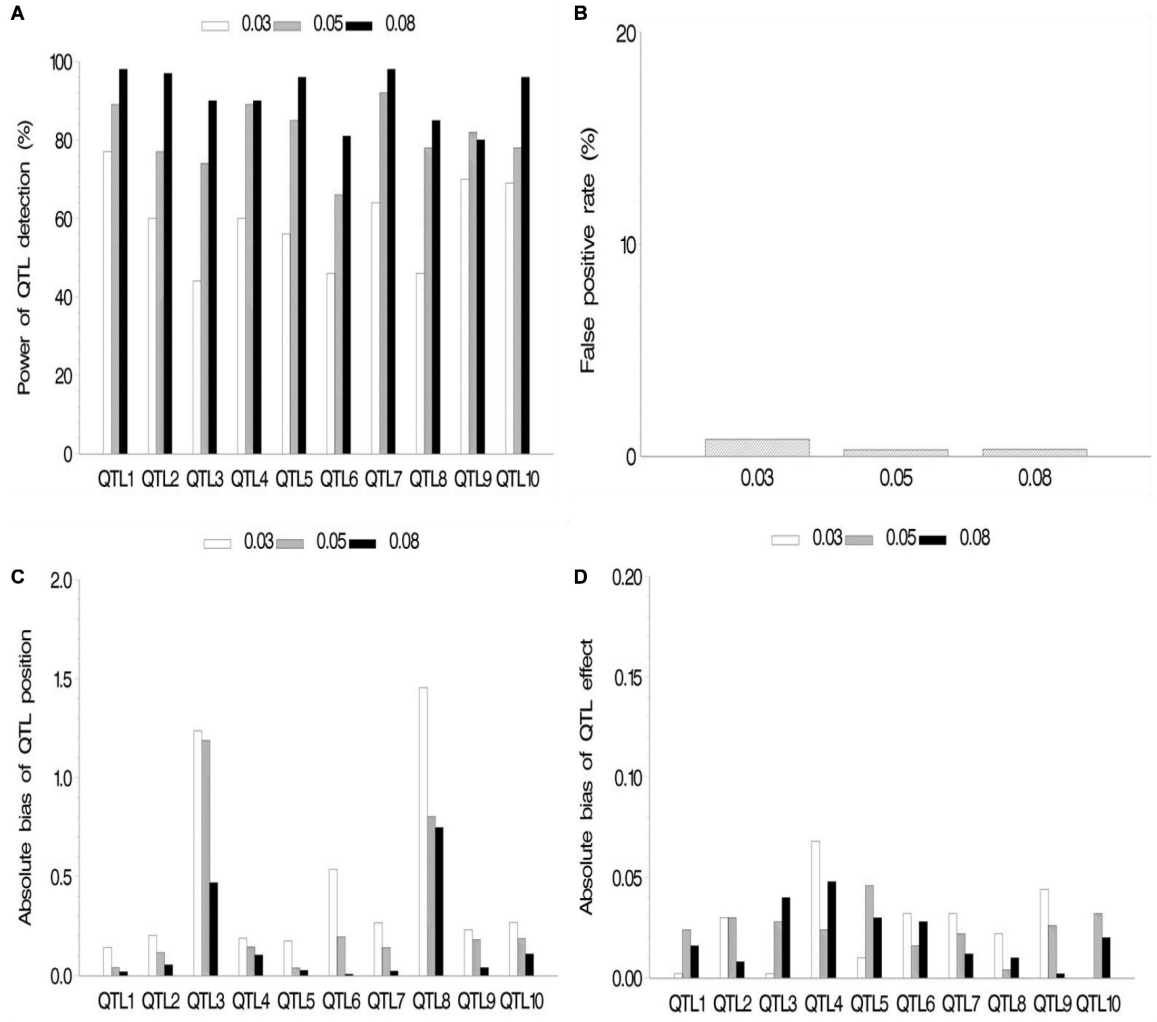

FIGURE 2 | Effect of quantitative trait locus (QTL) heritability on the new method. **(A)** Power of QTL detection, **(B)** false positive rate, **(C)** average of absolute bias between estimated and true positions, and **(D)** average of absolute bias between estimated and true effects.

and resolution for association mapping in the genetic dissection of complex traits. This is because the association mapping population has more historical recombination events and high linkage disequilibrium (LD), which can increase allelic diversity and mapping resolution. However, conventional association mapping is always confounded by population structure between diverse lines (Flint-Garcia et al., 2005; Yu et al., 2006). The NAM design promised to address these weaknesses and utilize the advantages of linkage and association mapping (Yu et al., 2008). Therefore, it is necessary to propose an optimal approach in the genetic analysis of complex traits in the NAM population.

In this study, we found that genetic heterogeneity was a common factor in the NAM population, which would confound the results of association mapping. Thus, we proposed a multi-locus association model for mapping QTL of complex trait in the NAM population. This model could discriminate the QTL effects across various subpopulations, which addressed the problem of genetic heterogeneity across subpopulations. Because of "p≫ n" in the new model, we proposed a matrix transform approach to shrink the information of independence indicator variables. A multi-locus mapping method, involving with E-Bayes (Xu and Jia, 2007: Xu, 2010) and LRT, were proposed in this study.

In genetic analysis of the NAM population, jointing all families as mapping population is more common than using a single family, such as joint linkage mapping (JLM; Buckler et al., 2009; Tian et al., 2011), JICIM (Li et al., 2011), NAM (Xavier et al., 2015), and GWAS with mixed linear model (Chen et al., 2019). Because it had higher mean prediction ability and performed better at more stringent significance threshold. However, Li et al. (2011) observed that joint multifamily analysis has less power and worse resolution than single family for rare QTL, which is identified in only one or few subpopulations. Ogut et al. (2015) showed that most robust QTLs were restricted to one family and were often not detected at high frequency by joint family analysis. In this study, we found that most rare

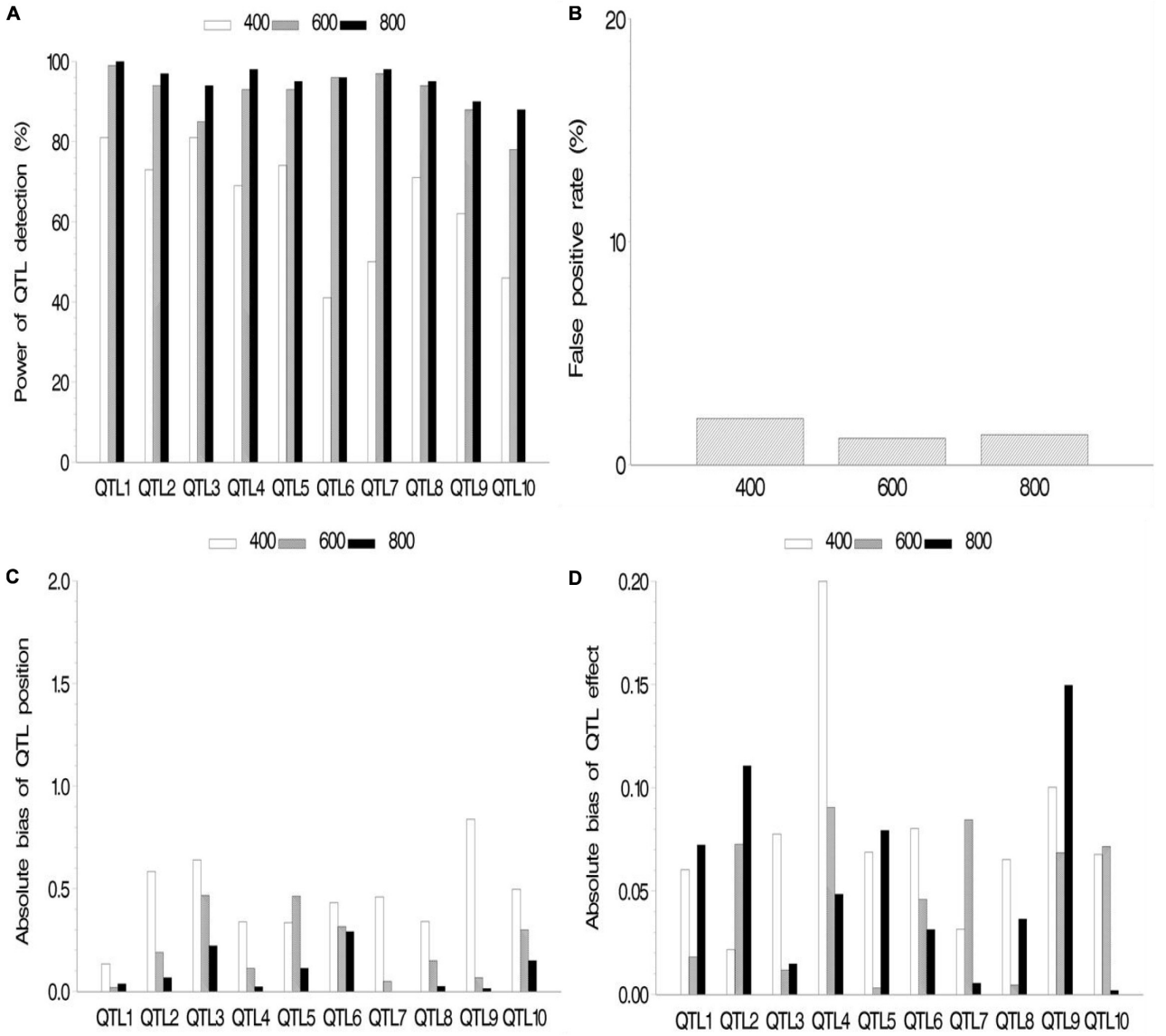

FIGURE 3 | Effect of sample size on the new method. **(A)** Power of quantitative trait locus (QTL) detection, **(B)** false positive rate, **(C)** average of absolute bias between estimated and true positions, and **(D)** average of absolute bias between estimated and true effects.

QTLs with large effect can be detected by our new method. For three flowering time traits, we can detect more than 90% of QTLs from the CIM method in the single NAM subpopulation. Besides, the new method can identify more small-effect QTLs than the CIM method.

In order to compare single family analysis (SF) with our new method, we conducted a series of simulations (more details about the simulation and results, see Supporting Information S4). In the simulation, 10 QTLs with five types of effects were assumed across five subpopulations. Stepwise regression was used for SF analysis, described by Buckler et al. (2009). The results showed that SF stepwise regression was powerful for large-effect QTLs rather than small-effect QTLs in the single-family NAM subpopulation. Because there are much less lines in the single-family NAM subpopulation than in the joint multifamily NAM subpopulations, enough precision for QTL detection cannot provided. However, our new method with multiple families had good power, not only for large- and small-effect QTLs but also for common and rare QTLs. On the one hand, joint multiple families increased population size (usually more than 20 times according to the NAM design) (Li et al., 2011). On the other hand, the new NAM model could discriminate QTL effects across various subpopulations, which controlled false positive signals from sample variance in nonrelated families. In addition, multi-locus GWAS methods are more powerful and robust in the detection of small-effect QTNs (Wang et al., 2016; Su et al., 2018; Wen et al., 2018; Zhang et al., 2019).

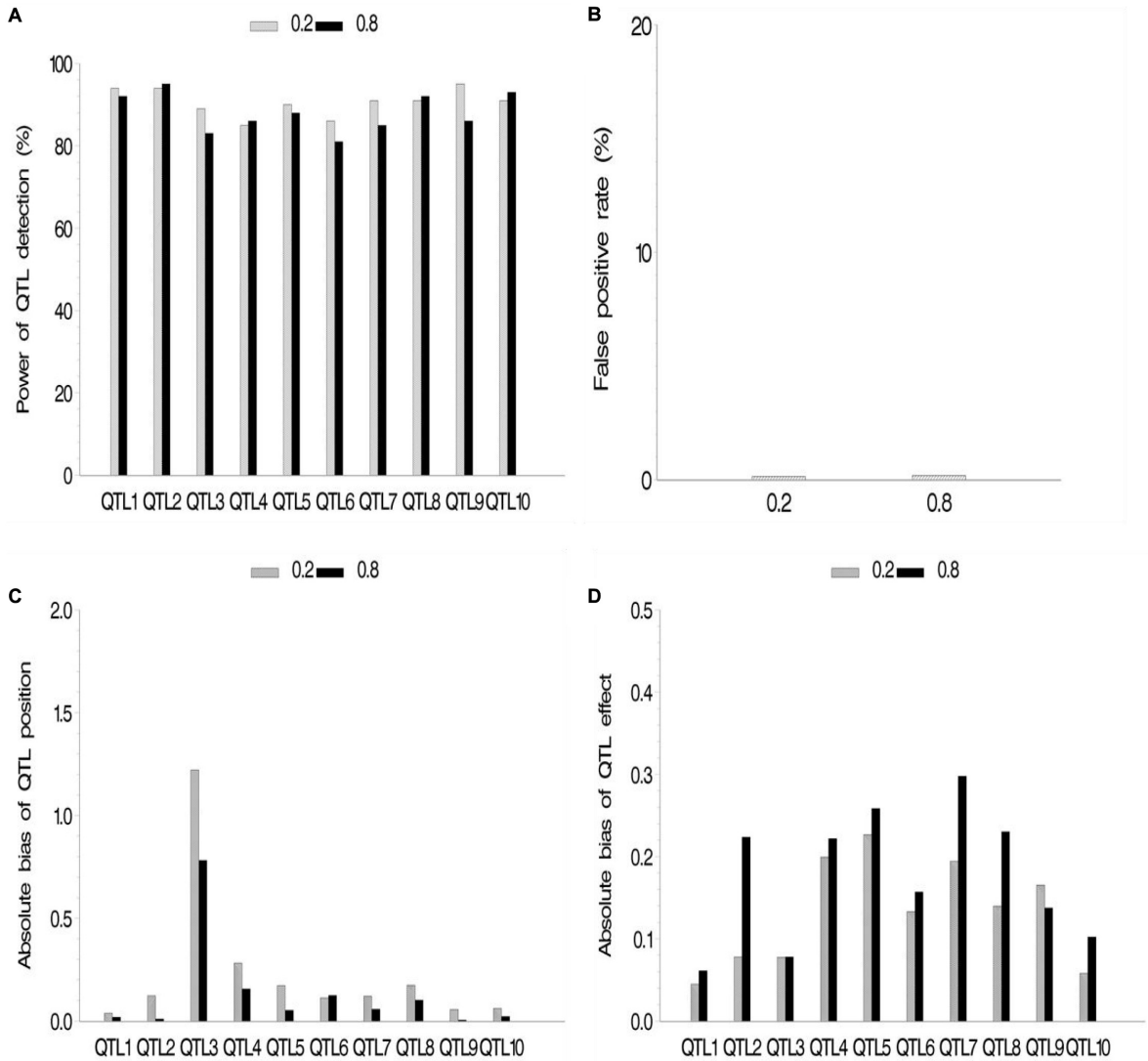

FIGURE 4 | Effect of quantitative trait locus (QTL) type on the new method. **(A)** Power of QTL detection, **(B)** false positive rate, **(C)** average of absolute bias between estimated and true positions, and **(D)** average of absolute bias between estimated and true effects.

Some GWAS software packages are available in the NAM population, such as Trait Analysis by Association, Evolution, and Linkage (TASSEL; Chen et al., 2019) and Jawamix5 (Long et al., 2013). With TASSEL, MLM method can capture the population structure and genetic relatedness of all the lines in the NAM population by Q and K matrices. Jawamix5 also provides a fast GWAS tool in structured populations using the mixed model, as well as stepwise regression in NAM design (Long et al., 2013). These GWAS software packages are very powerful in normal GWAS. However, they were not designed for the NAM population and did not involve the genetic heterogeneity. We have proved that genetic heterogeneity from parents contributed to the diverse effects of a QTL in different families (**Supplementary Table 1** and **Figure 1**). Therefore, the proper mapping methods are important, especially for the NAM population.

Joint linkage mapping (Buckler et al., 2009) and JICIM (Li et al., 2011) used the stepwise linear regression and linkage mapping to select marker effects nested within families and estimate QTL effects. It might lead to missing some large-effect QTLs identified only in one subpopulation (Ogut et al., 2015). In the NAM software (Xavier et al., 2015), a mixed linear model framework with EMMA algorithm (Kang et al., 2008) was used to map associated SNPs in multiparent population, such as the MAGIC population. Recently, this software was also used to detect QTLs in the NAM population (Sunil et al., 2020). In the genetic model of NAM software, the dimension will inflate to $k+1$ times (k families in NAM design), although marker effects can be estimated.

Our new method was designed for association mapping in the NAM population. Based on Monte Carlo

simulation experiments and real data analysis, some minor QTLs can be identified by the new method, indicating high QTL signal to noise ratio in the NAM mapping population. The new method gave a dimension reduction via matrix transformation, which can maintain the family information in the genetic model and reduce computational burden. Actually, this approach could be applied in genome-wide association studies (Lü et al., 2011). In this study, the new method was validated in the NAM mapping population. However, it is also suitable for MAGIC population, which is a large RIL population derived from multiple parents (Cavanagh et al., 2008). Thus, the new method is useful in genetic mating design.

DATA AVAILABILITY STATEMENT

The author selected the following statement: Publicly available datasets were analyzed in this study. This data can be found here: http://www.panzea.org (Buckler_etal_2009_Science_flowering_time_data-090807.zip).

AUTHOR CONTRIBUTIONS

SB performed the experiments, analyzed the data, and drafted the manuscript. WW and Y-MZ discussed the results. All authors designed and conceived the experiments.

FUNDING

This work was supported by the National Natural Science Foundation of China (31601375 and U1602261).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.590012/full#supplementary-material

Supplementary Figure 1 | Comparison of the new method in joint multi-family NAM sub-populations with the CIM method in single family for days to anthesis (DA).

REFERENCES

Bajgain, P., Rouse, M. N., Tsilo, T. J., Macharia, G. K., Bhavani, S., Jin, Y., et al. (2016). Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. *PLoS One* 11:e0155760. doi: 10.1371/journal.pone.0155760

Bouchet, S., Olatoye, M. O., Marla, S. R., Perumal, R., Tesso, T., Yu, J., et al. (2017). Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. *Genetics* 206, 573–585. doi: 10.1534/genetics.116.198499

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., et al. (2009). The genetic architecture of maize flowering time. *Science* 325, 714–718.

Cavanagh, C., Morell, M., Mackay, I., and Powel, L. W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr. Opin. Plant Biol.* 11, 215–221. doi: 10.1016/j.pbi.2008.01.002

Chen, Q., Yang, C. J., York, A. M., Xue, W., and Doebley, J. F. (2019). TeoNAM: a nested association mapping population for domestication and agronomic trait analysis in maize. *Genetics* 213, 1065–1078. doi: 10.1534/genetics.119.302594

Feng, J. Y., Zhang, J., Zhang, W. J., Wang, S. B., Han, S. F., and Zhang, Y. M. (2013). An efficient hierarchical generalized linear mixed model for mapping QTL of ordinal traits in crop cultivars. *PLoS One* 8:e59541. doi: 10.1371/journal.pone.0059541

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., et al. (2005). Maize association population: a high- resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064. doi: 10.1111/j.1365-313x.2005.02591.x

Fragoso, C. A., Moreno, M., Wang, Z., Heffelfinger, C., Arbelaez, L. J., Aguirre, J. A., et al. (2017). Genetic architecture of a rice nested association mapping population. *G3 Genes Genomes Genet.* 7, 1913–1926. doi: 10.1534/g3.117.041608

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

Li, H., Bradbury, P., Ersoz, E., Buckler, E. S., and Wang, J. (2011). Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. *PLoS One* 6:e17573. doi: 10.1371/journal.pone.0017573

Li, Y., Li, C., Bradbury, P. J., Liu, X., Lu, F., Romay, C. M., et al. (2016). Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. *Plant J. Cell Mol. Biol.* 2016, 391–402. doi: 10.1111/tpj.13174

Long, Q., Zhang, Q. R., Vilhjalmsson, B. J., Forai, P., Seren, U., and Nordborg, M. (2013). JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models. *Bioinformatics* 29, 1220–1222. doi: 10.1093/bioinformatics/btt122

Lü, H. Y., Liu, X. F., Wei, S. P., and Zhang, Y. M. (2011). Epistatic association mapping in homozygous crop cultivars. *PLoS One* 6:e17773. doi: 10.1371/journal.pone.0017773

Maurer, A., Draba, V., Jiang, Y., Schnaithmann, F., Sharma, R., Schumann, E., et al. (2015). Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16:290. doi: 10.1186/s12864-015-1459-7

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., et al. (2009). Supporting online material for: genetic properties of the maize nested association mapping population. *Science* 325, 737–741.

Morrell, P. L., Buckler, E. S., and Ross-Ibarra, J. (2012). Crop genomics: advances and applications. *Nat. Rev. Genet.* 13, 85–96.

Ogut, F., Bian, Y., Bradbury, P. J., and Holland, J. B. (2015). Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity* 114, 552–563. doi: 10.1038/hdy.2014.123

Peiffer, J. A., Flint-Garcia, S. A., De Leon, N., McMullen, M. D., Kaeppler, S. M., and Buckler, E. S. (2013). The genetic architecture of maize stalk strength. *PLoS One* 8:e67066. doi: 10.1371/journal.pone.0067066

Peiffer, J. A., Romay, M. C., Gore, M. A., Flint-Garcia, S. A., Zhang, Z., Millard, M. J., et al. (2014). The genetic architecture of maize height. *Genetics* 196, 1337–1356.

Rafalski, J. A. (2010). Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13, 174–180. doi: 10.1016/j.pbi.2009.12.004

Schmutzer, T., Samans, B., Dyrszka, E., Ulpinnis, C., Weise, S., Stengel, D., et al. (2015). Species-wide genome sequence and nucleotide polymorphisms from the model allopolyploid plant *Brassica napus*. *Sci. Data* 2:150072.

Song, Q., Yan, L., Quigley, C., Jordan, B. D., Fickus, E., Schroeder, S., et al. (2017). Genetic characterization of the soybean nested association mapping population. *Plant Genome* 10, 1–14.

Su, J. J., Ma, Q., Li, M., Hao, F. S., and Wang, C. X. (2018). Multi-locus genome-wide association studies of fiber-quality related traits in chinese early-maturity upland cotton. *Front. Plant Sci.* 9:1169. doi: 10.3389/fpls.2018.01169

Sunil, S. G., Wang, H., Yaduru, S., Pandey, M. K., Fountain, J. C., Chu, Y., et al. (2020). Nested-association mapping (NAM)-based genetic dissection uncovers candidate genes for seed and pod weights in peanut (*Arachis hypogaea*). *Plant Biotechnol. J.* 18, 1457–1471. doi: 10.1111/pbi.13311

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, 159–162. doi: 10.1038/ng.746

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444.

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Xavier, A., Xu, S., Muir, W. M., and Rainey, K. M. (2015). NAM: association studies in multiple populations. *Bioinformatics* 2015, 1–3.

Xu, S. (2010). An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* 105, 483–494. doi: 10.1038/hdy.2009.180

Xu, S., and Jia, Z. (2007). Genome-wide analysis of epistatic effects for quantitative traits in barley. *Genetics* 175, 1955–1963. doi: 10.1534/genetics.106.066571

Yi, N., and Banerjee, S. (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics* 181, 1101–1113. doi: 10.1534/genetics.108.099556

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* 17, 155–160. doi: 10.1016/j.copbio.2006.02.003

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics* 178, 539–551. doi: 10.1534/genetics.107.074245

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136, 1457–1468.

Zhang, Y. M., Jia, Z. Y., and Dunwell, J. M. (2019). Editorial: the applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 9:100. doi: 10.3389/fpls.2019.00100

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, Y. M., and Xu, S. (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* 95, 96–104. doi: 10.1038/sj.hdy.6800702

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

frontiers in Genetics

SYSTEMATIC REVIEW
published: 22 January 2021
doi: 10.3389/fgene.2020.587887

A Review of Statistical Methods for Identifying Trait-Relevant Tissues and Cell Types

Huanhuan Zhu[1†‡], *Lulu Shang*[1†] *and Xiang Zhou*[1,2*]

[1] *Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States,* [2] *Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, United States*

Genome-wide association studies (GWASs) have identified and replicated many genetic variants that are associated with diseases and disease-related complex traits. However, the biological mechanisms underlying these identified associations remain largely elusive. Exploring the biological mechanisms underlying these associations requires identifying trait-relevant tissues and cell types, as genetic variants likely influence complex traits in a tissue- and cell type-specific manner. Recently, several statistical methods have been developed to integrate genomic data with GWASs for identifying trait-relevant tissues and cell types. These methods often rely on different genomic information and use different statistical models for trait-tissue relevance inference. Here, we present a comprehensive technical review to summarize ten existing methods for trait-tissue relevance inference. These methods make use of different genomic information that include functional annotation information, expression quantitative trait loci information, genetically regulated gene expression information, as well as gene co-expression network information. These methods also use different statistical models that range from linear mixed models to covariance network models. We hope that this review can serve as a useful reference both for methodologists who develop methods and for applied analysts who apply these methods for identifying trait relevant tissues and cell types.

Keywords: trait-tissue relevance, epigenetic information, transcriptomic information, genetically regulated gene expression, gene co-expression network, eQTL information

OPEN ACCESS

Edited by:
Hailan Liu,
Sichuan Agricultural University, China

Reviewed by:
Qunfeng Dong,
Loyola University Chicago, United States
Guo-Bo Chen,
Zhejiang Provincial People's Hospital, China

***Correspondence:**
Xiang Zhou
xzhousph@umich.edu

†These authors have contributed equally to this work

‡Present address:
Huanhuan Zhu,
BGI-Shenzhen, Shenzhen, China

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

Received: *27 July 2020*
Accepted: *30 December 2020*
Published: *22 January 2021*

Citation:
Zhu H, Shang L and Zhou X (2021) A Review of Statistical Methods for Identifying Trait-Relevant Tissues and Cell Types. Front. Genet. 11:587887. doi: 10.3389/fgene.2020.587887

INTRODUCTION

Over the last one and half decades, genome-wide association studies (GWASs) have successfully identified and replicated many trait-relevant genetic variants in terms of single nucleotide polymorphisms (SNPs). However, most of these identified genetic variants reside outside protein-coding regions, making it challenging to understand the biological mechanism underlying these identified associations (Welter et al., 2014). Characterizing the biological mechanism underlying SNP associations is further complicated by the fact that the genetic effects of SNPs on complex traits are likely acted through a tissue-specific fashion. For example, many psychiatric disorders, such as bipolar disorder and schizophrenia, are consequences of dysfunctions of various genes, pathways, and regulatory elements in neuronal and glia cells, resulting from brain-specific genetic effects of polymorphisms (Lang et al., 2007; Uhlhaas and Singer, 2010; Fornito et al., 2015; Grunze, 2015; Xiao et al., 2017). Therefore, characterizing the function of variants in various brain tissues can help elucidate etiology of psychiatric disorders. However, for most complex traits, their

trait-relevant tissues and cell types are often unknown or uncertain. As a result, identifying trait-relevant tissues and cell types and characterizing the functions of genetic variants within the relevant tissues and cell types hold the key for better understanding of disease etiology and the genetic basis of phenotypic variation (Trynka et al., 2013, 2015; Kichaev et al., 2014; Pickrell, 2014; Farh et al., 2015; Finucane et al., 2015; Li and Kellis, 2016).

Many genomic studies have been carried out in parallel to GWASs to characterize the genetic and epigenetic landscape of the human genome. These genomic studies often collect samples from multiple different tissues or cell types and characterize genomic information in a tissue- or cell type-specific fashion. For example, the ENCODE (The ENCODE Project Consortium, 2012) and Roadmap (Kundaje et al., 2015) collect various epigenetic annotation measurements in the form of open chromatin accessibility, DNase I hypersensitive sites (DHSs), and histone modifications (e.g., H3K27me3 and H3K36me3) on 16 cell lines and 111 tissues. The epigenetic information measured from these projects allows for a functional characterization of the human genome. As another example, the GTEx project collects gene expression and genotype measurements from 54 human tissues on nearly 1,000 individuals using whole-genome sequencing, whole-exome sequencing, and bulk RNA sequencing (RNA-seq) (GTEx Consortium, 2015). By paring gene expression information with genotype information, GTEx allows for the study of tissue-specific gene expression and its genetic basis in the form of expression quantitative trait loci (eQTLs) mapping. Similarly, the CommonMind project collects gene expression, open chromatin accessibility and genotype information in the dorsolateral prefrontal cortex from up to 452 patients with schizophrenia and bipolar disorder as well as healthy controls (Fromer et al., 2016). Characterizing the cortex-specific transcriptomic and epigenetic profile in CommonMind can facilitate the investigation of the molecular mechanism underlying neuropsychiatric diseases. In addition, various single cell RNA-seq (scRNA-seq) studies are being performed to collect cell type-specific gene expression measurements on tens of thousands of cells from various tissues and organs (Bacher and Kendziorski, 2016). Such cell type-specific expression profiles can be used to understand how specific cell types may underlie complex traits (Watanabe et al., 2019). Finally, existing bulk and single cell gene expression studies also facilitate the characterization of gene co-expression pattern in a tissue- or cell type-specific fashion (GTEx Consortium, 2015; Bacher and Kendziorski, 2016; Shang et al., 2020b). Tissue- or cell type-specific gene co-expression provides invaluable information on the tissue or cell type basis of disease etiology (Shang et al., 2020b). Overall, various genomic studies have provided tissue- or cell type-specific information for inferring trait-relevant tissues and cell types.

With the increasing availability of different tissue- and cell type-specific genomic datasets, many statistical methods have been recently developed to integrate these genomic data with GWASs for identifying trait-relevant tissues and cell types. These various integrative methods differ in terms of the underlying statistical model and the particular genomic information they make use of. For example, the sLDSC (stratified LD score regression) converts tissue-specific epigenetic measurements into tissue-specific SNP functional annotations and estimates to what extent different tissue-specific functional annotations explain trait heritability (Finucane et al., 2015). The inferred SNP heritability due to tissue-specific annotation is treated as a quantitative measurement for trait-tissue relevance. sLDSC is a special case of MQS (minimal norm quadratic unbiased estimation for summary statistics) and effectively relies on a method of moments (MoM) to estimate SNP heritability based on linear mixed models (Zhou, 2017). While sLDSC and MQS were initially proposed to examine one SNP annotation at a time in the presence of multiple epigenetic annotations, SMART (scalable multiple annotation integration for trait-relevant tissue identification) (Hao et al., 2018) extends these methods to simultaneously incorporate multiple tissue-specific binary and/or continuous functional annotations to facilitate consistent trait-tissue inference (Liang and Zeger, 1986; Chen et al., 2004). SMART uses the generalized estimating equation (GEE) algorithm on the same linear mixed model to achieve such inference goal. Different from using epigenetic measurements, the LDSC-SEG (sLDSC applied to specifically expressed genes) uses tissue-specific transcriptomic annotations, allowing for the inference of trait-tissue relevance with transcriptomic data (Finucane et al., 2018). Similarly, RolyPoly (a regression-based polygenic model) relies on a similar linear mixed model as used in sLDSC/MQS/SMART and creates cell type-specific annotations based on scRNA-seq data (Calderon et al., 2017). In contrast, while using the tissue-specific bulk RNA-seq expression information, the deTS method (method of decoding tissue specificity) directly examines whether the tissue-specifically expressed genes tend to be trait-associated genes using standard enrichment analysis such as the Fisher's exact test to serve as evidence of trait-tissue relevance (Pei et al., 2019). Some methods can make use of the expression quantitative trait loci (eQTLs) information in detecting trait-relevant tissues and cell types. For example, NTCS (normalized tissue causality score) uses eQTLs to assess the genetic causality behind GWASs (Ongen et al., 2017) and eQTLEnrich tests whether eQTLs from a given tissue and/or cell type are significantly enriched for trait associations (Gamazon et al., 2018). Alternatively, other methods measure the trait-tissue relevance by evaluating the proportion of phenotypic variance explained by genetically regulated expression levels (GReX) in different tissues. For example, IGREX (impact of genetically regulated expression) (Cai et al., 2020) and RhoGE (Mancuso et al., 2017) obtain the predicted GReX in tissues and use the association evidence of tissue-specific GReX with the trait for inferring trait-relevant tissues. Finally, CoCoNet (composite likelihood-based covariance regression network model) (Shang et al., 2020b) integrates GWAS data with tissue- or cell type-specific gene co-expression patterns obtained from bulk or single cell gene expression studies based on a network model. In particular, CoCoNet expresses gene-level effect sizes for the given GWAS trait as a function of the tissue-/cell type-specific adjacency matrix and infers how a tissue is relevant to the given trait by examining how effective the tissue-specific gene co-expression

network is for predicting gene-level association pattern with the trait.

Despite the abundance of integrative methods developed for trait-tissue relevance inference, however, a comprehensive review is currently lacking for summarizing the technical details and benefits of each of the above methods. Previous reviews on tissue-trait relevance inference often focus on a limited number of methods that use only functional annotations (Cano-Gamez and Trynka, 2020). To fill this critical knowledge gap, we provide a systemic review on ten different integrative methods for trait-tissue relevance inference. These methods are organized into four main categories based on the tissue- or cell type-specific genomic information they reply on. For each method in turn, we describe the input genomic data types, the detailed statistical model and computational algorithm, the output for evaluating trait-tissue relevance, and the main results obtained in the original study. A summary of these methods is provided in **Table 1** and **Figure 1**, with a brief schematic illustration of each type of methods provided in **Figure 2**. We hope that this review can serve as a useful reference for practitioners who are interested in identifying the causal tissues/cell types of GWAS traits and understanding the SNP association with complex traits in a tissue-specific fashion, as well as for methodologists who develop computational methods for quantifying trait-tissue relevance.

METHODS BASED ON TISSUE-SPECIFIC SNP FUNCTIONAL ANNOTATIONS

Here, we describe the first category of methods for trait-tissue relevance inference. The first category of methods makes use of SNP functional annotations. Exemplary methods include sLDSC (Finucane et al., 2015) and SMART (Hao et al., 2018) that make use of epigenetic annotations; and LDSC-SEG (Finucane et al., 2018), deTS (Pei et al., 2019), and RolyPoly (Calderon et al., 2017) that make use of transcriptomic annotations. The key idea behind these methods is to estimate the contribution of tissue-/cell type-specific functional annotations to SNP heritability for the GWAS trait of interest.

Methods That Use Epigenetic Annotations

In parallel to trait mapping efforts, large-scale functional genomic studies have yielded a rich source of epigenetic annotations (The ENCODE Project Consortium, 2012; Akbarian et al., 2015; Kundaje et al., 2015; Stunnenberg et al., 2016). Various discrete and continuous epigenetic annotations are being developed to describe and characterize the biological function of genetic variants (Kellis et al., 2014; Carithers and Moore, 2015; Dixon et al., 2015). For example, we can now classify genetic variants based on their biochemical function as measured by histone modification, DNase I hypersensitive sites (DHSs), metabolomic QTL evidence, and/or a combination of all these measurements in the form of chromatin states (Pique-Regi et al., 2011; Ernst and Kellis, 2012; McVicker et al., 2013). Often times, these epigenetic annotations are tissue specific and/or cell type specific, allowing characterizing SNP functions in a tissue- or cell type-specific fashion. Paring such tissue-specific SNP epigenetic annotations with SNP association evidence with the GWAS trait allows us to infer trait-tissue relevance. Here, we introduce two methods, sLDSC and SMART, that make use of epigenetic information for trait-tissue relevance inference. In the present review, we simply refer to each tissue-specific epigenetic annotation (e.g., H3K4me1, H3K4me3, and H3K9ac) as a functional category.

sLDSC

The sLDSC (Finucane et al., 2015) estimates how a tissue-/cell type-specific functional annotation contributes to the SNP heritability of the GWAS trait as evidence for trait-tissue relevance inference. Specifically, for each examined tissue in turn, sLDSC first partitions SNPs into C different non-overlapping functional categories based on tissue-specific epigenetic annotations. We use H_c $(c = 1, \ldots, C)$ to denote the set of SNPs that belong to the c-th category. For example, C could be three, with H_1 = H3K4me1 that consists of SNPs that are inside or nearby H3K4me1 peaks in the examined tissue, H_2 = H3K4me3 that consists of SNPs that are inside or nearby H3K4me3 peaks, and H_3 = H3K9ac that consists of SNPs that are inside or nearby H3K9ac peaks. We denote χ_j^2 as the marginal chi-square statistics for the j-th SNP association with the trait. sLDSC considers the following model on the marginal chi-square statistic:

$$E\left[\chi_j^2\right] = 1 + N\sum_{c=1}^{C}\tau_c \ell\left(j,c\right), \tag{1}$$

where $\ell\left(j,c\right) = \sum_{j' \in H_c} r_{jj'}^2$ is the LD score of the j-th SNP with respect to category c, with $r_{jj'}^2$ being the R-squared value between j-th SNP and j'-th SNP that is in the set H_c; and τ_c represents the per-SNP heritability of category H_c. The total SNP heritability explained by the examined functional annotation H_c is defined as $h_g^2\left(c\right) = p_c\tau_c$ with p_c being the number of SNPs in category c. By replacing $E\left[\chi_j^2\right]$ with the observed GWAS marginal association statistic χ_j^2 and solve Equation (1), sLDSC can obtain the estimate of τ_c, $\hat{\tau}_c$, and subsequently $\hat{h}_g^2\left(c\right)$. With the standard error of $\hat{h}_g^2\left(c\right)$ estimated using a jackknife procedure (Quenouille, 1956), sLDSC can further compute a z-score $\hat{h}_g^2\left(c\right)/se\left(\hat{h}_g^2\left(c\right)\right)$ and a subsequent p-value as a measurement of the tissue/cell type relevance to the GWAS trait based on the functional annotation c. In the original paper, the sLDSC method is applied to analyze 17 complex diseases and traits using one functional annotation at a time. By analyzing cell type-specific functional annotations, sLDSC identified many cell type relevance to traits. Examples include the relevance of central nervous system cell types to body mass index, age at menarche, year of education, and smoking status.

SMART

sLDSC examines one functional annotation at a time. However, analyzing one epigenetic annotation at a time fails to incorporate the rich information contained in various other annotations that likely characterize other functionality of variants (Lu et al.,

TABLE 1 | A summary of statistical methods for trait-tissue relevance inference.

Genomic information	Method	GWAS inputs	Measurements	Strengths	Limitations	References
Epigenetic annotations	sLDSC	SNP-based Summary statistics	*p*-values	It extends the commonly used LDSC approach by partitioning SNPs into different functional categories and determining the contribution of each category to trait heritability; can test one annotation while controlling for other annotations in the model.	Examines one annotation at a time; relies on the standard linear mixed model that assumes a polygenic genetic architecture; uses method of moments for model fitting.	Finucane et al., 2015
	SMART	Either individual-level phenotype and genotype data or summary statistics	Posterior probabilities	It handles multiple binary and/or continuous annotations simultaneously; uses the computationally efficient GEE method to estimate and make inference on annotation coefficients.	Relies on the standard linear mixed model that assumes a polygenic genetic architecture.	Hao et al., 2018
Transcriptomic annotations	LDSC-SEG	SNP-based summary statistics	*p*-values	Same as the sLDSC model; effectively creates a gene level annotation by annotating SNPs in genes that are specifically expressed in a tissue to one and annotating the remaining SNPs to zero.	Model performance highly depends on the gene expression data, which is used to determine tissue specificity of gene expression and subsequently tissue specific SNP annotations; sensitive to gene expression correlation across cell and tissue types.	Finucane et al., 2018
	RolyPoly	SNP-based summary statistics	*p*-values	Similar to the sLDSC model; integrates scRNA-seq data with GWAS; jointly analyzes gene expression from multiple tissues or cell types; prioritizes trait-relevant cell types and genes.	Model performance highly depends on the gene expression data used; sensitive to gene expression correlation across cell and tissue types.	Calderon et al., 2017
	deTS	A list of trait-associated genes	*p*-values	Applicable when only a list of GWAS significant genes are available.	Model performance highly depends on the gene expression data; there is not a commonly accepted threshold for defining trait-associated genes, and different thresholds may result in different sets of genes and thus different enrichment results.	Pei et al., 2019
eQTL information	NTCS	A list of trait-associated and null SNPs	Ranking of tissues based on adjusted fold-enrichment	Rank genes in terms of their contribution to trait-tissue relevance.	No publicly available tools; model implementation is redundant and difficult to replicate.	Ongen et al., 2017
	eQTLEnrich	GWAS summary statistics	*p*-values	Both tissue-shared and tissue-specific regulatory effects of eQTLs are analyzed.	The adjusted fold-enrichment used for ranking tissues in eQTLEnrich is correlated with GWAS sample size.	Gamazon et al., 2018
Genetically regulated expression (GReX)	IGREX	Either individual-level phenotype and genotype data or summary statistics	*p*-values	Measures the phenotypic variance explained by GReX; can analyze both GWAS individual-level and summary data.	Uses REML for inference, which can be time consuming.	Cai et al., 2020

(Continued)

TABLE 1 | Continued

Genomic information	Method	GWAS inputs	Measurements	Strengths	Limitations	References
	RhoGE	SNP-based summary statistics	*p*-values	Measures the phenotypic variance explained by GReX.	Uses a two-stage regression for inference, which may fail to account for estimation uncertainty in the first stage.	Mancuso et al., 2017
Gene co-expression network	CoCoNet	Either individual-level phenotype and genotype data or summary statistics	Ranking of tissues based on log-likelihood	Incorporates tissue-specific gene co-expression networks constructed from either bulk or single cell RNA sequencing (RNAseq) studies with GWAS data; is scalable to tens of thousands of genes.	Currently only focuses on ranking tissues for a given disease.	Shang et al., 2020b

These methods make use of different genomic information (1st column), GWAS inputs (3rd column), and different measurements (4th column) for trait-tissue relevance inference, strengths (5th column), and limitations (6th column) of each method.

2016, 2017; He et al., 2017). For example, some annotations are designed to evaluate the function of a variant in determining the protein structure, while some other annotations are designed to quantify its ability to regulate gene expression. Even categories that belong to the same epigenetic annotation may characterize substantially different functions of a variant. For example, H3K4me1 is used to annotate enhancers while H3K4me3 is used to annotate promoters. Therefore, it is desirable to make use of multiple epigenetic annotations to obtain consistent and robust trait-tissue relevance inference results. A key step that facilitates the incorporation of multiple epigenetic annotations is the discovery that the data generating model underlying sLDSC is a standard linear mixed model and that sLDSC fits the linear mixed model using the method of moments (MoM) (Zhou, 2017). Indeed, sLDSC is practically a special case of MQS, which provides a unified framework for variance component estimation in linear mixed models (Zhou, 2017). Building upon the same linear mixed model that sLDSC and MQS use, SMART (Hao et al., 2018) was developed to incorporate multiple tissue-/cell type-specific epigenetic annotations for trait tissue/cell type inference. In particular, SMART allows for the incorporation of multiple tissue-specific binary and continuous epigenetic annotations. For example, a tissue-specific binary histone annotation can be an indicator that indicates whether the SNP resides inside the peak regions of the histone mark, while a tissue-specific continuous histone annotation can be an average of counts in the histone peak region. Importantly, because of its reliance on a data generative linear mixed model, SMART can be applied to handle either individual-level GWAS data or summary statistics. For individual-level GWAS data, SMART models the phenotype as

$$\mathbf{y} = \tilde{\mathbf{G}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_{\mathbf{y}}, \tag{2}$$

where $\mathbf{y}$ is a vector of phenotypes for N GWAS samples; $\tilde{\mathbf{G}}$ is an $N \times p$ genotype matrix measured from the same N samples and p genome-wide SNPs; $\boldsymbol{\gamma}$ is a p-vector of effect sizes; and $\boldsymbol{\varepsilon}_y \sim N\left(\mathbf{0}_N,\ \sigma_y^2\mathbf{I}_N\right)$ is the N-vector error term, where $\mathbf{0}_N$ represents an N-vector of zeros and $\mathbf{I}_N$ represents an N-dimensional identity matrix. The phenotype $\mathbf{y}$ and each column of the genotype matrix $\tilde{\mathbf{G}}$ are standardized to have zero mean and unit standard deviation, allowing us to ignore the intercept in Equation (2). SMART assumes that all SNPs are characterized by a set of s functional annotations. For the j-th SNP, we use a $(s+1)$-vector $\mathcal{F}_j = \left(1, F_{j1}, \ldots, F_{js}\right)^T$ to denote its annotation values across s functional epigenetic annotations, where the first value 1 corresponds to the intercept. Here, each of $F_{j1}, \ldots, F_{js}$ can either be a binary value or a continuous value. With the SNP annotations, SMART assumes that the SNP effect size γ_j follows a normal distribution with zero mean and SNP-specific variance that is a function of the annotation vector,

$$\gamma_j \sim N\left(0, \frac{\sigma_j^2}{p}\right),\ \ \boldsymbol{\sigma}_j^2 = \mathcal{F}_j\boldsymbol{\alpha}^*, \tag{3}$$

where $\boldsymbol{\alpha}^* = \begin{pmatrix}\alpha_0 \\ \boldsymbol{\alpha}\end{pmatrix}$ is a $(s+1)$-vector of coefficients that include an intercept α_0 and a s-vector of annotation coefficients $\boldsymbol{\alpha}$. To evaluate the joint contribution of multiple annotations to genetic effect sizes, SMART performs parameter inference using the generalized estimation equation (GEE) (Liang and Zeger, 1986). Use of GEE not only enables scalable computation, but also allows for the use of GWAS summary statistics based on the same model characterized by Equations (2) and (3). By applying GEE, SMART obtains point estimates $\hat{\boldsymbol{\alpha}}$ and their covariance matrix $Var\left(\hat{\boldsymbol{\alpha}}\right)$, which allow for the computation of the multivariate Wald statistic, $\hat{\boldsymbol{\alpha}}^T Var\left(\hat{\boldsymbol{\alpha}}\right)^{-1}\hat{\boldsymbol{\alpha}}$. The Wald statistic is further modeled as a mixture of two non-central chi-squared distributions for classifying tissues into trait-relevant and trait-irrelevant groups. An expectation-maximum (EM) algorithm is then applied to the chi-squared mixture to infer the posterior probability of a tissue being a trait-relevant tissue.

In the original paper, SMART analyzed 43 traits from 29 GWAS studies and obtained many trait-relevant tissues and cell types. For example, SMART identified the central nervous system (CNS) tissues to be the most trait-relevant for psychiatric disorders (e.g., schizophonia, Alzheimer's disease)

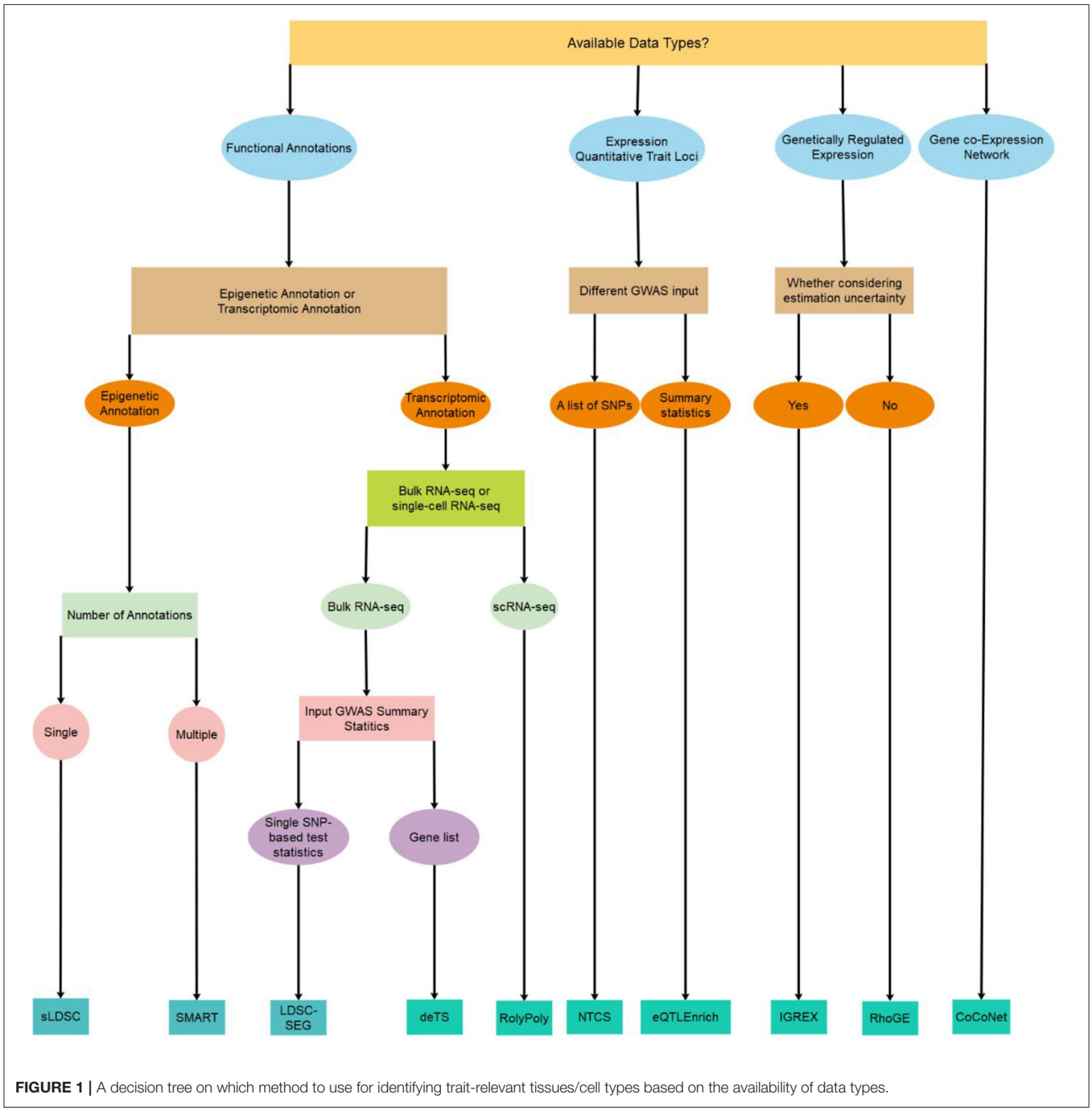

FIGURE 1 | A decision tree on which method to use for identifying trait-relevant tissues/cell types based on the availability of data types.

and neurological related traits (e.g., years of education, childhood BMI). These results are consistent with existing literature. For example, searching the trait-tissue pair schizophrenia-CNS on PubMed yielded 17,720 hits while searching for the trait-tissue pair Alzheimer-CNS yielded 34,395 hits, supporting their clear relevance. As another example, SMART identified the bone and connective tissues to be related to height and femur neck bone mineral density, and the blood/immune tissues to be related to immune diseases (e.g., Rheumatoid Arthritis, type 1 diabetes). These results are also in line with literature: PubMed search for height-BoneConnective yielded 13,644 hits and search for RA-BloodImmune yielded 6,868 hits, supporting their relevance.

Methods That Use Transcriptomic Annotations

Besides epigenomic studies, many gene expression studies have been carried out to characterize the transcriptomic landscape of various tissues and cell types (The ENCODE Project Consortium, 2012; GTEx Consortium, 2015; Kundaje et al., 2015). These tissue- and cell type-specific gene expression information can be

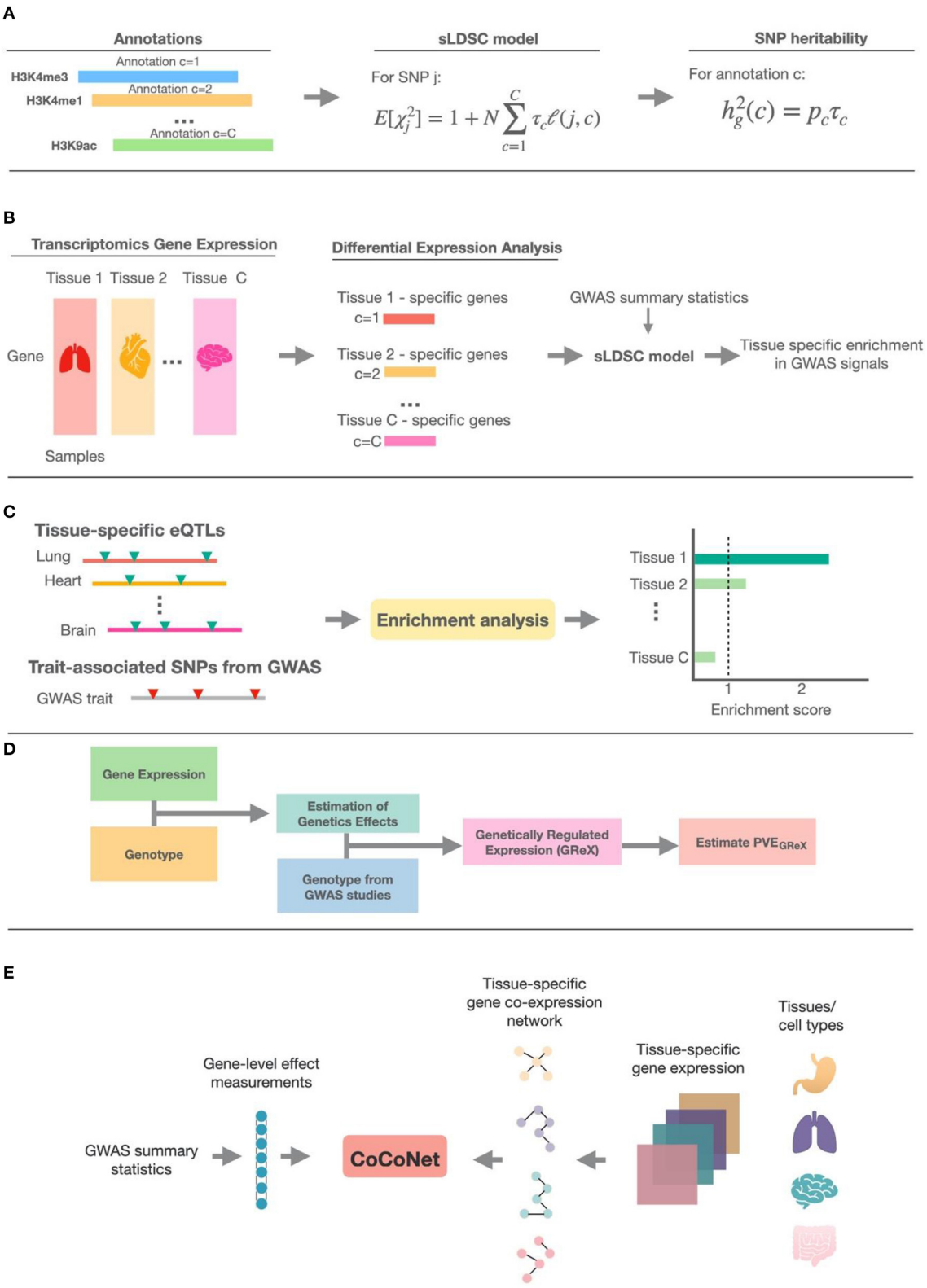

FIGURE 2 | The schematic illustration of methods in the five different categories. **(A)** The general schema of methods that make use of epigenetic annotation information; sLDSC is shown as the detailed example. **(B)** The general schema of methods that use tissue-specific transcriptomic annotation information; these

(Continued)

FIGURE 2 | methods first define specifically expressed genes (SEGs) based on differential expression analysis, then construct genomic annotations from the SEGs, and finally use sLDSC to perform trait-tissue relevance inference. **(C)** The schema of methods that test for enrichment of trait associations among eQTLs in each tissue. **(D)** The general schema of methods that obtain the estimated genetically regulated expression (GReX) and use the proportion of phenotypic variance explained by GReX (PVE_{GReX}) to measure the trait-tissue relevance. **(E)** The schema of methods that make use of tissue-specific gene co-expression networks; CoCoNet is shown as the detailed example.

invaluable for inferring trait-tissue relevance (Hu et al., 2011; Slowikowski et al., 2014; Pers et al., 2015; Gormley et al., 2016). In this section, we introduce three methods that make use of gene expression data in the form of transcriptomic annotations. These methods include LDSC-SEG (Finucane et al., 2018) and deTS (Pei et al., 2019) that make use of bulk RNA-seq expression data, and RolyPoly (Calderon et al., 2017) that makes use of single-cell RNA-seq expression data.

LDSC-SEG

LDSC-SEG consists of two separate steps. The first step of LDSC-SEG is a differential expression analysis on the gene expression data to identify a set of genes that are specifically expressed in certain tissues. These tissue specific genes are referred to either as specifically/differentially expressed genes (SEGs) or tissue-specific genes (TSGs). In the differential expression analysis, LDSC-SEG examines one gene at a time. For the given gene, LDSC-SEG contrasts the gene expression level of samples collected in a focal tissue (e.g., brain-cortex) with those of samples collected in all other tissues that are not in the same tissue category as the focal tissue (i.e., non-brain tissues). Because tissues within each tissue category tend to share similarly expressed genes, excluding the tissues in the same tissue category in the differential expression analysis step becomes the key to ensure robust detection of SEGs. Indeed, such differential expression analysis allows for the inclusion of as many genes as possible that are highly expressed in the focal tissues but not in tissues from other tissue categories. The SEG evidence for a gene is typically characterized by a t-statistic, with a higher value indicating that the gene is more specifically/differentially expressed in the focal tissue. With the differential expression analysis results, LDSC-SEG ranks all genes in a descending order based on their t-statistics. LDSC-SEG then defines SEGs as the top 10 percentage of all genes. The identification of SEGs allows LDSC-SEG to create a binary SNP annotation in a tissue specific fashion. In particular, for each tissue at a time, LDSC-SEG annotates the SNP to be one if the SNP resides within 100 kb of the transcription start site of any SEG and annotates it to be zero otherwise. With the tissue-specific binary annotation, LDSC-SEG then performs the second step of applying the sLDSC method described in the previous section to estimate the proportion of SNP heritability explained by each tissue-specific binary SNP annotation. The resulting test statistic from sLDSC is then served as a relevance evidence between the tissue and trait.

In real data applications, LDSC-SEG analyzed GWAS summary statistics for 48 diseases and traits and found significant tissue-/cell type-specific enrichments for 34 traits. Several of these findings recapitulate known biology. For example, immunological traits exhibit immune tissue-type enrichments; psychiatric traits exhibit strong brain-related tissue enrichments; and type II diabetes exhibits enrichments in the pancreas. LDSC-SEG also validated several recent genetic analyses results, including robust brain-specific enrichments for smoking status, years of education, body mass index, and age at menarche.

deTS

deTS also consists of two-steps. The first step of deTS also consists of a differential expression analysis as in the first step of LDSC-SEG. The only minor difference there is the definition of SEGs: while LDSC-SEG defines top 10% as SEGs, deTS defines top 5% as SEGs. However, the second step of deTS relies on an enrichment analysis rather than sLDSC. Specifically, deTS implements Fisher's exact approach to test whether the SEGs are enriched in the focal tissue or not. The Fisher's exact test builds upon a two-by-two contingency table, where the two rows represent the number of SEGs vs. the number of non-SEGs in the tissue, while the two columns represent the number of trait-associated genes vs. the number of non-trait-associated genes. Here, the trait-associated gene is defined based on a gene-level p-value threshold of 5×10^{-3}, where the p-value is calculated from a gene-based test (Lamparter et al., 2016). In the original study, deTS is applied to analyze GWAS summary statistics for 26 traits. deTS found that artery tissues were primarily associated with anthropometric trait, liver was primarily associated with metabolic traits, blood and spleen were primarily associated with immune-related traits, and brain tissues were primarily associated with neurodegenerative/neuropsychiatric diseases.

RolyPoly

RolyPoly (Calderon et al., 2017) is specifically developed for single cell expression studies. It consists of the same two steps as LDSC-SEG. In the first step, RolyPoly uses a slightly different approach than LDSC-SEG to define the SEGs. Specifically, for each tissue, RolyPoly ranks all genes in a descending order based on the normalized expression values and define the top 20% of genes as SEGs. Afterwards, RolyPoly creates a binary SNP annotation based on whether a SNP resides within a 10 kb window nearby the transcription start site of any SEGs. In the second step, RolyPoly applies the same linear mixed model as used in sLDSC for inference (Finucane et al., 2015). In real data analysis, RolyPoly identified significant relevance of oligodendrocytes and fetal replicating cells with schizophrenia.

METHODS BASED ON EXPRESSION QUANTITATIVE TRAIT LOCI INFORMATION

In recent years, expression mapping studies have succeeded in identifying many cis-acting genetic variants known as cis-eQTLs that are associated with gene expression levels (Schadt et al., 2003; Morley et al., 2004; Lappalainen et al., 2013; Battle et al., 2014). The identified eQTLs can help elucidate the molecular mechanisms underlying human disease associations and facilitate the identification of biological pathways underlying disease etiology. For example, it has been shown that the GWAS variants frequently colocalize and likely share functional effects with eQTLs (Nica et al., 2010; Nicolae et al., 2010; Grundberg et al., 2012; Shang et al., 2020a). Thus, at least some of these variants influence traits through regulatory effects. In addition, the identified eQTLs in multiple tissues and/or cell types can help interpret the GWAS results through linking non-coding genomic regions to gene functions and identifying causal tissues/cell types behind the genetic associations (Nica and Dermitzakis, 2008; Montgomery and Dermitzakis, 2011; Grundberg et al., 2012). In this section, we will introduce two methods, NTCS (Ongen et al., 2017) and eQTLEnrich (Gamazon et al., 2018), that make use of tissue- and cell type-specific eQTL information to infer the trait-relevant tissues and cell types that are behind genetic causality.

NTCS

For a given tissue, NTCS makes use of a list of significant eQTLs that are not in linkage disequilibrium (LD) with each other along with their colocalized GWAS variants. These eQTLs are obtained from a conditional eQTL mapping analysis, performed through, for example, FastQTL (Welter et al., 2014). The identified eQTLs are overlapped with common variants downloaded from the NHGRI-EBI GWAS catalog (Storey and Tibshirani, 2003) to obtain a list of eQTLs that have GWAS significance ($P <$ 5e−8). These eQTLs are denoted as real GWAS variants, GWAS variants, or GWAS-associated variants.

The NTCS method first uses the Regulatory Trait Concordance (RTC) (Nica et al., 2010) approach to detect colocalized variants between the GWAS study and the eQTL study while properly accounting for LD. The resulted RTC score is then converted to a probability value that measures the sharing between a GWAS variant and an eQTL in a tissue, or between two eQTLs in a pair of tissues based on Bayes' theorem:

$$P\left(shared\middle|RTC = rtc\right) = \frac{P\left(RTC = rtc\middle|shared\right)\cdot\pi_1}{P\left(RTC = rtc\middle|shared\right)\cdot\pi_1 + P\left(RTC = rtc\middle|not\ shared\right)\cdot\pi_0}, \quad (4)$$

where $P\left(shared\right) = \pi_1$ is a π_1 statistics and $\pi_0 = 1 - \pi_1$. When calculating the probability of sharing between the GWAS variants and eQTLs in a given tissue, the π_1 statistics is calculated from eQTL p-values in the tissue and GWAS variants. When calculating the probability of sharing between two eQTLs in a pair of tissues, the π_1 statistics is calculated from eQTL p-values in the two tissues. Both $P\left(RTC = rtc\middle|not\ shared\right)$ and $P\left(RTC = rtc\middle|shared\right)$ are estimated through simulations, where the RTC scores are simulated under both the null and alternative hypotheses. Specifically, for each coldspot that has colocalized GWAS and eQTL variants ($eQTL_{real}$), under the null hypothesis (H_0) where GWAS and eQTL are tagging two different variants, two hidden causal variants (GWAScausal and eQTLcausal) are randomly selected. Under the alternative hypothesis (H_1) where GWAS and eQTL are tagging the same variant, one hidden causal variant ($eQTL_{causal}$) is randomly selected. In both hypotheses, the GWAS and eQTL variants are randomly selected from the variants that are in linkage disequilibrium with the hidden causal variants with $r^2 \geq 0.5$. Afterwards, gene expression is simulated based on the $eQTL_{real}$ effect size. The RTC analyses are then performed under H_0 and H_1, each for 200 times. For each coldspot, the total 400 simulated RTC scores under H_0 and H_1 are merged and sorted to obtain a point probability. Finally, for each GWAS trait in each given tissue and each eQTL that colocalizes with a GWAS variant, NTSC defines a normalized GWAS variant-eQTL probability as the probability of the GWAS variant and eQTL tagging the same functional effect divided by the sum of the tissue-sharing probabilities for the eQTL in that tissue. Intuitively, tissue-specific eQTLs would more likely be a GWAS variant than tissue non-specific eQTLs that are shared across tissues. Therefore, for each GWAS trait in each given tissue, NTCS defines a normalized tissue causality score (NTCS) and a null NTCS as follows:

$$NTCS = \frac{1}{p_2}\times\sum_{j=1}^{p_1}\frac{P\left(SNP_j - eQTL_j\ shared|rtc\right)}{P\left(eQTL_j\ shared|rtc\right)}, \quad (5)$$

$$Null\ NTCS = \frac{p_1}{p_0 p_2}\times\sum_{j=1}^{p_0}\frac{P\left(null\ SNP_j - eQTL_j\ shared|rtc\right)}{P\left(eQTL_j\ shared|rtc\right)}, \quad (6)$$

where p_1 is the number of GWAS-associated variants for the trait; p_2 is the total number of eQTLs in a given tissue; p_0 is the number of GWAS-null variants; $P\left(SNP_j - eQTL_j\ shared|rtc\right)$ is the probability that a GWAS variant (i.e., SNP_j) and $eQTL_j$ tagging the same functional effect; and $P\left(eQTL_j\ shared|rtc\right)$ is defined in Equation (4). An enrichment metric is further defined as $\frac{NTCS}{null-NTCS}$. The tissues with an enrichment metric greater than one are likely the causal tissues for the diseases/traits. To create a p-value for testing trait-relevance of each tissue, NTCS first selects a *null* GWAS variant to match each of the GWAS variant, based on minor allele frequency and distance to the closest transcription start site. Afterwards, NTCS repeats the above enrichment metric calculation using the set of null GWAS variants, examines one tissue at a time, compares the tissue metric for the disease-associated variants to the metric observed under the null for that tissue, and calculates a corresponding p-value based on a Mann-Whitney test that compares the distribution containing each of the j-th elements in Equation (5) and (6) for the real GWAS and under the null. In the NTCS paper, NTCS method discovers that liver is the tissue most likely to be causal in most of the GWAS traits. Brain tissues are the top tissues relating to traits like schizophrenia, height, and age of onset of puberty.

eQTLEnrich

eQTLEnrich is a rank- and permutation-based method that aims to test for enrichment of trait associations among eQTLs in each tissue. For a given GWAS trait, for each of the tissues with eQTLs, eQTLEnrich first finds the most significant cis-eQTL per eGene, and then extracts the GWAS variant association p-values for each set of eQTLs. Afterwards eQTLEnrich tests for the enrichment of the distribution of GWAS p-values for each set of eQTLs in the corresponding tissue. The distribution of the GWAS p-values for each set of eQTLs is tested for enrichment of highly ranked trait associations compared to an empirical null distribution sampled from non-significant variant-gene expression associations.

Specifically, eQTLEnrich first computes the fold-enrichment for each GWAS-tissue pair. The fold-enrichment is defined as the fraction of eQTLs with GWAS variant $p < 0.05$ compared to expectation. Similarly, eQTLEnrich also computes fold-enrichment values for randomly sampled sets of non-significant variant-gene expression associations of equal size to the eQTL set, matching the distance of eQTL to TSS of the target gene, MAF, and number of proxy variants (at $r^2 \geq 0.5$), to account for LD. Then eQTLEnrich computes an enrichment p-value as the fraction of permutations with similar or higher fold-enrichment than the observed value. Finally, eQTLEnrich computes an adjusted fold-enrichment by dividing the fold-enrichment for a specific GWAS-tissue pair by the fold-enrichment of all non-significant variant-gene expression associations with GWAS $P < 0.05$ for the tissue-trait pair. The eQTLEnrich method is applied to analyze 18 complex diseases and traits on 44 GTEx tissues and identifies many trait-relevant tissues. Examples include the relevance of left heart ventricle and adipose visceral omentum to type I diabetes, ovary and artery coronary to coronary artery disease, and hippocampus to Alzheimer's disease.

METHODS BASED ON TISSUE-SPECIFIC GENETICALLY REGULATED EXPRESSION LEVELS

Here, we describe the third category of methods for trait-tissue relevance inference. The third category of methods use information from genetically regulated expression levels (GReX) that are constructed in a tissue specific fashion. GReX measures the part of gene expression levels that can be predicted by (cis-)SNPs (Gamazon et al., 2015). In a given tissue, GReX is constructed for each gene by fitting a prediction model that relates the gene expression level to the cis-SNPs. Common prediction models for GReX construction include elastic net (Zou and Hastie, 2005), BSLMM (Zhou et al., 2013), and DPR (Zeng and Zhou, 2017). Constructed GReX is often tested with the GWAS trait for association evidence through transcriptome-wide association studies (TWAS) (Gamazon et al., 2015; Gusev et al., 2016). Indeed, GReX of many genes have been identified to be associated with diseases and disease-related complex traits. In this section, we will introduce two methods, IGREX (Cai et al., 2020) and RhoGE (Mancuso et al., 2017), that rely on GReX to infer trait-tissue relevance. Both methods effectively are built upon the same model but rely on different algorithms for model inference.

Specifically, both methods consider two separate models, one for the gene expression study and the other for the GWAS. In the gene expression study, both methods examine one tissue and one gene at a time. For the m-th gene in the tissue, both methods consider the following linear model for modeling the relationship between gene expression and genotypes of cis-SNPs,

$$\mathbf{z}_m = \mathbf{G}_m \mathbf{w}_m + \boldsymbol{\varepsilon}_z, \tag{7}$$

where $\mathbf{z}_m$ is an n-vector of expression values measured from a focal tissue, with n being number of available samples in this tissue; $\mathbf{G}_m$ is an $n \times p$ genotype matrix for the same n samples and p cis-SNPs for the given gene; $\mathbf{w}_m$ is a p-vector of SNP effect sizes on the gene expression; and $\boldsymbol{\varepsilon}_z \sim N(\mathbf{0}_n, \sigma_z^2 \mathbf{I}_n)$ is the residual error term. The gene expression $\mathbf{z}_m$ and each column of genotype matrix $\mathbf{G}$ are standardized, allowing us to ignore the intercept term in Equation (7). The genetic effects on gene expression is assumed to follow a normal distribution *a priori*, with $\mathbf{w}_m \sim N(\mathbf{0}_p, \sigma_w^2 \mathbf{I}_p)$.

In the GWAS data, both methods consider the following regression model that relates the phenotype to genotype:

$$\mathbf{y} = \tilde{\mathbf{G}}_r \boldsymbol{\gamma} + \sum_{m=1}^{M} \beta_m \tilde{\mathbf{G}}_m \mathbf{w}_m + \boldsymbol{\varepsilon}_y, \tag{8}$$

where $\mathbf{y}$ and $\boldsymbol{\varepsilon}_y$ are defined as in Equation (2); $\tilde{\mathbf{G}}_m$ is the $N \times p$ genotype matrix for p cis-SNPs in the given gene; $\mathbf{w}_m$ is the same SNP effects on gene expression as defined in Equation (7); the scalar $\beta_m \sim N\left(0, \sigma_\beta^2\right)$ represents the genetic effect of GReX (i.e., $\tilde{\mathbf{G}}_m \mathbf{w}_m$) on $\mathbf{y}$ and can be interpreted as the causal effect of GReX on $\mathbf{y}$ (Yuan et al., 2019; Zhu and Zhou, 2020); and $\boldsymbol{\gamma} \sim N\left(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_q\right)$ is the q-length vector of alternative genetic effects; note that $\tilde{\mathbf{G}}_\gamma$ is not the same genotype matrix as $\tilde{\mathbf{G}}_m$, and the q SNPs in $\tilde{\mathbf{G}}_\gamma$ are those who show direct horizontal effects on $\mathbf{y}$, such as the trans-eQTLs and SNPs associated with alternative splicing events (Matlin et al., 2005).

Above, the proportion of phenotypic variance explained by GReX is calculated as

$$PVE_{GReX} = \frac{Var\left(\sum_m \beta_m \tilde{\mathbf{G}}_m \mathbf{w}_m\right)}{Var(\mathbf{y})}. \tag{9}$$

IGREX

IGREX (Cai et al., 2020) relies on a two-stage method to perform inference for the model defined in Equations (7) and (8). Specifically, IGREX first estimates the posterior distribution of genetic effects on expression based on Equation (7) and obtains the posterior distribution $\mathbf{w}_m | \mathbf{z}_m, \mathbf{G}_m \sim N\left(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)$ for each gene m. Afterwards, IGREX treats the posterior distribution $\mathbf{w}_m | \mathbf{z}_m, \mathbf{G}_m$ from Equation (7) as the prior distribution for Equation (8), and obtain the estimates of σ_β^2, σ_γ^2 and σ_y^2 using

either the method of moments (MoM) or REML. Finally, the estimate of PVE_{GReX} is obtained by

$$\widehat{PVE}_{GReX} = \frac{tr\left(\sum_m \hat{\sigma}_\beta^2 \tilde{\mathbf{G}}_m (\boldsymbol{\mu}_m \boldsymbol{\mu}_m^T + \boldsymbol{\Sigma}_m) \tilde{\mathbf{G}}_m^T\right)}{tr\left(\sum_m \hat{\sigma}_\beta^2 \tilde{\mathbf{G}}_m (\boldsymbol{\mu}_m \boldsymbol{\mu}_m^T + \boldsymbol{\Sigma}_m) \tilde{\mathbf{G}}_m^T + \hat{\sigma}_\gamma^2 \tilde{\mathbf{G}}_r \tilde{\mathbf{G}}_r^T + \hat{\sigma}_\gamma^2 I_N\right)}. \quad (10)$$

In the above two-step estimation procedure, IGREX relies on the posterior distribution $\mathbf{w}_m|\mathbf{z}_m, \mathbf{G}_m$ to account for estimation uncertainty associated with $\mathbf{w}_m$ in Equation (8). Given the point estimate $\widehat{PVE}_{GReX}$ and its standard error estimated by block jackknife (Quenouille, 1956), IGREX tests the tissue-specific null hypothesis that $H_0 : PVE_{GReX} = 0$ by using a simple z-test. While IGREX is presented based on individual level data, IGREX is also applicable for GWAS summary statistics using the same model defined above. In the original study, IGREX used the GTEx project as expression mapping study and GWAS data in both individual-level and summary statistics. IGREX identified several trait-relevant tissue types. For example, significant GReX components were observed in liver for both high-density lipoprotein and low-density lipoprotein, in brain-amygdala for bipolar disorder, in brain-spinal cord (cervical c-1) for coronary artery disease, and in spleen for height.

RhoGE

RhoGE (Mancuso et al., 2017) fits a similar model as defined in Equations (7) and (8) as IGREX, but with three differences. First, RhoGE uses only the posterior mean estimate $\boldsymbol{\mu}_m$ obtained from Equation (7) and subsequently ignores the uncertainty in the estimation of $\mathbf{w}_m$. Second, RhoGE is based on LDSC, and thus estimates the variance components σ_β^2 effectively using MoM. Third, RhoGE does not account for the horizontal pleiotropic effects $\tilde{\mathbf{G}}_r \boldsymbol{\gamma}$. Technically, RhoGE modifies the LDSC estimation procedure to use gene level summary statistics. Specifically, the gene-level statistic χ_m^2 is computed as $\hat{\mathbf{w}}_m^T \boldsymbol{\phi}_m \boldsymbol{\phi}_m^T \hat{\mathbf{w}}_m / \hat{\mathbf{w}}_m^T \mathbf{V}_m \hat{\mathbf{w}}_m$, where $\hat{\mathbf{w}}_m$ is obtained from the genomic best linear unbiased prediction (GBLUP) (de los Campos et al., 2013); $\boldsymbol{\phi}_m$ are the p-vector of SNP-based Wald statistics from the GWAS study; and $\mathbf{V}_m$ is an $p \times p$ LD matrix calculated from a reference panel. Afterwards, RhoGE follows the same inference procedure as in LDSC to estimate PVE_{GReX} and tests whether PVE_{GReX} is statistically significant from zero. The resulting test statistic is served as evidence for trait-tissue relevance inference. RhoGE analyzed GWAS summary statistics for 30 complex traits and found 108 significant trait-tissue pairs across 17 traits and 33 tissues, including BMI-brain, schizophrenia-brain, and high-density lipoprotein-heart.

METHODS BASED ON TISSUE-SPECIFIC GENE CO-EXPRESSION NETWORK

In this section, we introduce the fourth category of methods, which currently consists of only CoCoNet (Shang et al., 2020b), for trait-tissue relevance inference. CoCoNet performs trait-tissue relevance inference using tissue- or cell type-specific gene co-expression network information obtained from bulk or single cell gene expression studies. Gene co-expression networks characterize how genes are connected with each other and are coregulated together. Gene co-expression networks have been shown to be informative for predicting gene-level association effect sizes on diseases in GWASs and are often tissue and cell type specific (Chen et al., 2011; Hou et al., 2014; Jia and Zhao, 2014; Hao et al., 2018). Genes with high network connectivity have also been shown to be enriched for heritability of GWAS traits (Kim et al., 2019). Therefore, it is important to take advantage of tissue-specific gene connection information in tissue-specific gene co-expression networks to facilitate the inference of disease tissue relevance.

CoCoNet

CoCoNet (Shang et al., 2020b) first obtains an M-vector of gene-level effect sizes with the trait of interest from the GWAS, denoted as $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)^T$. In the gene expression study, CoCoNet examines one tissue at a time and for the given tissue constructs an M by M gene-gene adjacency matrix $\mathbf{A} = (a_{mm'})$ to represent the gene co-expression network there. The mm'-th element of the adjacency matrix $a_{mm'}$ is 1 if gene m is connected to gene m' in the network and 0 otherwise. a_{mm} is set to be 0 for any $1 \leq m \leq M$ to ensure the absence of self-loops (Urry and Sollich, 2013). CoCoNet then relies on a covariance regression network model (Lan et al., 2018) to model the relationship between $\mathbf{A}$ and $\boldsymbol{\theta}$

$$\boldsymbol{\theta} \sim N\left(\mathbf{1}_M \mu, \Sigma\left(\mathbf{A}\right)\right), \quad (11)$$

where μ is the intercept and $\Sigma(\mathbf{A})$ is the covariance of $\boldsymbol{\theta}$ as a function of the adjacency matrix $\mathbf{A}$. The covariance $\Sigma(\mathbf{A})$ is in a general form $\Sigma(\mathbf{A}) = \sum_{l=0}^{L} \sigma_l^2 \mathbf{A}^l$, where $\mathbf{A}^l = \left(a_{mm'}^{(l)}\right)$ is the l-th power of $\mathbf{A}$, and L is the maximum number of paths considered for linking between any two genes. For any integer l, $a_{mm'}^{(l)}$ is the number of l-paths linking from gene m to gene m' in the co-expression network, where an l-path is any path of length l. For example, when $l = 2$, $a_{mm'}^{(2)} = \sum_{h=1}^{M} a_{mh} a_{hm'}$, where $a_{mh} a_{hm'}$ is 1 only when there is a link connecting the three genes $m - h - m'$ and 0 otherwise. For $l \geq 1$, CoCoNet sets $a_{mm}^{(l)} = 0$. When $l = 0$, CoCoNet sets $\mathbf{A}^0 = \mathbf{I}$. In the real data application, CoCoNet suggests choosing L based on Bayesian Information Criterion (BIC) according to real data analysis.

Because of the computation burden associated with the model in Equation (11), CoCoNet relies on composite likelihood for approximate inference. In particular, the composite likelihood only needs to make an assumption that each pair $(\theta_m, \theta_{m'})$ follows a bivariate normal distribution, instead of making a strong assumption that the m-vector of θ jointly follows a multivariate Gaussian distribution. Specifically, for each pair of genes m and m', CoCoNet considers the composite likelihood $P(\theta_m, \theta_{m'}|\mu, \sigma_0^2, \sigma_1^2)$ as

$$\begin{pmatrix} \theta_m \\ \theta_{m'} \end{pmatrix} \sim BN\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \sum_{l=0}^{L} \sigma_l^2 \begin{pmatrix} a_{mm}^{(l)} & a_{mm'}^{(l)} \\ a_{mm'}^{(l)} & a_{m'm'}^{(l)} \end{pmatrix}\right), \quad (12)$$

where BN represents bivariate normal distribution. CoCoNet finally constructs the log composite likelihood as

$$loglik(\boldsymbol{\theta}) = \sum_{m=1}^{M} \sum_{m'>m}^{M} \log P\left(\theta_m, \theta_{m'} \middle| \mu, \sigma_0^2, \cdots, \sigma_L^2\right). \quad (13)$$

CoCoNet fits the above composite likelihood through a standard maximum likelihood inference procedure. Afterwards, CoCoNet calculates the maximum composite likelihood for each tissue and eventually ranks tissues by the corresponding log likelihoods. In the original study, the comparative results between CoCoNet and LDSC-SEG/RolyPoly in the original study suggest that tissue-specific gene co-expression network provides valuable trait-tissue relevance information, perhaps more so than the information provided by marginal tissue-specific gene expression pattern used in LDSC-SEG/RolyPoly. CoCoNet analyzed eight different disease GWASs that include four neurological disorders and four autoimmune disorders on 38 tissues obtained from GTEx, CoCoNet found that the top relevant tissues identified for neurological disorders are generally brain tissues, which are disease causing tissues. CoCoNet also found the top relevant tissues for autoimmune disorders to be intestinal tissues, which are disease-target tissues. In trait-cell type relevance identification, CoCoNet found GABAergic interneurons, oligodendrocyte precursor cells, astrocytes, and microglia are the top relevant cell types in Alzheimer's disease. CoCoNet also found both pyramidal neurons and various glia cells are selected as top relevant cell types in bipolar disorder.

DISCUSSION

We have presented a systematic review on existing statistical methods for trait-tissue relevance inference. Our review comes from a technical perspective and summarizes the input data types, detailed statistical model and inference algorithm, criteria for evaluating tissue/cell type relevance of a trait, as well as the main findings from these existing methods. Identifying trait-relevant tissues using these methods not only facilitates the understanding of disease etiology but also enables more powerful association analysis in future GWASs (Hao et al., 2018). For example, tissue-specific SNP annotations and their contributing weights to SNP heritability in the trait-relevant tissue can be used to construct more powerful SNP set tests in GWASs (Hao et al., 2018). In addition, the inferred trait-relevant tissues and/or cell types facilities the interpretation of TWAS analysis and improves the analysis power (Gamazon et al., 2015; Gusev et al., 2016).

Thus far, existing methods have primarily relied on *ad hoc* procedures to validate the inferred trait-tissue relevance results. For example, one would examine top trait-relevant tissues one by one and look for corresponding evidence in the literature to support such results. Manually cross checking with literature, however, requires domain knowledge and may yield biased results. Manual literature checking is also time consuming and the outcome results are not easy to quantify. To overcome the shortcomings of manual literature checking, Hao et al. (2018) provided a convenient approach to quantitatively validate trait-tissue relevance identified from real data applications in an unbiased fashion. Specifically, Hao et al. (2018) performs cross checking with previous literature quantitatively via PubMed search. The intuition behind Hao's approach is that, if a tissue is truly relevant to a given trait, then the number of previous biomedical researches would have been carried out on the tissue for the trait. Consequently, the relevance of a tissue to a trait can be measured by the number of previous publications on the trait-tissue pair. Therefore, for each trait-tissue pair, Hao et al. (2018) used the names of trait and tissue as input and counted the number of publications that contain the input values either in the abstract or in the title. For example, for the schizophrenia-CNS trait-tissue pair, they conducted the search by using "schizophrenia [Title/Abstract] AND (CNS [Title/Abstract] OR brain [Title/Abstract] OR central nervous system [Title/Abstract] OR neuron [Title/Abstract] OR glia [Title/Abstract])." By counting the number of previous publications on the trait-tissue pair, Hao et al. (2018) provides a somewhat ground truth for quantifying and comparing the inferred trait-tissue relevance results. For example, PubMed yielded 17,720 hits for the pair of schizophrenia-CNS, which covers 63.8% of all schizophrenia-tissue search results from the previous literatures, supporting the relevance between CNS and schizophrenia. By performing PubMed search, Hao et al. (2018) shows that certain histone modification marks often provide more information than others. A follow up study using similar PubMed search approach also shows that histone modifications are more informative in inferring trait-tissue relevance than using either the marginal expression information or gene co-expression network information extracted from gene expression studies (Shang et al., 2020b).

Existing methods are primarily developed to take advantage of one particular genomic information for trait-tissue relevance inference. As we summarized in the review, some methods make use of histone modification marks (for example, sLDSC and SMART) while some other methods make use of gene expression data (for example, LDSC-SEG and RolyPoly). However, different genomic information may contain complementary information for trait-tissue relevance inference. Indeed, Finucane et al. (2018) found that one function annotation may be more preferable than another. The same study thus proposed ways to combine two annotations together either by creating a joint synthetic annotation or by combining p-values from analyses of the two annotations separately. A follow up method, SMART, formally models multiple genomic annotations jointly with a multivariate statistical model to improve the accuracy of trait-tissue relevance inference (Hao et al., 2018). SMART found that substantial accuracy gain can be achieved by combining multiple genomic annotations than using one annotation at a time. Besides methodology development to directly incorporate multiple annotations for trait-tissue relevance inference, methods have also been developed to combine multiple annotations into a single, more interpretable and more informative annotation. For example, GenoSkyline creates synthetic annotation based on a variety of epigenetic annotations (Lu et al., 2016). An updated version of GenoSkyline, GenoSkyline-Plus, can now incorporate

both RNA-seq data and DNA methylation data in addition to epigenetic annotations to produce functional epigenetic annotations across 127 tissues and cell types (Lu et al., 2017). A similar method, FUMA, is a recently developed web-based platform that can annotate GWAS significant SNPs for functional consequences on genes, CADD scores, and chromatin states in 127 tissues and cell types (Watanabe et al., 2017). Similarly, in gene expression studies, while existing approaches use either the list of tissue-specifically expressed genes, tissue-specific gene expression levels, or tissue-specific gene co-expression pattern, combining the use of all the information together may have added benefits. Therefore, developing statistical methods to incorporate multiple genomic data types as well as multiple aspects of the same data type will likely yield more accurate tissue-trait relevance in the future. Beyond the scope of our review on trait-tissue relevance, we would add a word for GWAS. GWAS has been developed and used for nearly two decades and reported over 200,000 trait-SNP associations (GWAS catalog as of Dec 15, 2020). However, sample size is always a controversial issue. Current GWAS is toward larger and larger sample sizes in order to discover novel SNPs, however, the "overly-identified" SNPs are often lack of meaningful biological explanations. In contrast, small sample size typically cannot detect any signals. The first issue is now relatively well-studied, for example fine-mapping, gene-based test, etc. We think that the second issue is worth more investigations in the field of GWAS. In addition, factors that determine the phenotype/disease are complex and various, further questions include when and how, i.e., what, when, and how a factor/factors determines a phenotype/disease. We believe that all of the theoretical, computational and experimental work are very meaningful to explore the "truth" of how genome affects "us" and makes "us" different.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This study was supported by the National Institutes of Health (NIH) Grant R01HG009124 and the National Science Foundation (NSF) Grant DMS1712933.

REFERENCES

Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., et al. (2015). The psychencode project. *Nat. Neurosci.* 18, 1707–1712. doi: 10.1038/nn.4156

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* 17:63. doi: 10.1186/s13059-016-0927-y

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24. doi: 10.1101/gr.155192.113

Cai, M., Chen, L. S., Liu, J., and Yang, C. (2020). IGREX for quantifying the impact of genetically regulated expression on phenotypes. *NAR Genomics Bioinformatics*. 2:lqaa010. doi: 10.1093/nargab/lqaa010

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

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

Carithers, L. J., and Moore, H. M. (2015). *The Genotype-Tissue Expression (GTEx) Project*. New York, NY: Mary Ann Liebert, Inc.

Chen, M., Cho, J., and Zhao, H. (2011). Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.* 7:e1001353. doi: 10.1371/journal.pgen.1001353

Chen, W. M., Broman, K. W., and Liang, K. Y. (2004). Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and haseman-elston regression. *Genet. Epidemiol.* 26, 265–272. doi: 10.1002/gepi.10315

de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 9:e1003608. doi: 10.1371/journal.pgen.1003608

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336. doi: 10.1038/nature14222

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343. doi: 10.1038/nature13835

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235. doi: 10.1038/ng.3404

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

Fornito, A., Zalesky, A., and Breakspear, M. (2015). The connectomics of brain disorders. *Nat. Rev. Neurosci.* 16, 159–172. doi: 10.1038/nrn3901

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* 19, 1442–1453. doi: 10.1038/nn.4399

Gamazon, E. R., Segrè A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nat. Genet.* 50, 956–967. doi: 10.1038/s41588-018-0154-4

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Gormley, P., Anttila, V., Winsvold, B. S., Palta, P., Esko, T., Pers, T. H., et al. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat. Genet.* 48, 856–866. doi: 10.1038/ng.3598

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., et al. (2012). Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nat. Genet*. 44, 1084–1089. doi: 10.1038/ng.2394

Grunze, H. (2015). "Bipolar disorder," in *Neurobiology of Brain Disorders*, 655–673. doi: 10.1016/B978-0-12-398270-4.00040-9

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

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

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

He, Z., Xu, B., Lee, S., and Ionita-Laza, I. (2017). 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

Hou, L., Chen, M., Zhang, C. K., Cho, J., and Zhao, H. (2014). Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Hum. Mol. Genet*. 23, 2780–2790. doi: 10.1093/hmg/ddt668

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet*. 89, 496–506. doi: 10.1016/j.ajhg.2011.09.002

Jia, P., and Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum. Genet*. 133, 125–138. doi: 10.1007/s00439-013-1377-1

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., et al. (2014). Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6131–6138. doi: 10.1073/pnas.1318948111

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., et al. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*. 10:e1004722. doi: 10.1371/journal.pgen.1004722

Kim, S. S., Dai, C., Hormozdiari, F., van de Geijn, B., Gazal, S., Park, Y., et al. (2019). Genes with high network connectivity are enriched for disease heritability. *Am. J. Hum. Genet*. 104, 896–913. doi: 10.1016/j.ajhg.2019.03.020

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol*. 12:e1004714. doi: 10.1371/journal.pcbi.1004714

Lan, W., Fang, Z., Wang, H., and Tsai, C.-L. (2018). Covariance matrix estimation via network structure. *J. Bus. Econom. Stat*. 36, 359–369. doi: 10.1080/07350015.2016.1173558

Lang, U. E., Puls, I., Müller, D. J., Strutz-Seebohm, N., and Gallinat, J. (2007). Molecular mechanisms of schizophrenia. *Cell. Physiol. Biochem*. 20, 687–702. doi: 10.1159/000110430

Lappalainen, T., Sammeth, M., Friedländer, M. R., Ac't Hoen, P., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531

Li, Y., and Kellis, M. (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res*. 44:e144. doi: 10.1093/nar/gkw627

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22. doi: 10.1093/biomet/73.1.13

Lu, Q., Powles, R. L., Abdallah, S., Ou, D., Wang, Q., Hu, Y., et al. (2017). Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet*. 13:e1006933. doi: 10.1371/journal.pgen.1006933

Lu, Q., Powles, R. L., Wang, Q., He, B. J., and Zhao, H. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet*. 12:e1005947. doi: 10.1371/journal.pgen.1005947

Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet*. 100, 473–487. doi: 10.1016/j.ajhg.2017.01.031

Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol*. 6, 386–398. doi: 10.1038/nrm1645

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., et al. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–749. doi: 10.1126/science.1242429

Montgomery, S. B., and Dermitzakis, E. T. (2011). From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet*. 12, 277–282. doi: 10.1038/nrg2969

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., et al. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* 430, 743–747. doi: 10.1038/nature02797

Nica, A. C., and Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet*. 17, R129–R134. doi: 10.1093/hmg/ddn285

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

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

Ongen, H., Brown, A. A., Delaneau, O., Panousis, N. I., Nica, A. C., and Dermitzakis, E. T. (2017). Estimating the causal tissues for complex traits and diseases. *Nat. Genet*. 49, 1676–1683. doi: 10.1038/ng.3981

Pei, G., Dai, Y., Zhao, Z., and Jia, P. (2019). deTS: tissue-specific enrichment analysis to decode tissue specificity. *Bioinformatics* 35, 3842–3845. doi: 10.1093/bioinformatics/btz138

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun*. 6:5890. doi: 10.1038/ncomms6890

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet*. 94, 559–573. doi: 10.1016/j.ajhg.2014.03.004

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res*. 21, 447–455. doi: 10.1101/gr.112623.110

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* 43, 353–360. doi: 10.1093/biomet/43.3-4.353

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302. doi: 10.1038/nature01434

Shang, L., Smith, J. A., Zhao, W., Kho, M., Turner, S. T., Mosley, T. H., et al. (2020a). Genetic architecture of gene expression in European and African Americans: an eQTL mapping study in GENOA. *Am. J. Hum. Genet*. 106, 496–512. doi: 10.1016/j.ajhg.2020.03.002

Shang, L., Smith, J. A., and Zhou, X. (2020b). Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. *PLoS Genet*. 16:e1008734. doi: 10.1371/journal.pgen.1008734

Slowikowski, K., Hu, X., and Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* 30, 2496–2497. doi: 10.1093/bioinformatics/btu326

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Stunnenberg, H. G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., et al. (2016). The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell* 167, 1145–1149. doi: 10.1016/j.cell.2016.11.007

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., et al. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet*. 45, 124–130. doi: 10.1038/ng.2504

Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., et al. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet*. 97, 139–152. doi: 10.1016/j.ajhg.2015.05.016

Uhlhaas, P. J., and Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nat. Rev. Neurosci*. 11, 100–113. doi: 10.1038/nrn2774

Urry, M. J., and Sollich, P. (2013). Random walk kernels and learning curves for gaussian process regression on random graphs. *J. Mach. Learn. Res*. 14, 1801–1835. arXiv:1211.1328v2.

Watanabe, K., Mirkov, M. U., de Leeuw, C. A., van den Heuvel, M. P., and Posthuma, D. (2019). Genetic mapping of cell type specificity for complex traits. *Nat. Commun*. 10:3222. doi: 10.1038/s41467-019-11181-1

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun*. 8:1826. doi: 10.1038/s41467-017-01261-5

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Xiao, X., Chang, H., and Li, M. (2017). Molecular mechanisms underlying noncoding risk variations in psychiatric genetic studies. *Mol. Psychiatry* 22, 497–511. doi: 10.1038/mp.2016.241

Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., et al. (2019). Testing and controlling for horizontal pleiotropy with the probabilistic mendelian randomization in transcriptome-wide association studies. *bioRxiv* 2019:691014. doi: 10.1101/691014

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

Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann. Appl. Stat*. 11, 2027–2051. doi: 10.1214/17-AOAS1052

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

Zhu, H., and Zhou, X. (2020). Transcriptome-wide association studies: a view from Mendelian randomization. *Quant. Biol*. 17, 1–15. doi: 10.1007/s40484-020-0207-4

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 March 2021
doi: 10.3389/fgene.2021.612045

Subsampling Technique to Estimate Variance Component for UK-Biobank Traits

Ting Xu[1†], *Guo-An Qi*[2†], *Jun Zhu*[2], *Hai-Ming Xu*[2] *and Guo-Bo Chen*[3,4*]

[1] Department of Mathematics, Zhejiang University, Hangzhou, China, [2] Department of Agricultural and Biotechnology, Zhejiang University, Hangzhou, China, [3] Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Clinical Research Institute, Hangzhou, China, [4] Key Laboratory of Endocrine Gland Diseases of Zhejiang Province, Hangzhou, China

The estimation of heritability has been an important question in statistical genetics. Due to the clear mathematical properties, the modified Haseman–Elston regression has been found a bridge that connects and develops various parallel heritability estimation methods. With the increasing sample size, estimating heritability for biobank-scale data poses a challenge for statistical computation, in particular that the calculation of the genetic relationship matrix is a huge challenge in statistical computation. Using the Haseman–Elston framework, in this study we explicitly analyzed the mathematical structure of the key term $tr(\boldsymbol{K}^T\boldsymbol{K})$, the trace of high-order term of the genetic relationship matrix, a component involved in the estimation procedure. In this study, we proposed two estimators, which can estimate $tr(\boldsymbol{K}^T\boldsymbol{K})$ with greatly reduced sampling variance compared to the existing method under the same computational complexity. We applied this method to 81 traits in UK Biobank data and compared the chromosome-wise partition heritability with the whole-genome heritability, also as an approach for testing polygenicity.

Keywords: polygenicity, UK Biobank, subsampling estimator, effective number of markers, Haseman-Elston regression

OPEN ACCESS

Edited by:
Min Zhang,
Purdue University, United States

Reviewed by:
Cen Wu,
Kansas State University, United States
Wenjian Bi,
University of Michigan, United States

****Correspondence:***
Guo-Bo Chen
chenguobo@gmail.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: *30 September 2020*
Accepted: *18 January 2021*
Published: *05 March 2021*

Citation:
Xu T, Qi G-A, Zhu J, Xu H-M and Chen G-B (2021) Subsampling Technique to Estimate Variance Component for UK-Biobank Traits. Front. Genet. 12:612045. doi: 10.3389/fgene.2021.612045

INTRODUCTION

Given the increasing sample size and sequencing capability, high-throughput genetic data is presented as the standard input that challenges statistical computation. For example, in the estimation of heritability for complex traits using all markers concurrently, both (i) constructing the genetic relationship matrix [GRM, denoted as $\boldsymbol{K}$ and the mathematical expression can be seen in section Materials and Methods, with its computational cost $\mathcal{O}(MN^2)$] and (ii) the estimation of heritability using linear mixed model [$\mathcal{O}(N^3)$] are computationally expensive (Yang et al., 2010). In order to alleviate computational burden, various solutions have been proposed. Modified Haseman–Elston regression (HE) can be used to estimate heritability with reduced computational cost in the estimation step [$\mathcal{O}(N^2)$], but the construction of GRM is still needed (Chen, 2014). Using summary statistics, such as those estimated from the genome-wide association study (GWAS), rather than individual-level data, can provide a theoretical equivalence estimate of the heritability under the assumption that the source of summary statistics and the linkage disequilibrium (LD) reference are homogeneous (Bulik-Sullivan et al., 2015), if not always the case.

Even under the HE framework, given the availability of biobank-scale data, such as UK Biobank (UKB) data (Bycroft et al., 2018), the computational cost for GRM poses a challenge for heritability

estimation mentioned procedure above. In order to reduce the computational cost of GRM, recently a randomized estimation of heritability has been introduced by Wu and Sankararaman (2018), called randomized Haseman–Elston regression (RHE), a promising method that can be used for both single-trait and bi-trait analyses (Sankararaman, 2019). This method is built on a hybrid framework which can be applied to biobank-scale data, and a key innovation involved is a quick evaluation for $tr(\boldsymbol{K}^T\boldsymbol{K})$, the trace of the multiplication of GRM with its transpose. Direct computation of $tr(\boldsymbol{K}^T\boldsymbol{K})$ can be time-consuming, at the time cost of $\mathcal{O}(N^2M)$, but in RHE the numerical evaluation of $tr(\boldsymbol{K}^T\boldsymbol{K})$ can be realized via a randomization method expressed in quadric form. However, we found that the sampling variance of RHE in the original report is incorrect because of their wrong derivation (refer to Appendix A3 in Wu and Sankararaman's original report). In this study, we further investigate the statistical property of RHE, in particular the term about $tr(\boldsymbol{K}^T\boldsymbol{K})$, and its relevant extensions.

This report was written for three purposes. First, we found that the provided ***randomization estimate*** for $tr(\boldsymbol{K}^T\boldsymbol{K})$ is correct but with its sampling variance, which is proportional to $tr(\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{K})$, not properly treated in Wu and Sankararaman's original report. We derived and numerically validated the sampling variance of $tr(\boldsymbol{K}^T\boldsymbol{K})$. Second, recently a hybrid routine that can use either individual-level data and summary statistics has also been found (Zhou, 2017; Wu and Sankararaman, 2018), in which ***subsampling technique*** is used to evaluate $tr(\boldsymbol{K}^T\boldsymbol{K})$; however, its sampling variance was not available. We provided a similar method as subsampling but with availability of its analytical sampling variance. Third, we partitioned the heritability based on the ***effective number of markers*** and applied them in the partitioning of heritability for some complex traits in UKB.

MATERIALS AND METHODS

Genetic Relationship Matrix

For a homogenous unrelated sample, its genotypic matrix can be written as $\boldsymbol{X}$, a matrix of N rows—individuals, and M columns—coding the count of the reference allele for a biallelic locus. After standardization for each genotype $\tilde{x}_{kl} = \frac{x_{kl}-2p_l}{\sqrt{2p_lq_l}}$, in which $2p_l$ is the allele frequency and $\sqrt{2p_lq_l}$ the square root of the variance, we can define GRM as $\boldsymbol{K} = \frac{1}{M}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T$. Given $\boldsymbol{K}$, we can easily derive some characters of $\boldsymbol{K}$. Denote $\boldsymbol{K}_o$ as the off-diagonal elements, and it is easy to see that $\mathbb{E}(\boldsymbol{K}_o) = \frac{-1}{N-1}$, because the summation of the diagonal is $N-1$. $var(\boldsymbol{K}_o)$ is the sampling variance of the all off-diagonal elements.

Of note, $var(\boldsymbol{K}_o)$ relates to the concept, so-called ***effective number of markers***, denoted as M_e thereafter. As noticed, M_e is defined as the reciprocal of $var(\boldsymbol{K}_o)$. $M_e = \frac{1}{var(\boldsymbol{K}_o)} = \frac{M^2}{M+\sum_{l_1\neq l_2}^{M} E\left(\rho_{l_1l_2}\right)^2}$, in which $\mathbb{E}(\rho_{l_1l_2})$ is the expected Pearson's correlation between the l_1^{th} and l_2^{th} loci. Alternatively, $M_e = \frac{1}{E\left(\overline{\rho}_{l_1l_2}\right)^2}$. It is known that for a population, the averaged linkage disequilibrium across the genome is nearly a constant given the markers; in other words, M_e is a constant genetic parameter. The definition of M_e in this report allows researchers to calculate M_e based on a reference population of the same origin to the population in question. Similarly, $M_{e.c}$ represents the averaged LD for any pair of markers on the c^{th} chromosomes.

As the causal variants are hardly observed directly, their relationship with markers are surrogated by relationship between markers, as reflected in M_e. As M_e is a critical parameter in many genetic applications, a conceptional parameter is its involvement in genetic prediction (Dudbridge and Wray, 2013), or power calculation for the estimation of heritability (Visscher et al., 2014). In the estimation for variance components, as shown below, M_e is a key parameter.

Haseman–Elston Regression Framework for the Estimation of Heritability

Haseman–Elston regression (HE) has been initially proposed for the linkage analysis (Haseman and Elston, 1972). With its original kernel relatedness between sib pairs via linkage replaced by linkage disequilibrium for unrelated samples, the modified HE can be used for the estimation of heritability (Chen, 2014). Due to its clear mathematical property, HE has been found a bridge to connect and develop various parallel methods for the estimation of heritability, such as LD score regression that estimates heritability and uses summary statistics from GWAS (Bulik-Sullivan et al., 2015; Zhou, 2017).

However, LD score regression is based on various assumptions that may or may not be met in practice. LD score regression uses SNPs in a sliding window instead of all genome-wide SNPs to calculate LD scores, which will lose efficiency if heterogeneity exists between the reference population and the population that generates the GWAS summary statistics. If we directly use individual-level data, the time cost will be unaffordable, such as for the restricted maximum likelihood estimation method (REML); in contrast, a method of moment (MoM) can provide equivalent estimation for the heritability for complex traits.

We assume that

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{e};\ \boldsymbol{\beta} \sim \mathcal{N}\left(0, \frac{h^2}{M}\boldsymbol{I}\right);\ \boldsymbol{e} \sim\ \mathcal{N}\left(0, \sigma_e^2\boldsymbol{I}\right)$$

in which $\boldsymbol{y}$ is the standardized phenotype for a trait of interest, $\tilde{\boldsymbol{X}}$ is the standardized genotypic matrix of N individuals and M the biallelic markers, β is the additive effect associated with each marker, $\boldsymbol{e}$ is the residual, h^2 is the SNP heritability, and σ_e^2 is the residual variance. It is easy to know that $var\left(\boldsymbol{y}\right) = \mathbb{E}\left(\boldsymbol{y}\boldsymbol{y}^T\right) - \mathbb{E}\left(\boldsymbol{y}\right)\mathbb{E}\left(\boldsymbol{y}^T\right) = \frac{h^2}{M}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T + \sigma_e^2\boldsymbol{I} = h^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}$.

Estimation for Heritability *via* Modified Randomized Haseman–Elston Regression

Consequently, we extend the work by Wu and Sankararaman (2018); the moment estimator is to minimize

$$\mathcal{Q} = tr\left\{\left[\boldsymbol{y}\boldsymbol{y}^T - \left(h^2\boldsymbol{K} + \sigma_e^2\boldsymbol{I}\right)\right]^2\right\}$$

By taking the differentiation in terms of h^2 and σ_e^2, we have

$$\begin{cases} \frac{\partial \mathcal{Q}}{\partial h^2} = tr\left\{h^2 \boldsymbol{K}^T\boldsymbol{K} + \sigma_e^2 \boldsymbol{K} - \boldsymbol{y}\boldsymbol{y}^T\boldsymbol{K}\right\} = 0 \\ \frac{\partial \mathcal{Q}}{\partial \sigma_e^2} = tr\left\{h^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I} - \boldsymbol{y}\boldsymbol{y}^T\boldsymbol{I}\right\} = 0 \end{cases}$$

After algebra, we have the normal equations below:

$$\begin{bmatrix} tr\left(\boldsymbol{K}^T\boldsymbol{K}\right) & tr\left(\boldsymbol{K}\right) \\ tr\left(\boldsymbol{K}\right) & N \end{bmatrix} \begin{bmatrix} \hat{h}^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^T\boldsymbol{I}\boldsymbol{y} \end{bmatrix} \quad (1)$$

The estimator for $\hat{h}^2$ can be written as

$$\hat{h}^2 = \frac{\boldsymbol{y}^T\left(\boldsymbol{K} - \boldsymbol{I}\right)\boldsymbol{y}}{tr\left(\boldsymbol{K}^T\boldsymbol{K}\right) - N} \quad (2)$$

However, different computational strategies deal with the computational expensive part for both the numerator and the denominator. In particular, for the numerator, $\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}$ can be decomposed as $\boldsymbol{y}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\boldsymbol{y}/M$, and $\boldsymbol{y}^T\tilde{\boldsymbol{X}}$, equal to $\left(\tilde{\boldsymbol{X}}^T\boldsymbol{y}\right)^T$. Each element $\tilde{\boldsymbol{X}}_j^T\boldsymbol{y}$ of $\tilde{\boldsymbol{X}}^T\boldsymbol{y}$ is just the regression coefficient between the jth marker and $\boldsymbol{y}$ that can be computed via simple linear regression, or multivariate linear regression if covariates are included. It is easy to recognize that $\boldsymbol{y}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\boldsymbol{y}/M$ follows χ_1^2 after scaling by the sample size N, and a possible non-central parameter related to the heritability of the trait. Alternatively, we derive the mathematical expectation $\mathbb{E}\left(\boldsymbol{y}^T\boldsymbol{K}\boldsymbol{y}\right) = N\mathbb{E}\left(\chi_{1|h^2}^2\right) = N(1 + Nh^2\bar{r}^2)$, in which $\bar{r}^2$ is the averaged LD score between a marker to a causal variants in LD.

The denominator involves the trace of $\boldsymbol{K}^T\boldsymbol{K}$, a high-order function for GRM. Alternatively, according to the property of the trace of a matrix, it can be calculated that $tr(\boldsymbol{K}^T\boldsymbol{K}) = \sum_{i,j}^N K_{i,j}^2$, a summation of the square of each element in $\boldsymbol{K}$. From the first glance, it seems inevitable to compute $\boldsymbol{K}$, the computational cost of which is $\mathcal{O}(N^2M)$, a substantial cost given a large sample size, such as for UKB of about 500,000 samples (Bycroft et al., 2018). In order to have a proper estimate for $tr(\boldsymbol{K}^T\boldsymbol{K})$ but reduce computation cost, three methods are proposed for estimating $tr(\boldsymbol{K}^T\boldsymbol{K})$.

Estimating tr($\boldsymbol{K}^T\boldsymbol{K}$)

We present three methods in estimating $tr(\boldsymbol{K}^T\boldsymbol{K})$. Sampling method I has been proposed by Wu and Sankararaman, but we provide its correct sampling variance, which was incorrectly given in their original report (Wu and Sankararaman, 2018). Sampling method II derives the expectation of $tr(\boldsymbol{K}^T\boldsymbol{K})$ and estimates it in a reference population with the similar genetic origin of the population of question. Sampling method III slightly modifies method II if the reference population is big and yields smaller sampling variance of $tr(\boldsymbol{K}^T\boldsymbol{K})$ than that of method II.

Sampling Method I: The Randomized Estimator With Corrected Analytical Sampling Variance

Using randomized estimation, an unbiased estimator L_B is employed to estimate $tr(\boldsymbol{K}^T\boldsymbol{K})$ in RHE (Wu and Sankararaman, 2018). The rational for a randomized estimate is as below:

$$L_B = \frac{1}{B}\frac{1}{M^2}\sum_b^B tr(\boldsymbol{z}_b^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{z}_b) = tr(\boldsymbol{K}^T\boldsymbol{K})$$

In each iteration, a vector $\boldsymbol{z}$, of length N, is generated from the standard normal distribution. As long as $\boldsymbol{z}$ has been generated B time and B is large enough, it is guaranteed to approach $tr(\boldsymbol{K}^T\boldsymbol{K})$. As $\boldsymbol{z}_b^T\boldsymbol{X}\boldsymbol{X}^T$ can be calculated easily, the computational cost is $\mathcal{O}(NMB)$. Then, L_B can be plugged into a normal equation (Equation 2).

In Wu and Sankararaman's original report, the sampling variance of L_B was given as $var\left(L_B\right) = 2tr\left(\boldsymbol{K}^T\boldsymbol{K}\right)/B$, which was incorrect, and the correct one should have been

$$\begin{aligned} var\left(L_B\right) &\equiv Var\left(\frac{1}{B}\sum\nolimits_{b=1}^B \boldsymbol{z}_b^T\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{z}_b\right) \\ &= \frac{1}{B^2}\sum\nolimits_{b=1}^B Var\left(\boldsymbol{z}_b^T\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{z}_b\right) \\ &= \frac{1}{B^2}\sum\nolimits_{b=1}^B 2tr\left(\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{K}\right) = \frac{2tr\left(\boldsymbol{K}^4\right)}{B} \end{aligned} \quad (3)$$

The derivation of the penultimate step uses the quadratic variance calculation formula. The sampling variance of L_B is proportional to $tr(\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{K})$, the computational cost of which is likely to be infeasible for biobank-scale data. However, its practical sampling variance can be estimated from B iterations above $var\left(L_B\right) = \sum_{j=1}^B \left(L_{B_j} - \overline{L}_B\right)^2/B$.

Sampling Method II: Estimating $tr(\boldsymbol{K}^T\boldsymbol{K})$ by Subsampling

In an alternative route, we bypass the direct computation of $tr(\boldsymbol{K}^T\boldsymbol{K})$. It is shown that $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right) = N^2/M_e + N$ for unrelated samples (see **Supplementary Notes**). N is the sample size, a known parameter; we only need to estimate M_e. As noted above, M_e can be estimated by subsampling a proportion of the study population (**Figure 1**) or by a reference population of the same origin with the population of study (Zhou, 2017). Thus, we can estimate M_e using a small proportion of the sample, as long as we can estimate $\hat{M}_e$; we can easily get the estimator of $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right)$. We define a new L_S estimator: $L_S \equiv N^2/\hat{M}_e + N$. It is an unbiased estimate (see **Supplementary Notes**). Suppose the sample size of subsample is s, there are $s^2/2$ off-diagonal elements and it takes $\mathcal{O}(s^2M/2)$ time to calculate $\hat{M}_e$. The sampling variance of L_S is, using the Delta method, $\left(L_s^{'}\right)^2 var\left(\hat{M}_e\right) = \frac{N^4}{M_e^4}\sigma_{\hat{M}_e}^2$, in which $L_S^{'} = -\frac{N^2}{M_e^2}$ the first derivative of L_S and $\sigma_{\hat{M}_e}^2$ the sampling variance for $\hat{M}_e$. $\sigma_{\hat{M}_e}^2$ is not directly known but can be directly estimated in the third method proposed below.

Sampling Method III: Estimating $tr(\boldsymbol{K}^T\boldsymbol{K})$ *via* Shotgun Randomization

However, $\sigma_{\hat{M}_e}^2$ in method II is not analytical probably because each individual will be involved s times in the estimation of variance. With slightly modification, we developed a new

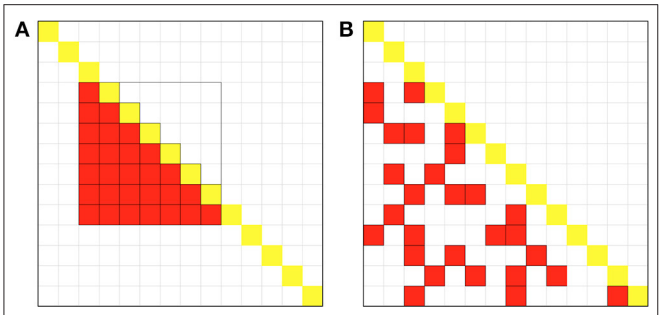

FIGURE 1 | Schematic diagram of the sampling methods for the two estimators. The large square represents the entire genetic relationship matrix (GRM) with sample size $N = 14$, and each small an element in GRM, and the yellow the diagonal element of the matrix. **(A)** Schematic diagram of sampling methods for the L_S estimator. The red squares in the lower triangular of GRM represent the pairs of samples for L_S. This figure shows that when $B = 2$ (the parameter for the L_B estimator given by Wu and Sankararaman, 2018), we set subsample size $s = \sqrt{2BN} \approx 7$ for L_S to guarantee the same computational cost. **(B)** Schematic diagram of sampling methods for the L_T estimator. Equally total number ($n = BN = 28$) of GRM elements in the lower triangular in red represents the pair of the samples for L_T. The elements are drawn as if they are randomly, as shot by a gun, to reduce correlation between individuals.

estimator L_T (lower triangle shotgun sampling estimator) to estimate $tr(\boldsymbol{K}^T\boldsymbol{K})$ in RHE. The difference in sampling schemes between methods II and III can be visualized as **Figure 1**. Given the whole GRM, method II samples a square matrix of size $s \times s$ after rearrangement and calculates half elements, whereas method III randomly samples $n = s^2/2$ elements in the whole GRM without replacement so as to reduce overlapping of samples (**Figure 1**). This sampling idea is similar to the shotgun method in the first-generation DNA sequencing technology, so we call the method III shotgun sampling estimator.

Given a random subset of n elements $A \subseteq \{1, 2, \cdots, N(N-1)/2\}$, we define

$$L_T \equiv N + \frac{N^2}{n}\sum\nolimits_{i=1}^{n} K_{o_{A_i}}^2$$

It can be proved that L_T is an unbiased estimator of $tr(\boldsymbol{K}^T\boldsymbol{K})$ with its sampling variance $N^4 var(\boldsymbol{K}_o{}^2)/n$, which can be estimated by $N^4 var(\boldsymbol{K}_{o_{A_i}}^2)/n$ (see **Supplementary Notes**). Therefore, we can get the unbiased estimate of $tr(\boldsymbol{K}^T\boldsymbol{K})$ and its sampling variance at the same time in one step. It does not need to calculate all the elements in $\boldsymbol{K}_o$ but the corresponding pairs of the individuals, and to calculate the mean of the product of all their genetic values. Therefore, each item in the summation can be computed in $\mathcal{O}(M)$, and the total running time is $\mathcal{O}(nM)$.

The Estimation of Variance Components and Its Sampling Variance

If we replace $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right)$ with its subsampling estimators, we can get the synthesized estimator for heritability

$$\hat{h}^2 = \frac{N\mathbb{E}\left(\chi_{1|h^2}^2\right) - N}{N^2/\hat{M}_e} = \hat{M}_e h^2 \bar{r}^2$$

where $\mathbb{E}\left(\chi_{1|h^2}^2\right)$ is the mean of χ_1^2 for each SNP with or without the adjustment of covariates. Using the Delta method, we show in **Supplementary Notes** that the variance of $\hat{h}^2$ can be formulated as

$$\sigma_{h^2} \approx \frac{2M_e}{N^2} + {M_e}^2 \frac{\sigma^2_{\boldsymbol{K}_o{}^2}}{n}(h^2)^2$$

and each item in the above formula is estimable, then we can get the variance estimator of the variance component

$$\hat{\sigma}_{h^2} = \frac{2\hat{M}_e}{N^2} + \hat{M}_e^2 \frac{\sigma^2_{\boldsymbol{K}_{o_{A_i}}{}^2}}{n}(\hat{h}^2)^2 \tag{4}$$

Except for $\hat{h}^2$, all other parts involved are independent to the phenotype, so given a specific sample of question, the estimator has a linear relationship with the square of the estimated heritability.

Genetic Partitioning of Heritability

Yang et al. (2010) estimated the chromosome-wise partitioned heritability and found that the heritability of complex trait, such as human height is proportional to the length of the chromosome, that is, proportional to the number of causal variants. Some researchers gave more weight to large effects to explain heritability and to study polygenicity (O'Connor et al., 2019; Yang and Zhou, 2020). In this report, we instead calculated heritability based on $\hat{M}_e$ and compared the chromosome-wise partition heritability with the whole-genome heritability

$$\hat{h}_C^2 = \sum_{c=1}^{22} \frac{N\mathbb{E}\left(\chi_{1|h^2}^2\right) - N}{N^2/\hat{M}_{e.c}} = \sum_{c=1}^{22} \hat{M}_{e.c} h_c^2 \bar{r}_c^2 \tag{5}$$

in which $\hat{M}_{e.c}$ is the effective of markers for the c^{th} chromosome and $\bar{r}_c^2$ is the averaged squared correlation between a casual variant and a marker on the c^{th} chromosome. Under the assumption of polygenicity, $\hat{h}_C^2 = \frac{h^2}{\hat{M}_e}\sum_{c=1}^{22}\bar{r}_c^2$, and the ratio between $\frac{h^2}{h_C^2} = \frac{\bar{r}^2}{\sum_{c=1}^{22} r_C^2}$. As both $\bar{r}^2$ and $\sum_{c=1}^{22}\bar{r}_c^2$ are unknown, we use $\frac{1}{M_e}$ and $\frac{1}{\sum_{c=1}^{C} M_{e.c}}$ as the surrogates for $\bar{r}^2$ and $\sum_{c=1}^{22}\bar{r}_c^2$.

By breaking the GRM of the whole genome $\boldsymbol{K}$ in Equation (1) into the GRMs for 22 autosomes, we can also estimate the chromosome heritability jointly in one model. This method has to inverse a 23×23 matrix. Under the assumption that the genotype of each chromosome contains the same N individuals, the inversed matrix is completely upon N and $M_{e.c}$, so a computation cost linear to 23, without bothering the conventional matrix inversion procedure, a computation cost of 23^3, can be written down analytically. In particular, the c^{th} diagonal element of the inverse matrix is $M_{e.c}/N^2$, and the last column/row is $-M_{e.c}/N^2$. For more details, please see **Supplementary Notes**.

Heritability for the Weighted Genetic Relationship Matrix

Given the definition of the weighted GRM

$$\boldsymbol{K}_w = \frac{\sum_{l=1}^{M}(x_{il}-2p_l)(x_{jl}-2p_l)}{\sum_{l=1}^{M}2p_lq_l}$$

we can get an estimator of the weighted heritability as well as its variance estimator based on weighted GRM through a similar derivation

$$\hat{h}_w^2 = \frac{N\frac{\sum_{l=1}^{M}2p_lq_l\chi^2_{1|h^2,l}}{\sum_{l=1}^{M}2p_lq_l}-N}{N^2/\hat{M}_{ew}},\ \hat{\sigma}^2_{h_w^2} = \frac{2\hat{M}_{ew}}{N^2}+\hat{M}^2_{ew}\frac{\sigma^2_{\boldsymbol{K}_{o_{A_i}}{}^2}}{n}(\hat{h}_w^2)^2$$

where $\hat{M}_{ew}$ is the estimation of M_e for $\boldsymbol{K}_w$, and $\chi^2_{1|h^2,l}$ is the square of the z-score for the l^{th} SNP with or without the adjustment of covariates. The weighted chromosome-wise partition heritability can be expressed as

$$\hat{h}^2_{Cw} = \sum_{c=1}^{22}\frac{N\frac{\sum_{l=1}^{M_c}2p_lq_l\chi^2_{1|h^2,l}}{\sum_{l=1}^{M_c}2p_lq_l}-N}{N^2/\hat{M}_{ew.c}}$$

where $\hat{M}_{ew.c}$ is the estimation of weighted M_e for the c^{th} chromosome and M_c is the number of SNP of the c^{th} chromosome.

Connection to Other Estimators

The BOLT-LMM method (Loh et al., 2015) might be the most widely used method in the field of heritability estimation for large-scale data. Theoretically, the computational complexity of BOLT-LMM is $\mathcal{O}(PMN^{1.5})$, where P is the number of iterations for convergence. In the L_T estimator, the subsample size $n \ll N^{1.5}$, so our calculation time is less than BOLT-LMM in theory. In terms of actual calculation, the L_B estimator used less calculation time to get an accuracy similar to BOLT-LMM (Wu and Sankararaman, 2018); the variance of our estimators is about an order of magnitude smaller than L_B under the same calculation time. Thus, our method is better than BOLT-LMM in calculation accuracy and time. In terms of memory, the memory complexity of BOLT-LMM is $\mathcal{O}(NM/4)$, while the memory of our subsampling estimators is proportional to M and the subsample size s in the L_S estimator, which generally does not exceed 10% of the total sample size.

Given the availability of the estimators and their sampling variances, it is able to evaluate the statistical power of the estimators and estimate the sample size for the given type I and type II error rates. Under the null hypothesis $h^2 = 0$, the sampling variance for the additive variance component can be reduced to $\hat{\sigma}_{h^2} \approx \frac{2\hat{M}_e}{N^2}$, which are equivalent to that of REML (Visscher et al., 2014). It is consequently known that the statistical power of the presented method will be equivalent to REML. In contrast, the original Haseman–Elston regression has doubled sampling variances where $\hat{\sigma}_{h^2} \approx \frac{4\hat{M}_e}{N^2}$ (Chen, 2014), because the original HE regression only uses the off/upper-diagonal of the matrix, as presented in the numerator above. The connection to LD score regression is obviously too; here, the whole M_e can be seen as a genome LD score, rather than being partitioned into genomic bins.

RESULTS

Simulation Results for the Evaluation of $tr(K^TK)$

In the simulation and in the real data, we compared the mean and variance of the three estimators L_B, L_S, and L_T, and the results are as presented in **Figure 2**. We took $n = BN$ and $s = \sqrt{2BN}$ to make sure the three estimators are under the equal computational cost of $\mathcal{O}(NMB)$ (see **Figure 1** for an example). In the simulation, we set the genotype in two ways: (1) The minor allele frequency (MAF) of each SNP was randomly generated from a uniform distribution between 0.03 and 0.5, and two levels of LD (linkage disequilibrium, in terms of Lewontin's D', the normalized LD parameter) strength were set as 0–0.2 (weak LD) and 0.6–0.8 (strong LD) with the SNP number $M = 2{,}000$, 5,000 and sample size $N = 500$, 1,000, and 2,000, respectively. (2) The real genotype data consisted of 12,980 adjacent markers on chromosome 22 of 2,000 randomly sampled unrelated white British individuals in UKB. B was set from 5 to 50 and repeated 100 times for each to assess the mean and variance of the three estimators.

Across different parameter settings (sample sizes, number of loci, MAF, and LD), it yielded a similar pattern for the evaluated results of $tr(\boldsymbol{K}^T\boldsymbol{K})$. We chose $M = N = 2{,}000$ and strong LD for detailed presentation, and the rest were shown in **Supplementary Figures 1, 2**). **Figure 2** shows that all the three estimators were unbiased. The variance of each of these estimators, as expected, was inversely proportional to B. The real sampling variance of L_B was several times larger than the analytical incorrect result given in Wu and Sankararaman's study (refer to Appendix A3 in their original report) but was consistent with $2tr\left(\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{K}\right)/B$, just the corrected one as derived in this study (Equation 3). The sampling variance of L_T was about an order of magnitude smaller than that of L_B. The simulation results in the real data shown in **Supplementary Figure 3** were consistent with **Figure 2**.

Real Data Analysis for $tr(K^TK)$

We compared the performance of the three estimators L_B, L_S, and L_T in UKB. After quality control, 525,460 autosome SNPs with MAF > 0.01 for 278,788 unrelated British white individuals, whose pairwise genetic relationship coefficient <0.0125, were included for analysis. We set $B = 5$, 10, 20, 40, 60, 80, and 100, and calculated each of the three estimators 100 times to get the mean and the variance for each B. We compared the means of the three estimators with the expected value of $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right) = N^2/\hat{M}_e + N$, where $\hat{M}_e$ was estimated from subsamples; given with $\hat{M}_e \approx 87{,}351$, $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right)$ was expected to be 1,168,573 for each of the three estimators.

The calculation was performed on an Intel(R) Xeon(R) Bronze 3104 CPU @ 1.70-GHz server cluster, and about

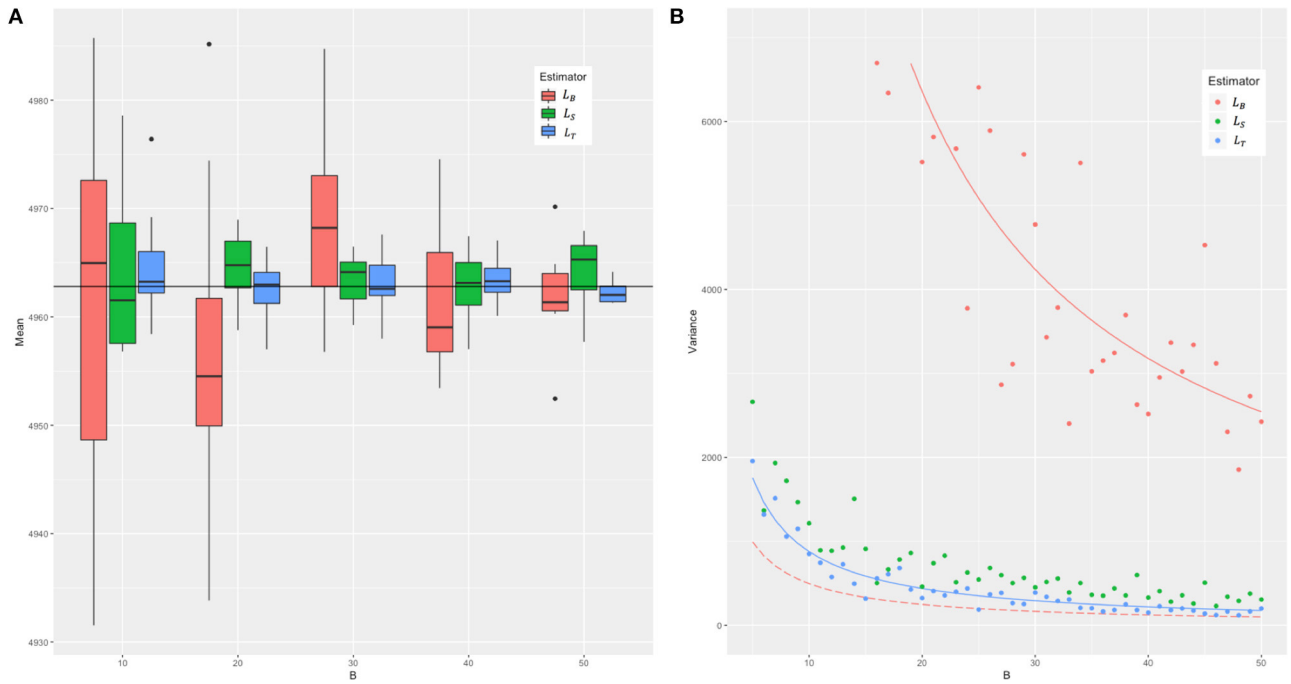

FIGURE 2 | The means and the sampling variances of three estimators in simulation. The genotype data were constituted by 2,000 individuals and 2,000 markers with strong LD (from 0.6 to 0.8 in Lewontin's measure D'). B is the parameter related to the iteration for L_B. For each B, the number of samples of the L_S and L_T estimators was adjusted to ensure that the three estimators have the same computational cost. **(A)** The boxplots of means and **(B)** the variance of 100 random experiments of three estimators for each B. In **(A)**, we combined 10 cases with $B = 5$–14 into the first group, and so on, and the last five cases with $B = 46$–50 were combined into the fifth group. The red boxes represent the means of L_B, green the means of L_S, and blue the means of L_T, respectively. The black horizontal line represents the real value of $tr\,(\boldsymbol{K}^T\boldsymbol{K})$. In **(B)**, the solid red line represents the true sampling variance of the L_B estimator derived in this study, and in contrast the long-dash red line the sampling variance of the L_B estimator incorrectly given by Wu and Sankararaman (2018). The solid blue line represents the theoretical sampling variance of L_T. The solid green line is not given because we cannot get the theoretical sampling variance of L_S yet.

30 threads were allocated for each calculation. The actual calculation time of the three estimators were basically the same (see **Supplementary Table 1**) and conformed to the theoretical calculation complexity $\mathcal{O}(NMB)$. The variances of the three estimators L_B, L_S, and L_T for $tr(\boldsymbol{K}^T\boldsymbol{K})$ are listed in **Table 1**. In particular, between the randomized estimator and the subsampling estimators, there was a huge difference between their variances. Under the real data, the sampling variance of L_B was large, while the sampling variances of the other two estimators were smaller and the variance of L_T was about half that of L_S. The variances of each of the three estimators decreased with the increasing B, consistent with the simulation.

Chromosome-Wise Partitioning for Heritability

Equation (2) presents how heritability is estimated using all 22 autosomes, and Equation (5) offers an alternative method by summation of chromosome-wise estimation for heritability. For ease of comparison, we only estimated heritability for 81 traits as demonstrated by Ge et al. (2017). We used the first two principal components as the covariates to control the possible population stratification; other covariates were adjusted upon the traits. The chromosome-wise partition heritability was calculated by the summation of the heritability estimated for $M_{e.c}$ for each chromosome (**Table 2**), and the whole-genome heritability was calculated from the GRM of the whole genome.

The estimated heritability of some selected UKB traits is listed in **Table 3** (see **Supplementary Table 2** for all the 81 traits). The heritability of all traits was basically very similar to Ge et al.'s result and within the error range. Several physiological traits, such as height and weight had high heritability, while social traits that were more affected by social factors, such as the duration of certain activities, showed lower heritability. This result was consistent with the mainstream conclusion. The left part of Equation (4) for variance estimators of the whole-genome heritability ($\frac{2\hat{M}_e}{N^2}$) contributed a large part of the total variance (about 0.0017 in 0.002 for $N = 270{,}000$). Although the variances of the $\hat{L}_B$, $\hat{L}_S$, and $\hat{L}_T$ were several times different, they influenced little on the variance of estimated heritability.

In the comparison of the two kinds of heritability for each trait, all the chromosome-wise partition heritability was higher than the whole-genome heritability except for the trait of the age diabetes diagnosed (the explanation of this exception is given below). For a certain polygenic trait, the heritability attributed to each chromosome was proportional to $\hat{M}_{e.c}$ according to the heritability estimation formula (Equation 5). Since the LD score between chromosomes could be considered as 0, this causes the $\hat{M}_e$ of the whole genome to be diluted by a large number

TABLE 1 | The sampling variance of the three estimators.

Estimator	$B = 5$	$B = 10$	$B = 20$	$B = 40$	$B = 60$	$B = 80$	$B = 100$
L_B	13,918,476	6,970,471	3,501,313	1,689,667	1,214,649	875,951	781,079
L_S	4,185,873	982,406	501,486	267,914	138,344	142,868	186,764
L_T	1,587,955	787,997	389,442	217,566	130,818	8,6461	81,037

B represents the iterations taken by L_B. We took sample size $s = \sqrt{2BN}$ for L_S and $n = BN$ for L_T in each step to guarantee the three estimators having the equal computational cost of $\mathcal{O}(NMB)$, where N is the total sample size.

TABLE 2 | $\widehat{M}_{e.c}$ and $\widehat{M}_{ew.c}$ of each autosome.

Autosome	Number of markers	$\widehat{M}_{e.c}$	$\widehat{M}_{ew.c}$
1	41,805	10,333.95	5,531.40
2	42,087	10,131.61	5,410.58
3	35,488	8,377.99	4,557.51
4	33,248	8,168.11	4,567.46
5	31,855	7,772.11	4,200.29
6	36,643	1,217.21	522.52
7	28,868	6,996.85	3,882.00
8	27,244	5,878.64	2,941.71
9	23,120	6,172.40	3,423.64
10	26,242	5,978.38	3,607.96
11	26,119	4,978.77	2,835.29
12	25,041	6,204.85	3,385.99
13	18,065	4,988.62	2,802.15
14	17,040	4,492.59	2,458.88
15	16,555	3,911.11	2,174.81
16	18,570	4,448.96	2,461.84
17	17,140	3,868.71	2,040.32
18	15,837	4,561.48	2,549.61
19	13,998	3,151.14	1,816.42
20	13,997	3,800.48	2,080.39
21	7,949	2,223.92	1,231.52
22	8,549	2,114.08	1,240.90

$\widehat{M}_{e.c}$ is the estimation of the effective of markers (M_e) for the c^{th} autosome, and $\widehat{M}_{ew.c}$ is the estimation of weighted M_e for the c^{th} autosome.

of blank LD, so the overall $\hat{M}_e$ was smaller than the average $\hat{M}_{e.c}$ of each chromosome. In order to eliminate the influence of blank LD to see the contribution of effects of causal variants to heritability, **Figure 3** shows the relationship of these two estimations of the heritability for the 81 traits. The slope of the solid gray line in the figure represents the ratio of the whole-genome $\hat{M}_e$ to $\sum_{i=1}^{c} \hat{M}_{e.c}$, a ratio of 0.729. This figure was to capture traits that do not meet the assumptions of polygenic assumption—or fitness of the model. If a trait were purely polygenic, the point representing this trait would be expected just along the solid gray line. However, the points were mostly distributed above the line, indicating that the effect size of causal variants was not evenly distributed on the chromosomes. In particular, the trait of the age diabetes was diagnosed, the Manhattan plot of which showed many statistically significant SNPs concentrated on the major histocompatibility complex (MHC) region on chromosome 6. They all belong to MHC, which is related to many human traits. Obviously, these loci breached the polygenic assumption underlying. After deleting these loci, we reestimated the two kinds of heritability, and all the traits were basically close to the solid gray line and were closer compared with **Figure 3** (see **Supplementary Figure 4**). This shows that the model assumptions were basically valid, and the estimated value of heritability had a certain degree of reliability.

Alternatively, the chromosome-wise partitioning heritability could be estimated jointly by fitting 22 autosomes altogether. It was basically the same as those calculated singly but slightly lower than the latter. It was because when calculating the heritability of chromosomes jointly, we set N for the whole genome in Equation (5), but smaller N were taken in the equation for estimating the heritability of each chromosome singly, as fewer individuals met the quality control standards for a single chromosome. We mentioned in Method that the fast estimation of joint heritability should meet the precondition that N of each chromosome are equal. The heritability estimated by the two methods will be strictly equal if this precondition holds (see **Supplementary Notes**). For traits with large sample sizes, this precondition could be met well, and the heritability estimated by the two methods was almost the same.

We also estimated the weighted chromosome-wise partition heritability and the weighted whole-genome heritability for these traits (see **Supplementary Figure 5**). In general, the weighted estimation of heritability was similar to that without weight.

DISCUSSION

In this study, we corrected the erroneous variance of the L_B estimator and proposed another two unbiased estimators of $tr(\boldsymbol{K}^T\boldsymbol{K})$, which was the most time-consuming term in RHE (Wu and Sankararaman, 2018). Instead of plotting the running time and accuracy of different methods like most articles, we used a different experimental design to make a special comparison with the L_B estimator. We borrowed the sampling size parameter B in L_B and adjusted the sample size of our estimators so that the theoretical calculation time of the three estimators was the same under different sample size parameter B. Under the same time complexity, our results showed better stability with smaller variances. In other words, under the same accuracy requirements, our method could greatly reduce the computation cost.

TABLE 3 | Estimation of heritability for some traits in UK Biobank.

Field ID	Field name	*N*	$\widehat{h}^2_{Chr}$	$\widehat{h}^2_{Gen}$	$\widehat{h}^2_{se}$
3786	Age asthma diagnosed	31,535	0.271	0.265	0.013
2754	Age at first live birth	100,951	0.281	0.217	0.004
2976	Age diabetes diagnosed	12,628	0.231	0.618	0.033
2139	Age first had sexual intercourse	255,880	0.064	0.051	0.002
21001	Body mass index (BMI)	277,223	0.360	0.282	0.002
4079	Diastolic blood pressure, automated reading	259,815	0.199	0.161	0.002
894	Duration of moderate activity	231,311	0.045	0.034	0.002
914	Duration of vigorous activity	166,696	0.032	0.025	0.003
874	Duration of walks	267,826	0.055	0.040	0.002
20150	Forced expiratory volume in 1-s (FEV1), best measure	207,848	0.257	0.207	0.002
20151	Forced vital capacity (FVC), best measure	207,848	0.314	0.252	0.002
2149	Lifetime number of sexual partners	253,460	0.011	0.008	0.002
20127	Neuroticism score	226,198	0.160	0.133	0.002
20161	Pack years of smoking	81,555	0.275	0.201	0.005
102	Pulse rate, automated reading	259,815	0.217	0.170	0.002
21021	Pulse wave arterial stiffness index	92,137	0.045	0.033	0.005
1299	Salad/raw vegetable intake	278,142	0.079	0.057	0.002
20015	Sitting height	277,231	0.528	0.444	0.002
1160	Sleep duration	278,142	0.089	0.070	0.002
50	Standing height	277,508	0.895	0.729	0.002
4080	Systolic blood pressure, automated reading	259,812	0.194	0.155	0.002
48	Waist circumference	277,649	0.278	0.219	0.002
1528	Water intake	278,142	0.098	0.075	0.002
21002	Weight	277,325	0.372	0.299	0.002
23102	Whole body water mass	273,248	0.468	0.386	0.002

N is the sample size, $\widehat{h}^2_{Chr}$ is the chromosome-wise partition heritability calculated by adding the heritability of each chromosome, $\widehat{h}^2_{Gen}$ is the whole-genome heritability calculated from the GRM of the whole genome, and $\widehat{h}^2_{Se}$ is the standard error of the whole-genome heritability.

We noted that Wu and Sankararaman further reduced the calculation time in matrix multiplication by introducing the mailman algorithm (Liberty and Zucker, 2009), which could also be used in our calculation by writing our estimators in the form of multiplication of genetic matrix and a random vector with multinoulli distribution. From these perspectives, our estimators were superior substitutions of the L_B estimator in Haseman–Elston regression.

We also gave the sampling variance of the subsampling estimator, which could be calculated by one sampling without additional calculation. As a result, the variance estimator of the heritability could be easily derived. Although the variance of the L_B estimator $2tr\left(\boldsymbol{K}^T\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{K}\right)/B$ could also be derived by the subsampling method (beyond the scope of this study), its time complexity greatly exceeded the calculation of $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right)$ as far as we know.

The variance of L_S was always slightly larger than that of L_T. This was because L_T randomly extracted nearly uncorrelated elements in the lower triangular matrix $\boldsymbol{K}_o$, while L_S extracted all elements in a triangle of $\boldsymbol{K}_o$ (after reordering the individuals). Although their sampling variance was approximately equal to the population variance $var(\boldsymbol{K}_o)$, the sampling variance of L_T was relatively smaller because it uses less related individuals.

One possible drawback of the L_T estimator relies on a much larger reference population than that of L_S. When the reference sample size is small, it is obvious that L_T becomes L_S. Therefore, the L_T estimator can make full use of large sample size, such as that of UKB. Although the difference between the variances of these two estimators is small, and the difference in the final heritability estimation is even slight, we still provide a novel and simple subsampling idea, which can be used in many situations involving large samples.

In the early analysis of heritability, both GRM and Haseman–Elston regression were applied to related individuals under the context of linkage analysis using sibling data. Under linkage, relatedness is actually related to the concept of identity by descent (IBD). However, with the increasing amount of data, the significance and application range of GRM and HE have been expanded. The unrelated individuals we emphasize here are mainly to distinguish from the linkage analysis of pedigree data. There is no problem in the estimation of heritability with related individuals, as demonstrated below. The

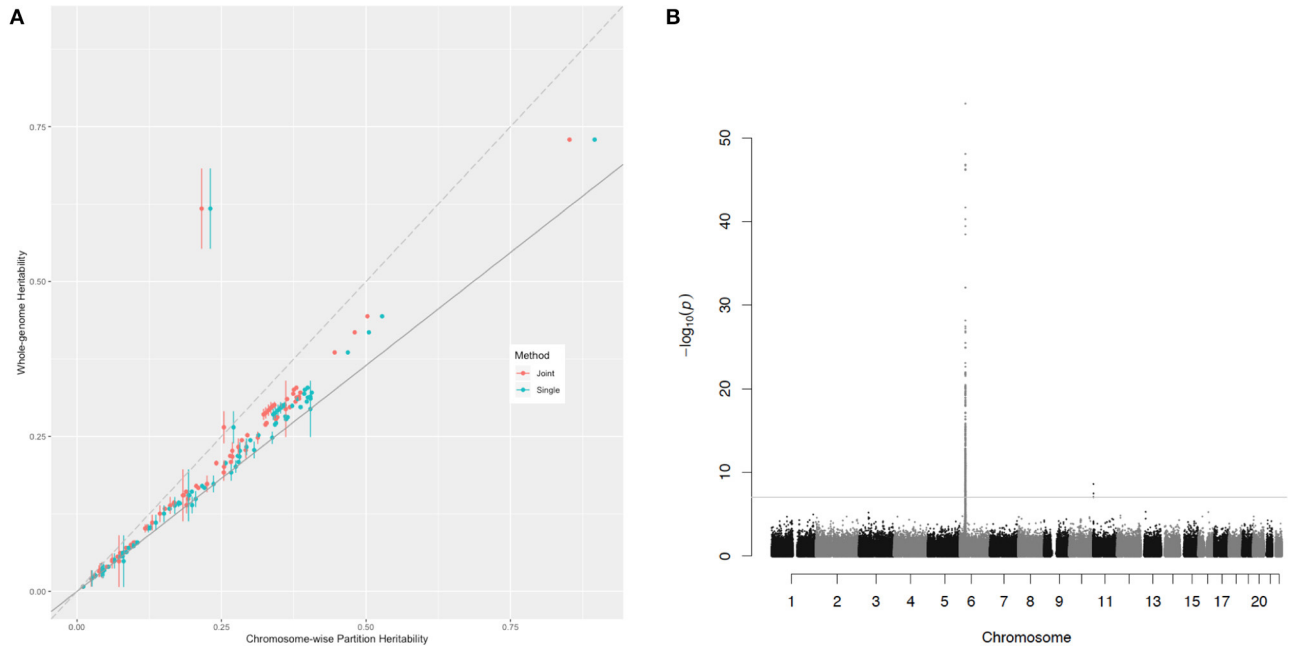

FIGURE 3 | The relationship of the chromosome-wise partition heritability and the whole-genome heritability. **(A)** Each dot represents a trait in the UKB dataset listed in **Supplementary Table 2**. The horizontal axis represents their chromosome-wise partition heritability, and the vertical line at each point is their error bar; the vertical axis represents their whole-genome heritability. The red color represents the chromosome-wise partition heritability calculated jointly, and the green color represents the chromosome-wise partition heritability calculated singly. The long-dash line crosses the origin has a slope of 1. The slop of the solid line is 0.729, the ratio of $\frac{\hat{M}_e}{\sum_{c=1}^{22}\hat{m}_{e,c}}$. **(B)** The Manhattan plot for the trait age diabetes diagnosed, and the threshold is for genome-wise threshold for $\alpha = 0.05$ after Bonferroni correction.

expression $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right) = N^2/M_e + N$ still holds true for related samples (see **Supplementary Notes**), which was confirmed in simulation (**Supplementary Table 3**). We have expanded the sample to all UKB British Whites, which included extra 131,850 individuals, totaling sample size $N = 410{,}638$, of various possible relatedness with the 278,788 unrelated samples and reestimated the heritability. The results are listed in **Supplementary Table 4**. In general, the heritability increased compared to the previous results of the unrelated set, but negligible. It shows that our estimators are basically applicable among a more realistic population even containing partially related individuals but leave some concerns in theoretical soundness.

Using modified Haseman–Elston regression to estimate heritability is becoming more and more popular in summary statistics. We further explored an important connection between Haseman–Elston regression and M_e, the effective number of independent SNPs, which is also a critical concept in quantitative genetics. We found that M_e plays a pivotal role in the estimation of variance components and heritability. As long as we get the estimation of M_e, we can easily get the estimation of its corresponding variance components.

Although we used only individual-level data to estimate heritability in this report, the nature of M_e allows researchers to estimate heritability based on a reference population of the same origin to the population in meta-analysis. However, the existence of family structure will make M_e shrink (see **Supplementary Table 3**; the expansion of trace means the shrinkage of M_e), and different family structures make it shrink differently, leading to inaccurate meta-analysis. Therefore, we do not recommend using our method in samples with various related individuals, but it raises a very interesting question for the estimation theory using mega-scale family trees (Kaplanis et al., 2018; Shor et al., 2019).

Due to the statistical property of M_e, we can easily extend M_e to the dominant model and use the same method to obtain both additive and dominant heritability, as long as their codes for the count of the reference allele are orthogonal, as discussed (Vitezica et al., 2017; Álvarez-Castro and Crujeiras, 2019). We can also extend M_e to estimate a genetic correlation for a pair of traits, in which $tr\left(\boldsymbol{K}^T\boldsymbol{K}\right) = N_1N_2/\hat{M}_e + N_o$, where N_o is the overlap sample size between a pair of cohorts, which have N_1 and N_2 individuals, respectively.

URLs: The related source code, https://github.com/GuoanQi1996/LT-Estimator.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://biobank.ctsu.ox.ac.uk/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and

institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TX developed the method and wrote the manuscript. G-AQ analyzed the data. JZ and H-MX revised the manuscript. G-BC designed the study and improved the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31771392 to G-BC, 31671570 and 31871707 to H-MX) and Zhejiang Provisional People's Hospital Research Startup Fund (ZRY2018A004 to G-BC).

ACKNOWLEDGMENTS

We thank the participants of UK Biobank for making this work possible (application 41376).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.612045/full#supplementary-material

REFERENCES

Álvarez-Castro, J. M., and Crujeiras, R. M. (2019). Orthogonal decomposition of the genetic variance for epistatic traits under linkage disequilibrium—applications to the analysis of Bateson-Dobzhansky-Müller incompatibilities and sign epistasis. *Front. Genet.* 10:54. doi: 10.3389/fgene.2019.00054

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

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

Chen, G. B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman–Elston regression. *Front. Genet.* 5:107. doi: 10.3389/fgene.2014.00107

Dudbridge, F., and Wray, N. R. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 9:e1003348. doi: 10.1371/annotation/b91ba224-10be-409d-93f4-7423d502cba0

Ge, T., Chen, C. Y., Neale, B. M., Sabuncu, M. R., and Smoller, J. W. (2017). Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* 13:e1006711. doi: 10.1371/journal.pgen.1006711

Haseman, J. K., and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* 2, 3–19. doi: 10.1007/BF01066731

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., et al. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360, 171–175. doi: 10.1126/science.aam9309

Liberty, E., and Zucker, S. W. (2009). The mailman algorithm: a note on matrix–vector multiplication. *Inform. Process. Lett.* 109, 179–182. doi: 10.1016/j.ipl.2008.09.028

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284–290. doi: 10.1038/ng.3190

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* 105, 456–476. doi: 10.1016/j.ajhg.2019.07.003

Sankararaman, S. (2019). "Fast estimation of genetic correlation for Biobank-scale data," in *Research in Computational Molecular Biology.* ed L. J. Cowen (Washington, DC: Springer), 322.

Shor, T., Kalka, I., Dan, G., Erlich, Y., and Weissbrod, O. (2019). Estimating variance components in population scale family trees. *PLoS Genet.* 15:e1008124. doi: 10.1371/journal.pgen.1008124

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G. B., Lee, S. H., Wray, N. R., et al. (2014). Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* 10:e1004269. doi: 10.1371/journal.pgen.1004269

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics* 206, 1297–1307. doi: 10.1534/genetics.116.199406

Wu, Y., and Sankararaman, S. (2018). A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics* 34, i187–i194. doi: 10.1093/bioinformatics/bty253

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

Yang, S., and Zhou, X. (2020). Accurate and scalable construction of polygenic scores in large biobank data sets. *Am. J. Hum. Genet.* 106, 679–693. doi: 10.1016/j.ajhg.2020.03.013

Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann. Appl. Stat.* 11:2027. doi: 10.1214/17-AOAS1052

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership