MACHINE LEARNING IN GENOME-WIDE ASSOCIATION STUDIES

EDITED BY: Ting Hu, Ryan Urbanowicz and Christian Darabos
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-229-6
DOI 10.3389/978-2-88966-229-6

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: researchtopics@frontiersin.org

MACHINE LEARNING IN GENOME-WIDE ASSOCIATION STUDIES

Topic Editors:
Ting Hu, Memorial University of Newfoundland, Canada
Ryan Urbanowicz, University of Pennsylvania, United States
Christian Darabos, Dartmouth College, United States

Citation: Hu, T., Urbanowicz, R., Darabos, C., eds. (2020). Machine Learning in Genome-Wide Association Studies. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-229-6

Table of Contents

frontiers in Genetics

EDITORIAL
published: 30 October 2020
doi: 10.3389/fgene.2020.593958

Editorial: Machine Learning in Genome-Wide Association Studies

Ting Hu[1,2]*, *Christian Darabos*[3] *and Ryan Urbanowicz*[4]

[1] Memorial University of Newfoundland, St. John's, NL, Canada, [2] Queen's University, Kingston, ON, Canada, [3] Dartmouth College, Hanover, IN, United States, [4] University of Pennsylvania, Philadelphia, PA, United States

Keywords: GWAS—genome-wide association study, machine learning, complex diseases, gene-gene interaction, epistasis

Editorial on the Research Topic

Machine Learning in Genome-Wide Association Studies

INTRODUCTION

Genome-wide association studies (GWAS) are used to detect genetic variants that explain common human diseases in populations. The initial GWAS achieved notoriety by successfully identifying thousands of genes associated with a variety of genetic disorders. However, these identified genes have been most successful in establishing individual associations with Mendelian diseases and explaining only a small portion of the heritability. Complex diseases are likely better explained by multiple interacting genetic and environmental variants. Such non-linear, non-additive gene-gene interaction effects, i.e., epistasis, render traditional one-gene-at-a-time analysis methods ineffective for GWAS. Instead, powerful machine learning algorithms that can detect and characterize high-order interactions among multiple genetic variants are needed.

The focus of this Special Topic Issue is to examine the novel design and application of machine learning algorithms in detecting interacting genetic variants for GWAS in six included articles.

Liu et al. proposed a deep-learning framework using convolutional neural networks to predict the quantitative traits from single nucleotide polymorphisms (SNPs) and to investigate genotypic contributions to the trait using saliency maps. The authors evaluated the performance of the proposed approach using both simulation and experimental soybean datasets. The results showed that deep learning modeling can bypass the imputation of missing values and achieve more accurate results for predicting quantitative phenotypes than well-established statistical methods. The authors claim their approach effectively and efficiently identifies significant SNPs and SNP combinations associated with GWAS data.

Zhang et al. presented circLGB, a machine learning-based framework to discriminate circRNA from other lncRNAs. This approach combined commonly used sequence-derived features and three new ones; adenosine to inosine (A-to-I) deamination, A-to-I density, and internal ribosome entry site. circLGB categorizes circRNAs by utilizing a LightGBM classifier with feature selection. In addition, the authors apply circMRT, another ensemble machine learning framework to systematically predict the regulatory information for circRNA, including their interactions with microRNA, RNA binding protein, and transcriptional regulation. Feature sets including sequence-based features, graph features, genome context, and regulatory information features were modeled in circMRT. Experiments on publicly available datasets and lab generated ones showed that the proposed algorithms outperform the available state-of-the-art methods.

In a review article by Nicholls et al., the authors discussed the landscape of ML applications in GWAS by following three components: selected models, input features, and output model performance. The authors focused particularly on the prioritization of complex disease-associated

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

***Correspondence:**
Ting Hu
ting.hu@mun.ca

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

Received: *12 August 2020*
Accepted: *15 September 2020*
Published: *30 October 2020*

Citation:
Hu T, Darabos C and Urbanowicz R (2020) Editorial: Machine Learning in Genome-Wide Association Studies. Front. Genet. 11:593958. doi: 10.3389/fgene.2020.593958

loci and explored the contributions ML has made toward reaching the GWAS end-game with consequent wide-ranging translational impact.

Leem et al. have proposed a permutation method for GWAS, i.e., ENhanced Permutation tests via multiple Pruning (ENPP). ENPP prunes the features in each permutation round if they were determined to be non-significant. Their simulation study showed that the ENPP method could remove about 50% of the features, at the first permutation round, and by the 100th permutation round, 98% of the features were removed. Only 7.4% of the compute time was required, compared to the original unpruned permutation approach. In addition, they applied this approach to a real data set of ∼300 K SNPs, to find the association with a non-normal distributed phenotype.

Arabnejad et al. designed a machine learning algorithm, i.e., Nearest-neighbor Projected-Distance Regression (NPDR), in order to detect complex multivariate effects for GWAS. NPDR used a regression formalism that allowed statistical significance testing and efficient control for multiple testing. In addition, the regression formalism provided a mechanism for NPDR to adjust for population structure, which was applied to GWAS data of Systemic Lupus Erythematosus (SLE). The authors also tested NPDR on benchmark simulated genetic variant data with epistatic effects, main effects, imbalanced data for case-control design, and continuous outcomes. NPDR identified potential epistatic and other effects that influence the complex SLE disorder.

Lastly, in the article by Ni et al., ∼300 K stomach tissue-specific eSNPs with gastric cancer (GC) risk in three GWAS datasets were investigated. The authors conducted a gene-based analysis to calculate the cumulative effect of eSNPs through a sequence kernel association combined test and Sherlock integrative analysis. At the SNP-level, they identified two novel variants associated with GC risk. Gene-based analyses identified 2 novel susceptibility genes for GC which were significantly overexpressed in GC tissues than in their adjacent tissues and the high expression level of these two genes was associated with an unfavorable prognosis of GC patients. Co-expression genes with these two novel genes in normal stomach tissues were significantly enriched in several cancer-related pathways.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 November 2019
doi: 10.3389/fgene.2019.01091

Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean

Yang Liu[1,2], *Duolin Wang*[2,3], *Fei He*[3,4], *Juexin Wang*[2,3], *Trupti Joshi*[1,3,5] *and Dong Xu*[1,2,3]*

[1] Institute of Data Science and Informatics, University of Missouri, Columbia, MO, United States, [2] Department of Electrical Engineer and Computer Science, University of Missouri, Columbia, MO, United States, [3] Christopher S. Bond Life Science Center, University of Missouri, Columbia, MO, United States, [4] Department of Computer Science and Information Technology, Northeast Normal University, Changchun, China, [5] Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO, United States

Genomic selection uses single-nucleotide polymorphisms (SNPs) to predict quantitative phenotypes for enhancing traits in breeding populations and has been widely used to increase breeding efficiency for plants and animals. Existing statistical methods rely on a prior distribution assumption of imputed genotype effects, which may not fit experimental datasets. Emerging deep learning technology could serve as a powerful machine learning tool to predict quantitative phenotypes without imputation and also to discover potential associated genotype markers efficiently. We propose a deep-learning framework using convolutional neural networks (CNNs) to predict the quantitative traits from SNPs and also to investigate genotype contributions to the trait using saliency maps. The missing values of SNPs are treated as a new genotype for the input of the deep learning model. We tested our framework on both simulation data and experimental datasets of soybean. The results show that the deep learning model can bypass the imputation of missing values and achieve more accurate results for predicting quantitative phenotypes than currently available other well-known statistical methods. It can also effectively and efficiently identify significant markers of SNPs and SNP combinations associated in genome-wide association study.

Keywords: genomic selection, deep learning, genome-wide association study, soybean, genotype contribution

OPEN ACCESS

Edited by:
Ting Hu,
Memorial University of Newfoundland, Canada

Reviewed by:
Dusanka Savic Pavicevic, University of Belgrade, Serbia
Valentino Ruggieri, Centre for Research in Agricultural Genomics (CRAG), Spain

***Correspondence:**
Dong Xu
xudong@missouri.edu

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

Received: *22 July 2019*
Accepted: *09 October 2019*
Published: *22 November 2019*

Citation:
Liu Y, Wang D, He F, Wang J, Joshi T and Xu D (2019) Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean. Front. Genet. 10:1091. doi: 10.3389/fgene.2019.01091

INTRODUCTION

The marker-assisted selection (MAS) strategy has made significant improvements in phenotype prediction for quantitative traits in breeding, assuming that genotype markers have significant associations with their phenotypes. The genome-wide association study (GWAS) has also been applied to select those phenotype-associated genetic variants. Genomic selection (GS) is one type of MAS strategy, using single-nucleotide polymorphisms (SNPs) to predict breeding values (BVs) or quantitative phenotypes. The strategy has been widely applied in i) major crops (Jannink et al., 2010), such as soybeans [*Glycine max*], rice [*Oryza sativa*], and maize [*Zea mays*] (Zhao et al., 2012; Spindel et al., 2015; Xavier et al., 2018); ii) crops with long life cycles, such as oil palm [*Elaeis guineensis* Jacq.] (Cros et al., 2015) and domesticated animals like Holstein dairy cattle (Schaeffer 2006;

Verbyla et al., 2009). Traditional statistical methods, such as the best linear unbiased prediction (BLUP), Bayesian A, B, Cπ, and Bayesian LASSO (BL) (Hayes and Goddard, 2001; Pérez et al., 2010; Endelman, 2011) have been widely utilized for modeling genotype effects and predicting phenotypes. These statistical methods usually assume that genotype random effects follow a prior distribution such as Gaussian, and the contribution of each genotype to the associated phenotype is considered as an independent feature. This prior assumption requires sufficiently large training samples to cover the overall population structure and to make it true. However, in practice, the individual genotype effect is unknown and may not strictly follow a certain distribution. In addition, SNPs may also have interactions with other SNPs that contribute to complex diseases or traits (Wang et al., 2015) as seen due to the epistasis effects.

Missing values in a genotype matrix represent another challenge for statistical methods, wherein these missing values are usually screened out during preprocessing or filled with values through imputation (Howie et al., 2009; Marchini and Howie 2010; Rutkoski et al., 2013). Imputation is a computational process for estimating missing values in genotypes from a template population. Several methods have been developed for genomic imputation with or without the reference genome information. The calculated mean, expectation–maximization (EM) algorithm is provided in the R package rrBLUP (Endelman 2011); random forest (RF) is provided in missForest (Stekhoven and Bühlmann, 2011), and a hidden Markov model (HMM)-based method is applied in BEAGLE (Browning et al., 2018) and MaCH (Li et al., 2010) with the reference genome. The imputation accuracy is highly dependent on observed non-missing genotypes and the missing rate of the whole population, which directly affects the performance of the phenotype prediction model (Rutkoski et al., 2013; Xavier et al., 2016). To develop a phenotype prediction model through statistical approaches, the genotype matrix is required to be imputed together and then divided into training and testing datasets for model training and testing. To some extent, the testing set is not totally independent from the training set, since the training set may contain genotypes estimated from the testing set under this circumstance. Inaccurate imputation methods may also introduce errors and uncertainty and further affect biomarker selection. Therefore, these imputation approaches may not be effective in inferring informative genetic markers hidden in the entire genome.

Recently, deep learning has been applied in computational biology (Angermueller et al., 2016), with the introduction of noncoding variant function prediction (Zhou and Troyanskaya, 2015), protein localization prediction (Alipanahi et al., 2015; Zhang N et al., 2018), protein secondary structure prediction (Spencer et al., 2015), and protein post-translational modification site prediction (Wang D et al., 2017; Wang et al., 2018). In genotype association studies, deep learning has also been used to identify SNP interactions (Uppu et al., 2016), classify genomic variants (Liang et al., 2016). DeepGS, an ensemble of convolutional neural network (CNN) (Krizhevsky et al., 2012) and rrBLUP have been used to predict phenotypes using imputed SNPs (Ma et al., 2018), and a simple dense neural network (DNN) is used on genotype-by-sequencing (GBS) data (Montesinos-López et al., 2018). For these phenotype prediction problems, CNN can capture spatial information from raw sequencing reads or genomic variants without feature engineering. To some extent, the CNN also resolves the local epistasis effect as the convolving process is considering interactions among neighboring SNPs within different ranges of the kernel window. However, the above deep learning methods have not effectively addressed the problem of missing values, and they all treat the deep learning models as black boxes without discussing the effective SNP markers. In particular, none of them have explored the internal features associated with the traits through attention mechanisms, which is an approach developed for visualization of the black box of deep learning architecture. The saliency map (Simonyan et al., 2013) of deep learning was first introduced for visualizing image features in classification and now plays a major role in image segmentation and image style transfer (Gatys et al., 2016). This strategy can evaluate the contribution of each input component to differentiate output categories.

In this paper, we propose an independent deep CNN (Szegedy et al., 2017) model to predict phenotypes from SNPs, which contains dual-stream of CNNs and can take either an imputed or non-imputed genotype matrix as the input. We also applied the saliency map deep learning visualization approach to select significant associated biomarkers from our trained model. To the best of our knowledge, this is the first study to apply a saliency map for a GWAS. The comparison results with traditional statistical methods (rrBLUP, Bayesian ridge regression (BRR), Bayesian A, and BL) and existing deep learning used several evaluation metrics on both simulation and experimental data, which indicate that our proposed deep learning model serves as a robust and efficient architecture in selecting germplasms and discovering genotype–phenotype relationships.

MATERIALS AND METHODS

Dataset

We used an experimental soybean dataset and a simulation dataset as the benchmark to evaluate the performance of our deep learning model, as summarized in **Table 1**.

Soybean Dataset: The soybean dataset from the soynam project was generated using a nested association panel (Xavier et al., 2015; Song et al., 2017). The soybean dataset contains more than 5,000 recombination inbred lines (Rils) and 4,236 common SNPs between imputed data and raw quality assured data. The genotype and phenotype data were available in the

TABLE 1 | Summary of soybean experimental dataset.

Dataset	Trait	Environment	Sample (*N*)	Heritability	Reference
SoyNAM	Yield	2013 Illinois	5,001	0.512	(Xavier et al., 2015)
	Protein	2012 Illinois	5,128	0.545	
	Oil	2012 Illinois	5,128	0.617	
	Moisture	2012 Illinois	5,128	0.582	
	Height	2013 Illinois	5,138	0.667	

"SoyNAM" R Package (Xavier et al., 2015). We selected five traits from the 2013 and 2012 Illinois Location. Missing genotypes in the soybean dataset were imputed using the MaCH software (Li, et al., 2010) based on the HMM Approach. The imputation method applied on the soybean dataset was discussed in Xavier et al. (2016), who found it to have the best performance in imputing accuracy and phenotype predicting ability.

Simulation Dataset: The simulation dataset was constructed using Hypred (Technow, 2011), which simulates 10,000 F2 recombined individuals with 5,000 SNPs. We assigned quantitative trait locus (QTL) every 500 SNPs at SNP index position 100, 600, 1100, 1600, 2100, 2600, 3100, 3600, 4100, and 4600. No missing value was included in the simulation set.

The genotype matrix used as inputs for the three datasets was coded into 0, 1, or 2 to represent homozygous, heterozygous, and reference homozygous, respectively, and missing genotypes were coded as −1 for genotypes without imputation.

Narrow-Sense Heritability

The narrow-sense heritability of each trait is calculated based on the BRR model from the R package SoyNAM. It is defined as the ratio of phenotypic variance due to additive genotypes as follows:

$$h^2 = \frac{V_g}{V_g + V_e}$$

where V_g is the phenotypic variance and V_e is the residual variance estimated from a BRR model.

Deep Learning Architecture

Genotype Coding With One-Hot

Three genotypes (0, 1, 2) and missing values (−1) are first encoded using one-hot binary coding and serve as the input vector. Using one-hot coding, each marker is represented by a four-dimensional vector with 1 at the index for one genotype and the rest of them are set at 0 as shown in the far left inset of **Figure 1**. For example, three genotypes [AA, Aa, aa] are represented as [0, 1, 0, 0], [0, 0, 0, 1], and [0, 0, 1, 0], respectively. The missing genotype is represented as [1, 0, 0, 0]. Encoded genotypes serve as input to our model.

Genotype Processing Blocks

Our dual-stream CNN-based deep network contains three building blocks as shown in **Figure 1**, i.e., the input processing block, the feature processing block, and the output processing block.

Input Processing Block: This block contains an input layer, a dual-CNN layer, which contains two parallel CNN streams (Szegedy et al., 2017) and a sum-up layer to combine the parallel CNN streams. The input layer contains one-hot encoded genotypes, and subsequently the encoded genomics makers are simultaneously passed to the dual-CNN layer. We applied the idea of residual learning (He et al., 2016) in this dual-CNN layer, which was first introduced for image recognition and classification to solve the vanishing gradient problem. The residual connection is a shortcut connection from a previous layer and was added to identity mapping used to form a residual mapping. This approach has been applied in predicting protein backbone torsion angles and protein contact maps (Wang S et al.,

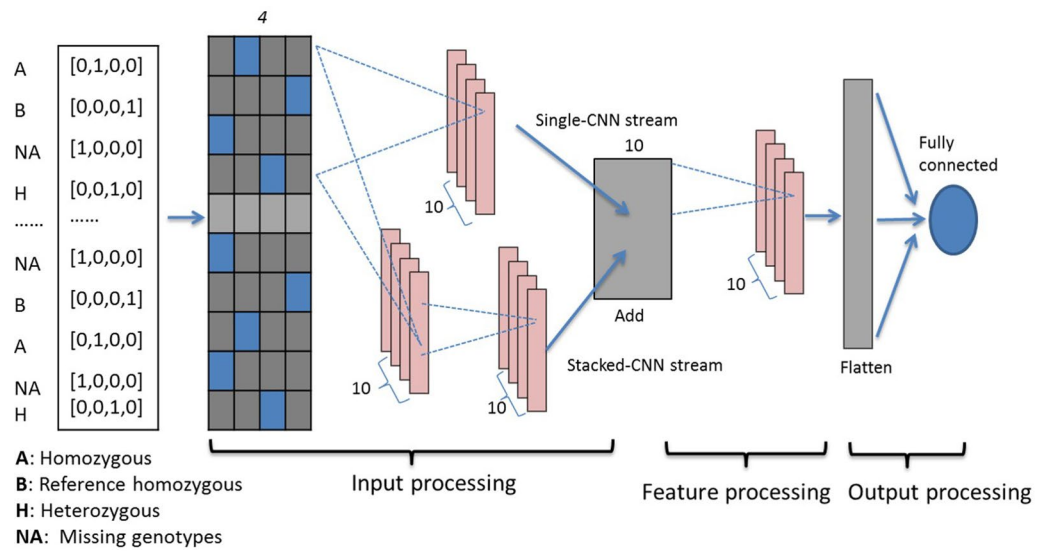

FIGURE 1 | Dual-stream CNN model structure. Genotypes are one-hot coded and passed to the input processing block, which contains two streams of CNNs. The first stacked-CNN stream contains two feed-forward CNN layers with kernel sizes 4 and 20. The second single-CNN stream contains one CNN layer with kernel size 4, followed by an add-up layer to aggregate outputs from the two streams. The feature processing block contains another single convolution layer with kernel size 4. Processed features are then passed to the output processing block, which contains a flatten layer and a fully connected dense layer. CNN, convolutional neural network.

2017; Fang et al., 2018). In the dual-CNN layer, the single-CNN stream served as a residual connection to the other stacked-CNN stream. The stacked-CNN contains two stacks of 1D convolutional layer with different kernel sizes, 4 and 20; and the single-CNN stream contains one convolutional layer with kernel size 4. The sum-up layer is used to aggregate the outputs from previous dual-CNN layer, and it is the element sum of both.

In order to optimize the kernel sizes, we used the affinity propagation (AP) (Frey and Dueck, 2007) clustering method on the genotype features to help guide us in selecting convolution sizes in this block. AP divided genotypes into clusters without assigning a number of clusters. The algorithm estimates the cluster center as the "exemplar" from data points. Real-time messages were exchanged between data points until a set of exemplars and clusters emerges through minimizing negative Euclidean distance. This clustering algorithm has been applied in computer vision and regulates transcript gene identification (Vlasblom and Wodak, 2009). We conducted AP clustering on 4,236 SNPs from the soybean dataset and repeated the process 100 times. Python package "sklearn" was used for AP cluster estimates (Pedregosa et al., 2011). We recorded sizes of clusters from 100 runs and tested kernel sizes using the number of genotypes clustered together. We aimed to capture short-range and long-range marker effects at various scales across the genome (Xu and Taylor, 2009; Brodie et al., 2016) so that small and large convolving sizes were used in our model. We finalized 4 and 20 as our convolving kernel sizes for stacked-CNN stream and 4 for the single-CNN stream.

Feature Processing Block: After completing our work on the input processing block, we determined that the aggregated sim-up outputs with different kernel sizes had more powerful representations of important genotypes than with a single kernel size. Hence, another convolution layer with a small kernel size 4 was added to integrate all the outputs and to further process genotype features in this block.

Output Processing Block: After completing our work on the feature processing block, a flattened layer was added to convert the convolution layer into a flattened layer. The flattened layer integrates the extracted features from the previous feature processing blocks, and features are passed to the last dense output layer, which contains a single neuron to represent the final predicted phenotypes.

Activation Function

We used the inverse square root unit (ISRU) (Carlile et al., 2017) activation functions in the model, which is defined as follows:

$$\boldsymbol{Y} = \frac{\boldsymbol{x}}{\sqrt{\mathbf{1} + \boldsymbol{a}\boldsymbol{x}^2}}$$

The ISRU function was applied to add constraint of the predicted phenotype value and to speed up the model learning process. The activation function is bound to the range $\left(-\frac{1}{\sqrt{a}}, \frac{1}{\sqrt{a}}\right)$. Thus, we estimated a according to the maximum observed absolute phenotype values, which are 0.5, 0.03, 0.02, 0.02, and 0.02 for grain yield, height, moisture, oil, and protein of the soybean dataset, respectively.

Model Training for Overfitting Control

It is important for the deep learning model to avoid overfitting because of the small training population of our datasets and because the total sample size is much smaller than the number of genotypes used as features. To reduce the effect of overfitting, we added dropout layers (Srivastava et al., 2014) after convolutional layers with a dropout ratio of 0.75. We then applied the L2 regularization on the cost function of mean square error (MSE) between estimated and predicted phenotypes:

$$MSE = \frac{\sum_{i=1}^{i=n} \left(Y_i - Y_i^{'}\right)^2}{n}$$

We also monitored the mean absolute error (MAE) on our validation set and stopped the model training process as soon as the observed MAE stopped decreasing enough to confirm cessation. Hyperparameters, such as batch size and learning rate, were tuned by Hyperas (Pumperla, 2019). The deep learning models were implemented using Keras 2.1.1 on a workstation with GPU NVidia GTX 1080 Ti.

SNP Contribution Using Saliency Map

We defined saliency values based on the idea of saliency map (Simonyan et al., 2013) to measure individual marker effects and their associations with quantitative GWAS trait. In the phenotype prediction problem, saliency values can be interpreted as scores to indicate effects of markers inside a window at length of a decided convolution kernel size from our deep learning model. The saliency values can guide extracting meaningful SNPs that show high-order marker effects correlated with phenotypes. In our deep model, given a genotype matrix X (n*p) of n individuals and p genotypes, the phenotype value was estimated as follows:

$$\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{X} + \boldsymbol{b}$$

where W represents the trained weight of each genotype and b is the model bias. In this case, after training the model, we can retrieve the output from the last output layer and calculate gradients w with respect to each input genotype using independent testing set as below:

$$\boldsymbol{w} = \frac{\boldsymbol{\partial}(\boldsymbol{Y})}{\boldsymbol{\partial}\boldsymbol{X}}$$

Since our genotypes were coded into one-hot vectors with four dimensions as the model inputs, we define the saliency value of each genotype as the maximum absolute value of gradients among those four coding channels. Therefore, to calculate the saliency value SV of a single genotype whose index is i and is coded in the c-dimension of one-hot vector, we use the following function:

$$\boldsymbol{SV}_i = \boldsymbol{MAX}(\boldsymbol{ABS}(\boldsymbol{w}_{i,\mathbf{c}}))$$

We then calculate the median saliency value of whole populations, and this population median value is used as a measurement of our SNP contribution.

Model Performance With Cross-Validation

Phenotype Prediction Accuracy

To measure our dual-stream CNN deep learning model performance, we calculated the Pearson correlation coefficient (PCC) between genomic predicted phenotypes and observed phenotype values of the testing dataset. We compared our deep learning model with four statistical models (rrBLUP, BRR, Bayesian A, and BL) and three deep learning models using the same training, validating, and testing datasets. The rrBLUP was implemented using the "mixed.solve" function from the "rrBLUP" package (Endelman, 2011) based on the maximum-likelihood (ML) estimation. BRR, Bayes A, and BL were implemented using the "wgr" function from the "SoyNAM" package (Xavier et al., 2015) based on the Monte Carlo Markov chain (MCMC) strategy with 4,000 iterations and 500 burn-ins.

The three compared deep learning models were a dense network (Montesinos-López et al., 2018) using several dense layers, the deepGS (Ma et al., 2017) a feed-forward three layer convolutional neural work, and a single-stream CNN that only contains the stacked-CNN layers from our proposed model. Hyperparameters were adopted from published codes.

Snp Contribution Accuracy

To measure the performance of our saliency value associated with the genotype contribution, we compared our results with a standard GWAS method using "gwas2" function from "NAM" R package based on the empirical Bayesian model (Xavier et al., 2015) that the significance of each genotype marker was evaluated through the Wald statistical test value.

Ten-Fold Cross-Validation

All soybean individuals were first split into 10 equal folds, in which eight folds formed the training set. One fold was assigned as the validation set, and the remaining one fold was employed to test the model performance. We repeated the same process 10 times, and the average PCC from the 10 calculations was reported to measure model performance.

RESULTS AND DISCUSSION

Model Performance and Comparison With Other Methods

Dual-Stream CNN Model Improves Performance on Low Heritability Phenotypes

By using deep learning, missing genotypes can be coded using the one-hot binary coding method and can be treated as a category of genotype through computation. We coded both raw and imputed genotype matrix with a one-hot vector with four channels and applied the same deep learning architecture on them. The comparison of average PCC using existing statistical and deep learning methods is shown in **Table 2**. Missing value is not accepted by statistical methods, and hence, we only show results of imputed genotypes of statistical methods. The singleCNN network has similar PCC to statistical methods, and our dual-stream CNN outperforms statistical model and singleCNN using same imputed genotypes. Among the five traits, PCC of trait yield increases from 0.41 to 0.43, moisture increase from 0.38 to 0.412 and oil increase from 0.388 to 0.412 that is better than height and protein increasing from 0.458 to 0.465 and from 0.392 to 0.402.

Compare to singleCNN, performance of proposed dualCNN increases by adding a parallel single-CNN stream to the stacked-CNN stream. The add-up layer then integrates feature maps from both CNN streams, and this is necessary due to the loss of important features through convolving process with different kernel sizes, and it strengthens the signal of genotype features.

Predicting Phenotype With Imputed vs Non-Imputed Genotype Using Deep Learning

All four deep learning based methods have higher PCC on non-imputed than imputed genotypes (**Table 2**). The soybean dataset has ~25% missing genotypes in the quality assured raw datasets. One reason deep learning model has higher predicting ability on raw datasets may be because the imputation process fills most missing genotypes with reference alleles, and it deflates the effects of different genotypes. Imputation methods assimilate missing genotype effects based on non-missing genotypes, which may compromise the prediction ability of selected quantitative traits.

TABLE 2 | Average Pearson correlation coefficient of five traits from cross-validation.

	Yield	Protein	Oil	Moisture	Height
dualCNN (imp/non-imp)	0.434/0.452	0.402/0.619	0.412/0.668	0.426/0.463	0.465/0.615
DeepGS (imp/non-imp)	0.347/0.391	0.231/0.506	0.344/0.531	0.024/0.310	0.357/0.452
Dense (imp/non-imp)	0.359/0.449	0.357/0.603	0.401/0.657	0.370/0.427	0.434/0.612
singleCNN (imp/non-imp)	0.422/0.463	0.380/0.573	0.392/0.627	0.370/0.449	0.442/0.565
rrBLUP	0.412	0.392	0.39	0.413	0.458
BRR	0.422	0.392	0.39	0.413	0.458
Bayes A	0.419	0.393	0.388	0.415	0.458
Bayesian LASSO	0.419	0.394	0.388	0.416	0.458

CNN, convolutional neural network; BRR, Bayesian ridge regression.

Our dualCNN outperforms single-stream CNN and followed by a dense network (Montesinos-López et al., 2018) and then the DeepGS (Ma et al., 2017) for this soybean dataset (**Figure 2**) with lowest training loss on validation set. DualCNN, singleCNN, and the dense network have close performance on high heritability traits of oil and height, and our dualCNN has better performance in the other three low heritability traits yield, protein, and moisture on both imputed and non-imputed dataset. The dense network is better than deepGS for this soybean dataset, probably because the deepGS with more parameters is easier to be over-trained than the dense network. The DeepGS has a convolution layer of kernel size 18 that is not fit for the soybean SNP distribution of whole genome, while the dense network does not contain convolution layer, and each SNP was treated as a feature contribute independently to associated phenotype. But this dense network may also fail to integrate neighbor SNP associations within the convolution kernel.

Effects of Training Population on Model Performance

The training population size is a major factor in both machine learning and statistical approaches, and it directly affects predicting performance (Xavier et al., 2016; Cericola et al., 2017). Good training data will be able to represent the whole population structure and to satisfy the prior assumption of genotype effects for statistical models. **Figure 3** shows the average PCC of five traits predicted on the testing set trained with different sizes of training sets. For soybean dataset, the dualCNN reaches a higher PCC than the other four statistical models and was less affected by the training population size in low heritability traits as yield, moisture, and protein. As long as the training size reached 1,500, our model showed a higher performance than statistical models. The whole genome regression (BRR, BayesA, and BayesLASSO from the NAM package) had a better performance than the rrBLUP package, since the former applies Gibbs resampling and MCMC to update regression coefficients.

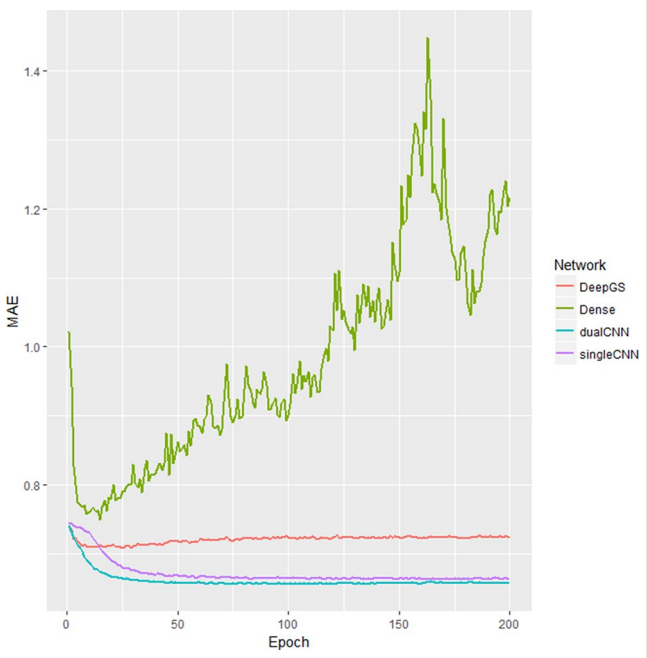

FIGURE 2 | Training loss different deep learning models. The x-axis is number of epochs; the y-axis is the training the loss of mean absolute error (MAE) of validation dataset. The singleCNN (purple), dualCNN (blue), and Dense (green) network are conserved, and DeepGS is overfitting after 20 epochs, and our dualCNN has the lowest training loss. CNN, convolutional neural network.

Comparison of Genotype Contribution Between Saliency Map and GWAS

We compared our deep learning saliency value against GWAS results through Manhattan plot using a simulation and an experimental dataset (**Figure 4**). Their calculated saliency values and Wald test score are available at **Supplemental Table 1**. For the two datasets, we observed a similar curve pattern from both saliency values and the GWAS Wald test score. In the experimental dataset, we compared the top three SNPs according to their significance and discussed potential markers discovered using our method. The top ranked SNPs and their relative position in the other measurement were plotted in red. Since the soybean linkage disequilibrium extent region of a significant SNP ranges from ~20 to ~100 kb, we located the closest gene within the 20-kbp region centered with the identified SNPs and annotated genes with Gene Ontology (GO) (Ashburner et al., 2000), protein family (PFAM) (Bateman et al., 2004) using Soybase Gbrowser (Grant et al., 2009) and SoyKB (Joshi et al., 2012; Joshi et al., 2013) according to gene model "Glyma.Wm82.a1.v1.1" (Schmutz et al., 2010). Gene annotations and literature reports indicate those markers, and their nearby regions are highly associated with their traits. Several novel markers and regions were detected and are listed as follows:

Simulation: Both saliency values and GWAS results showed the same three peaks on the simulation dataset in **Figure 4**. The three peaks were correlated with the QTLs assigned at the SNP index positions of 2100, 4100, and 4600. It strongly indicates that the saliency approach can find similar SNPs with statistical GWAS models.

Grain Yield: For soybean grain yield, we identified SNPs Gm01_28793495, Gm07_36725068, and Gm15_15220084, with the highest saliency value as shown in **Figure 4**. The top SNPs from GWAS, Gm19_10774629, and Gm19_40740547 also have high saliency value and locate in the same haplotype block with a linkage disequilibrium r^2=0.9766. Potential genes Glyma15g18430 and Glyma15g18450 are close to SNP Gm15_15220084. Glyma15g18430 reported by Won Oh et al. (2014) has differentially changed soybean root proteins with gibberellic acid treatment under flooding stress. It belongs to the glycosyl hydrolases family (PF01301) and involves in

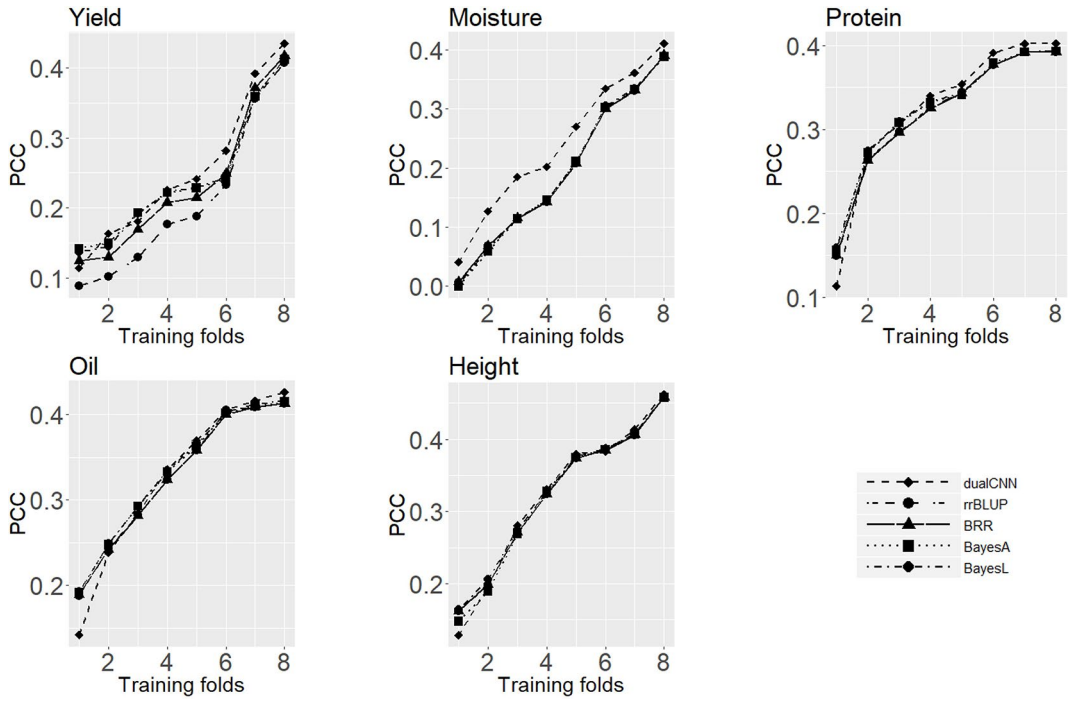

FIGURE 3 | Average Pearson correlation coefficient of five traits using different sizes of training dataset. The x-axis is number of folds of training data; the y-axis is the average Pearson correlation coefficient from cross-validation.

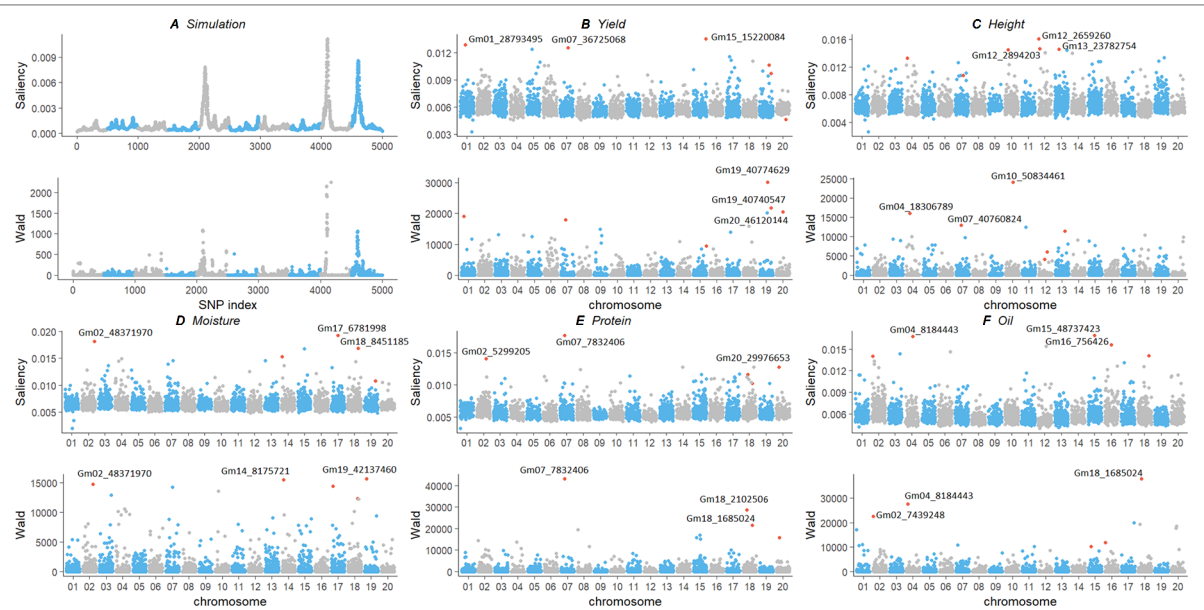

FIGURE 4 | Comparison of genotype contribution using saliency map and GWAS Wald test of simulation **(A)** and experimental soybean dataset with five traits **(B–F)**. The x-axis is the index of SNPs in the genotype matrix; the y-axis is the saliency and Wald test results. Top ranked SNPs were plotted in red. GWAS, genome-wide association study; SNP, single-nucleotide polymorphism.

carbohydrate metabolic process (GO: 0005975). Glyma15g18450 is associated with plant flowering (Jung et al., 2012) has biological process of flower development (GO: 0009908) and leaf morphogenesis (GO: 0009965).

Plant Height: For soybean plant height, saliency and Wald test value were plotted in **Figure 4**. One region on chromosome 12 is most significant from the saliency value but not present in the GWAS results; thus, we investigated

the closest gene, Glyma12g04400, of the SNP Gm12_2894203. This gene belongs to the putative snoRNA binding domain (PF01798,GO:0003677) and is reported by Komatsu et al. (2012; 2014) with differential protein change under flooding stress. The region around SNP Gm12_2659260, from 26624**kb to 26629**kb, is reported in a 302 resequencing soybean dataset (Zhou et al., 2015) as a copy number variation signal that is associated with plant height. Two SNP Gm12_2894203 and Gm12_2659206 are in the same haplotype block with r^2=0.9510. The closest region of SNP Gm13_23782754 is reported as a QTL region associated with plant height (Zhang X et al., 2018). Both saliency and GWAS identified SNP Gm04_18306789 and Gm10_50834461, and close gene Glyma10g44500 is associated with salt tolerance (Pantalone et al., 1997) and is involved in lipid transport (GO: 0006869).

Moisture: The most significant SNP Gm17_6781998 and Gm18_8451185 from saliency values also present in the GWAS results in **Figure 4**. The closest gene Glyma17g09165 belongs to the protein kinase domain (PF00069) and is involved in the biological process in response to cold, wounding, salt stress, and mannitol stimulus, that is, GO: 0009409, GO: 0009611, GO: 0009651, and GO: 0010555, respectively. Gene Glyma18g09550 belongs to seed storage family (PF00234) with lipid transport (GO: 0006869). Both methods identified SNP Gm02_48371970, and the closest gene Glyma02g43602 is response to fungus, chitin, and fatty acid (GO: 0009620, GO: 0010200, GO: 0071398).

Protein: For the soybean protein content, saliency value and Wald test score were plotted in **Figure 4**. The SNPs Gm02_5299205 and Gm20_29976653 are only present in the saliency value, and the former is in gene region of Glyma02g06650. The region around both SNPs may associated with protein content in chromosome 2 (Akond et al., 2012) and chromosome 20 (Hwang et al., 2014). Both saliency value and Wald test score indicate SNP Gm07_7832406 as the most significant one, and it is a missense mutation in the coding sequence region of gene Glyma07g09400. This gene belongs to the PP-loop family (PF01170) with molecular functions of ATP binding, ligase activity, and forming carbon–nitrogen bonds (GO: 0000166, GO: 0005524). This could also be a new marker associated with protein QTL region (Jun et al., 2008).

Oil: For SoyNAM protein content, saliency value identified a potential novel SNP Gm15_48737423, and it is inside the gene region on Glyma15g41600 **Figure 4**. It belongs to the pyridocal-phosphate-dependent enzyme protein family (PF00291) and involves a sulfur amino acid metabolic process, a cysteine biosynthetic process, and a cell wall modification (GO:0000096, GO: 0006535, GO: 0042545). This gene was reported by Prince et al. (2015) with an association with potential root QTL, and it was also reported as a putative β-substituted alanine synthase isoform by Yi et al. (2010). A new marker around region Gm16_756426 also detected associated with oil content (Jun et al., 2008). The common SNP Gm04_8184443 is close to gene Glyma04g09900, and this gene belongs to the protein tyrosine kinase family (PF07714), which involves the protein phosphorylation process and the oligopeptide transport process (GO: 0006468, GO: 0006857).

SUMMARY

In this paper, we proposed a deep learning of dual-stream CNN method to accurately predict phenotypes using SNP markers that can avoid missing genotype imputation. We also proposed using saliency map approach to measure SNPs associated with the selected traits, which helps to determine important markers and QTL regions. We have explored several different deep learning architectures, such as the fully connected DNN, deepGS, single-stream CNN, as well as several statistical approaches. We have found the two-stream CNN structure has best predicting performance on real experimental datasets, especially with low heritability quantitative traits, and it less relies on the structure of training population. To our knowledge, we are the first to use saliency value as a measurement of SNP contribution. By using CNN, the saliency map calculates the genotype effect not only as a single marker but also through convolving with their neighboring SNPs, which helps detect important trait associated regions.

Computing efficiency is also important for machine learning problems. It may not be fair to compare computing efficiency of a deep learning model applicable on GPU with statistical models on CPU, but GPU-based deep learning models actually outperformed most R-based genomics selection packages with much less computing time. Our dual-stream CNN model costs around 10 minutes, and statistical regressions cost more than 3 hours to train the model and test results for the soybean dataset. Taking the advantage of GPU computing and progress in the state-of-art deep learning technique, we expect this deep learning approach to be useful in accurately predicting phenotypes and detecting meaningful genomic markers in a more efficient way. In the future, we will continue improving our model and studying effects of genotype interactions on phenotypes explicitly. We will also work with biologists to interpret underlying biological significance of the prediction results. It is recommended to use deep learning on a large population of high-dimensional genotype and low-heritability phenotypes in phenotype prediction and biomarker selection.

DATA AVAILABILITY STATEMENT

The deep learning model, results, and datasets used can be found at https://github.com/kateyliu/DL_gwas.

SoyNAM dataset can be found at https://cran.r-project.org/web/packages/SoyNAM/index.html.

AUTHOR CONTRIBUTIONS

YL: designing the experiments, modeling, summing up, and writing the manuscripts. FH and DW: performing discussing and revising experiments. JW: generating simulation data. TJ and DX: advising and revising the project.

FUNDING

This work was partially supported by National Institutes of Health (award R35-GM126985).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01091/full#supplementary-material

SUPPLEMENTAL TABLE1 | Each column in the table represents: SNP: SNP ID wald_protein: Wald test value of protein sigma2_protein: estimated residual variance of protein eff_protein: estimated allele effect of protein saliency_protein: saliency value of protein wald_yield: Wald test value of yield sigma2_yield: estimated residual variance of yield eff_yield: estimated allele effect of yield saliency_yield: saliency value of yield wald_oil: Wald test value of oil sigma2_oil: estimated residual variance of oil eff_oil: estimated allele effect of oil saliency_oil: saliency value of oil wald_height: Wald test value of height sigma2_height: estimated residual variance of height eff_height: estimated allele effect of height saliency_height: saliency value of height wald_moisture: Wald test value of moisture sigma2_moisture: estimated residual variance of moisture eff_moisture: estimated allele effect of moisture saliency_moisture: saliency value of moisture.

REFERENCES

Akond, A. M., Ragin, B., Bazzelle, R., Kantartzi, S. K., Meksem, K., and Kassem, M. A. (2012). Quantitative trait loci associated with moisture, protein, and oil content in soybean [Glycine max (L.) Merr.]. *J. Agric. Sci.* 4 (11), 16. doi: 10.5539/jas.v4n11p16

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33 (8), 831. doi: 10.1038/nbt.3300

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Systems Biol.* 1 2 (7), 878. doi: 10.15252/msb.20156651

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25 (1), 25. doi: 10.1038/75556

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32 (suppl_1), D138–D141. doi: 10.1093/nar/gkh121

Brodie, A., Azaria, J. R., and Ofran, Y. (2016). How Far SNP May Causative Genes Be? *Nucleic Acids Res.* 44 (13), 6046–6054. doi: 10.1093/nar/gkw500

Browning, B. L., Zhou, Y., and Browning, S. R. A. (2018). One-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103 (3), 338–348. doi: 10.1016/j.ajhg.2018.07.015

Carlile, B., Delamarter, G., Kinney, P., Marti, A., and Whitney, B. (2017). *Improving Deep Learning by Inverse Square Root Linear Units (ISRLUs)*. arXiv preprint arXiv:1710.09967.

Cericola, F., Jahoor, A., Orabi, J., Andersen, J. R., Janss, L. L., and Jensen, J. (2017). Optimizing training population size and genotyping strategy for genomic prediction using association study results and pedigree information. A case of study in advanced wheat breeding lines. *PloS One* 12 (1), e0169606. doi: 10.1371/journal.pone.0169606

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., et al. (2015). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theor. Appl. Genet.* 128 (3), 397–410. doi: 10.1007/s00122-014-2439-z

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4 (3), 250–255. doi: 10.3835/plantgenome2011.08.0024

Fang, C., Shang, Y., and Xu, D. (2018). Prediction of protein backbone torsion angles using deep residual inception neural networks. *IEEE/ACM Transactions Comput. Biol. Bioinformatics*. 16 (3), 1020–1028. doi: 10.1109/TCBB.2018.2814586

Frey, B. J., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* 315 (5814), 972–976. doi: 10.1126/science.1136800

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423. doi: 10.1109/CVPR.2016.265

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38 (suppl_1), D843–D846. doi: 10.1093/nar/gkp798

Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (4), 1819–1829.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proc. IEEE Conference Comp. Vision Pattern Recognition.*, 770–778. doi: 10.1109/CVPR.2016.90

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PloS Genet.* 5 (6), e1000529. doi: 10.1371/journal.pgen.1000529

Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., and Cregan, P. B. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15 (1), 1. doi: 10.1186/1471-2164-15-1

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings Funct. Genomics* 9 (2), 166–177. doi: 10.1093/bfgp/elq001

Joshi, T., Fitzpatrick, M. R., Chen, S., Liu, Y., Zhang, H., Endacott, R. Z., et al. (2013). Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. *Nucleic Acids Res.* 42 (D1), D1245–D1252. doi: 10.1093/nar/gkt905

Joshi, T., Patil, K., Fitzpatrick, M. R., Franklin, L. D., Yao, Q., Cook, J. R., et al. (2012). Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC Genomics* 13 (1), S15. doi: 10.1186/1471-2164-13-S1-S15

Jun, T. H., Van, K., Kim, M. Y., Lee, S. H., and Walker, D. R. (2008). Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica* 162 (2), 179–191. doi: 10.1007/s10681-007-9491-6

Jung, C. H., Wong, C. E., Singh, M. B., and Bhalla, P. L. (2012). Comparative genomic analysis of soybean flowering genes. *PloS One* 7 (6), e38250. doi: 10.1371/journal.pone.0038250

Komatsu, S., Hiraga, S., and Nouri, M. Z. (2014). Analysis of flooding-responsive proteins localized in the nucleus of soybean root tips. *Mol. Biol. Rep.* 41 (2), 1127–1139. doi: 10.1007/s11033-013-2959-7

Komatsu, S., Kuji, R., Nanjo, Y., Hiraga, S., and Furukawa, K. (2012). Comprehensive analysis of endoplasmic reticulum-enriched fraction in root tips of soybean under flooding stress using proteomics techniques. *J. Proteomics* 77, 531–560. doi: 10.1016/j.jprot.2012.09.032

Krizhevsky, A., Sutskever, I., and Hinton, G. E., (2012). Imagenet classification with deep convolutional neural networks, in *Advances in neural information processing systems*. 1097–1105.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34 (8), 816–834. doi: 10.1002/gepi.20533

Liang, Z., Huang, J. X., Zeng, X., and Zhang, G. (2016). DL-ADR: a novel deep learning model for classifying genomic variants into adverse drug reactions. *BMC Med. Genomics* 9 (2), 48. doi: 10.1186/s12920-016-0207-4

Ma, W., Qiu, Z., Song, J., Cheng, Q., and Ma, C. (2017). DeepGS: Predicting phenotypes from genotypes using Deep Learning. *bioRxiv* 241414. doi: 10.1101/241414

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248 (5), 1307–1318. doi: 10.1007/s00425-018-2976-9

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11 (7), 499. doi: 10.1038/nrg2796

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3: Genes Genomes Genet.* 8 (12), 3813–3828. doi: 10.1534/g3.118.200740

Pantalone, V. R., Kenworthy, W. J., Slaughter, L. H., and James, B. R. (1997). Chloride tolerance in soybean and perennial Glycine accessions. Euphytica 97, 2, 235–239. doi: 10.1023/A:1003068800493

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Pérez, P., de los Campos, G., Crossa, J., and Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian

linear regression package in R. *Plant Genome* 3 (2), 106–116. doi: 10.3835/plantgenome2010.04.0005

Prince, S. J., Song, L., Qiu, D., dos Santos, J. V. M., Chai, C., Joshi, T., et al. (2015). Genetic variants in root architecture-related genes in a Glycine soja accession, a potential resource to improve cultivated soybean. *BMC Genomics* 16 (1), 132. doi: 10.1186/s12864-015-1334-6

Pumperla M. (2019). Hyperas: A very simple wrapper for convenient hyperparameter optimization. v 0.4.1. https://github.com/maxpumperla/hyperas.

Rutkoski, J. E., Poland, J., Jannink, J. L., and Sorrells, M. E. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes Genomes Genet.* 3 (3), 427–439. doi: 10.1534/g3.112.005363

Schaeffer, L. (2006). Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* 123 (4), 218–223. doi: 10.1111/j.1439-0388.2006.00595.x

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *nature* 463 (7278), 178. doi: 10.1038/nature08670

Simonyan, K., Vedaldi, A., and Zisserman, A., (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv* 1312, 6034.

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 (2). doi: 10.3835/plantgenome2016.10.0109

Spencer, M., Eickholt, J., and Cheng, J. (2015). A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM Transactions Comput. Biol. Bioinf. (TCBB)* 12 (1), 103–112. doi: 10.1109/TCBB.2014.2343960

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., et al. (2015). Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PloS Genet.* 11 (2), e1004982. doi: 10.1371/journal.pgen.1004982

Srivastava, N., et al. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Machine Learn. Res.* 15 (1), 1929–1958.

Stekhoven, D. J., and Bühlmann, P. (2011). MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28 (1), 112–118. doi: 10.1093/bioinformatics/btr597

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *In,* #Thirty-First AAAI Conference on Artificial Intelligence. p. 12. 2017.

Technow, F. R. (2011). *Package hypred: Simulation of Genomic Data in Applied Genetics*. Stuttgart, Germany: University of Hohenheim, Institute of Plant Breeding, Seed Science and Population Genetics.

Uppu, S., Krishna, A., and Gopalan, R. P. A. (2016). Deep learning approach to detect SNP interactions. *JSW* 11 (10), 965–975. doi: 10.17706/jsw.11.10.965-975

Verbyla, K. L., Hayes, B. J., Bowman, P. J., and Goddard, M. E. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res.* 91 (5), 307–311. doi: 10.1017/S0016672309990243

Vlasblom, J., and Wodak, S. J. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinf.* 10 (1), 99. doi: 10.1186/1471-2105-10-99

Wang, J., Joshi, T., Valliyodan, B., Shi, H., Liang, Y., Nguyen, H., et al. (2015). A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC Genomics* 16 (1), 1011. doi: 10.1186/s12864-015-2217-6

Wang, D., Liang, Y., and Xu, D. (2018). Capsule network for protein post-translational modification site prediction. *Bioinformatics*. 35 (14), 2386–2394. doi: 10.1093/bioinformatics/bty977

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., et al. (2017). MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 33 (24), 3909–, 3916. doi: 10.1093/bioinformatics/btx496

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PloS Comput. Biol.* 13 (1), e1005324. doi: 10.1371/journal.pcbi.1005324

Won Oh, M., Nanjo, Y., and Komatsu, S. (2014). Analysis of soybean root proteins affected by gibberellic acid treatment under flooding stress. *Protein Peptide Letters* 21 (9), 911–947. doi: 10.2174/0929866521666140403122602

Xavier, A., Beavis, W. D., Specht, J. E., Diers, B., Muir, W. M., and Rainey, K. M. (2015). SoyNAM: Soybean nested association mapping dataset. *R package version*, 1.

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., et al. (2018). Genome-Wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3: Genes Genomes Genet.* 8 (2), 519–529.

Xavier, A., Muir, W. M., and Rainey, K. M. (2016). Assessing predictive properties of genome-wide selection in soybeans. *G3: Genes Genomes Genet.* g3, 116.032268. doi: 10.1534/g3.116.032268

Xavier, A., Muir, W. M., and Rainey, K. M. (2016). Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. *BMC Bioinf.* 17 (1), 55. doi: 10.1186/s12859-016-0899-7

Xu, Z., and Taylor, J. A. (2009). SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* 37 (suppl_2), W600–W605. doi: 10.1093/nar/gkp290

Yi, H., Ravilious, G. E., Galant, A., Krishnan, H. B., and Jez, J. M. (2010). From sulfur to homoglutathione: thiol metabolism in soybean. *Amino Acids* 39 (4), 963–978. doi: 10.1007/s00726-010-0572-9

Zhang, N., Rao, R., Salvato, F., Havelund, J., Møller, I., Thelen, J., et al. (2018). MU-LOC: A machine-learning method for predicting mitochondrially localized proteins in plants. *Front. Plant Sci.* 9, 634. doi: 10.3389/fpls.201800634

Zhang, X., Wang, W., Guo, N., Zhang, Y., Bu, Y., Zhao, J., et al. (2018). Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genomics* 19 (1), 226. doi: 10.1186/s12864-018-4582-4

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., et al. (2012). Accuracy of genomic selection in European maize elite breeding populations. *Theor. Appl. Genet.* 124 (4), 769–776. doi: 10.1007/s00122-011-1745-y

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12 (10), 931. doi: 10.1038/nmeth.3547

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33 (4), 408. doi: 10.1038/nbt.3096

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

REVIEW
published: 15 April 2020
doi: 10.3389/fgene.2020.00350

Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci

Hannah L. Nicholls[1,2], Christopher R. John[2,3], David S. Watson[2,4], Patricia B. Munroe[1,5], Michael R. Barnes[1,2,5,6] and Claudia P. Cabrera[1,2,5]**

[1] Clinical Pharmacology, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, [2] Centre for Translational Bioinformatics, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, [3] Centre for Experimental Medicine and Rheumatology, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, [4] Oxford Internet Institute, University of Oxford, Oxford, United Kingdom, [5] NIHR Barts Biomedical Research Centre, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, [6] The Alan Turing Institute, British Library, London, United Kingdom

OPEN ACCESS

Edited by:
Ryan Urbanowicz, University of Pennsylvania, United States

Reviewed by:
Le Shu, University of California, Los Angeles, United States
Richa Gupta, DNAnexus, Inc., United States

***Correspondence:**
Michael R. Barnes
m.r.barnes@qmul.ac.uk
Claudia P. Cabrera
c.cabrera@qmul.ac.uk

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

Received: *19 December 2019*
Accepted: *23 March 2020*
Published: *15 April 2020*

Citation:
Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR and Cabrera CP (2020) Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Front. Genet. 11:350. doi: 10.3389/fgene.2020.00350

Genome-wide association studies (GWAS) have revealed thousands of genetic loci that underpin the complex biology of many human traits. However, the strength of GWAS – the ability to detect genetic association by linkage disequilibrium (LD) – is also its limitation. Whilst the ever-increasing study size and improved design have augmented the power of GWAS to detect effects, differentiation of causal variants or genes from other highly correlated genes associated by LD remains the real challenge. This has severely hindered the biological insights and clinical translation of GWAS findings. Although thousands of disease susceptibility loci have been reported, causal genes at these loci remain elusive. Machine learning (ML) techniques offer an opportunity to dissect the heterogeneity of variant and gene signals in the post-GWAS analysis phase. ML models for GWAS prioritization vary greatly in their complexity, ranging from relatively simple logistic regression approaches to more complex ensemble models such as random forests and gradient boosting, as well as deep learning models, i.e., neural networks. Paired with functional validation, these methods show important promise for clinical translation, providing a strong evidence-based approach to direct post-GWAS research. However, as ML approaches continue to evolve to meet the challenge of causal gene identification, a critical assessment of the underlying methodologies and their applicability to the GWAS prioritization problem is needed. This review investigates the landscape of ML applications in three parts: selected models, input features, and output model performance, with a focus on prioritizations of complex disease associated loci. Overall, we explore the contributions ML has made towards reaching the GWAS end-game with consequent wide-ranging translational impact.

Keywords: machine learning, artificial intelligence, genome-wide association study, genomics, candidate gene, clinical translation, deep learning, data science

INTRODUCTION

A genome-wide association study (GWAS) examines a genome-wide set of genetic variants in a group of individuals to identify variants associated with a trait or phenotype. The goal of GWAS is to identify variants which show a statistically significant association with a phenotype. This enables guided functional investigation of the most likely causal variants and genes driving the genetic association, thus pinpointing genes and pathways of interest for disease diagnosis, drug discovery, and precision medicine.

As GWAS studies have scaled up to discover ever more disease variants (Evangelou et al., 2018; Giri et al., 2019; Nalls et al., 2019) it has become impractical to perform functional investigation on all disease relevant loci. This limitation arises in part due to variability in reporting of GWAS results, some studies report loci which have been independently replicated in a different cohort (the gold standard approach), and others do not. This reporting can question the confidence of some discovered loci, calling for a balance between stringent *p*-values to correct for multiple testing and false discovery, and conservative correction leading to false negative association. A compounding factor is also the need to differentiate causal variants or genes from other genes associated by linkage disequilibrium (LD), thus confounding the detection of causal genes within a locus – making it unclear which variants and genes warrant further analysis and potential functional study. This range of issues undermines the robustness of GWAS, and challenges the validity of downstream analyses and biological hypothesis development, critically undermining some of the major motivators for performing GWAS in the first place, such as target validation (Hurle et al., 2016). Ultimately this highlights the need for computational solutions to improve the signal to noise ratio of GWAS results and to highlight genes and variants that are most likely to be causal.

Machine learning (ML) has been one emerging branch of computational applications (alongside network analysis and tools such as text-mining) built to enhance GWAS performance and downstream interpretation (Seyyedrazzagi and Navimipour, 2017; Raj and Sreeja, 2018). Machine learning algorithms build mathematical models that are learnt from training data in order to make predictions or decisions. Machine learning consists of supervised, unsupervised, and reinforcement learning methods, with supervised and unsupervised learning being the most commonly implemented with GWAS data. Supervised learning provides ML algorithms with labeled training data and aims to infer a mapping function from the input variables to the output variable – or label for classification tasks (**Figure 1**). This mapping function may then be used to predict the labels of new "testing" data. Unsupervised learning, by contrast, has no response variable. Instead, the algorithm must attempt to find patterns in the data, such as clusters or outliers. When tailored for understanding GWAS data, ML predictions can provide an improved statistical foundation of evidence to support or improve GWAS results. For instance, ML in GWAS has been applied to identify loci, increase the statistical power of GWAS (Mieth et al., 2016), detect epistatic interactions (Leem et al., 2014), improve polygenic risk scoring produced from GWAS (Pare et al., 2017), and prioritize genes and variants on post-GWAS analysis (Vitsios and Petrovski, 2019). Here we will focus on the ML applications developed for post-GWAS prioritization.

The growth of GWAS over the past decade has identified thousands of associated loci, in September 2019 the NHGRI-EBI GWAS catalog contained 161,525 variant-trait associations from 4,298 publications[1]. Thousands of variant associations can now be found within a single complex disease, such is the case for inflammatory bowel diseases (IBD) with 1,829 variant associations and schizophrenia with 3,069 variant associations (see text footnote 1). In the case of blood pressure (BP) with 5,148 associations (see text footnote 1) 2,293 genes are implicated (Evangelou et al., 2018; Giri et al., 2019), these represent almost 10% of the known gene complement and 5.82% of the genome by LD alone. These results represent an important insight into the complex systems regulating BP and offer a basis for a better understanding of BP biology and the personalization of hypertension treatment. However, this knowledge still has great potential to confound understanding. Based on the simplifying assumption that each locus is driven by only one gene (whereas gene cluster associations are also possible), if we subtract 901 loci reported by Evangelou et al. (2018) from 2,605 genes mapping to these loci, 65.4% of "associated" genes can be expected to be unrelated to BP. This level of signal to noise, still presents a considerable problem to the formulation of an efficient follow up strategy.

Individual GWAS loci have already shown the potential for large scale prioritization by providing novel biological insights and potential drug targets and drug repositioning opportunities (Sanseau et al., 2012). For example, a GWAS on BP found associations in the *SLC5A1* gene. The association of *SLC5A1* with BP and its role as a target of a type 2 diabetes drug, canagliflozin, highlights the opportunity to re-purpose drugs for treating hypertension (Evangelou et al., 2018). Currently, research has shown only 38% of essential hypertension patients have effective treatment (Banegas et al., 2011). Similarly, IBD and schizophrenia both currently have lacking treatment options alongside their thousands of associations (Danese, 2012; Leucht et al., 2013) – suggesting that a path to improved therapeutics for complex diseases may lie within the associated loci and the biological functions contained within them.

Defining functional impact of associated variants is a unique challenge in itself, but it is subsumed by a greater problem. Although it is possible to predict functional impact with some confidence in coding regions and to a lesser extent in non-coding regions, differentiating variants and inferring causality is very challenging without further laboratory investigation. For example, BP associations found in several *SMAD* family genes and the *TGF*β gene, which collectively participate in the TGFβ pathway, led to the suggestion that these may affect sodium transport in the kidney and ventricular remodeling (Evangelou et al., 2018). However, multiple genes impacting the same pathway raise the question of which gene should be functionally investigated first. Usually the evidence is not strong enough to warrant laboratory investigation of all the associated genes in a

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

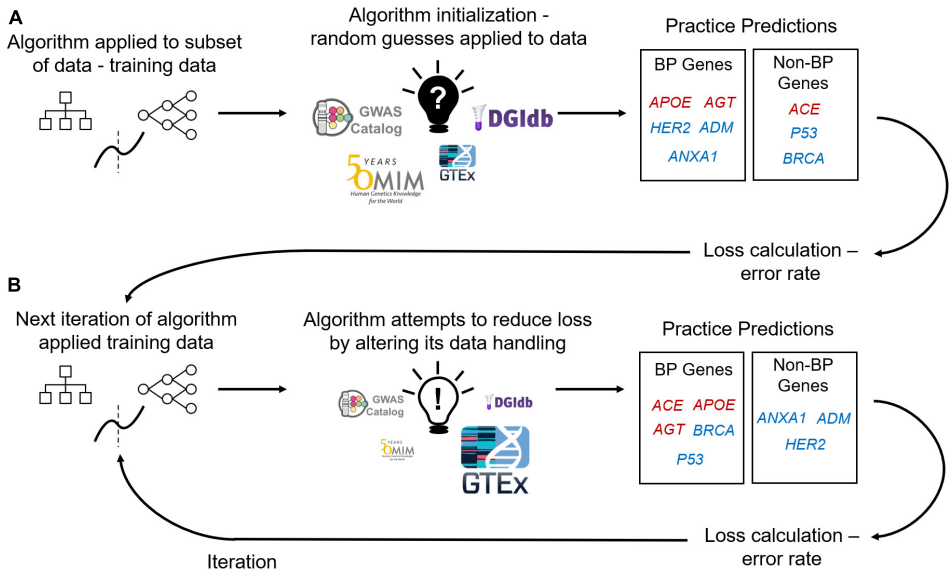

FIGURE 1 | Supervised Machine Learning Algorithm Training. **(A)** Data containing labeled genes (e.g., genes labeled as causal or non-causal for blood pressure – BP) and columns of features describing those genes are input into a machine learning algorithm. Machine learning algorithms firstly initialize themselves by applying their rules to a subset of the data (deemed training data) and its features at random. E.g., an algorithm's first practice iteration can involve assigning feature importance at random (importance denoted by size of feature image). The algorithm uses its feature initialization to classify genes into either affecting BP (red genes) or not affecting BP (blue genes). Algorithms then use the practice predictions to calculate loss (an error rate) and iterate over the data again with applying the previous iteration's loss calculation to adjust feature handling **(B)**. With using the loss calculations the algorithm aims to improve predictive performance with each training iteration.

particular pathway. The follow-up GWAS laboratory studies to date have developed without a standardized method for selecting causal genes and consequently they are likely to be susceptible to personal or "cherry picking" bias. These issues highlight the need for a pipeline that methodically triages variants and genes based on their likelihood of affecting a trait. Only then, will there be consistency in follow-up of genetic results using functional analysis with minimized risk of investigating false positives or low impact genes. The standardized *in silico* identification of the most likely causal genes at a genome scale may be an opportunity to gain higher level systems insights into trait biology. This in turn may help to fine-tune ML algorithms, as seen with research using ML variant prioritization as a feature fed into gene prioritization (Khan et al., 2018).

The development of systematic prioritization post-GWAS using ML has been researched as early as 2007 (Lewinger et al., 2007). Since then several computational methods for prioritizing GWAS associated loci have been developed with growing attention on ML applications (Fridley et al., 2011; Gagliano et al., 2015; Raj and Sreeja, 2018; Wu et al., 2018). ML for prioritizing GWAS results has used common models (**Figure 2**) such as logistic regression, decision tree classifiers such as – e.g., gradient boosting machines (GBM) and random forests (Wang et al., 2013; Oh et al., 2017), – and support vector machines (SVM; Vitsios and Petrovski, 2019), with more recent advances including deep learning models (Khan et al., 2018; Zhou et al., 2018).

An increasing number of studies are investigating how ML can be tailored to locus prioritization across diseases, but the ML pipelines for GWAS prioritization are mainly limited by the range and quality of training data. In order for ML models to present reliable guide-posts for post-GWAS research, a critical assessment of developing methods is needed – as the most recent systematic and literature reviews of post-GWAS prioritization cover few ML studies in comparison to other prioritization methods (Seyyedrazzagi and Navimipour, 2017; Raj and Sreeja, 2018). Here we will review the current landscape of ML applications for post-GWAS prioritization, and how ML can aid reaching the end-game for GWAS, which we define as a state where all common population variation with impact on a trait is identified; providing solid biological insights and mechanisms with reliable translational capability.

MACHINE LEARNING MODELS

GWAS prioritization as a classification problem has been approached using both simplistic and complex models (**Table 1**) depending on the problem requirements and data available. Primarily five types of models have been implemented: logistic regression, SVM, random forest, gradient boosting, and deep neural networks (**Figure 2**), each with varying advantages and disadvantages (**Table 2**). Logistic regression is a commonly applied statistical method that when used with categorical variables can be contemplated as a generalized linear model. In a logistic regression, it is typical to apply a regularization term – e.g., L1 (the sum of the absolute value of feature weights) and L2 (the sum of squared feature weights) – that introduce some

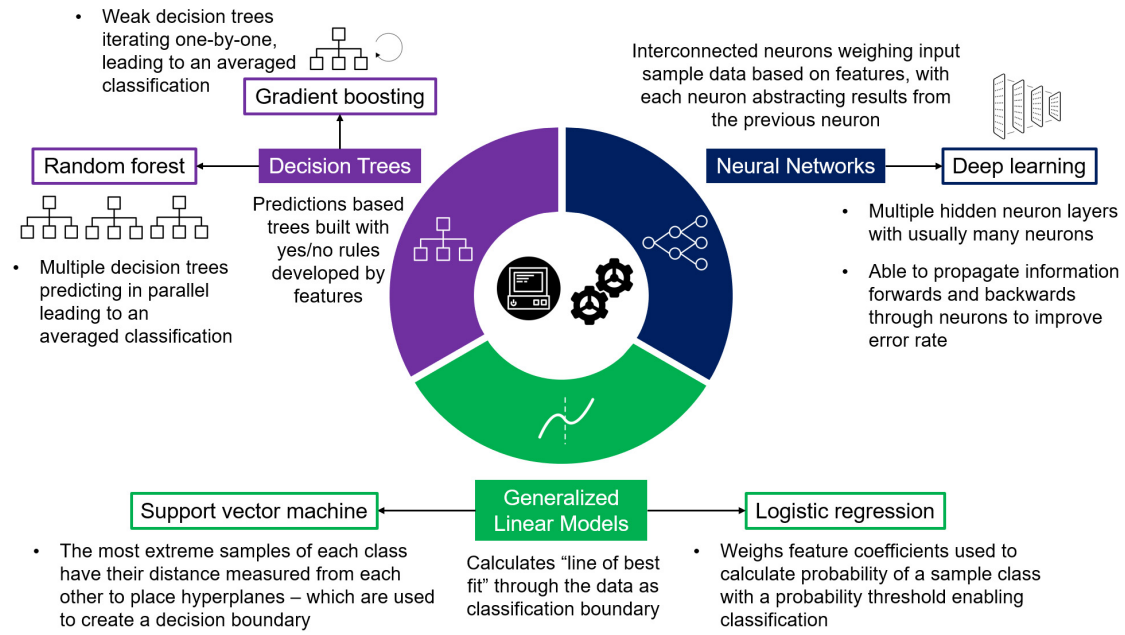

FIGURE 2 | Supervised Machine Learning Models. Diagram detailing three machine learning model bases used in supervised learning, each providing varying algorithms most commonly used in post-GWAS prioritization.

bias while reducing variance, thereby improving predictive ability (Demir-Kavuk et al., 2011). Isakov et al. (2017) used elastic net logistic regression (Zou and Hastie, 2005) which combines L1 and L2 penalties to prioritize IBD genes. This method performs both variable selection (L1), and shrinks coefficient sizes to reduce variance (L2) (Ogutu et al., 2012). Regularized logistic regression with elastic net aims to minimize the "curse of dimensionality" – where data has a larger number of features than samples – which is a particular blight on GWAS. For example, Isakov et al. (2017) used data consisting of 314 positive genes and 1,736 negative genes each annotated with 1,027 features. By applying logistic regression with elastic net they could then select the best data for their models (309 features selected which are predominantly from biological ontologies). However, due to the growing size in genetic data, and the broader range of features becoming available to describe genes and variants, the increased computational demand requires more advanced models.

Seven out of 19 ML models for post-GWAS prioritization curated in this review (**Table 1**) are ensemble models, namely random forests and gradient boosting. Ensemble methods combine multiple models to improve performance and are ideal for heterogenous GWAS data. Deo et al. (2014) developed a GBM (OPEN – Objective Prioritization for Enhanced Novelty) for prioritizing causal genes in multiple diseases. They used data comprising of more than 40,000 genomic features from public databases [Gene ontology (GO), Mouse Phenotype database, Human Phenotype Ontology (HPO), and Online Mendelian Inheritance in Man (OMIM)] aiming to benefit from unbiased features. GBM is a tree-based model, with tree branches performing yes/no decisions leading to a sample's classification (Natekin and Knoll, 2013). GBM operates one tree at a time, attempting to optimize with each tree. Deo et al. (2014) made accurate predictions with GBM identifying genes affecting cardiovascular disease (CVD) related traits. Performance was measured by the area under the receiver operating characteristic curve (AUROC), with values ranging between 0.75 and 0.9 across traits (Deo et al., 2014). The model's consistently high scores are due to the ensemble methods providing the opportunity for predictive mistakes to be removed in aggregate, due to multiple models testing different hypotheses and taking an average, expanding the representational space of a classification problem (Dietterich, 2000). This is seen with gradient boosting across research, with the model known for reducing bias and variance and offering improved accuracy (Natekin and Knoll, 2013). However, there is also a need to benchmark model performance, as whilst ensemble models are reliable, a singular approach into a novel classification problem provides a risk of unnoticed overfitting – which is also a known issue for gradient boosting depending on regularization techniques used.

Vitsios and Petrovski (2019) built a semi-supervised learning framework in which they benchmarked seven models (random forest, extremely randomized trees, GBM, extreme gradient boosting, SVM, deep neural networks, and a stacking classifier using all models) to prioritize genes for three diseases – amyotrophic lateral sclerosis, chronic kidney disease and epilepsy. In total they used data containing more than 1,200 features describing tens of thousands of genes for each disease. They found that random forest was the top-performing classifier, with this ensemble model consisting of multiple decision trees predicting in parallel (Breiman, 2001). Gradient boosting was the second most accurate, showing the high performance of

TABLE 1 | Curation of machine learning studies applied to post-GWAS prioritization of variants and genes.

PMID	Models description	Methods description	Data and assessment descriptions
30692607*	LR; Genes – Crohn's disease	Uses backward stepwise regression to build significant expression datasets (with emphasis on epigenetic data) to give prediction in combination with genotype data. Expression data reduces the uncertainty of smaller effect loci shown in fine-mapping and prioritization was followed-up with protein network analyses for validation	10-fold cross validation (2,000 genes per fold)
25935003*	LR; Genes – Crohn's disease	Combines GWAS results with gene expression features and whether genes are associated with other autoimmune diseases to better identify disease-related genes. More powerful for prioritizing rare missense variants	Cross-validation performed. 50:50 training:testing ratio. Training iterated 500 times. 54 Crohn's disease genes used as positively labeled training genes
29407288	SVM, LASSO, classification-regression trees; Variants – major depressive disorder and adverse drug response (duloxetine)	Models used features selected by LASSO regression and classified variants based on a clinical depression scoring defining drug response and remission	Dataset size: 186 patients. Nested 5-fold cross-validation. 80:20 training:testing ratio
21317188*	SVM, RF; Variants – arthritis and T1D	Compares support vector machine and random forest performance to chi-squared ranking	Dataset size: 452,176 T1D SNPs 63 arthritis SNPs
31779641	RF; Variants – intronic variants associated to cellular sensitivity to clofarabine-induced cytotoxicity	Focuses on integrating splicing data features with other types. Validates model prioritization with laboratory follow-up – limited by technical noise during laboratory work	3-fold cross-validation. Training data size: 6,676 variants. Testing data size: 1,222 variants
24564704*	Parallel RF Regression; Variants – brain structure and function. Alzheimer's disease GWAS	Designed to run on large Hadoop clusters, including those available through cloud computing. Multivariate applications not available on Hadoop	Each tree bootstraps to form training data (63.2%) with out-of-bag samples for test data. 500 simulated datasets
28592878*	RF Hyper-ensemble; Non-coding variants – curated mendelian diseases	Addresses class imbalance via resampling using simultaneous oversampling of minority class and undersampling of majority class. HyperSMURF can detect disease variants nearby to non-disease variants	10-fold cross-validation partitioning variants into chromosomal bands so no variants had same location, gene or disease in training and testing. GWAS total size approximately 2,000 variants
25633252*	GB; Genes – cardiovascular diseases and traits	Explored prioritization of 38 phenotypes (predominantly cardiovascular). Each tree within model updates a log-odds of disease association per gene. GWA-prediction assigns scores to genes in loci based on reasonings (transcription sites, experimental evidence, etc.) to identify likely positives which are used in training for phenotypes with GWAS training data	Six rounds of 8-fold cross-validation. Seventy percent of loci as positive training examples with matching numbers of negative samples
30591030*	LR and DL; Genes and variants – schizophrenia and autism	Performed variant prioritization which fed into gene prioritization. Variant prioritization used eQTL and pathogenic scoring data features. Gene prioritization used the variant rank in combination with genotypic data. Used to prioritize an individual's variants and genes and can be re-applied to GWAS data	10-fold cross-validation on four training and test sets
28795970	LR with elastic net, RF, SVM with polynomial kernel, extreme GB; Genes – inflammatory bowel diseases	All genes in dataset were annotated with 1,027 features. 16,390 genes scored and classified, with prediction as a score between 0 and 1. Models evaluated separately and together in combined performance score	5-fold cross-validation repeated 10 times. Training data: 314 positive genes and 1,736 negative genes
30013180*	DL – ExPecto; Variants – publicly available GWAS for four immune diseases	Data profiling >140 million promoter-proximal mutations allowed for deep learning to predict variant effect, with effect feeding into the prioritization of SNPs	Dataset size: 390,085 variants. Whole-chromosome holdout of chromosome 8 with 990 genes – using these genes for testing
30859622	LR with stochastic gradient descent, SVM, RF, K-Nearest Neighbors; Genes – colorectal cancer	Used a network approach – collecting both global and local data to create an epistasis network. Topology of the network was then used as features in machine learning, with different types of feature selection compared, to prioritize genes biologically relevant to colorectal cancer	Dataset size: 185,180 SNPs. Training on 90% of the dataset with 10-fold cross validation

(Continued)

TABLE 1 | Continued

PMID	Models description	Methods description	Data and assessment descriptions
doi: 10.1101/655449*	SVM, RF, extra trees, GB, extreme GB, DNN and a stacking classifier with four base classifiers (RF, extra trees, GB and SVM) followed by a DNN in the second layer. Genes – chronic kidney disease, amyotrophic lateral sclerosis, epilepsy	Models applied with positive-unlabeled learning – stochastic semi-supervised learning. Explored combinational impact of all models, and chose best performing model for each disease. There was a dependency on existing patterns – beneficial for finding new causal associated genes which may impact known mechanisms	10-fold cross validation. Gene samples: 25,000 for chronic kidney disease, 17,000 for epilepsy and 79,500 for amyotrophic lateral sclerosis
21687685	Bayesian latent variable model; Variants – ovarian GWAS	Used features about a SNP to estimate a latent quality score, with SNPs prioritized based on the posterior probability distribution of the rankings of latent quality scores. Incorporated the uncertainty of the ranking into the prioritization via probability calculation	NA
23369106*	Genetic algorithm; Variants – select OMIM diseases	Algorithm estimates feature weights to characterize SNPs related to an input dataset of genes, biological processes or GWAS results. Users can select features and assign a custom relevance and model relies on data mining of public data	Leave one out cross validation – single disease in the set used to validate (repeats for each disease)
29874547*	Network representation learning (random walk); Genes – Parkinson's, RA, Crohn's, Ulcerative Colitis, CAD, T2D	Unsupervised model learns embeddings of genes from multiple gene networks and develops hierarchical statistical model to integrate the learned embeddings of genes with GWAS summary data. Gene-level *p*-values infer each gene's posterior probability of association, which is in turn used for gene prioritization. Lack of direct biological interpretations available for the learned embeddings of genes	NA
21977986*	Multi-task learning ProDiGe; Genes – 265 diseases and 936 associations	Model learns from positive and unlabeled examples. The model shared information across diseases to improve the predictive performance for diseases with minimal positive labeled genes. The information shared is weighed depending on similarity of one disease to another	Training set: at least one known disease gene in training data. Training data per disease >11 genes. Leave one out validation on select diseases
26504140*	Unsupervised model – bayes classifier – GenoWAP; Variants – schizophrenia and Crohn's disease	Unsupervised learning – integrates GenoCanyon (their previous model) functional prediction and GWAS *p*-values. Reduce noises caused by linkage disequilibrium and rescues marginal signals in GWASs with insufficient sample sizes	NA
27058395*	Unsupervised model – bayes classifier – Genoskyline; Variants – schizophrenia and coronary artery disease	Successor of GenoWAP model, building from it by using annotations integrating tissue-specificity. Customizable with researchers able to input many feature annotations. Whilst tissue-specific it also lacked data from all tissue types	NA

**Software/code available; LR, logistic regression; RF, random forest; GBM, gradient boosting machine; SVM, support vector machine; DL, deep learning; DNN, deep neural networks; ET, extra trees; GWAS; genome wide association study; SNP; single nucleotide polymorphism; CAD, coronary artery disease; T1D, type 1 diabetes; T2D, type 2 diabetes.*

tree-based ensemble classification. However, the AUCs between all algorithms were deemed too similar to conclude one model out-performed all others across datasets. These results were also supported by comparison with a combined framework using all models in prioritization, the stacking classifier, ensuring the highest reliability in the chosen classifier for each disease (Vitsios and Petrovski, 2019). Kafaie et al. (2019) aimed to prioritize genes associated with colorectal cancer comparing various models (SVM, random forest, logistic regression with stochastic gradient descent, and K−nearest neighbors). They found that logistic regression was the highest performing ML model – emphasizing that a classification problem may require simpler solutions.

Besides ensemble learning and logistic regression, SVM is also consistently used within studies performing benchmark comparisons (Roshan et al., 2011; Isakov et al., 2017; Maciukiewicz et al., 2018; Vitsios and Petrovski, 2019). SVM aims to plot a decision boundary between groups by measuring hyperplanes – based on the distances between the most extreme samples of each classification group (Smola and Scholkopf, 2004; **Figure 2**). However, within benchmarking studies, SVM has not shown itself to be a top-performing model. For example, Vitsios and Petrovski (2019) found it had the lowest AUC (0.83, only slightly lower than the top-performing random forest at 0.85) of their seven models, while Kafaie et al. (2019) found SVM performed better than random forest yet worse than logistic regression. The varying performance of SVM also highlights the importance of input data, as Kafaie et al. (2019) were one of the only studies to focus on comparing feature selection methods as well as models. Kafaie et al. (2019) found SVM performed well given certain features, whilst in comparison logistic regression had a more stable high performance regardless of external selection, emphasizing the value of logistic regression's internal feature selection via regularization.

TABLE 2 | Comparison of machine learning model performance. Comparison of the most common models used in post-GWAS prioritization including performance metrics, comparing metrics of each model's highest performance score per study.

Models	PMID	Best performance	Model advantages and disadvantages
Logistic regression	25935003 28795970	0.94 (AUC) – Crohn's disease 0.775 (ROC) – inflammatory bowel diseases	Advantages: - Easy to implement - Efficient to train - High interpretability - Can act as a benchmark for exploring more complex algorithms Disadvantages: - Difficulty recognizing complicated data patterns - Difficulty handling large datasets
Random forest	28592878 31779641 21317188 28795970 doi: 10.1101/655449	0.635 (AUCROC) – curated Mendelian diseases 0.96 (AUCROC) – cellular sensitivity to clofarabine-induced cytotoxicity 0.81 (AUC) – T1D 0.80 (ROC) – inflammatory bowel diseases 0.85 (AUC) – average between all diseases	Advantages: - It can handle large data with higher dimensions - Ensemble method reduces overfitting by several models testing multiple hypotheses Disadvantages: - Many parameters to tune, affecting computational efficiency - Ensemble method lows interpretability
Gradient boosting	28795970 doi: 10.1101/655449 25633252	0.783 (ROC) – inflammatory bowel diseases 0.848 (AUC) – average between all diseases 0.959 (ROC) – HCM	Advantages: - High power performance - Flexible with several parameter tuning options - Ensemble method reduces overfitting by several models testing multiple hypotheses Disadvantages: - Reliance on high quality training data - Many parameters to tune, affecting computational efficiency
Support vector machine	28795970 29407288 doi: 10.1101/655449	0.786 (ROC) – inflammatory bowel diseases 0.66 (Accuracy) – major depressive disorder and adverse drug response (duloxetine) 0.832 (AUC) – average between all diseases	Advantages: - Computationally efficient - It handle can handle large data and high dimensions Disadvantages: - Does not provide class probabilities - Difficulty to interpret
Deep neural network	30013180	0.815 (AUCROC) – lymphoblastoid expression	Advantages: - Recognizes patterns in large complex data - High power performance - Able to handle noisy data Disadvantages: - Difficulty to interpret - Computationally expensive requiring GPUs for high power performance

AUC, area under curve; GPU, graphics processing unit; ROC, receiver operating characteristic; T1D, type 1 diabetes; HCM, hypertrophic cardiomyopathy.

Deep learning has also been explored for prioritization, this method can increase sensitivity in larger datasets due to the methods ability to incrementally capture abstract representations of high-level information. In general, this is beneficial for GWAS prioritization where the data is growing dramatically in size and heterogeneity with increasing annotations post-GWAS, and also has few labeled samples (known disease causing variants/genes) for supervised learning. Deep learning becomes advantageous in this scenario as it identifies complex patterns via supervised and unsupervised learning from large datasets (Najafabadi et al., 2015) and can be applied for further insights into GWAS data. However, whilst deep learning enables the consideration of millions of parameters, its application to date has mostly flourished in image classification and natural language processing (Zeng et al., 2018; Aung et al., 2019; Hampe et al., 2019), requiring an investment in its development and benchmarking with traditional models for developing GWAS application. A deep neural network (ExPecto) applied by Zhou et al. (2018) prioritized causal variants for immune-related diseases using sequence-based features. This dataset contained more than 140 million promoter-proximal mutations, and allowed for the unidirectional flow of information from base-sequence to functional predictions which enabled variant prioritization. To approach this large dataset ExPecto applies spatial transformation to the data, weighting transformations based on transcription start site distances. This was performed on a tissue-specific basis of over 200 tissues (Zhou et al., 2018), providing hundreds of features for the model to process. ExPecto is also able to perform pattern recognition and prioritization of rare and unobserved variants. However, whilst models are selected based on their

suitability to the data, performance can also be dependent on class balance and data quality available.

Predominantly, ML studies use cross-validation to ensure a reliable estimate of model performance. However, with GWAS data commonly lacking functionally validated disease causing variants and genes, there are minimal learning opportunities for supervised models. Oversampling or undersampling techniques can be used to address class imbalance. Schubach et al. (2017) developed a hyper-ensemble model (hyperSMURF) using random forests with imbalance-awareness by using both under- and oversampling. By balancing the training data classes, and exposing the base learners in the hyper-ensemble to different training datasets, the random forests are able to diversify their understanding of the data, improving accuracy regardless of data size. Using hyperSMURF they prioritized thousands of GWAS variants annotated with 1,842 features. Their sampling techniques created balanced training data, where the original GWAS data had a 1:700 label imbalance. However, oversampling techniques develop synthetic samples based on example data points to increase the minority class size, which can create overfitting. Schubach et al. (2017) addressed this by preventing example variants of the same location/gene to occur in the training and test sets, minimizing the oversampling bias.

Whilst only one post-GWAS prioritization study has focused on class imbalance (Schubach et al., 2017), several have targeted data quality with a focus on data labeling. For example, positive unlabeled learning is semi-supervised learning with only positive labeled examples, a common occurrence for GWAS data where only a few causal genes have been functionally validated. For positive unlabeled learning overfitting is avoided using approaches such as classing unlabeled samples as negative and bootstrapping random samples. Vitsios and Petrovski (2019) applied positive labels to disease genes from the HPO with further validating clinician confirmation, and treated any unlabeled genes as negative samples. They then conducted random sampling of positive and unlabeled samples, aiming to equalize the ratios of the positive and negative genes to expose their models to a balanced dataset. Mordelet and Vert (2011) also applied their model (ProDiGie) using positive unlabeled learning. Whilst they only had minimal positive samples per disease, the model shared information across diseases – enabling it to use information from causal genes for closely related diseases in prioritization. Despite these benefits, positive unlabeled learning is limited by prior knowledge of known causal genes, leading to potential false negatives, and unlikely scenarios for a model to prioritize genes in novel mechanisms.

Overall, there is a need for benchmarking in order to select the model best suited to the data, and for post-GWAS prioritization the optimal model currently varies across diseases without a one-size-fits-all winner. An optimal model also hinges on data size and quality for reliability and performance, with studies varying in data size and choice of features – from using hundreds of selected features (Isakov et al., 2017) to others exploring tens of thousands (Deo et al., 2014). Further *in silico* methods need to address these aspects of ML, the lack of functionally validated associated genes at the disposal of ML, and how features are used in order to build a model tailored to post-GWAS prioritization.

FEATURE CURATION

To fine-tune a model, researchers must perform data curation and feature quality control to achieve the best possible performance. GWAS associations are typically annotated to a wide range of biological annotations. Biological features range from eQTL (expression quantitative trait loci), RNA, epigenetic, and protein data to describe a variant or gene's functionality. For example, several studies use eQTL data, providing tissue-specific and population-specific insight, with researchers noting the use of eQTLs can improve the ability for models to distinguish single causal genes within a locus (Deo et al., 2014). For example, Ning et al. (2015) built a logistic regression for prioritizing Crohn's disease associated genes. They found that integration of eQTL data with GWAS data provided an overlap of information between the two that strengthened model performance. Furthermore, the cataloging of eQTLs mapped to non-coding RNA provides a better insight into how non-coding RNA affects gene expression (Branco et al., 2018), increasing the strength of regulatory information at the disposal of ML models. The growing integration of related biological features suggest this will provide clearer insight for models to be able to pinpoint the most likely disease causing genes in a locus (Branco et al., 2018; Dai et al., 2019).

Other features used by studies are those provided by GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. Merelli et al. (2013) built SNPranker using terms from GO and KEGG to prioritize variants across diseases via a genetic algorithm. This model focuses on user-guided optimization, which is beneficial as SNPranker also takes features from data mining, allowing the researcher to adjust feature weights to minimize bias. Merelli et al. (2013) also focus on sharing information across ontologies, illuminating similar genes to those with known functional causality, indicating that this can grow the causal gene list (Merelli et al., 2013). Despite this possibility of increasing the model's training data, expanding a list of causal genes based on known biological processes alone is likely to create susceptibility to bias and weaker model performance – as the model is then less able to prioritize loci within novel systems which may be affecting a phenotype.

The use of other biological features, e.g., RNA and epigenomic features, has also grown in recent years. These features may provide further insights into associated loci located in non-coding regions. For example, researchers developed and combined models GenoWAP and GenoSkyline (Lu et al., 2016a; Casas et al., 2005). Both methods use unsupervised learning – GenoWAP performs GWAS prioritization and GenoSkyline integrates tissue-specific and epigenomic annotations for predicting tissue-specific functional regions. They found these annotations showed both functional and non-functional tissue-specific variants were enriched, suggesting LD between variants in both regions (Lu et al., 2016b). For example, one schizophrenia associated locus within an intergenic region, upstream of *MMP16*, had high prioritization by GenoWAP in brain tissue (Lu et al., 2016b). This result is then augmented as GenoSkyline predicted that this locus plays a role in the functional regions downstream, offering new targets for further

research. However, they concluded that their results can be improved with cell-specific data. Since GenoWAP, the ExPecto tool was built to make cell-type-specific predictions with high accuracy which it uses for variant prioritization, providing a novel method for generating cell-specific data *in silico* (Zhou et al., 2018). Whilst this method of predicting cell-specific data is disadvantageous to manual curation, the systematic collection of cell-specific data is in development and standardized resources have not been widely applied to post-GWAS analysis. Methods such as ExPecto provide a starting point for cell-specific curation and also a potential benchmark for the manual curation as it develops.

Alongside general biological characterization, disease-specific data is gradually increasing, further enabling accurate prioritization of GWAS associated loci. Vitsios and Petrovski (2019) for example prioritized chronic kidney disease genes, using annotations from the Chronic Kidney Disease database among their features to improve stratification. Algorithmic scorings are also used for prioritization (e.g., Eigen, CADD, DANN, GWAVA, DeepSea). These scorings predict pathogenicity of variants based on their expected functional consequences, and have been used to aid variant prioritization, however, to the best of our knowledge this is only been demonstrated by Khan et al. (2018), requiring further exploration into their benefits as features in ML prioritization.

Beyond data collection, studies also need to consider feature importance and feature selection to gain an understanding of models "under-the-hood" This is often a part of why researchers choose L1 regularized logistic regression, which automatically performs feature selection. Several studies have used logistic regression, such as Isakov et al. (2017) with using the elastic net, who found positive feature coefficients (predicting causal genes) were highest for immune and inflammatory response features from GO. Recently Gettler et al. (2019) also used logistic regression – as part of their gene prioritization regression model (GPRM) – to prioritize genes for Crohn's disease. While Gettler et al. (2019) do not discuss the impact of feature importance, they note that GO enrichment analysis showed immune and inflammatory genes were significantly enriched. This enrichment is to be expected from an autoimmune disease, however, it also suggests validation for the feature importance found by Isakov et al. (2017). Maciukiewicz et al. (2018) applied L1 logistic regression to identify significant features, and followed-up with SVM for predicting causal variants for duloxetine response in major depressive disorder. They found a non-coding RNA annotation had the largest positive coefficient. However, unlike the study of IBDs, Maciukiewicz et al. (2018) is the first prioritization study to focus on their drug response phenotype, requiring further work to validate feature importance and begin to suggest how that may fit into biological understanding of GWAS results. There is also work focused primarily on improving feature selection for GWAS data (Szymczak et al., 2016; Nembrini et al., 2018). For example, random forests provide feature importance measures and have been investigated by Szymczak et al. (2016). They developed a recurrent relative variable importance measure from random forest to rank important variants in GWAS. This focus on feature importance developed a useful tool for highlighting loci deserving of functional-follow up and could be used to reduce false positive GWAS results (Szymczak et al., 2016). The only other study investigating feature importance in prioritization has been SNPranker, with Merelli et al. (2013) finding epigenetic features (namely enhancer, CpG islands, and DNase cluster data) had the highest importance for default prioritization. Additionally to a model's internal feature weightings, permutation is also able to provide feature importance, doing so for any model by shuffling feature values and viewing model error rate. Vitsios and Petrovski (2019) use permutation via the boruta algorithm, which creates synthetic features from random permutation to weigh the importance of original features and remove any unimportant annotations. For all studies incorporating feature selection or importance they note an improvement in model performance or understanding of their predictive reasoning.

PRIORITIZATION OF VARIANTS AND CANDIDATE GENES

Prioritization methods post-GWAS have had development for several models that aim to be applicable for multiple diseases – e.g., ExPecto (Zhou et al., 2018), GenoWAP (Lu et al., 2016b), HyperSMURF (Schubach et al., 2017), and SNPRanker (Merelli et al., 2013). For example, ExPecto used all publicly available GWAS data for prioritizing variants for Crohn's disease, ulcerative colitis, Behçet's disease, and hepatitis B virus (Zhou et al., 2018). On prioritization they found highly ranked variants were also most likely to be replicated across GWAS. For Crohn's disease the top prioritized variant by ExPecto was rs1174815 (Zhou et al., 2018), yet neither the variant or gene (*IRGM*) has been highly prioritized by any other study focusing on Crohn's disease. In comparison with other model rankings for Crohn's disease loci, there are only a handful of genes that have been highly prioritized in more than one study. An example of this is *GSDMB*, a gasdermin gene known to affect apoptosis in epithelial cells. GPRM prioritized this gene, alongside ExPecto prioritizing a variant in *GSDMB* (rs58989791) (Zhou et al., 2018; Gettler et al., 2019). This prioritization has aligned with experimental work recently focusing on *GSDMB* in IBDs, finding an increase in the gene's expression may have a developmental role for IBDs (Rana and Pizarro, 2019). Another disease that has been prioritized by multiple studies is Alzheimer's disease, for which models consistently prioritize *APOE* (Mordelet and Vert, 2011; Wang et al., 2013; Deo et al., 2014). However, this questions model training in these studies, as *APOE* has been reported as affecting Alzheimer's disease as early as 1993 (Schmechel et al., 1993).

An issue with prioritizing variants and genes is the ability to ascertain if the model predictions are accurate. Schubach et al. (2017) address this by prioritizing regulatory variants for both mendelian diseases and complex diseases, for which the mendelian disease variants had been validated with a biomedical literature review. They found hyperSMURF consistently out-performed other methods (Eigen, GWAVA, CADD, and DeepSea) on both mendelian and GWAS data, suggesting minimized risk of overfitting and the potential for

ML to be able to generalize across datasets. In terms of performance metrics, Schubach et al. (2017) also explore multiple measurements – F1 score, AUROC, precision, recall, and the area under the precision-recall curve (AUPRC) – however, other studies primarily use AUROC. Whilst AUROC is an excellent metric in many cases, it can be highly misleading for imbalanced datasets like those commonly found in GWAS prioritization (Jeni et al., 2013; Saito and Rehmsmeier, 2015). Precision-recall curves are a popular alternative in cases of extreme class imbalance, with Schubach et al. (2017) applying these in combination with other metrics in a particularly rigorous approach. Studies focused on addressing imbalanced data are important for developing reliable GWAS applications, and continuing to focus on imbalance-aware approaches will reinforce the reliability of model predictions as much as possible *in silico*.

In order to establish model capability past performance metrics, a prioritized variant or gene's causality can be evidenced with functional follow-up. For example, Lin et al. (2019) developed RegSNPs-Intron which was a random forest prioritizing intronic variants associated to cellular sensitivity to clofarabine-induced cytotoxicity – with the model primarily relying on splicing data. After prioritization they performed ASSET-seq (ASsay for Splicing using ExonTrap and sequencing), which measures the impact of splicing on an intronic variant. They found 63 out of 82 experimentally tested variants had a significant splicing impact in multiple cell lines (Lin et al., 2019), suggesting further directions for functional study and validating the RegSNPs-Intron's prioritization. Zhou et al. (2018) also performed experimental follow-up, looking at their top prioritized variants with a luciferase assay. This confirmed prioritized variants affect regulatory activity – e.g., variant rs381218 prioritized to affect chronic hepatitis B virus had a significant change in reporter activity, predicted also by ExPecto to impact *HLA-DOA* (Zhou et al., 2018). These functional results improve the interpretation of potential regulatory roles for prioritized loci by validating prioritizations *in vitro*, enabling hypotheses produced by ML to be confirmed and further expanded upon.

PAST AND PRESENT CARDIOVASCULAR MACHINE LEARNING PRIORITIZATION

ML approaches for post-GWAS prioritization have been applied over the last decade, with applications providing the projected outputs expected from GWAS with biological insights and translational results. In 2014, Deo et al. (2014) applied OPEN to prioritize 38 phenotypes, many of which were CVD traits. CVD is a particularly appropriate example to investigate, due to its high powered GWAS with thousands of associated loci, presenting a large benefit to gain from ML prioritization. To the best of our knowledge, this is the only ML study that includes CVD traits. OPEN was applied to prioritize BP associated loci, for which several of its highly ranked genes have since been studied in laboratory experiments and leading to insights on biological mechanisms with possible translational impacts. *NPR3* was the second prioritized gene to affect BP by Deo et al. (2014). At the time of prioritization GWAS was one line of evidence showing a relationship between *NPR3* and BP, however, Ren et al. (2018) focused on this gene's functional roles in vascular smooth muscle. They found variants at this locus were associated with reduced *NPR3* mRNA and changes to chromatin structure, supporting a regulatory role leading to increases in vascular smooth muscle proliferation and suggesting a mechanism which can be a therapeutic target for BP. Overall with examining the top ten prioritized BP genes by Deo et al. (2014) (*ANTXR2*, *NPR3*, *MECOM*, *PLCE1*, *ENPEP*, *PDGFRA*, *CACNB2*, *ARID5B, MRVI1*, and *GUCY1B3*) eight of the associations have been validated by GWAS and mechanisms characterized by experimental work and indicate effects on BP (Rippe et al., 2017; Takeuchi et al., 2018; Giri et al., 2019; Kichaev et al., 2019) – only *ANTXR2*[2] and *PDGFRA*[3] have not been validated in recent BP GWAS. The gene *GUCY1B3*, ranked tenth by Deo et al. (2014), and *JAG1* (ranked 11th) have consistently been studied in relation to BP and nitric oxide regulation (Rippe et al., 2019). Rippe et al. (2017) identified both genes as affecting Notch pathway signaling in the aorta of mice, rats and humans – this study provided further insight into each gene's activity across species. Interestingly, variants at *MRVI1* (ranked eighth) have been found to be genome-wide significant in an arterial stiffness GWAS (Fung et al., 2019), implying a possible relation to BP and opportunity for follow-up investigation such as with colocalization analyses (Kanduri et al., 2019).

OPEN also ranked genes without high prioritization but have since been demonstrated to be important to BP regulation and have clinical significance (Deo et al., 2014). An example of this is uromodulin (*UMOD*), which Deo et al. (2014) prioritized approximately in the middle of their rankings of hundreds of associated genes affecting BP. *UMOD* has been replicated in GWAS (Evangelou et al., 2018) and is a target currently being tested in a clinical trial for its interaction with NKCC2 in hypertension – as *UMOD* genotypes of increased or decreased expression affect salt sensitivity in the kidney and a person's propensity for hypertension.

Aside from BP, Deo et al. (2014) also report success for other cardiac conditions that have additional evidence and support today. *FLNC* was prioritized as affecting left ventricular diameter. Deo et al. (2014) investigated *FLNC* further in a zebrafish model, finding knocked down *FLNC* showed cardiac abnormalities and hypertrophy, and also found one dilated cardiomyopathy patient (who had no known dilated cardiomyopathy gene mutations) with a splice-site mutation in *FLNC*. This work aligned with *FLNC* gaining functional cardiovascular research attention, with its role in cardiomyopathies also being first discovered in 2014 (Valdes-Mas et al., 2014). This result validates OPEN's high performance for cardiomyopathies (AUCROCs of 0.88 and 0.96), with its performance ranging from 0.75 to 0.9 for all other cardiac traits. Notably, Deo et al. (2014) used known causal genes as their training examples for cardiomyopathies, unlike the use of GWAS associated genes in the training data for other phenotypes, implying the benefit of using well-curated input data.

[2]https://www.ebi.ac.uk/gwas/genes/ANTXR2
[3]https://www.ebi.ac.uk/gwas/genes/PDGFRA

The insights into the functions of prioritized genes since 2014 indicate the potential of ML for guiding hypothesis generation, but also outline examples of the experimental work ahead for validating the biological mechanisms of such ranked genes in order to confidently identify drug targets post-GWAS. With 451 BP associated genes gathered by Deo et al. (2014) in comparison to 2,993 validated associated genes in 2019 (Evangelou et al., 2018; Giri et al., 2019), this suggests that re-running OPEN now with updated data would provide interesting results detailing which genes have withstood the test of time in terms of maintaining their ranking.

DISCUSSION

Machine learning is advancing rapidly but its applications in GWAS are still in their infancy with respect to becoming gold standard methods producing consistently validated biological insights. This review has focused on post-GWAS ML prioritization methodologies ranging from model selection and input features, to performance assessment and output prioritization results. For model selection several studies explore only one algorithm without comparison. Studies using benchmarking comparisons with several models offer a form of standardization for selection, contributing to research transparency which is crucial for work justifying investment in functional study. Recent studies are more frequently incorporating benchmarking comparison showing the development of robust methodology in this field (Isakov et al., 2017; Kafaie et al., 2019; Vitsios and Petrovski, 2019).

The feature curation also needs improved interpretation of selected features and their importance, as current work highlights the need to account for bias within biological features, and the requirement for continued upkeep of biological data. This interlinks with a broader demand for standardized use of recently discovered datatypes, as prioritization studies differ in their resources, hindering the interpretation of model performance. For example with growing epigenomics resources, Cazaly et al. (2019) note this is leading to research using varying standardization methods. How that data is collected and recorded then also affects the reliability of ML methods and comparison of model performances. This point can also be made for models such as ExPecto or iMEGES firstly applying variant prediction which feeds into gene prioritization as a feature (Khan et al., 2018; Zhou et al., 2018), as there is a risk of the predicted features overfitting, and those features then not being reproducible.

There are also datatypes, such as clinical datasets and wider ranges of omics data which are underrepresented in ML prioritization studies. Studies focus on genomic features, however, the contributions of transcriptomic, epigenomic and proteomic data are less frequently investigated. This lack is contrasted by studies solely integrating wide-ranging omics data to calculate GWAS prioritization scores (Ayalew et al., 2012; Ciesielski et al., 2014) – and identifies potential for collaboration with ML to improve data integration methods. To date ML studies highlight the benefits of multi-omic integration, but few directly investigate that need (Merelli et al., 2013; Dai et al., 2019). Building this multi-omic range of data could improve accuracy and provide information specifying not only the most likely causal genes, but the biological functions contributing to their causality. With current data and research there is a disconnect between prioritization of genes and identification of the mechanism that links a feature to gene/variant causality, which could benefit hypothesis specification in functional work.

As high quality disease-specific data becomes increasingly available to fine-tune model training, ML models may become more efficient in the prioritization of heterogenous data to identify the most likely causal disease genes. However, reliance on specific annotations presents a challenge for the prioritization of novel genes and hence novel mechanisms without prior knowledge. More generally, models including data mining features are also susceptible to this issue, as they contribute to a bias for prioritizing already characterized genes in known disease pathways. These already researched genes may be highly ranked not due to impactful biological knowledge but simply due to having a wealth of study. Overall how feature curation is implemented is a key factor to the developing success of ML applications for GWAS, especially when considering imbalanced data where positively labeled disease genes and variants are limited. This highlights the need for high quality gene annotation and disease resources – if features are not accurately researched and curated, the potential for models to accurately prioritize GWAS results will be diminished, ultimately ML methods are limited by the quality and quantity of input training data.

When comparing output prioritizations there is a need to appraise the quality of the training data, understanding which genes/variants are included and how they might impact prediction. For example with the prioritization of *APOE* by models for Alzheimer's disease (Mordelet and Vert, 2011; Wang et al., 2013; Deo et al., 2014), it could also be argued that this validates the model performance, as this gene is expected to be prioritized. However, the studies prioritizing Alzheimer's disease genes do not provide their training and testing data to explore this further (Mordelet and Vert, 2011; Wang et al., 2013; Deo et al., 2014), showing the need to improve reproducibility. More recent studies prioritizing different phenotypes are beginning to provide both their data and source code, such as Khan et al. (2018), enabling the development of more accessible and reliable tools. This development is essential for applications to be used and interpreted by non-computer scientists and for the output biological findings to have a traceable reasoning as to why they were prioritized.

On investigating OPEN's prioritizations and comparing them with more recent research, it emphasizes the potential for post-GWAS ML to give GWAS results a wider-impact contribution to complex diseases. The accuracy of the model across multiple diseases identifies the possibility that one model can be applied to several diseases successfully. Furthermore, the early prioritization of diagnostic genes such as *FLNC* shows the power of ML which, when combined with functional follow-up building biological insights, can lead into translational impacts. However, OPEN also showed genes which upon recent review were mis-prioritized (*UMOD* and *ANTXR2*). This ranking may be due to Deo et al. (2014) using GWAS results as part of their training data with

them also noting that their features may be too weak to prioritize genes a part of novel mechanisms for a pathology (Deo et al., 2014). These misjudged genes highlight flaws applicable for all ML models, with reliance on current biological data, requiring that data to be high quality for reliable loci prioritization.

For future applications ML can learn from work such as Deo et al. (2014) in combination with more recent work on larger datasets, e.g., Zhou et al. (2018). Research can develop models aiming to be applied across diseases, and re-used by other researchers, with consideration for the size of present GWAS data, varying datatypes, and feature importance. Doing so could then lead to more accessible, reusable models – for example with source code or web-interfaces that are useable by a wider range of GWAS researchers – and create more globally implemented ML applications for GWAS prioritization, thus accelerating researchers towards the post-GWAS endgame of understanding disease.

With the creation of accessible models, a role for ML prioritization in personalizing medicine can develop. For example, ML could potentially be used to augment genetic risk scores, identifying which genes contribute to a person's high risk score, and offering more information at the disposal of clinicians. To build ML tools to a clinically acceptable standard, however, requires comparison with other prioritization methods and ensuring model interpretability. One of the most common other methods used in post-GWAS prioritization is network analysis. This method builds networks ranging from the gene to protein level, enabling a flow of information from GWAS to protein and metabolic pathways (Leal et al., 2019). However, studies note that gene networks can contain noise, and the analysis is confounded by its aggregation of GWAS data to the gene level, causing a loss of variant information (Wu et al., 2018; Leal et al., 2019). Machine learning offers an improvement for this with data integration, that can preserve variant information, and with the ability to handle noisy data. Another method identifying causality is Mendelian randomization, although in some cases this can provide a clear illustration of risk, such as the link between homocysteine concentration and stroke risk (Casas et al., 2005), it is limited to high risk variants and independent variables (Haycock et al., 2016). In comparison, unlike other computational methods, the choices ML models make for prioritization are not always clearly available to be understood by the user. However, ML has also been applied in combinational approaches with network modeling (Kafaie et al., 2019) and Mendelian randomization for causal inference (Hemani et al., 2017) to overcome the disadvantages of a singular method. Hybrid approaches such as these highlight the many avenues of ML research to be explored for developing optimal GWAS prioritization. Aside from method comparison, improving data curation, and model benchmarking, the interpretability of models is a critical challenge for future research, and one of the largest obstacles for GWAS prioritization by ML to gain widespread reliable use. Developing model interpretability will involve a strong understanding of not only a model's mechanics but of feature importance and known disease causing genes given in model training – requiring an interdisciplinary effort to explore the potential of ML post-GWAS prioritization in full.

KEY CONCEPTS

Supervised learning: Models learn from labeled training data. Labeled positive and negative examples in training allow a model to practice decision-making before being assessed on new "testing" data.

Unsupervised learning: Models learn from unlabeled data. The models recognize patterns between samples that can identify clusters or outliers.

Semi-supervised learning: Models use both labeled and unlabeled data during training to perform pattern recognition. This is usually with a larger amount of unlabeled data than labeled data and enables techniques such as positive unlabeled learning.

Overfitting: When a model performs well on training data but poorly on test data. Some amount of overfitting is inevitable, but extreme cases can render a model useless.

Cross-validation: A procedure for assessing generalization error. Data are split into k subsets (or folds) of roughly equal size. Train k separate models with each fold held out once for testing. Average error across the k trials is reported.

Class imbalance: When the ratio of positive to negative labels is far from one, creating less opportunity for a model to learn from the minority class. Imbalance-aware methods perform undersampling or oversampling of majority and minority classes, respectively, to balance the dataset.

Sensitivity: The number of true positive samples correctly classified by a model. Also known as the true positive rate or recall.

Specificity: The number of true negative samples correctly classified by a model. Also known as the true negative rate or selectivity.

Precision: The ratio of true positives to declared positives. Also known as the positive predictive value, and equal to the complement of the false discovery rate.

AUROC: Area under the receiver operating characteristic curve, which illustrates the tradeoff between sensitivity and specificity. Can be interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

AUPRC: Area under the precision-recall curve, which illustrates the tradeoff between precision and recall; useful when classes are imbalanced.

CONCLUSION

Machine Learning is gradually proving itself to be a valuable tool for post-GWAS analysis, as methodology and high quality training data iterates, ML is showing increasingly optimized performance for prioritizing loci. It has begun to output results which have been validated by showing clinical impact. For complex diseases such as CVD, its ability to generate hypotheses has streamlined functional work that has led to biological insights – enabling the unraveling of how the predominantly non-coding associated loci may affect cardiovascular health. However, before ML models can consolidate their role in the

post-GWAS analyses, research needs to address several aspects ranging from performance (including model benchmarking and fine-tuning), reproducibility, and accessibility. There also needs to be greater comparison between ML and other prioritization methods in order to understand ML's place in the post-GWAS pipeline and enable GWAS to truly provide the projected biological insights and translational capability that it has so long promised.

AUTHOR CONTRIBUTIONS

HN, PM, MB, and CC outlined and drafted the manuscript. All authors contributed and provided critical review of the manuscript.

FUNDING

HN is funded by the British Heart Foundation as part of their Ph.D. program at Barts and The London School of Medicine and Dentistry. CC and MRB are funded by the National Institute for Health Research (NIHR) as part of the portfolio of translational research of the NIHR Biomedical Research Centre at Barts and The London School of Medicine and Dentistry. PM acknowledges support from the National Institute for Health Research (NIHR).

ACKNOWLEDGMENTS

We would like to dedicate this manuscript to our friend and colleague, Chris John.

REFERENCES

Aung, N., Vargas, J. D., Yang, C., Cabrera, C. P., Warren, H. R., Fung, K., et al. (2019). Genome-wide analysis of left ventricular image-derived phenotypes identifies fourteen loci associated with cardiac morphogenesis and heart failure development. *Circulation* 140, 1318–1330. doi: 10.1161/CIRCULATIONAHA.119.041161

Ayalew, M., Le-Niculescu, H., Levey, D. F., Jain, N., Changala, B., Patel, S. D., et al. (2012). Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol. Psychiatry* 17, 887–905. doi: 10.1038/mp.2012.37

Banegas, J. R., Lopez-Garcia, E., Dallongeville, J., Guallar, E., Halcox, J. P., Borghi, C., et al. (2011). Achievement of treatment goals for primary prevention of cardiovascular disease in clinical practice across Europe: the EURIKA study. *Eur. Heart J.* 32, 2143–2152. doi: 10.1093/eurheartj/ehr080

Branco, P. R., de Araujo, G. S., Barrera, J., Suarez-Kurtz, G., and de Souza, S. J. (2018). Uncovering association networks through an eQTL analysis involving human miRNAs and lincRNAs. *Sci. Rep.* 8:15050. doi: 10.1038/s41598-018-33420-z

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5–32. doi: 10.1023/A:1010933404324

Casas, J. P., Bautista, L. E., Smeeth, L., Sharma, P., and Hingorani, A. D. (2005). Homocysteine and stroke: evidence on a causal link from mendelian randomisation. *Lancet* 365, 224–232. doi: 10.1016/S0140-6736(05)17742-3

Cazaly, E., Saad, J., Wang, W., Heckman, C., Ollikainen, M., and Tang, J. (2019). Making sense of the epigenome using data integration approaches. *Front. Pharmacol.* 10:126. doi: 10.3389/fphar.2019.00126

Ciesielski, T. H., Pendergrass, S. A., White, M. J., Kodaman, N., Sobota, R. S., Huang, M., et al. (2014). Diverse convergent evidence in the genetic analysis of complex disease: coordinating omic, informatic, and experimental evidence to better identify and validate risk factors. *BioData Min* 7:10. doi: 10.1186/1756-0381-7-10

Dai, Y., Pei, G., Zhao, Z., and Jia, P. (2019). A convergent study of genetic variants associated with Crohn's disease: evidence from GWAS, gene expression, methylation, eQTL and TWAS. *Front. Genet.* 10:318. doi: 10.3389/fgene.2019.00318

Danese, S. (2012). New therapies for inflammatory bowel disease: from the bench to the bedside. *Gut* 61, 918–932. doi: 10.1136/gutjnl-2011-300904

Demir-Kavuk, O., Kamada, M., Akutsu, T., and Knapp, E. W. (2011). Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinform.* 12:412. doi: 10.1186/1471-2105-12-412

Deo, R. C., Musso, G., Tasan, M., Tang, P., Poon, A., Yuan, C., et al. (2014). Prioritizing causal disease genes using unbiased genomic features. *Genome Biol.* 15:534. doi: 10.1186/s13059-014-0534-8

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Mult. Classifier Syst.* 1857, 1–15.

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* 50, 1412–1425. doi: 10.1038/s41588-018-0205-x

Fridley, B. L., Iversen, E., Tsai, Y. Y., Jenkins, G. D., Goode, E. L., and Sellers, T. A. (2011). A latent model for prioritization of SNPs for functional studies. *PLoS One* 6:e20764. doi: 10.1371/journal.pone.0020764

Fung, K., Ramirez, J., Warren, H. R., Aung, N., Lee, A. M., Tzanis, E., et al. (2019). Genome-wide association study identifies loci for arterial stiffness index in 127,121 UK biobank participants. *Sci. Rep.* 9:9143. doi: 10.1038/s41598-019-45703-0

Gagliano, S. A., Ravji, R., Barnes, M. R., Weale, M. E., and Knight, J. (2015). Smoking gun or circumstantial evidence? Comparison of statistical learning methods using functional annotations for prioritizing risk variants. *Sci. Rep.* 5:13373. doi: 10.1038/srep13373

Gettler, K., Giri, M., Kenigsberg, E., Martin, J., Chuang, L. S., Hsu, N. Y., et al. (2019). Prioritizing Crohn's disease genes by integrating association signals with gene expression implicates monocyte subsets. *Genes Immun.* 20, 577–588. doi: 10.1038/s41435-019-0059-y

Giri, A., Hellwege, J. N., Keaton, J. M., Park, J., Qiu, C., Warren, H. R., et al. (2019). Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* 51, 51–62. doi: 10.1038/s41588-018-0303-9

Hampe, N., Wolterink, J. M., van Velzen, S. G. M., Leiner, T., and Išgum, I. (2019). Machine learning for assessment of coronary artery disease in cardiac ct: a survey. *Front. Cardiovasc. Med.* 6:172. doi: 10.3389/fcvm.2019.00172

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

Hemani, G., Bowden, J., Haycock, P., Zheng, J., Davis, O., Flach, P., et al. (2017). Automating mendelian randomization through machine learning to construct a putative causal map of the human phenome. *bioRxiv*[Preprint]

Hurle, M. R., Nelson, M. R., Agarwal, P., and Cardon, L. R. (2016). Impact of genetically supported target selection on R&D productivity. *Nat. Rev. Drug Discov.* 15, 596–597. doi: 10.1038/nrd.2016.187

Isakov, O., Dotan, I., and Ben-Shachar, S. (2017). Machine learning-based gene prioritization identifies novel candidate risk genes for inflammatory bowel disease. *Inflamm. Bowel Dis.* 23, 1516–1523. doi: 10.1097/MIB.0000000000001222

Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013). "Facing imbalanced data recommendations for the use of performance metrics," in *Proceeding of the International Conference on Affective Computing and Intelligent Interaction (ACII)* (Geneva: IEEE), 245–251. doi: 10.1109/ACII.2013.47

Kafaie, S., Chen, Y., and Hu, T. (2019). A network approach to prioritizing susceptibility genes for genome-wide association studies. *Genet. Epidemiol.* 43, 477–491. doi: 10.1002/gepi.22198

Kanduri, C., Bock, C., Gundersen, S., Hovig, E., and Sandve, G. K. (2019). Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 35, 1615–1624. doi: 10.1093/bioinformatics/bty835

Khan, A., Liu, Q., and Wang, K. (2018). iMEGES: integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes. *BMC Bioinform.* 19(Suppl. 17):501. doi: 10.1186/s12859-018-2469-7

Kichaev, G., Bhatia, G., Loh, P. R., Gazal, S., Burch, K., Freund, M. K., et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* 104, 65–75. doi: 10.1016/j.ajhg.2018.11.008

Leal, L. G., David, A., Jarvelin, M.-R., Sebert, S., Männikkö, M., Karhunen, V., et al. (2019). Identification of disease-associated loci using machine learning for genotype and network data integration. *Bioinformatics* 35, 5182–5190. doi: 10.1093/bioinformatics/btz310

Leem, S., Jeong, H. H., Lee, J., Wee, K., and Sohn, K. A. (2014). Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. *Comput. Biol. Chem.* 50, 19–28. doi: 10.1016/j.compbiolchem.2014.01.005

Leucht, S., Cipriani, A., Spineli, L., Mavridis, D., Orey, D., Richter, F., et al. (2013). Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *Lancet* 382, 951–962. doi: 10.1016/S0140-6736(13)60733-3

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., and Thomas, D. C. (2007). Hierarchical bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* 31, 871–882. doi: 10.1002/gepi.20248

Lin, H., Hargreaves, K. A., Li, R., Reiter, J. L., Wang, Y., Mort, M., et al. (2019). RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol.* 20:254. doi: 10.1186/s13059-019-1847-4

Lu, Q., Powles, R. L., Wang, Q., He, B. J., and Zhao, H. (2016a). 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

Lu, Q., Yao, X., Hu, Y., and Zhao, H. (2016b). GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics* 32, 542–548. doi: 10.1093/bioinformatics/btv610

Maciukiewicz, M., Marshe, V. S., Hauschild, A. C., Foster, J. A., Rotzinger, S., Kennedy, J. L., et al. (2018). GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *J. Psychiatr. Res.* 99, 62–68. doi: 10.1016/j.jpsychires.2017.12.009

Merelli, I., Calabria, A., Cozzi, P., Viti, F., Mosca, E., and Milanesi, L. (2013). SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. *BMC Bioinform.* 14(Suppl. 1):S9. doi: 10.1186/1471-2105-14-S1-S9

Mieth, B., Kloft, M., Rodriguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suarez, C., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci. Rep.* 6:36671. doi: 10.1038/srep36671

Mordelet, F., and Vert, J. P. (2011). ProDiGe: prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinform.* 12:389. doi: 10.1186/1471-2105-12-389

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *J. Big Data* 2:1. doi: 10.1186/s40537-014-0007-7

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 18, 1091–1102. doi: 10.1016/S1474-4422(19)30320-5

Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot.* 7:21. doi: 10.3389/fnbot.2013.00021

Nembrini, S., Konig, I. R., and Wright, M. N. (2018). The revival of the gini importance? *Bioinformatics* 34, 3711–3718. doi: 10.1093/bioinformatics/bty373

Ning, K., Gettler, K., Zhang, W., Ng, S. M., Bowen, B. M., Hyams, J., et al. (2015). Improved integrative framework combining association data with gene expression features to prioritize Crohn's disease genes. *Hum. Mol. Genet.* 24, 4147–4157. doi: 10.1093/hmg/ddv142

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc.* 6(Suppl. 2):S10. doi: 10.1186/1753-6561-6-S2-S10

Oh, J. H., Kerns, S., Ostrer, H., Powell, S. N., Rosenstein, B., and Deasy, J. O. (2017). Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. *Sci. Rep.* 7:43381. doi: 10.1038/srep43381

Pare, G., Mao, S., and Deng, W. Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci. Rep.* 7:12665. doi: 10.1038/s41598-017-13056-1

Raj, M. R., and Sreeja, A. (2018). Analysis of computational gene prioritization approaches. *Procedia Comput. Sci.* 143, 395–410. doi: 10.1016/j.procs.2018.10.411

Rana, N., and Pizarro, T. T. (2019). Elucidating the expression and role of epithelial-derived gasdermin-B (GSDMB) in the context of chronic intestinal inflammation. *FASEB J.* 33(Suppl. 1):496.28. doi: 10.1096/fasebj.2019.33.1_supplement.496.28

Ren, M., Ng, F. L., Warren, H. R., Witkowska, K., Baron, M., Jia, Z., et al. (2018). The biological impact of blood pressure-associated genetic variants in the natriuretic peptide receptor C gene on human vascular smooth muscle. *Hum. Mol. Genet.* 27, 199–210. doi: 10.1093/hmg/ddx375

Rippe, C., Albinsson, S., Guron, G., Nilsson, H., and Sward, K. (2019). Targeting transcriptional control of soluble guanylyl cyclase via NOTCH for prevention of cardiovascular disease. *Acta Physiol. (Oxf)* 225:e13094. doi: 10.1111/apha.13094

Rippe, C., Zhu, B., Krawczyk, K. K., Bavel, E. V., Albinsson, S., Sjolund, J., et al. (2017). Hypertension reduces soluble guanylyl cyclase expression in the mouse aorta via the Notch signaling pathway. *Sci. Rep.* 7:1334. doi: 10.1038/s41598-017-01392-1

Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K., and Hakonarson, H. (2011). Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Res.* 39:e62. doi: 10.1093/nar/gkr064

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10:e0118432. doi: 10.1371/journal.pone.0118432

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., et al. (2012). Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* 30, 317–320. doi: 10.1038/nbt.2151

Schmechel, D. E., Saunders, A. M., Strittmatter, W. J., Crain, B. J., Hulette, C. M., Joo, S. H., et al. (1993). Increased amyloid beta-peptide deposition in cerebral cortex as a consequence of apolipoprotein E genotype in late-onset Alzheimer disease. *Proc. Natl. Acad. Sci. U.S.A* 90, 9649–9653. doi: 10.1073/pnas.90.20.9649

Schubach, M., Re, M., Robinson, P. N., and Valentini, G. (2017). Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Sci. Rep.* 7:2959. doi: 10.1038/s41598-017-03011-5

Seyyedrazzagi, E., and Navimipour, N. J. (2017). Disease genes prioritizing mechanisms: a comprehensive and systematic literature review. *Netw. Model. Anal. Health Inform. Bioinform.* 6:13. doi: 10.1007/s13721-017-0154-9

Smola, A. J., and Scholkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. doi: 10.1023/b:stco.0000035301.49549.88

Szymczak, S., Holzinger, E., Dasgupta, A., Malley, J. D., Molloy, A. M., Mills, J. L., et al. (2016). r2VIM: a new variable selection method for random forests in genome-wide association studies. *BioData Min.* 9:7. doi: 10.1186/s13040-016-0087-3

Takeuchi, F., Akiyama, M., Matoba, N., Katsuya, T., Nakatochi, M., Tabara, Y., et al. (2018). Interethnic analyses of blood pressure loci in populations of East Asian and European descent. *Nat. Commun.* 9:5052. doi: 10.1038/s41467-018-07345-0

Valdes-Mas, R., Gutierrez-Fernandez, A., Gomez, J., Coto, E., Astudillo, A., Puente, D. A., et al. (2014). Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. *Nat. Commun.* 5:5326. doi: 10.1038/ncomms6326

Vitsios, D., and Petrovski, S. (2019). Stochastic semi-supervised learning to prioritise genes from high-throughput genomic screens. *bioRxiv*[Preprint]

Wang, Y., Goh, W., Wong, L., Montana, G., and Alzheimer's Disease Neuroimaging Initiative, (2013). Random forests on hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. *BMC Bioinform.* 14(Suppl. 16):S6. doi: 10.1186/1471-2105-14-S16-S6

Wu, M., Zeng, W., Liu, W., Lv, H., Chen, T., and Jiang, R. (2018). Leveraging multiple gene networks to prioritize GWAS candidate genes via network representation learning. *Methods* 145, 41–50. doi: 10.1016/j.ymeth.2018.06.002

Zeng, W., Wu, M., and Jiang, R. (2018). Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics* 19(Suppl. 2):84. doi: 10.1186/s12864-018-4459-6

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* 50, 1171–1179. doi: 10.1038/s41588-018-0160-6

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 June 2020
doi: 10.3389/fgene.2020.00509

Enhanced Permutation Tests via Multiple Pruning

Sangseob Leem[1†], Iksoo Huh[2†] and Taesung Park[1]*

[1] Department of Statistics, Seoul National University, Seoul, South Korea, [2] College of Nursing and Research Institute of Nursing Science, Seoul National University, Seoul, South Korea

Big multi-omics data in bioinformatics often consists of a huge number of features and relatively small numbers of samples. In addition, features from multi-omics data have their own specific characteristics depending on whether they are from genomics, proteomics, metabolomics, etc. Due to these distinct characteristics, standard statistical analyses using parametric-based assumptions may sometimes fail to provide exact asymptotic results. To resolve this issue, permutation tests can be a way to exactly analyze multi-omics data because they are distribution-free and flexible to use. In permutation tests, *p*-values are evaluated by estimating the locations of test statistics in an empirical null distribution generated by random shuffling. However, the permutation approach can be infeasible when the number of features increases, because more stringent control of type I error is needed for multiple hypothesis testing, and consequently, much larger numbers of permutations are required to reach significance. To address this problem, we propose a well-organized strategy, "ENhanced Permutation tests via multiple Pruning (ENPP)." ENPP prunes the features in every permutation round if they are determined to be non-significant. In other words, if the feature statistics from the permuted datasets exceed the feature statistics from the original dataset, beyond a predetermined threshold, the feature is determined to be non-significant. If so, ENPP removes the feature and iterates the process without the feature in the next permutation round. Our simulation study showed that the ENPP method could remove about 50% of the features at the first permutation round, and, by the 100th permutation round, 98% of the features had been removed and only 7.4% of the computation time with the original unpruned permutation approach had elapsed. In addition, we applied this approach to a real data set (Korea Association REsource: KARE) of 327,872 SNPs to find association with a non-normally distributed phenotype (fasting plasma glucose), interpreted the results, and discussed the feasibility and advantages of the approach.

Keywords: permutation test, multiple hypothesis testing, pruning, big multi-omics data, GWAS

OPEN ACCESS

Edited by:
Christian Darabos, Dartmouth College, United States

Reviewed by:
Tiejun Tong, Hong Kong Baptist University, Hong Kong
Gil Speyer, Arizona State University, United States

***Correspondence:**
Taesung Park
tspark@stats.snu.ac.kr

[†]These authors have contributed equally to this work

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

Received: *24 December 2019*
Accepted: *27 April 2020*
Published: *25 June 2020*

Citation:
Leem S, Huh I and Park T (2020) Enhanced Permutation Tests via Multiple Pruning. Front. Genet. 11:509. doi: 10.3389/fgene.2020.00509

INTRODUCTION

Unlike typical big data, big data in bioinformatics consists of huge numbers of features and relatively small numbers of samples. For example, the data from genome-wide association studies (GWAS) contain at least thousands of samples and several hundred thousands of single nucleotide polymorphisms (SNPs) (Manolio, 2010). In the case of transcriptomic analysis for

finding differently expressed genes, tens of thousands of genes are tested from only hundreds of samples at most (McLachlan et al., 2005). In epigenomic data, such as DNA methylation, the number of features (e.g., CpG sites) varies from tens of thousands to several million according to profiling techniques and their resolution (Bibikova et al., 2011; Adusumalli et al., 2014). Moreover, not only large numbers of features but also various characteristics of the features are important points to be considered. For example, in genomic data, such as SNPs, a feature is represented as a count of a minor allele at a genomic locus in each individual. In transcriptome data sets, gene expression levels are represented as continuous and positive real values measured from microarray spot intensities. In the case of epigenomics data, the DNA methylation levels of loci can be provided as a ratio between read counts of C and read counts of C and T. In addition, proteomics and metabolomics data provide marker intensities from mass-spectrometry-based approaches. Therefore, detecting association between phenotypes and biomarkers using standard statistical approaches may sometimes be inaccurate, as many of these are based on parametric assumptions that require specific properties of the features. Although several remedies have been proposed in terms of parametric approaches (Thygesen and Zwinderman, 2004; Lin et al., 2008; Park and Wu, 2016), they are naturally asymptotic ones and still possibly have type 1 error inflation or low power.

As an alternative to these issues, the permutation test (Pitman, 1937; Annis, 2005) has become a popular approach for analyzing multi-omics data because it can be used regardless of the shape of distribution of the biomarkers' expression and uses a simple algorithm. In the permutation test, a p-value is assessed through evaluating the relative rank of the observed test statistic in an empirical null distribution of the test statistic generated by random shuffling. The permutation test has already been used in some omics analysis. For example, in GWAS, the permutation test is used for adjusting for multiple tests (Browning, 2008), considering biological structures (Pahl and Schäfer, 2010), and identifying gene-gene interactions (Ritchie et al., 2001; Greene et al., 2010). In next-generation sequencing data analysis, rare variants have been identified by permutation test for association with a phenotype (Madsen and Browning, 2009) and as a significance test of structural models (Lee et al., 2016; Kim et al., 2018). In integration analysis of multi-omics data, the permutation test is used for finding edges in the integrated network (Jeong et al., 2015) and significance testing of an aggregated unit with a structure (Kim et al., 2018). In metagenome studies, the permutation test is used for testing differences between distances of groups (Chen et al., 2012), finding differentially abundant operational taxonomic units (Anderson, 2005), and finding differentially abundant genomic features (Paulson et al., 2011).

However, a major obstacle to the permutation test is its large computation time, because the smallest p-value that a permutation test can reach is inversely proportional to the permutation time. Therefore, if a data set has a large number of features, it requires a large number of permutations to detect significantly associated features because larger numbers of features require more stringent type 1 error control in terms of multiple hypothesis testing correction. For example, if a researcher wants to test an association between 5.0×10^5 SNPs and a specific phenotype, the p-value threshold will be 1.0×10^{-7} [$0.05/(5.0 \times 10^5)$ by Bonferroni correction]. To achieve such a stringent p-value threshold, the number of permutations must be at least $1.0 \times 10^7 - 1$ for each SNP, and the total computation time for all features is impractical. Considering that only significant features are of general interest to researchers, pruning insignificant features can be a way to resolve the issue.

Therefore, in this study, we propose a well-organized strategy, ENhanced Permutation tests via multiple Pruning (ENPP). The key idea of ENPP is simple. When the number of features is large, the p-value threshold is very low due to multiple testing correction. In most cases, if a feature is reported to be significant, its observed test statistic value should be more extreme than those from permuted data sets. On the other hand, if a feature has more than a set number of instances of having larger statistics from permuted data sets, it can be regarded as a feature with significantly less chance of being significant, and ENPP prunes the feature during a certain permutation round. In other words, ENPP specifically removes non-significant features and continues the permutation procedures with the remaining features, which can then be candidates for a predetermined significance level. This approach can reduce total permutation time to a feasible level compared to ordinary permutation approaches that conduct the same number of permutation tests on all features. Herein, we show that ENPP can remove about 50% of features in the first permutation round and requires, at the 100th permutation round, only 7.4% of the computation time needed for the unpruned permutation approach. This relative proportion of computation time becomes smaller as the iteration time increases. In addition, we applied our approach to a real data set (Korea Association REsource: KARE) (Cho et al., 2009) containing 327,872 SNP features and a non-normally distributed phenotype (fasting plasma glucose, FPG) for validation of our approach in terms of feasibility and usefulness.

MATERIALS AND METHODS

Data Set

For real data analysis, we chose a Korean GWAS data set collected since 2007 by The Korean Association REsource (KARE) project (Cho et al., 2009). In this project, all participants were recruited from either of two region-based cohorts (rural Ansung and urban Ansan). The total number of participants was 10,038 (5,018 from Ansung and 5,020 from Ansan), and they were all genotyped, using genomic DNA from peripheral blood, using the Affymetrix (Santa Clara, CA, United States) Genome-Wide Human SNP array 5.0, containing 500,568 SNPs. For quality control, we followed the same process used in a previous study (Oh et al., 2016). As a result, we finally obtained 8,842 individuals and 327,872 SNPs, and the processed data set was used in our real data analysis. The study was reviewed and approved by the Institutional Review Board of Seoul National University (IRB No. E1908/001-004).

ENPP Approach

Suppose that there are N samples, each with a dependent variable Y, and J features $X_1, \ldots, X_J$, representing features from a multi-omics data set. In general, for a significance test of association between a specific X_j and Y, the null distribution of the test statistic S consists of test statistics from permuted data sets, and we call the statistics s_r, where $r = 1,2,\ldots,R$, with R denoting the total number of permutation rounds for the feature. Then, the observed value, s_{obs} (i.e., the original value of the test statistic, S) is compared to the null distribution of S, and the significance is assessed by the proportion of s_r values more extreme than s_{obs}. For exact generation of the null distribution, $N!$ iterations are required. However, when $N!$ is too large, R iterations of random shuffling (R) ($R << N!$) are generally used for assessing computational feasibility in terms of Monte-Carlo estimation. A finding that a s_{obs} value is larger than the simulated s_r values implies that the test is more supportive of the alternative hypothesis, and the p-value is then calculated by the following equation:

$$P_{perm} = \frac{1 + \sum_{r=1}^{r=R} I(s_{obs} \leq s_r)}{R + 1}, \quad (1)$$

where I(·) is an indicator function, and +1 in the numerator and denominator can be omitted.

When the number of features is multiple, the p-value threshold should be adjusted for a multiple testing comparison. For example, a typical p-value threshold is 0.05, and, if there are 1,000 features for association tests, then the p-value threshold becomes 0.05/1,000, for the Bonferroni correction. In other words, when a feature has a p-value smaller than this adjusted p-value threshold it is reported as significant. Therefore, the possibility of $I(\cdot) = 1$ (more extreme than the observed statistic value) is extremely low for this feature. On the other hand, if $I(\cdot) = 1$ frequently appears in a feature, the p-value of the feature may be closer to 1, meaning that it may not be significant and would therefore be of no interest to researchers. Let p_{raw} be an unadjusted p-value threshold (e.g., 0.05) and p_{adj} be an adjusted p-value threshold, for each feature, after the multiple testing correction (e.g., 0.05/J by Bonferroni correction). p_{adj} is then the significance level for which we need to detect significant features, and the decision of whether or not to prune a feature, in any specific round, is based on the hypothesis that:

$$\mathrm{H0} : \mathrm{p} = \mathrm{p}_{adj}, \text{ and } \mathrm{H1} : \mathrm{p} > \mathrm{p}_{adj}, \quad (2)$$

where p implies the true p-value from the permutation approach. In the hypothesis, the significance level for the test needs to be determined, and we call the threshold p_{prun}. For the hypothesis test, a binomial test can be used, and, based on p_{adj} and p_{prun}, we can set an integer C_{prun} that satisfies p_{prun} in a permutation round. Therefore, C_{prun} is a variable that depends on permutation numbers, while p_{adj} and p_{prun} are fixed values for the whole pruning process. Consequently, using this rule, EPNN counts in how many cases a feature has a more extreme test statistic than its observed test statistic value in each permutation round. If a feature is equal to or greater than C_{prun} in a round, it is removed from the next permutation round. The following is a detailed explanation of the parameter determination.

Let us assume that $\mathrm{p}_{adj} = 5 \times 10^{-5}$, which is equivalent to a threshold Bonferroni correction with 1,000 features, and $\mathrm{p}_{prun} = \mathrm{p}_{adj}$. In addition, if we let $p_{k|r}$ denote a probability of observing at least a number k of test statistics values more extreme than the observed test statistics at the rth permutation round, then $p_{k|r} = \sum_{t=k}^{t=r} \binom{r}{t} p_{adj}^t (1 - p_{adj})^{r-t}$. Therefore, if the p-value of a feature is significant, then $p_{k|r}$ should be equal to or smaller than p_{prun}. As an illustration, consider the first permutation round. Based on a setting of $\mathrm{p}_{adj} = 5 \times 10^{-5}$, two probabilities, $p_{0|1}$, $p_{1|1}$, are given. Because we set $p_{prun=}p_{adj}$, $p_{0|1}$ will be 1 and $p_{1|1}$ will be p_{adj}, implying that $C_{prun} = 1$ is in the first round. For the second round, there are three probabilities, $p_{0|2}$, $p_{1|2}$ and $p_{2|2}$, that can be easily computed. In this case, $p_{1|2} = 1 \times 10^{-4} > p_{prun}\, p_{2|2} = 10^{-9} < p_{prun}$. Therefore C_{prun} will be 2 for the second round. In this manner, we can obtain C_{prun} for all permutation rounds conducted. We will show the properties of the parameters in the next section.

RESULTS

Simulation Analysis

In this section, we evaluated the advantages of ENPP compared to a strict permutation approach, including its need for only very few counts for rejecting and removing non-significant features. As a consequence of this attribute, ENPP can greatly reduce total computation time to a feasible level compared to an unpruned permutation approach. To show the desired properties, we artificially generated data sets whose features did not associate with a feature. When the Bonferroni threshold was applied and $p_{raw} = 0.05$, the first example had $p_{adj}^1 = 0.05/1{,}000$ and the second example had $p_{adj}^2 = 0.05/(5 \times 10^5)$. In addition, we also assumed that $p_{prun} = p_{adj}$ for both examples.

Distribution of C_{prun}

Firstly, we investigated the distribution of C_{prun} values according to each permutation round for p_{adj}^1, and p_{adj}^2, respectively. Using the formula described in the methods, C_{prun} values were calculated for $r = 1,2,\ldots, 10{,}000$, and the resulting values are shown in **Figure 1A**, which also shows that the values of C_{prun} for p_{adj}^1 are at most 6 in the 10,000th round. This implies that the threshold is not hard to satisfy and that we can reduce a large proportion of the number of features at each permutation round. In the case of p_{adj}^2, C_{prun} becomes smaller (**Figure 1B**). In detail, C_{prun} is 1 for $i = 1, 2$ for $\mathrm{i} \in [2, 4, 473]$, and 3 for $\mathrm{i} \in [4, 474, 10, 000]$, implying that smaller p_{adj} values provide smaller C_{prun} values, although p_{prun} is proportional to p_{adj}.

Pruning Rates and Computational Efficiency in Each Permutation Round

Based on the C_{prun} values calculated above, we also evaluated the pruned proportion of the total features for each permutation

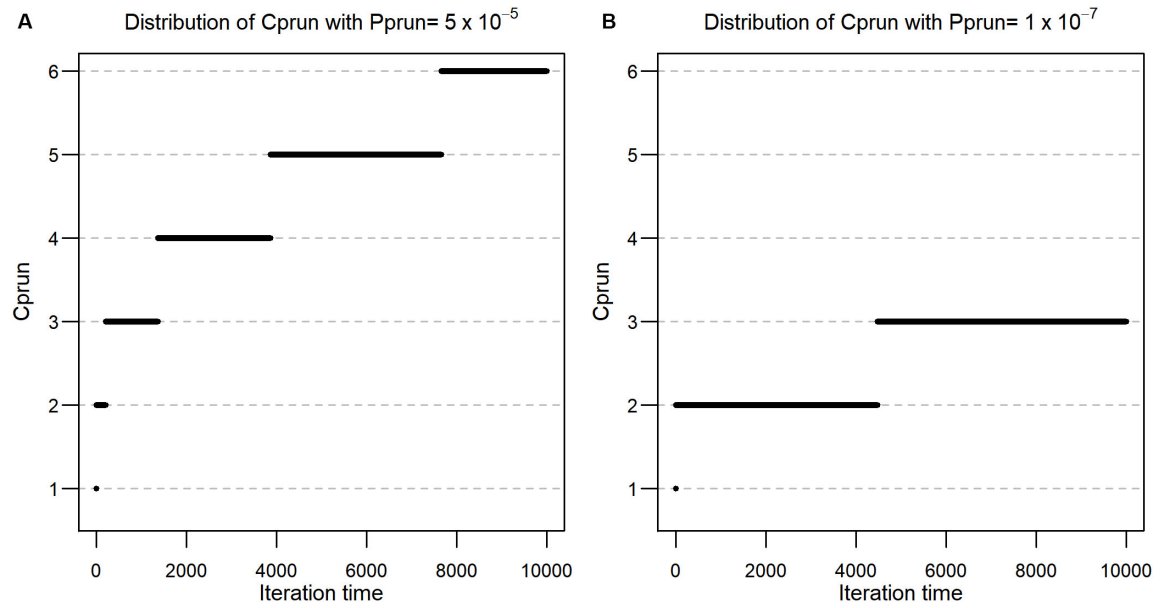

FIGURE 1 | (A) Distribution of C_{prun} with $P_{prun} = 5 \times 10^{-5}$. **(B)** Distribution of C_{prun} with $P_{prun} = 1 \times 10^{-7}$.

round. Suppose that the p-value of a feature has a uniform distribution, meaning that the feature has no association with a phenotype. In this setting, the pruned proportion of features depends only on C_{prun}. For example, at the first round, for $C_{prun}(1) = 1$, the proportion of pruned features will be $\int_0^1 pdp = \frac{1}{2}$. At the second round, for $C_{prun}(2) = 2$, no pruning will happen, because the event that $C_{prun}(1) = 1$ includes the event that $C_{prun}(2) = 2$. At the third round of permutation, for $C_{prun}(3) = 2$, the expected pruning proportion after the permutation will be:

$$\int_0^1 (1-p)p^2 dp = \int_0^1 (p^2 - p^3)dp = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

In other words, at the first permutation, $\frac{1}{2}$ of the features are expected to be pruned, and $\frac{1}{12}$ of the features are additionally pruned after the third round. In this manner, the expected proportions of remaining features after pruning from 1 to 10,000 permutation rounds are calculated using the C_{prun} values (**Figure 1**), and the results are described in **Figure 2**. Because the cumulative pruning proportion is not easily derived by numerical calculation, we estimated the proportion by simulation using variables from a Bernoulli distribution, with the probability for success taken from a uniform distribution U(0,1). In **Figure 2A**, only about 2% of features remain after the 100th permutation round in both p_{prun} settings, thus greatly reducing the number of tests for the data set at the round. However, as C_{prun} becomes different, the remaining proportions also become different. For example, at the 1000th permutation round, 0.3% of total features remained for p^1_{adj} and 0.2% for p^2_{adj}. The ratio between the two proportions became larger at the 10,000th permutation round, with 0.057% for the former, p^1_{adj}, and 0.028% for the latter, p^2_{adj}. These results reflect the differences of C_{prun} provided in **Figure 1**.

We next assessed computational efficiency by comparing the total permutation time for ENPP to that for the original, unpruned permutation test. The efficiency is represented as a ratio between the number of tests in the original unpruned permutation approach and the cumulative number of tests in the ENPP approach. The total permutation time for a given permutation round in ENPP is calculated by accumulating all permutation times of earlier permutation rounds. Therefore, larger computational efficiencies imply a large timesaving advantage for ENPP analysis. For example, during the first round, there is no reduction of permutation time, but for the second and third permutation rounds, ENPP needs only $\frac{1}{2}$ the computations compared to the original unpruned permutation tests, and $\frac{5}{12}$ the permutations are needed for the fourth round. Therefore, computational efficiency will be $\frac{1}{1} = 1$ for the first permutation round, and $\frac{1+1}{1+\frac{1}{2}} = \frac{4}{3}$, $\frac{1+1+1}{1+\frac{1}{2}+\frac{1}{2}} = \frac{3}{2}$, $\frac{1+1+1+1}{1+\frac{1}{2}+\frac{1}{2}+\frac{5}{12}} = \frac{48}{29}$. for the second, third, and fourth permutation rounds, respectively. The Inverse Computational Efficiency (ICE) for each permutation round is summarized in **Figure 2B**. In **Figure 2B**, ICE does not seem to decrease as fast as the remaining proportion, as shown in **Figure 2A**, due to the fact that permutation times of precedent rounds accumulate in estimating computational efficiency. Compared to the ordinary unpruned permutation test, only about 7.4% of the computation time is needed at the 100th permutation round in both settings, because they have the same numbers for C_{prun} and the same resulting remaining proportions. However, as in the remaining proportion of features, ICE became more different in terms of ratios between the two settings as the permutation round progresses. For example, at the 1000th permutation round, ICE is 1.3% for $p_{prun} = 5 \times 10^{-5}$ and

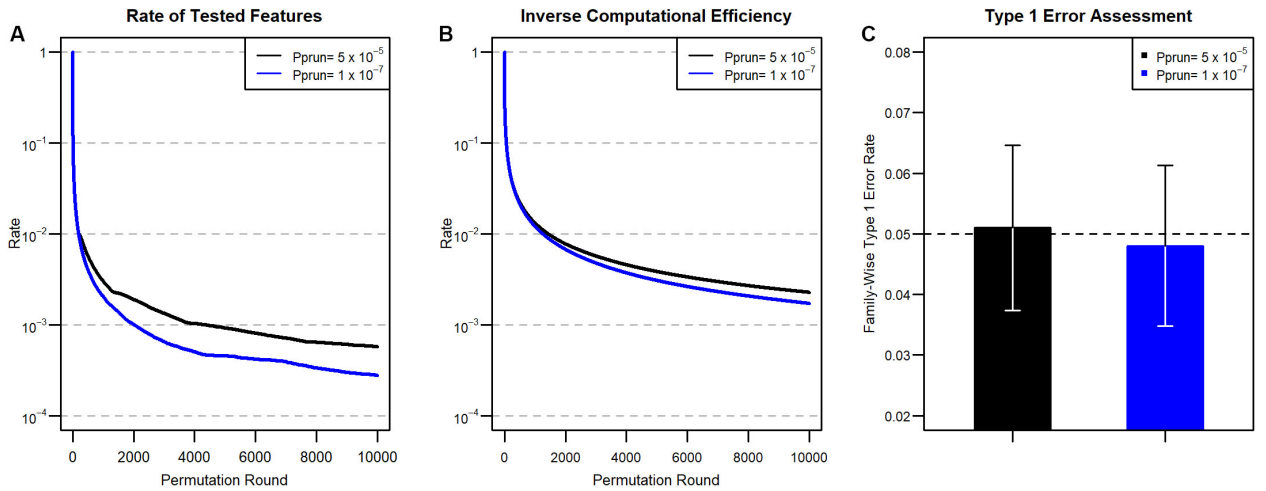

FIGURE 2 | (A) Proportion of tested features at each round after pruning for the two P_{prun} values in **Figure 1**. **(B)** Inverse of computational efficiency (ICE) for 2A. **(C)** Type 1 error results. We divided the number of false-positive features by 1,000 to obtain the family-wise type 1 error rate. 95% confidence intervals of the estimated type 1 errors are also provided.

1.2% for $p_{prun} = 1 \times 10^{-7}$. However, in the 10,000 iteration, 0.23% is needed for the former, p_{prun} while 0.17% is needed for the latter, p_{prun}. Thus, the overall computational efficiency improves as the iteration round progresses because the remaining rate of the features grows smaller, and smaller p_{prun} requires less computation.

On the other hand, we assessed the type 1 error rate of non-associated features from the ENPP approach. For p^1_{adj} and p^2_{adj}, we generated 10^6 and 5×10^8 non-associated features from the Bernoulli distribution so that the expected numbers of features with type 1 error are 50 in both settings. We first applied the pruning process to the non-associated features and then the full permutation approach to the remaining unpruned features. After the full permutation approach had been applied, we counted how many non-associated features were found significant at the given significance levels. The type 1 error rates are summarized in **Figure 2C**, showing that the ENPP approach controls the type 1 error well.

Real Data Analysis

We next applied our approach to a real genome-wide data set (Korea Association REsource: KARE), which has 327,872 SNPs from each of 8,842 individuals (Cho et al., 2009). In order to detect significant SNP features at the Bonferroni significance level in the data set, the ordinary permutation approach (without ENPP) requires at least $(1/0.05) \times 327{,}872^2 = 2.15 \times 10^{12}$, a computationally impractical number of tests. Therefore, using a pruning approach for this data set becomes inevitable when the permutation approach is used. For the application of ENPP, we set $p_{raw} = 0.05$ and $p_{prun} = p_{adj} = 0.05/327{,}872 = 1.52 \times 10^{-7}$, and the corresponding C_{prun} is calculated and described in **Figure 3A**. Here, we set the number of iterations to 100,000 because simulation analysis found that the remaining proportion of features was 3.7×10^{-5} at the 100,000th round and the corresponding expected count of remaining features was $3.7 \times 10^{-5} \times 327{,}872 = 12.13$ if all features were assumed not to associate with a phenotype. We selected fasting plasma glucose (FPG) as a phenotype because its distribution is very highly skewed (skewness = 5.32) and the skewness is still high (=2.71) (Kim, 2013) even after log-transformation. Consequently, we expected that this property may produce results that differ between a parametric approach and a permutation approach. For the association analysis, we used age, gender, and living regions as covariates, and we assumed that the genotype of the SNP features has an additive effect on the phenotype. As a test statistic for the permutation test, we used a t-statistic for the genotype effect.

Based on the expected remaining proportion of the features, we found ICE to be 2.4×10^{-4} at the 100,000th permutation round (**Figure 3C**), meaning that we needed only 24 times more computation compared to the parametric linear regression approach. This number of permutation tests can be done in a few days, even in a single thread. After implementing the 100,000th iteration of ENPP with the real data set, we plotted the number of remaining features (**Figure 3B**) and the ICE (**Figure 3C**) in each round. Those results showed that 46 SNP features remained and that the computational efficiency was 3.7×10^{-4}, implying that some SNP features were candidates for significant features. For each of 46 SNP features, we implemented a $3 \times 10^7 - 1$ permutation test to provide a p-value not only for Bonferroni correction but also for a genome-wide significance of 5×10^{-8} (Xu et al., 2014). After implementation of the test, we found that five SNP features passed the Bonferroni threshold, and two SNPs also passed for genome-wide significance (**Table 1**). On the other hand, the parametric approach found four SNPs for Bonferroni correction, and two SNPs passed genome-wide significance. However, only three SNPs overlapped for the former threshold, and one SNP overlapped for the latter one. To determine substantial differences of p-values between the two approaches, we used an exact binomial test (Clopper and

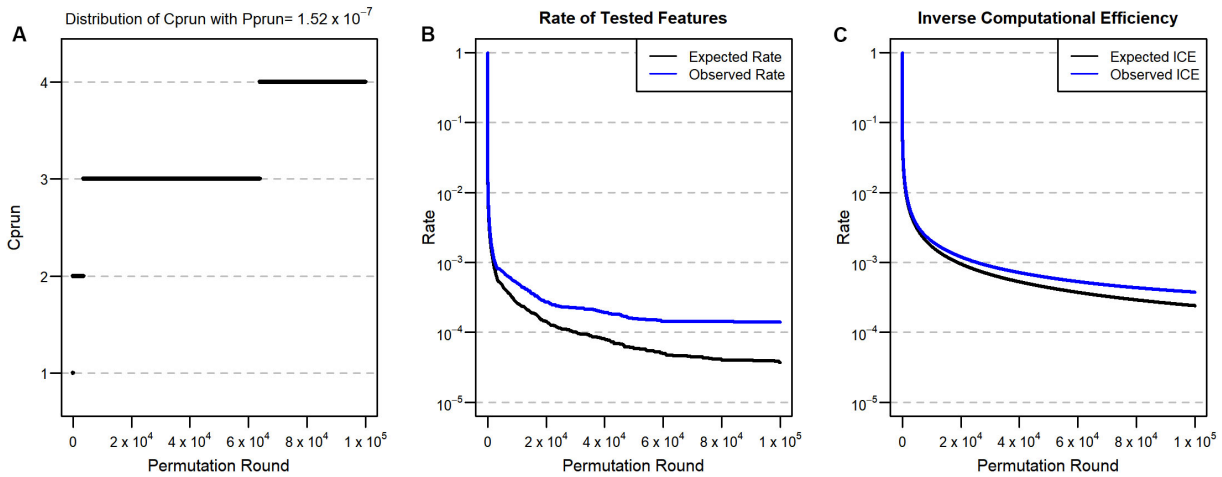

FIGURE 3 | (A) Distribution of C_{prun} for the real data set with $P_{prun} = 1.52 \times 10^{-7}$. **(B)** Proportion of tested features after pruning. **(C)** ICE for $P_{prun} = 1.52 \times 10^{-7}$. Black lines are expected values from the simulation, and blue lines are observed values from the real data analysis.

TABLE 1 | 6 SNPs selected from either parametric (linear regression) or non-parametric (ENPP) tests at a Bonferroni significance level $p = 1.52 \times 10^{-7}$.

CHR	SNP id	MAF	*P*-value from linear regression	*P*-value from permutation	*P*-value from comparison between the two values
6	rs9348440T	0.478	1.63×10^{-7}	1.33×10^{-7}	1
6	rs6456368C	0.480	1.54×10^{-7}	1.00×10^{-7}	0.640
6	rs10946398C	0.479	8.35×10^{-8}	6.67×10^{-8}	1
6	rs7754840C	0.479	4.93×10^{-8}	3.33×10^{-8}	1
6	rs9460546G	0.481	5.45×10^{-8}	3.33×10^{-8}	1
16	rs7197218G	0.014	4.81×10^{-8}	7.33×10^{-7}	$<2.2 \times 10^{-16}$

Here, we provide information for SNP features such as chromosome, SNP id, and minor allele frequency (MAF) and the p-values from both tests. In the last column of the table, we also include the results of an exact binomial test for permutation results based on the null hypothesis that the p-value of the permutation test is the same as the results from the parametric approach.

Pearson, 1934) that regarded *p*-values from the parametric approach as a null hypothesis *p*-value for the permutation results. From the test, we found that only one SNP (rs7197218G in chromosome 16) showed a significant difference between the two results (**Table 1**). This SNP showed a more conservative result from the permutation approach; this result may come from type 1 error inflation in the parametric test in the presence of very low minor allele frequency and large differences of variance between FPG values with and without the minor allele (Zimmerman, 2004).

DISCUSSION

For the analysis of multi-omics data, the permutation test has been popularly used because it is non-parametric and flexible to use. However, the main drawback of this approach is that it may require such a large number of tests as to make it infeasible, especially for data sets with large numbers of features and a Bonferroni-corrected significance level. To resolve this issue, we proposed a well-organized strategy, ENhanced Permutation tests via multiple Pruning (ENPP), for enhanced permutation tests, using the idea of pruning. ENPP investigates the features at every permutation round and removes them if they have less chance of being significant. Our empirical study showed that the ENPP method could remove about 50% of the number of features at the first permutation round and required only 7.4% of the total computation time at the 100th permutation round as is needed by an unpruned approach. Moreover, in real data analysis, on a data set of 327,872 SNP features, our approach was found to greatly reduce computational burden to a feasible level, and the analysis results seemed more reliable than the results from a parametric approach because they were not affected by a specific assumption of a null distribution. Interestingly, we found that the number of tests conducted in the ENPP process was much smaller than the number in the final evaluation of the 46 SNP features to obtain precise *p*-values. In the pruning process of real GWAS data, about 1.2×10^7 permutations were needed, while in parallel, the full permutation analysis required about 1.4×10^9 iterations. Since the pruning process and the full permutation process are performed on each feature independently, they can easily be parallelized. We believe that parallelism has a large impact on the full permutation process because the full permutation process seems to take much more computing time than the pruning process in our real data analysis. Therefore, with the help of parallel computing, our ENPP approach can easily handle,

without computational burden, larger data sets such as human methylation data with 2×10^7 CpG site features.

Our EPNN algorithm is also flexible for pruning processes. Researchers can modify p_{adj} and p_{prun} as they want. In this study, we set $p_{adj} = p_{prun}$, with p_{adj} from a Bonferroni correction, and conducted 100,000 ENPP permutations. These settings could be interpreted with the number of expected significant features and the number of tests of the features, considering that summation of the actual significance level, calculated for C_{prun}, from the first round to the 100,000th round is 2.66×10^{-3}, and it admits $0.05/(2.66 \times 10^{-3}) \approx 18$ truly significant features at the Bonferroni threshold. In other words, if there are 18 or fewer significant features, at $p = 1.52 \times 10^{-7}$, we can control the probability of falsely pruning any significant features under 0.05. This assumption of the number of the significant features is reasonable, considering that only a few features may satisfy Bonferroni cutoff in general and that our analysis results in both parametric and permutation approaches found only four or five SNPs, respectively. In addition, researchers may sometimes be interested not only in features for a specific Bonferroni significance level but also in a *p*-value distribution of whole features. For this purpose, ENPP can be applied after some number of unpruned permutation rounds, such as 100, so that more precise *p*-values can be obtained, even for non-significant features, and the results can be used in false discovery rate (FDR) approaches (Benjamini and Hochberg, 1995) or in combining *p*-value approaches for some group-wise testing such as gene- or pathway-wise significance tests (Subramanian et al., 2005). Our ENPP approach will help many researchers achieve precise *p*-values in a feasible time, even for datasets with a large number of features. A brief R script for performing ENPP is provided for SNPs at http://statgen.snu.ac.kr/software/ENPP. This will enable more accurate decisions based on the statistical results.

DATA AVAILABILITY STATEMENT

The data will be publicly distributed by the Distribution Desk of the Korea Biobank Network (https://koreabiobank.re.kr/), to whom data requests should be directly made. Any inquiries should be sent to admin@koreabiobank.re.kr.

ETHICS STATEMENT

The study was reviewed and approved by the Institutional Review Board of Seoul National University (IRB No. E1908/001-004). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SL, IH, and TP developed the algorithm. SL conducted the simulation study and wrote the manuscript. IH conducted real data analysis and wrote the manuscript. TP supervised the whole research project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

REFERENCES

Adusumalli, S., Omar, M. F. M., Soong, R., and Benoukraf, T. (2014). Methodological aspects of whole-genome bisulfite sequencing analysis. *Brief. Bioinform.* 16, 369–379. doi: 10.1093/bib/bbu016

Anderson, M. (2005). *PERMANOVA: A fortran Computer Program For Permutational Multivariate Analysis Of Variance*. New Zealand: University of Auckland.

Annis, D. H. (2005). *Permutation, Parametric, And Bootstrap Tests Of Hypotheses*. Milton Park: Taylor & Francis.

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

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288–295. doi: 10.1016/j.ygeno.2011.07.007

Browning, B. L. (2008). PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinform.* 9:309. doi: doi: 10.1186/1471-2105-9-309

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., and Wu, G. D. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28, 2106–2113. doi: 10.1093/bioinformatics/bts342

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H.-J., et al. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* 41:527. doi: 10.1038/ng.357

Clopper, C. J., and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404–413. doi: 10.1093/biomet/26.4.404

Greene, C. S., Himmelstein, D. S., Nelson, H. H., Kelsey, K. T., Williams, S. M., Andrew, A. S., et al. (2010). Enabling personal genomics with an explicit test of epistasis. *Biocomputing* 2010, 327–336. doi: 10.1142/9789814295291_0035

Jeong, H.-H., Leem, S., Wee, K., and Sohn, K.-A. (2015). Integrative network analysis for survival-associated gene-gene interactions across multiple genomic profiles in ovarian cancer. *J. Ovar. Res.* 8:42.

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor. Dent. Endod.* 38, 52–54.

Kim, Y., Lee, S., Choi, S., Jang, J.-Y., and Park, T. (2018). Hierarchical structural component modeling of microRNA-mRNA integration analysis. *BMC Bioinform.* 19:75. doi: 10.1186/s12859-018-2070-0

Lee, S., Choi, S., Kim, Y. J., Kim, B.-J., T2d-Genes Consortium, Hwang, H., et al. (2016). Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics* 32, i586–i594. doi: 10.1093/bioinformatics/btw425

Lin, S. M., Du, P., Huber, W., and Kibbe, W. A. (2008). Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* 36:e11. doi: 10.1093/nar/gkm1075

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

Manolio, T. A. (2010). Genome-wide association studies and assessment of the risk of disease. *New Engl. J. Med.* 363, 166–176.

McLachlan, G., Do, K.-A., and Ambroise, C. (2005). *Analyzing Microarray Gene Expression Data*. Hoboken, NJ: John Wiley & Sons.

Oh, S., Huh, I., Lee, S. Y., and Park, T. (2016). Analysis of multiple related phenotypes in genome-wide association studies. *J. Bioinform. Comput. Biol.* 14:1644005. doi: 10.1142/s0219720016440054

Pahl, R., and Schäfer, H. (2010). PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics* 26, 2093–2100. doi: 10.1093/bioinformatics/btq399

Park, Y., and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* 32, 1446–1453. doi: 10.1093/bioinformatics/btw026

Paulson, J. N., Pop, M., and Bravo, H. C. (2011). Metastats: an improved statistical method for analysis of metagenomic data. *Genome Biol.* 12:17.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Suppl. J. R. Statist. Soc.* 4, 119–130.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). ., Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147. doi: 10.1086/321276

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

Thygesen, H. H., and Zwinderman, A. H. (2004). Comparing transformation methods for DNA microarray data. *BMC Bioinform.* 5:77. doi: 10.1186/1471-2105-5-77

Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., and Greenwood, C. M. T. (2014). Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* 38, 281–290. doi: 10.1002/gepi.21797

Zimmerman, D. W. (2004). Inflation of type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicologica* 25, 103–133.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 July 2020
doi: 10.3389/fgene.2020.00679

Integration of GWAS and eQTL Analysis to Identify Risk Loci and Susceptibility Genes for Gastric Cancer

Jing Ni[1,2], Bin Deng[3], Meng Zhu[1,2], Yuzhuo Wang[1,2], Caiwang Yan[1,2], Tianpei Wang[1,2], Yaqian Liu[1,2], Gang Li[4], Yanbing Ding[3] and Guangfu Jin[1,2]**

[1] Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China, [2] Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing, China, [3] Department of Gastroenterology, Affiliated Hospital of Yangzhou University, Yangzhou, China, [4] Department of General Surgery, Jiangsu Institute of Cancer Research, Jiangsu Cancer Hospital, The Affiliated Cancer Hospital of Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Ting Hu,
Memorial University of Newfoundland, Canada

Reviewed by:
Zhaohui Steve Qin,
Emory University, United States
Enrique Medina-Acosta,
State University of the North Fluminense Darcy Ribeiro, Brazil

***Correspondence:**
Yanbing Ding
ybding@yzu.edu.cn
Guangfu Jin
guangfujin@njmu.edu.cn

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

Received: 20 February 2020
Accepted: 03 June 2020
Published: 10 July 2020

Citation:
Ni J, Deng B, Zhu M, Wang Y, Yan C, Wang T, Liu Y, Li G, Ding Y and Jin G (2020) Integration of GWAS and eQTL Analysis to Identify Risk Loci and Susceptibility Genes for Gastric Cancer. Front. Genet. 11:679. doi: 10.3389/fgene.2020.00679

Genome-wide association studies (GWAS) have identified several susceptibility loci for gastric cancer (GC), but the majority of identified single-nucleotide polymorphisms (SNPs) fall within the non-coding region and are likely to exert their biological function by modulating gene expression. To systematically estimate expression-associated SNPs (eSNPs) that confer genetic predisposition to GC, we evaluated the associations of 314,203 stomach tissue-specific eSNPs with GC risk in three GWAS datasets (2,631 cases and 4,373 controls). Subsequently, we conducted a gene-based analysis to calculate the cumulative effect of eSNPs through sequence kernel association combined test and Sherlock integrative analysis. At the SNP-level, we identified two novel variants (rs836545 at 7p22.1 and rs1892252 at 6p22.2) associated with GC risk. The risk allele carriers of rs836545-T and rs1892252-G exhibited higher expression levels of *DAGLB* ($P = 3.70 \times 10^{-18}$) and *BTN3A2* ($P = 3.20 \times 10^{-5}$), respectively. Gene-based analyses identified *DAGLB* and *FBXO43* as novel susceptibility genes for GC. *DAGLB* and *FBXO43* were significantly overexpressed in GC tissues than in their adjacent tissues ($P = 5.59 \times 10^{-7}$ and $P = 3.90 \times 10^{-6}$, respectively), and high expression level of these two genes was associated with an unfavorable prognosis of GC patients ($P = 1.30 \times 10^{-7}$ and $P = 7.60 \times 10^{-3}$, respectively). Co-expression genes with these two novel genes in normal stomach tissues were significantly enriched in several cancer-related pathways, including P53, MAPK and TGF-beta pathways. In summary, our findings confirm the importance of eSNPs in dissecting the genetic basis of GC, and the identified eSNPs and relevant genes will provide new insight into the genetic and biological basis for the mechanism of GC development.

Keywords: gastric cancer, eSNP, genome-wide association study, gene-based analysis, Sherlock integrative analysis

INTRODUCTION

Gastric cancer (GC) is the fifth most common neoplasm and second leading cause of cancer-related death globally. There were approximately one million newly diagnosed GC cases and 780,000 deaths in 2018 (Bray et al., 2018). Approximately half of the new GC cases and deaths worldwide occur in China, indicating a major public health burden (Chen et al., 2016). A large twin cohort study in Nordic countries suggested that up to 22% interindividual variability in GC risk could be explained by genetic factors (Mucci et al., 2016). In the past decade, we and other groups have reported a number of susceptibility loci for GC through genome-wide association study (GWAS), which only explain a fraction of GC heritability (Abnet et al., 2010; Shi et al., 2011; Wang et al., 2017; Park et al., 2019). Moreover, the vast majority of disease-related variants discovered by GWAS fall within intergenic or non-coding regions, which may regulate the expression of target genes and influence the process of pathogenesis (Maurano et al., 2012).

Expression quantitative trait locus (eQTL) analysis has been conducted to provide prior weights for the statistical analysis of new susceptibility single-nucleotide polymorphism (SNP) discovery and prioritize SNPs or genes for further functional experiments (Li et al., 2013). Integration of GWAS and eQTL can help us dissect genetic mechanism of multiple diseases (Guo et al., 2018; Heinrichs et al., 2018). The Genotype-Tissue Expression (GTEx) project has established the largest comprehensive public database with whole-genome and transcriptome sequencing data across 53 normal human tissues from nearly 1,000 individuals, making it better to dissect the effects and molecular mechanism of functional variations.

In a given gene, several variants modulate its expression level in stomach tissue. These expression-associated SNPs (eSNPs) may synergistically regulate the expression of the target gene. Thus, collections of multiple genetic variants, rather than individual highly significantly associated eSNPs, may account for the putative role of the novel gene in predisposition to GC. Pathway-based analysis evaluates the cumulative effect of multiple SNPs from the same gene set. Utilizing this approach, several novel genes and biological pathways enriched with significantly disease-associated SNPs were identified (Cheng et al., 2016; Yao et al., 2016; Walsh et al., 2019). Generally, most studies select the representative SNPs by their proximity to a specific gene, which inevitably obscures the genetic effect between the candidate gene and disease. Accordingly, incorporating functional eSNPs into the pathway analysis is appealing because of its ability to explore the mechanism of complex diseases. Through evaluating the cumulative effect of 322,324 eSNPs in Caucasian individuals, scientists found that the autoimmune thyroid disease pathway and JAK-STAT pathway were involved in basal cell carcinoma pathogenesis (Zhang et al., 2012). Moreover, a similar strategy was also applied to obtain biological insight into the development of lung cancer and type 2 diabetes (Zhong et al., 2010; Wang et al., 2018). During the preparation of the manuscript, another similar computational method called loci2path was reported (Xu et al., 2020).

Considering the fact that regulatory causal variants confer to GC risk by affecting their target gene expression, we initially conducted genome-wide screening of 389,207 potential eSNPs in stomach tissues from the GTEx database. We then evaluated the associations of 314,203 eSNPs shared in three GWAS datasets with GC risk. In addition, we performed a gene-based analysis to calculate the cumulative effect of eSNPs and identify additional susceptibility genes that might help provide new insight into the mechanism of GC.

MATERIALS AND METHODS

eSNP Analysis

Expression-associated SNPs in stomach tissues were derived from the GTEx v7 database (Stomach.allpairs.txt.gz). Genotyping was performed using Illumina HumanOmni 5 M and 2.5 M. Transcriptome dataset was generated by Affymetrix Expression Array or Illumina TruSeq RNA sequencing. A total of 237 stomach tissues with both genotype and expression data were available. Linear regression analysis was applied to evaluate the association between genetic variants and expression levels of genes within 1 Mb distance. As a result, a total of 636,426 *cis*-eQTL gene (eGene) pairs were defined with a false discovery rate (FDR) P-value < 0.05. After excluding indels, duplicated and non-biallelic eSNPs, there were 389,207 eSNPs remained.

GC GWAS Datasets

Three existing GC GWAS datasets were used in the current study, including 2,631 cases and 4,373 controls. Of them, NJ-GWAS, and BJ-GWAS were previously conducted by our group (Shi et al., 2011). All subjects recruited from Nanjing (550 cases and 1,155 controls) and Beijing (456 cases and 1,118 controls) were genotyped with Affymetrix Genome-Wide Human SNP Array 6.0. Another GC GWAS dataset named SX-GWAS was approved and downloaded from the dbGap (accession number: phs000361.v1.p1; Abnet et al., 2010). All participants (1,625 cases and 2,100 cancer-free individuals) recruited from Shanxi and Linxian were genotype using the Illumina 660W-Quad chips. The basic characteristics of study participates were shown in **Supplementary Table S1**.

Quality Control and Imputation for GWAS

We performed a standard quality control procedure for these three GWAS by excluding samples with lower call rates, sex discordance, or excessive heterozygosity. Then, we excluded eSNPs with a call rate $< 95\%$, minor allele frequency (MAF) <0.01, or $P < 1 \times 10^{-6}$ for Hardy-Weinberg equilibrium. Imputation was performed with SHAPEIT v2 (Delaneau et al., 2011) and IMPUTE2 (Howie et al., 2009) with the 1000 Genomes Project (Phase III integrated variant set release, across 2,504 samples) as reference. We selected eSNPs with INFO score ≥ 0.4 for further association analysis.

Association Analysis

For each eSNP, unconditional logistic regression was conducted to calculate odds ratios (ORs), and 95% confidence intervals

(CIs). We performed genetic association analysis assuming an additive effect model with adjustment for age, sex, smoking, alcohol consumption, and top ten principal components (PCs) in NJ-GWAS and BJ-GWAS. Since the smoking and drinking status were not available in the SX-GWAS dataset, we took age, sex and top ten PCs as covariates. Subsequently, a meta-analysis with the fixed-effects model was conducted to pool the results from each GWAS by using the GWAMA software (Magi and Morris, 2010). I^2 indicates the percentage of the effect estimates variability which can be attributed to heterogeneity, and an I^2 value of ≥75% represents high heterogeneity. We filtered significant eSNPs on linkage disequilibrium (LD; $r^2 < 0.1$), from which, the index eSNPs with the lowest p value in each LD block were obtained. All statistical analyses were conducted by using PLINK 1.9 and R language (version 3.5.0). Regional association plots were generated in LocusZoom.

Variance Explained

The phenotypic variance explained by genetic variants was estimated using the fixed-effects model in the single-variant analysis as previously described (Lee et al., 2012). Variants identified in the present study and those published in previous GWAS (**Supplementary Table S2**) were used to calculate the respective variances by assuming the 5-year prevalence of GC to be 32.43/100,000, 42.43/100,000, and 52.43/100,000 in China[1].

In silico Functional Annotation

We used ANNOVAR (Wang et al., 2010) to generate gene-based annotation and then described the distribution of all these eSNPs. We extracted candidate SNPs in strong LD ($r^2 \geq 0.6$) with the index variant based on the 1000 Genomes Phase 1 Asian individuals from the online HaploReg v4.2 tool (Ward and Kellis, 2012). According to the available data from ENCODE (Ward and Kellis, 2012) and Roadmap (Bernstein et al., 2010) we predicted regulatory elements (promoter, enhancer, etc.) through histone modification markers (H3K4me3, H3K4me1, and H3K27ac) and chromatin state segmentation in the stomach tissues and DNase I hypersensitivity sites (DHS) in 125 cell types. Other bioinformatics annotation tools, including RegulomeDB (**Supplementary Table S3**; Boyle et al., 2012) CADD (Kircher et al., 2014) GWAVA (Ritchie et al., 2014) and PINES (Corneliu et al., 2018) were also used to decipher the potential functional variants.

Gene-Based and Pathway Analysis

Gene-based analysis was performed using the sequence kernel association combined test (SKAT-C), which calculates the combined effect of common variants toward a particular phenotype (Ionita-Laza et al., 2013). Pathway analysis was conducted in merged dataset by the adaptive rank truncated product (ARTP) method with 10,000 permutations, which utilizes highly efficient permutations to analyze the association between genes within a pathway and diseases (Yu et al., 2009). All analyses were implemented in R package "SKAT" and "ARTP." Human-derived gene sets were cataloged by and obtained from the Molecular Signatures Database (MSigDB, version 6.2). Finally, a total of 1,077 pathways with 5,155 related genes were derived from KEGG ($n = 186$), Reactome ($n = 674$), and BioCarta ($n = 217$). The Benjamini-Hochberg method was applied to correct multiple testing, setting the threshold for significance at 5% FDR. In addition, genes were considered significant when they had P-values < 0.05 in at least two GWAS datasets.

Sherlock Integrative Analysis

We used Sherlock integrative analysis for further validation (He et al., 2013). Sherlock uses a Bayesian statistical method to calculate the individual Bayes factor for each eSNP, and their sum constitutes the final Log Bayes factor (LBF) score for each gene. The larger LBF score represents the higher probability that the gene is associated with GC. If an eSNP is significantly associated with GC, a positive score would be assigned. Otherwise, a negative LBF score would be given. The P threshold for statistical significance was set to 1.0×10^{-3}.

Differential Expression Analysis

We downloaded the normalized expression data and clinical information of individuals with GC from The Cancer Genome Atlas database. Differential expression analyses were performed in 32 paired gastric tumor and adjacent normal tissues.

Co-expression and Gene-Set Enrichment Analysis

The expression data of 23,424 genes in 237 normal stomach tissues were obtained from the GTEx v7 database. We conducted genome-wide expression correlation analysis to identify co-expression genes with the linear regression model. Gene-set enrichment analysis (GSEA) of the KEGG pathway gene set collection was implemented in R package "clusterProfiler" (Yu et al., 2012). All genes were pre-ranked according to the Pearson correlation coefficients calculated by the co-expression analysis. Then, gene sets were considered significantly enriched if the FDR was <0.05 after 100,000 permutations.

RESULTS

Individual eSNP Associated With GC Risk

As shown in the workflow chart (**Figure 1**), 389,207 eSNPs were found to be significantly associated with their surrounding gene expression levels (FDR < 0.05) in 237 stomach tissue samples from the GTEx database. Among them, 319,656, 321,098, and 322,370 eSNPs passed the quality control in NJ-GWAS, BJ-GWAS, and SX-GWAS, respectively. A total of 314,203 shared eSNPs were included in the genetic association analysis, and the association results of 307,676 variants without heterogeneity between studies ($I^2 < 75.0\%$) were shown in **Figure 2A**. Most of the eSNPs were located within intronic (48.21%) or intergenic (32.60%), and 8.19% had a RegulomeDB score less than 3 (**Figure 2B**). After LD pruning, we identified a total of 1,222 index eSNPs at $P < 0.05$. Among them, 4 eSNPs were retained after multiple testing correction (FDR < 0.05; **Table 1**).

[1]https://gco.iarc.fr/today/online

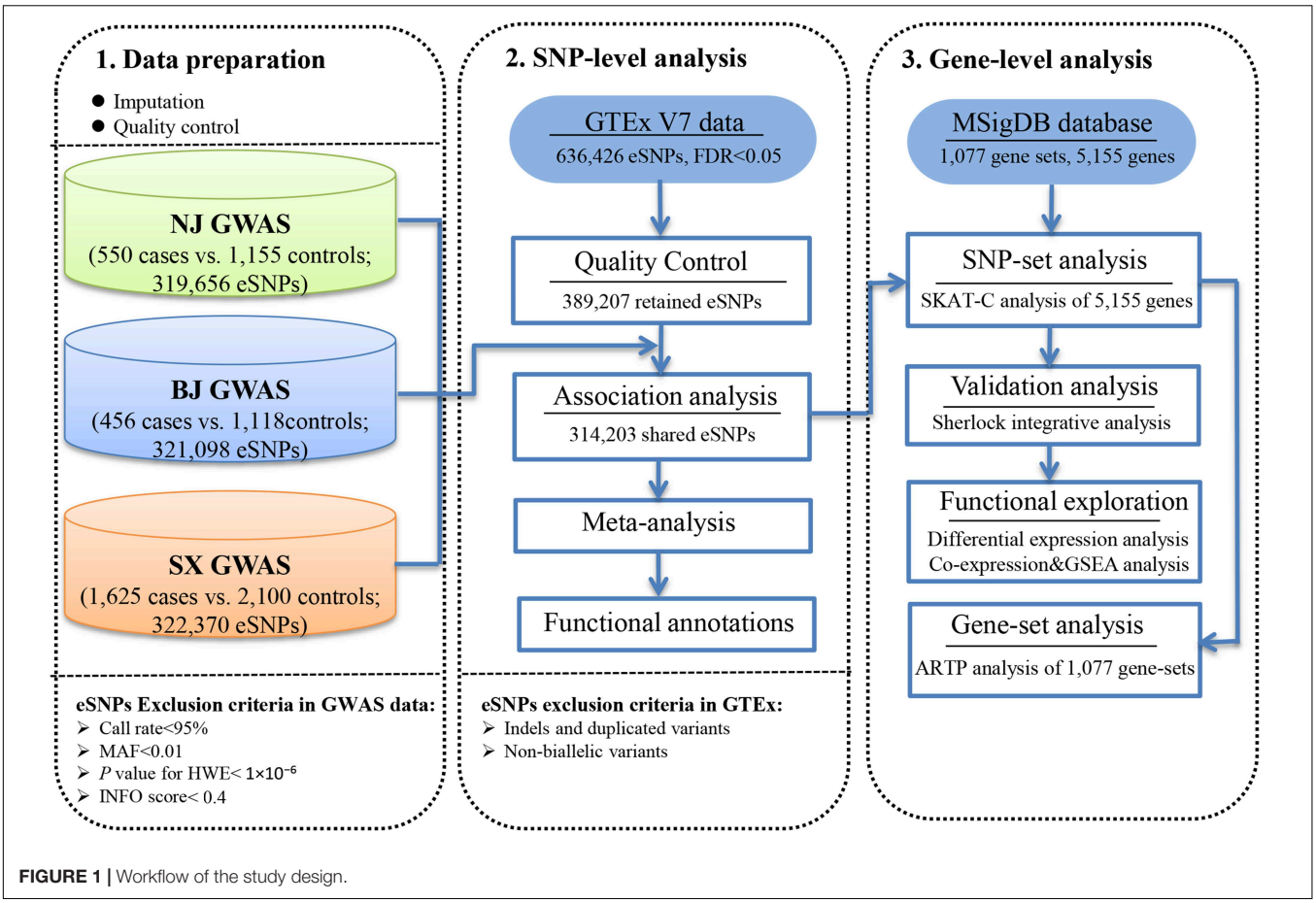

FIGURE 1 | Workflow of the study design.

Region plots of these four significant variants were depicted in **Supplementary Figure S1**.

The two most strongly risk-associated variants (rs6676150 at 1q22 and rs12217597 at 10q23.33) in known loci achieved genome-wide association significance ($P = 4.29 \times 10^{-10}$ and $P = 1.74 \times 10^{-8}$, respectively), which correlated with the expression level of *THBS3* and *NOC3L*, respectively, (**Figures 3A,B**). Moreover, these two variants were in strong LD with previously reported index SNPs (**Supplementary Table S4**). Of note, we found that two novel variants at 7p22.1 (rs836545), and 6p22.2 (rs1892252) were significantly associated with GC risk (per *T* allele OR = 1.23, 95% CI: 1.12–1.35, and $P = 7.46 \times 10^{-6}$; per *G* allele OR = 1.41, 95% CI: 1.20–1.66, and $P = 2.43 \times 10^{-5}$, respectively). Meanwhile, the risk alleles rs836545-T and rs1892252-G were correlated with higher expression levels of *DAGLB* ($P = 3.70 \times 10^{-18}$) and *BTN3A2* ($P = 3.20 \times 10^{-5}$), respectively (**Figures 3C,D**). A total of 63 candidate SNPs in strong LD ($r^2 \geq 0.6$) with rs836545 were extracted by using the HaploReg v4.2 tool (**Supplementary Table S5**). We found that the rs836545 site located within an active enhancer in three cell types, and the variant allele was predicted to alter the binding of four regulatory motifs; however, the chromatin status in stomach tissue was quiescent. As depicted in **Supplementary Figure S2**, we focused on the region nearby the promoter of *DAGLB* containing two variants in perfect LD (rs3828944 and rs4724806 at a 25 bp distance, pairwise $r^2 = 1.00$), where histone markers and chromatin state signatures exhibited a strong transcriptional activity as well as DNase-seq evidence for transcription factor binding. Using a combination of annotation tools, we proposed that rs3828944 might be the most promising functional variant in this region. We did not observe any variants in LD with the rs1892252 by HaploReg. Nevertheless, our previous study have observed a tumor-promoting role of *BTN3A2* that was remotely regulated by rs1679709 at 6p22.1 (Zhu et al., 2017).

Variance Explained by Independent eSNPs

Based on the eSNPs identified in present study and those reported by previous GWAS, we estimated the proportion of phenotypic variance explained by a liability threshold model assuming a GC prevalence of 32.43/100,000, 42.43/100,000, and 52.43/100,000 (**Table 2**). These four identified eSNPs showed 0.58, 0.60, and 0.62%, respectively, while nineteen of these GWAS-reported SNPs accounted for 1.14, 1.19, and 1.23% of the total phenotypic variance at the respective prevalence. In total, all these variants associated with susceptibility to GC showed 1.30, 1.35, and 1.39% of the phenotypic variance, respectively.These two novel eSNPs

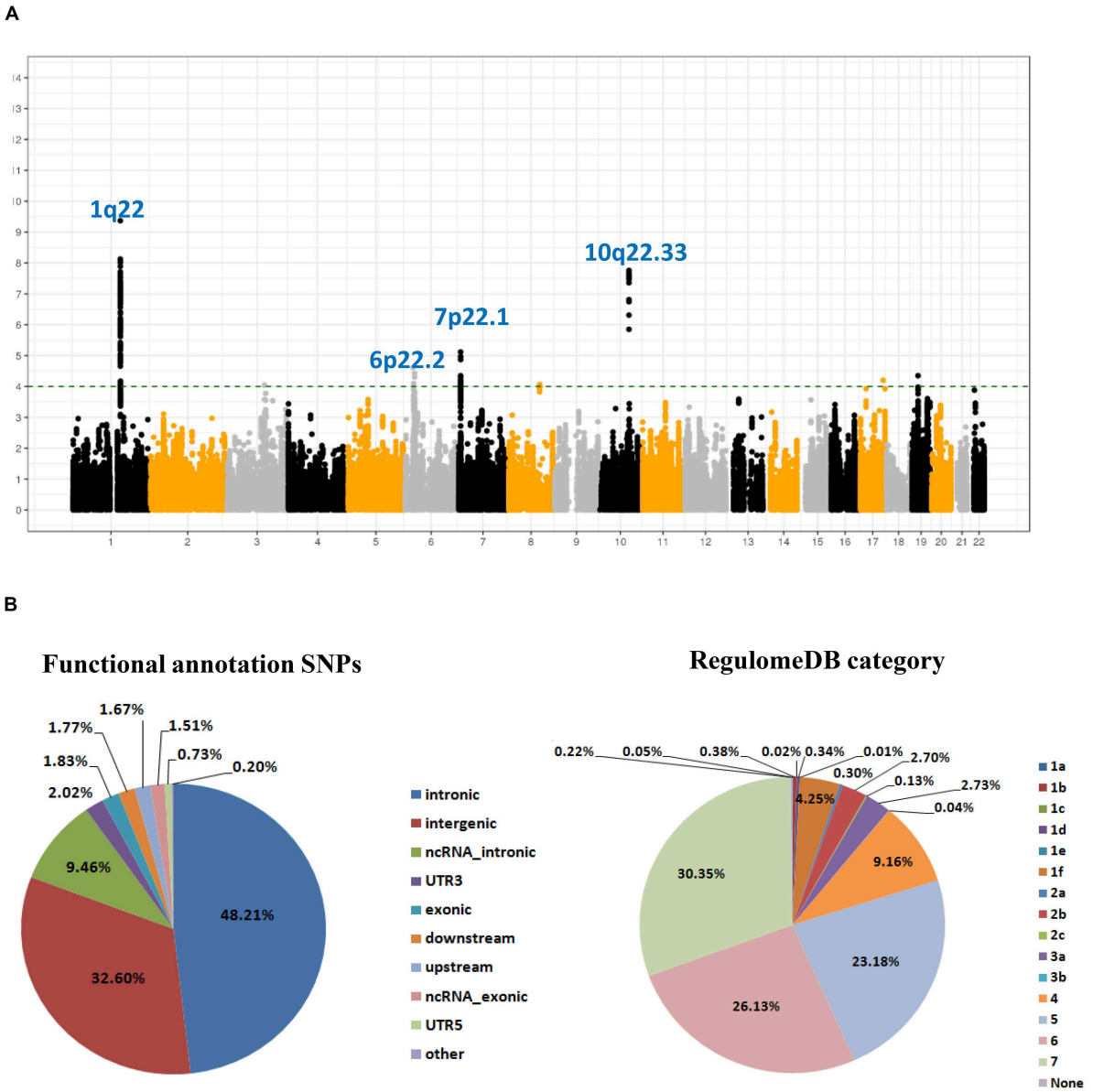

FIGURE 2 | SNP-based associations with GC in the GWAS meta-analysis. **(A)** Manhattan plot of *P* value for each expression-related SNPs (eSNPs) highlighting key chromosomal regions. The associations [-$\log_{10}$ (*P*) values, *Y*-axis] are plotted against genomic position (*X*-axis by chromosome and the chromosomal position of NCBI build 37). The green horizontal line corresponds to a *P* value threshold of 1.00×10^{-4}; **(B)** Pie charts showing the distribution of functional annotation and Regulome DB score (a categorical sore range from 1a to 7, indicating biological indicating biological evidence of a SNP being a regulatory element, with a low score denoting a higher likelihood of a SNP being regulatory) for 307,676 eSNPs without heterogeneity between studies.

(rs836545 and rs1892252) showed approximately 12.37% (0.49%/3.96%) of the phenotypic variance owing to known genetic variations.

Susceptibility Genes Associated With GC Risk and Pathway Analysis

At the gene level, 302 (5.97%) of 5,055 pathway genes were associated with GC risk at a nominal *P*-value < 0.05. Five protein-coding genes, including *THBS3* ($P = 2.65 \times 10^{-8}$), *GBA* ($P = 1.29 \times 10^{-6}$), *GPR27* ($P = 1.59 \times 10^{-5}$), *AMDHD1* ($P = 2.65 \times 10^{-5}$), and *FBXO43* ($P = 1.26 \times 10^{-4}$), were significantly related to GC susceptibility in the pooled dataset after correction for multiple testing (FDR < 0.05; **Table 3**). Two genes (*THBS3* and *GBA*) were located in known susceptibility locus (1q22), while the other three genes (*GPR27* at 3p13, *AMDHD1* at 12q23.1, and *FBXO43* at 8q22.2) were identified as novel GC susceptibility genes. At the pathway level, there were no significant pathways after multiple testing correction. However, 23 pathways reached a less stringent threshold ($P < 0.05$), which was predominantly

TABLE 1 | Associations of four significant expression-related SNPs (eSNPs) with GC risk under the additive genetic model.

SNP	Region	Alleles[a]	NJ-GWAS	BJ-GWAS	SX-GWAS	Fixed-effect meta-analysis		
			OR(95% CI)[b]	OR(95% CI)[b]	OR(95% CI)[c]	OR(95% CI)	*P* value	FDR[d]
rs6676150	1q22	G/C	0.67 (0.52–0.86)	0.79 (0.54–1.17)	0.55 (0.65–0.76)	0.67 (0.59–0.76)	4.29×10^{-10}	3.41×10^{-6}
rs12217597	10q23.33	T/C	1.05 (0.85–1.29)	1.31 (0.95–1.81)	1.28 (1.45–1.64)	1.33 (1.21–1.47)	1.74×10^{-8}	6.92×10^{-2}
rs836545	7p22.1	C/T	1.10 (0.91–1.33)	1.37 (1.04–1.81)	1.13 (1.26–1.41)	1.23 (1.12–1.35)	7.64×10^{-6}	2.03×10^{-2}
rs1892252	6p22.2	C/G	1.69 (1.33–2.16)	1.60 (1.05–2.43)	0.70 (0.89–1.14)	1.41 (1.20–1.66)	2.43×10^{-5}	4.83×10^{-2}

[a]*Reference allele/effect allele.* [b]*Adjusted for age, gender, smoking, drinking and top ten principal components (PCs).* [c]*Adjusted for age, gender and top ten PCs.* [d]*FDR was corrected by Benjamini-Hochberg procedure.*

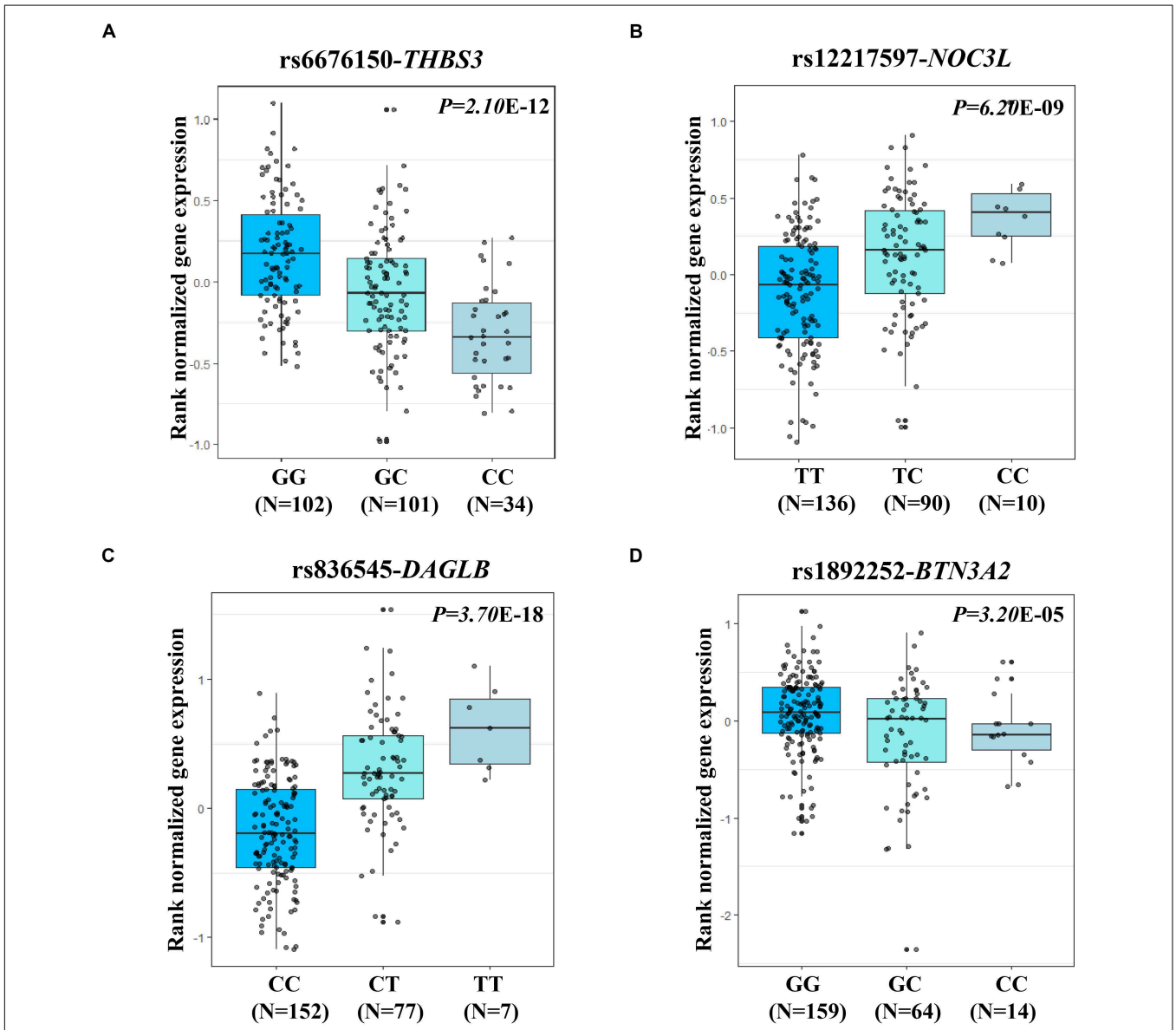

FIGURE 3 | eQTL analysis shown the associations of four expression-related SNPs (eSNPs) and its related genes in stomach tissues from GTEx. The small gray dot represents the individual log2 gene expression value. **(A)** eQTL analysis (rs6676150, risk allele C) for the expression of *THBS3* ($P = 2.10 \times 10^{-12}$); **(B)** eQTL analysis (rs12217597, risk allele C) for the expression of *NOC3L* ($P = 6.20 \times 10^{-9}$); **(C)** eQTL analysis (rs836545, risk allele T) for the expression of *DAGLB* ($P = 3.70 \times 10^{-18}$); and **(D)** eQTL analysis (rs1892252, risk allele G) for the expression of *BTN3A2* ($P = 3.20 \times 10^{-5}$).

TABLE 2 | Heritability estimated from variants associated with GC risk.

Model[a]	h^2(SE) observed scale	h^2(SE) liability scale		
		Prevalence (32.43/100,000)	Prevalence (42.43/100,000)	Prevalence (52.43/100,000)
SNPs identified by previous GWAS (n = 19)[a]	3.50%	1.14%	1.19%	1.23%
The identified eSNPs (n = 4)[b]	1.80%	0.58%	0.60%	0.62%
The identified eSNPs in unknown loci (n = 2)[c]	0.49%	0.16%	0.16%	0.17%
Combination (n = 21)[d]	3.96%	1.30%	1.35%	1.39%

[a]Variants reported by previous GWAS studies. [b]Significant eSNPs identified by the present study. [c]Significant novel eSNPs identified by the present study. [d]Consist of 19 GWAS reported SNPs and 2 novel identified eSNPs by present study.

TABLE 3 | Significant GC-associated protein-coding genes predicted by sequence kernel association combined test (SKAT-C).

Region	Gene	eSNP tested[a]	$P_{NJ-GWAS}$	$P_{BJ-GWAS}$	$P_{SX-GWAS}$	$P_{combined}$	FDR[b]
Known region							
1q22	*THBS3*	79	4.98×10^{-2}	2.75×10^{-1}	7.89×10^{-6}	2.65×10^{-8}	3.00×10^{-5}
1q22	*GBA*	14	2.56×10^{-2}	7.14×10^{-2}	8.56×10^{-4}	1.29×10^{-6}	1.11×10^{-3}
Unknown region							
3p13	*GPR27*	51	4.97×10^{-8}	9.99×10^{-1}	6.03×10^{-3}	1.59×10^{-5}	9.30×10^{-3}
12q23.1	*AMDHD1*	75	1.17×10^{-6}	2.83×10^{-3}	1.54×10^{-1}	2.65×10^{-5}	1.36×10^{-2}
8q22.2	*FBXO43*	28	3.81×10^{-1}	5.00×10^{-3}	1.61×10^{-3}	1.26×10^{-4}	4.97×10^{-2}

[a]Number of eSNPs mapped to each gene. [b]False discovery rate in the combined dataset.

TABLE 4 | Top GC-related protein-coding genes predicted by Sherlock integrative analysis.

Region	Gene	LBF[a]	P[b]	Supporting SNP[c]	P_{GWAS}[d]	P_{eQTL}[e]
Known region						
1q22	*THBS3*	7.31	2.45×10^{-5}	rs2049805	2.82×10^{-8}	1.85×10^{-9}
10q23.33	*NOC3L*	7.18	2.45×10^{-5}	rs12220125	2.09×10^{-9}	2.79×10^{-9}
1q22	*GBA*	6.87	3.43×10^{-5}	rs12034326	1.38×10^{-5}	2.90×10^{-6}
Unknown region						
8q22.2	*FBXO43*	5.79	9.31×10^{-5}	rs2453641	9.39×10^{-5}	3.45×10^{-6}
7p22.1	*DAGLB*	5.60	1.32×10^{-4}	rs4724806	1.08×10^{-5}	3.44×10^{-18}
19p13.11	*HAPLN4*	4.18	7.99×10^{-4}	rs2905421	4.48×10^{-5}	4.62×10^{-8}
19q13.43	*ZNF329*	4.17	8.08×10^{-4}	rs157375	3.34×10^{-4}	4.53×10^{-6}

[a]LBF (logarithm of Bayes factor) is to assess whether a gene is associated with GC through integrating the GWAS signal and eQTL. The larger LBF score represents the higher probability that the gene is associated with GC. For example, a LBF of 7.31 means that a gene is more likely [1495 times, (exp(7.31) = 1495] to be associated with GC than no association. [b]P-value from Sherlock integrative analysis. [c]eSNP with the highest LBF. [d]P-value from expression quantitative trait analysis. [e]P-value from meta-analysis of three GC GWAS datasets.

related to metabolism and transcription. Details are shown in **Supplementary Table S6**.

Sherlock Integrative Analysis Prioritizes Seven Risk Protein-Coding Genes

We integrated genetic associations from the meta-analysis of three GC GWAS (a total of 307,676 eSNPs with no heterogeneity) with stomach eQTL from the GTEx database. Sherlock integrative analysis identified seven top GC susceptibility genes whose expression might confer GC risk ($P < 1.0 \times 10^{-3}$; **Table 4**). Compared with the abovementioned results, this new approach validated five genes consisting of three known genes (*THBS3*, *NOC3L*, and *GBA*) and two novel genes (*FBXO43* and *DAGLB*).

Differential Expression Analysis and GSEA

We compared the expression level of *DAGLB* and *FBXO43* in 32 paired tissue samples of patients with GC. Both mRNA levels of the two genes were remarkably unregulated in tumors than in their adjacent normal tissues ($P = 5.59 \times 10^{-7}$ and $P = 3.90 \times 10^{-6}$, respectively; **Supplementary Figures S3A,B**). The Kaplan-Meier plotter online tool revealed that high expression level of *DAGLB* or *FBXO43* was associated with an unfavorable prognosis in patients with GC (*DAGLB*, HR = 1.77, 95%CI: 1.43–2.20, and $P = 1.30 \times 10^{-7}$; *FBXO43*, HR = 1.39, 95%CI: 1.09–1.78, and $P = 7.60 \times 10^{-3}$; **Supplementary Figures S3C,D**). To identify the potential function of these two genes in GC tumorigenesis, we conducted GSEA on the

correlation coefficients from co-expression analysis with 23,424 genes in 237 normal stomach tissues. We observed that co-expression genes with *DAGLB* or *FBXO43* were significantly enriched in several classical cancer-related pathways, including MAPK, WNT, JAK-STAT, and P53 signaling (all FDR < 0.05; **Supplementary Tables S7, S8**).

DISCUSSION

In the current study, we conducted a genome-wide scan with 2,631 cases and 4,373 controls to systematically explore the associations of 314,203 *cis*-eSNPs with GC risk, and then we incorporated the association signals with eQTL data to identify more risk genes for GC. Hitherto, this is the most extensive overview of the role of eQTL related variants in GC susceptibility. Of interest, we discovered two independent novel eSNPs associated with GC risk, which together captured nearly 12.37% of the phenotypic variance explained by all identified genetic loci. Synthesizing the results of single SNP association and gene-based analyses, we identified *DAGLB* and *FBXO43* as novel susceptibility genes for GC. Differential expression analysis and GSEA also highlighted the tumorigenicity of *DAGLB* and *FBXO43*.

At the individual eSNP level, we discovered two novel risk loci (rs836545 at 7p22.1 and rs1892252 at 6p22.2). The risk T allele of rs836545 increased the expression level of *DAGLB* in stomach tissues. As supporting evidence, it was shown that *DAGLB* was significantly elevated in GC tissues than in adjacent normal tissues. Moreover, Sherlock integrative analysis also confirmed that *DAGLB* was a promising susceptibility gene for GC. *DAGLB*, which encodes diacylglycerol lipase beta, has been widely studied in lipid mechanism. In *DAGLB* knockout mice, *DAGL*β inhibition can reduce 2-arachidonoylglycerol and arachidonic acid and eicosanoids in macrophages (Hsu et al., 2012). A recent GWAS reported a novel variant with HDL-C levels by modifying expression of *DAGLB* (Zhou et al., 2018). To the best of our knowledge, metabolism of lipids, especially arachidonic acid, has been proved to be an important regulator in the process of inflammation and cancer (Walduck et al., 2009). Using *In silico* analysis, we identified that rs3828944 (in perfect LD with rs836545, r^2 = 0.97) located in the promoter region of *DAGLB* was mapped with the center of DHS peaks in 125 cell types and within regions harboring histone marks (H3K4me1, H3K4me3, and H3K27ac) in stomach tissues or mucosae. These convergent lines of evidence implied that the risk T allele of rs3828944 at 7p22.1 might confer GC risk though enhancing the expression of *DAGLB*. For rs1892252 at 6p22.2, the risk allele rs1892252-G showed increased expression of *BTN3A2*, which was greatly overexpressed in GC tissues. A recent GWAS have reported that rs1892252-C was a risk allele for schizophrenia (OR = 1.12, 95%CI: 1.09–1.15, $P = 7.0 \times 10^{-13}$; Ikeda et al., 2019). Intriguingly, our group has previously verified that the rs1679709 at 6p22.1 remotely regulated *BTN3A2* expression by modulating its enhancer activity and deletion of *BTN3A2* inhibited proliferation, migration, and invasion of GC cells (Zhu et al., 2017). *BTN3A2*, an isoform of BTN3 family, participates in regulating immune signal in T and natural killer cells (Messal et al., 2011). Besides, *BTN3A2* also plays an important role in activating the phosphoantigen-mediated Vγ9Vδ2 T cells toward the development of pancreatic ductal adenocarcinoma (PDAC), implicating it as a promising immunotherapeutic target for the treatment of PDAC (Benyamine et al., 2017).

As mentioned above, only one candidate susceptibility gene was found based on single eSNP analysis. Therefore, collections of multiple genetic variants, rather than individual highly significantly associated eSNPs, may account for a putative role of the novel gene in predisposition to GC. From the results of the SKAT-C and Sherlock integrative analyses, we identified another new risk gene, *FBXO43*, also known as *EMI2*, which is a member of F-box protein family that influences the state of meiosis via translational regulation (Tan et al., 2018). A previous study has shown that the mRNA level of *FBXO43* is dramatically upregulated in hepatocellular carcinoma tissues than in normal tissues, and elevated *FBXO43* expression indicates a poor prognosis in patients with hepatocellular carcinoma (Tang et al., 2008). Consistent with the observation, *FBXO43* was overexpressed in GC tissues and associated with poor prognosis in patients with GC. Co-expression genes with *FBXO43* in normal stomach tissue were predominantly involved in several important signal transduction pathways, including MAPK, TGF-beta, WNT, and P53 signaling.

In conclusion, our findings highlighted the importance of eSNPs in dissecting genetic basis of GC. We discovered two novel eSNPs, rs836545 at 7p22.1, and rs1892252 at 6p22.2, which were significantly associated with susceptibility to GC. Furthermore, we integrated eQTL data with GWAS association signal to identify *FBXO43* and *DAGLB* as new GC risk genes. These susceptible eSNPs, together with candidate genes, will provide new insight into the genetic and biological basis for the mechanism of GC development.

DATA AVAILABILITY STATEMENT

eSNPs were derived based on stomach tissues from the GTEx database (V7 release; https://gtexportal.org/home/datasets).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of Nanjing Medical College. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

GJ and MZ designed and performed the research. CY and JN prepared the tables and figure. JN wrote the manuscript. TW and YW analyzed the data. YL, YD, BD, and GL collected the samples and

information. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China (81872702); National Major Research and Development Program (2016YFC1302703); Key Research and Development Program of Jiangsu Province (BE2019698); and Jiangsu Province "333" project (BRA 2018057). Project funded by China Postdoctoral Science Foundation (2019TQ0157).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00679/full#supplementary-material

REFERENCES

Abnet, C. C., Freedman, N. D., Hu, N., Wang, Z., Yu, K., Shu, X. O., et al. (2010). A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* 42, 764–767. doi: 10.1038/ng.649

Benyamine, A., Loncle, C., Foucher, E., Blazquez, J. L., Castanier, C., Chretien, A. S., et al. (2017). BTN3A is a prognosis marker and a promising target for Vgamma9Vdelta2 T cells based-immunotherapy in pancreatic ductal adenocarcinoma (PDAC). *Oncoimmunology* 7:e1372080. doi: 10.1080/2162402X.2017.1372080

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048. doi: 10.1038/nbt1010-1045

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Cheng, T. D., Ambrosone, C. B., Hong, C. C., Lunetta, K. L., Liu, S., Hu, Q., et al. (2016). Genetic variants in the mTOR pathway and breast cancer risk in African American women. *Carcinogenesis* 37, 49–55. doi: 10.1093/carcin/bgv160

Corneliu, A., Bodea, A., and Alex, B. (2018). Phenotype-specific information improves prediction of functional impact for noncoding variants. *bioRxiv* [preprint]. doi: 10.1101/083642

Delaneau, O., Marchini, J., and Zagury, J. F. (2011). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785

Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M., et al. (2018). A comprehensive cis-eQTL analysis revealed target genes in breast cancer susceptibility loci identified in genome-wide association studies. *Am. J. Hum. Genet.* 102, 890–903. doi: 10.1016/j.ajhg.2018.03.016

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., et al. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* 92, 667–680. doi: 10.1016/j.ajhg.2013.03.022

Heinrichs, S. K. M., Hess, T., Becker, J., Hamann, L., Vashist, Y. K., Butterbach, K., et al. (2018). Evidence for PTGER4, PSCA, and MBOAT7 as risk genes for gastric cancer on the genome and transcriptome level. *Cancer Med.* 7, 5057–5065. doi: 10.1002/cam4.1719

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Hsu, K. L., Tsuboi, K., Adibekian, A., Pugh, H., Masuda, K., and Cravatt, B. F. (2012). DAGLbeta inhibition perturbs a lipid network involved in macrophage inflammatory responses. *Nat. Chem. Biol.* 8, 999–1007. doi: 10.1038/nchembio.1105

Ikeda, M., Takahashi, A., Kamatani, Y., Momozawa, Y., Saito, T., Kondo, K., et al. (2019). Genome-wide association study detected novel susceptibility genes for schizophrenia and shared trans-populations/diseases genetic effect. *Schizophr. Bull.* 45, 824–834. doi: 10.1093/schbul/sby140

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* 92, 841–853. doi: 10.1016/j.ajhg.2013.04.015

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892

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

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., et al. (2013). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front. Genet.* 4:103. doi: 10.3389/fgene.2013.00103

Magi, R., and Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 11:288. doi: 10.1186/1471-2105-11-288

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Messal, N., Mamessier, E., Sylvain, A., Celis-Gutierrez, J., Thibult, M. L., Chetaille, B., et al. (2011). Differential role for CD277 as a co-regulator of the immune signal in T and NK cells. *Eur. J. Immunol.* 41, 3443–3454. doi: 10.1002/eji.201141404

Mucci, L. A., Hjelmborg, J. B., Harris, J. R., Czene, K., Havelick, D. J., Scheike, T., et al. (2016). Familial risk and heritability of cancer among twins in nordic countries. *JAMA* 315, 68–76. doi: 10.1001/jama.2015.17703

Park, B., Yang, S., Lee, J., Woo, H. D., Choi, I. J., Kim, Y. W., et al. (2019). Genome-wide association of genetic variation in the PSCA gene with gastric cancer susceptibility in a korean population. *Cancer Res. Treat.* 51, 748–757. doi: 10.4143/crt.2018.162

Ritchie, G. R., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nat. Methods* 11, 294–296. doi: 10.1038/nmeth.2832

Shi, Y., Hu, Z., Wu, C., Dai, J., Li, H., Dong, J., et al. (2011). A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat. Genet.* 43, 1215–1218. doi: 10.1038/ng.978

Tan, J., Fu, L., Chen, H., Guan, J., Chen, Y., and Fang, J. (2018). Association study of genetic variation in the autophagy lysosome pathway genes and risk of eight kinds of cancers. *Int. J. Cancer* 143, 80–87. doi: 10.1002/ijc.31288

Tang, W., Wu, J. Q., Guo, Y., Hansen, D. V., Perry, J. A., Freel, C. D., et al. (2008). Cdc2 and Mos regulate Emi2 stability to promote the meiosis I-meiosis II transition. *Mol. Biol. Cell* 19, 3536–3543. doi: 10.1091/mbc.E08-04-0417

Walduck, A. K., Weber, M., Wunder, C., Juettner, S., Stolte, M., Vieth, M., et al. (2009). Identification of novel cyclooxygenase-2-dependent genes in *Helicobacter pylori* infection *in vivo*. *Mol. Cancer* 8:22. doi: 10.1186/1476-4598-8-22

Walsh, N., Zhang, H., Hyland, P. L., Yang, Q., Mocci, E., Zhang, M., et al. (2019). Agnostic pathway/gene set analysis of genome-wide association data identifies associations for pancreatic cancer. *J. Natl. Cancer Inst.* 111, 557–567. doi: 10.1093/jnci/djy155

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

Wang, Y., Wu, W., Zhu, M., Wang, C., Shen, W., Cheng, Y., et al. (2018). Integrating expression-related SNPs into genome-wide gene- and pathway-based analyses identified novel lung cancer susceptibility genes. *Int. J. Cancer* 142, 1602–1610. doi: 10.1002/ijc.31182

Wang, Z., Dai, J., Hu, N., Miao, X., Abnet, C. C., Yang, M., et al. (2017). Identification of new susceptibility loci for gastric non-cardia adenocarcinoma: pooled results from two Chinese genome-wide association studies. *Gut* 66, 581–587. doi: 10.1136/gutjnl-2015-310612

Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917

Xu, T., Jin, P., and Qin, Z. S. (2020). Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. *Bioinformatics* 36, 690–697. doi: 10.1093/bioinformatics/btz669

Yao, S., Haddad, S. A., Hu, Q., Liu, S., Lunetta, K. L., Ruiz-Narvaez, E. A., et al. (2016). Genetic variations in vitamin D-related pathways and breast cancer risk in African American women in the AMBER consortium. *Int. J. Cancer* 138, 2118–2126. doi: 10.1002/ijc.29954

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., et al. (2009). Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* 33, 700–709. doi: 10.1002/gepi.20422

Zhang, M., Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Ding, J., et al. (2012). Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. *Hum. Genet.* 131, 615–623. doi: 10.1007/S00439-011-11047-810.1007/s00439-011-1107-5

Zhong, H., Yang, X., Kaplan, L. M., Molony, C., and Schadt, E. E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* 86, 581–591. doi: 10.1016/j.ajhg.2010.02.020

Zhou, D., Zhang, D., Sun, X., Li, Z., Ni, Y., Shan, Z., et al. (2018). A novel variant associated with HDL-C levels by modifying DAGLB expression levels: an annotation-based genome-wide association study. *Eur. J. Hum. Genet.* 26, 838–847. doi: 10.1038/s41431-018-0108-4

Zhu, M., Yan, C., Ren, C., Huang, X., Zhu, X., Gu, H., et al. (2017). Exome array analysis identifies variants in SPOCD1 and BTN3A2 that affect risk for gastric cancer. *Gastroenterology* 152, 2011–2021. doi: 10.1053/j.gastro.2017.02.017

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: 21 July 2020
doi: 10.3389/fgene.2020.00655

Identifying Circular RNA and Predicting Its Regulatory Interactions by Machine Learning

Guishan Zhang[1], Yiyun Deng[1], Qingyu Liu[1], Bingxu Ye[2], Zhiming Dai[3,4], Yaowen Chen[2] and Xianhua Dai[1,5]**

[1] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China, [2] Key Laboratory of Digital Signal and Image Processing of Guangdong Provincial, College of Engineering, Shantou University, Shantou, China, [3] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, [4] Guangdong Province Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China, [5] Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China

Circular RNA (circRNA) is a closed long non-coding RNA (lncRNA) formed by covalently closed loops through back-splicing. Emerging evidence indicates that circRNA can influence cellular physiology through various molecular mechanisms. Thus, accurate circRNA identification and prediction of its regulatory information are critical for understanding its biogenesis. Although several computational tools based on machine learning have been proposed for circRNA identification, the prediction accuracy remains to be improved. Here, first we present circLGB, a machine learning-based framework to discriminate circRNA from other lncRNAs. circLGB integrates commonly used sequence-derived features and three new features containing adenosine to inosine (A-to-I) deamination, A-to-I density and the internal ribosome entry site. circLGB categorizes circRNAs by utilizing a LightGBM classifier with feature selection. Second, we introduce circMRT, an ensemble machine learning framework to systematically predict the regulatory information for circRNA, including their interactions with microRNA, the RNA binding protein, and transcriptional regulation. Feature sets including sequence-based features, graph features, genome context, and regulatory information features were modeled in circMRT. Experiments on public and our constructed datasets show that the proposed algorithms outperform the available state-of-the-art methods. circLGB is available at http://www.circlgb.com. Source codes are available at https://github.com/Peppags/circLGB-circMRT.

Keywords: circular RNA, long non-coding RNA, microRNA, RNA binding protein, transcriptional regulation, machine learning

OPEN ACCESS

Edited by:
Ting Hu,
Memorial University of Newfoundland, Canada

Reviewed by:
Firoz Ahmed,
Jeddah University, Saudi Arabia
Xiaoyong Pan,
Shanghai Jiao Tong University, China

****Correspondence:***
Yaowen Chen
ywchen@stu.edu.cn
Xianhua Dai
issdxh@mail.sysu.edu.cn

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

Received: *28 October 2019*
Accepted: *29 May 2020*
Published: *21 July 2020*

Citation:
Zhang G, Deng Y, Liu Q, Ye B, Dai Z, Chen Y and Dai X (2020) Identifying Circular RNA and Predicting Its Regulatory Interactions by Machine Learning. Front. Genet. 11:655. doi: 10.3389/fgene.2020.00655

INTRODUCTION

Circular RNA (circRNA) constitutes a unique class of RNAs that is characterized by the presence of a covalently closed cyclic structure without a poly adenylated tail (Lasda and Parker, 2014). During pre-mRNA splicing, the 5′ and 3′ termini of exons can be covalently ligated to form circRNAs (Barrett et al., 2015; Wang and Wang, 2015). Owing to their circular structure and lack of free ends (Awasthi et al., 2018), circRNAs have greater stability and are more conserved across species than linear RNAs (Jeck et al., 2013). Although the functions of most circRNAs are still elusive, they have been shown to act as sponges to microRNAs (miRNAs; Hansen et al., 2013; Panda, 2018) and may

potentially sponge RNA binding proteins (RBPs; Memczak et al., 2013). In addition, circRNAs can also be involved in transcriptional regulation (TR) and alternative splicing (Zhang et al., 2013; Conn et al., 2017). circRNAs may even have translation potential (Li et al., 2018). circRNAs play crucial roles in gene regulation and the development of many complex diseases. Moreover, circRNAs have a promising potential as biomarkers of diseases due to their stability and relation to diseases (Zhang et al., 2018).

Circular RNAs have some different attributes from other long non-coding RNAs (lncRNAs), such as back-splicing (Xiong et al., 2015). Unlike lncRNA, which can be effectively recognized from other little non-coding RNAs (e.g., miRNA, siRNA, and snoRNA) according to the transcript size, it is scarcely possible to distinguish circRNA from different lncRNAs based on simple features (Xiong et al., 2015). Moreover, it is hard to classify circRNAs from other lncRNAs due to the low expression levels of almost all lncRNAs. To date, several machine learning-based methods have been developed for circRNA detection. For example, PredcircRNA (Pan and Xiong, 2015) identifies circRNAs by utilizing a multiple kernel learning-based (MKL) framework. This tool incorporates diverse sequence features including basic sequence features, graph features, conservation scores as well as features of transposable element (ALU), tandem repeats, ORF length, ORF proportion, and single nucleotide polymorphism (SNP) density (ATOS) to train and test the model. Hierarchical extreme learning machine (H-ELM; Chen et al., 2018) extracts identical features and discriminates circRNAs by performing a H-ELM algorithm with feature selection. circDeep (Chaabane et al., 2019) distinguishes circRNAs by integrating a reverse complement matching descriptor, an asymmetric convolutional neural network (CNN) combined with bidirectional long short-term memory sequence descriptor and a conservation descriptor for extracting high level abstract features of a given RNA sequence. When evaluating the performance on the published dataset proposed by Pan and Xiong (2015), circDeep achieves an improvement of over 12% in terms of accuracy (ACC) compared with PredcircRNA and H-ELM (with values of 0.778 vs. 0.789). However, there is still room for improving the performance. Thus, novel computational methods and comprehensive exploration of informative sequence features affecting back-splicing are required.

Technological obstacles for understanding the regulation and functions of circRNAs occur at various levels. Take suppression strategy as an example, it usually uses loss and gain functions to annotate gene function [i.e., RNAi (Boutros and Ahringer, 2008) and CRISPR/Cas9-mediated genome editing (Shalem et al., 2015)]. However, this technique does not have adequate ability to achieve specificity or high efficacy in targeting circRNAs. Therefore, decoding the regulatory interactions of circRNAs can greatly expand the understanding of their functions. Thanks to the development of high-throughput sequencing, alongside the advance of bioinformatics technology, a great number of circRNAs loci have been discovered in human genomes. Several databases and resources are available for describing the circRNAs regulatory interactions, which can facilitate research on miRNA, RBP, and TR interacting with specific circRNAs. For instance, Circ2Traits (Ghosal et al., 2013) predicts interactions between the disease-associated miRNAs and circRNAs. CircNet (Liu et al., 2016) provides circRNA–miRNA–gene regulatory networks and tissue-specific circRNA expression profiles. CircInteractome (Dudekula et al., 2016) explores circRNAs interacting with miRNAs. Besides, it identifies RBPs binding to circRNA junctions. CIRCpedia v2 (Dong et al., 2018) provides a comprehensive circRNA annotation from over 180 RNA-seq datasets across six different species. ENCORI (Li et al., 2014) identifies the miRNA–ceRNA, miRNA–ncRNA, and protein–RNA interaction networks. TRCirc (Tang et al., 2018) provides a resource for efficient retrieval, browsing and visualization of TR information of circRNAs. The availability of these databases speeds up the exploration of circRNAs biogenesis and the function analysis.

Machine learning has made impressive advances in the area of bioinformatics such as molecular interactions prediction. The machine learning-based predictors require considerable domain expertise to design the feature extractor. For example, Muppirala et al. (2011) proposed support vector machine (SVM) and random forest (RF)-based methods to predict the RNA–RBP interactions using sequence composition. Previous studies suggested that incorporating informative features can boost the predictive power (Ahmed et al., 2009, 2013; Wang L. et al., 2019). For instance, Ahmed et al. (2009) proposed SVM-based methods to predict guide strand of miRNAs and human Dicer cleavage sites (Ahmed et al., 2013). In their work, they found adding secondary structure information contributes to the improvement of ACC compared with considering sequence only. Owing to the non-coding nature of circRNA, the relationship between structure and function in it is stronger than in linear RNAs. There is increasing evidence that RNA secondary structure promotes exon skipping RNA circularization (Pervouchine, 2019) and alternative splicing (Buratti and Baralle, 2004). Besides, a quantitative characterization of the relationship between primary sequence and structure of circRNAs contributes to our understanding of how their function emerges. Inspired by this, incorporating secondary structure features may achieve better performance than considering primary sequence for circRNAs regulatory interactions prediction. Recently, machine learning-based identification of circRNAs coordinated regulatory interaction has been gradually applied in the bioinformatics field. For example, CircRNAs Interact with Proteins (CRIP) integrates CNN and a recurrent neural network to predict circRNA–RBP binding sites (Zhang et al., 2019). Wang Z. et al. (2019) proposed a multiple CNNs-based method to identify cancer-specific circRNA–RBP binding sites considering only nucleotide sequences. Ju et al. (2019) applied a hybrid LSTM-CNN-CRF (a long short-term memory network, CNN network and a conditional random field) model to identify RBP-binding sites on circRNAs (Ju et al., 2019). Lei and Fang (2019) proposed GBDTCDA, a gradient boosting decision tree (GBDT) regression model with multiple biological data to predict circRNA-disease associations (Lei and Fang, 2019). To the best of our knowledge, no machine learning-based tool has been proposed to systematically predict the regulatory information of circRNAs, including their interactions with miRNA, RBP, and TR.

In this study, we introduce two machine learning-based methods, circLGB and circMRT to combine both sequence and structure information, to identify circRNAs from other lncRNAs and to predict their regulatory interactions, respectively. circLGB extracts the commonly used features and three new features including adenosine to inosine (A-to-I) deamination, A-to-I density as well as internal ribosome entry site (IRES), and in turn, distinguishes circRNA by utilizing a LightGBM classifier. We propose a two-step feature optimization strategy to select the most discriminative features. circLGB achieves superior performance on the public and our datasets compared to the state-of-the-art methods. circMRT integrates sequence-based features, graph features, genome context and regulatory information for predicting circRNA interacting with miRNA, RBP, and TR. We first propose three classifiers to predict circRNA–miRNA, circRNA–RBP and circRNA–TR interactions, respectively. Each classifier extracts the abovementioned sequence features and predicts the regulatory interaction by applying an ensemble machine learning algorithm with optimal features. Then, the outputs of all three classifiers are fused by a union operator to predict the coordinated regulatory interaction of the candidate circRNA. As far as we know, circMRT is currently a comprehensive computational platform that predicts the regulatory information of circRNA using machine learning.

MATERIALS AND METHODS

Data Collection and Pre-processing

circlncRNA Datasets

We downloaded the human circRNAs from the circBase (Glazar et al., 2014) database. Taking circRNA isoforms into consideration and removing the transcripts which were shorter than 200 nt, we obtained 79,987 positive samples. Besides, we also downloaded the annotated human lncRNAs from LNCipedia (Volders et al., 2013). This database provides basic transcript information, gene structure and several statistics (e.g., miRNA binding sites and secondary structure) for each transcript. After excluding the overlapped circRNAs in circBase and deepBase (Zheng et al., 2016), we obtained 127,432 lncRNAs transcripts. We randomly selected 21,882 circRNAs and the same number of lncRNAs to construct our circlncRNA dataset. The determination of the sample size is given in **Supplementary Material** (**Supplementary Figure S1**).

CIRCdeep Dataset

We used a dataset available in Chaabane et al. (2019) (hereafter referred to as CIRCdeep). This dataset contains 32,914 human circRNAs and 19,683 lncRNAs. circRNAs were downloaded from the circRNADb (Chen et al., 2016) database. Transcripts shorter than 200 nt were removed. Negative data was collected from the GENCODE (Harrow et al., 2012) database. The annotated lncRNAs in GENCODE have three validation levels for RNA annotation, namely validation, manual annotation, and automated annotation. Only validated or manually annotated transcripts were chosen. CIRCdeep dataset can be downloaded at https://github.com/UofLBioinformatics/circDeep.

circMI, circRBP, circTR Datasets and the Independent Test Set

circRNA–miRNA and circRNA–RBP interactions were downloaded from the ENCORI database[1]. Additionally, circRNA–TR interactions were extracted from the TRCirc database[2]. After removing the duplicates and getting the full-length sequence and basic sequence information from circBase database, we built datasets circMI, circRBP, and circTR for training the classifiers of circRNA–miRNA, –RBP and –TR, respectively. To be specific, we randomly selected 1046 full-length circRNAs interacting with miRNAs to construct the positive data of circMI dataset. We collected 1036 and 2172 entire circRNAs which have interactions with RBPs and TRs, being used as positive data of circRBP and circTR datasets, respectively. Note that, there is no overlap among these three positive samples. We randomly selected 1046 circRNAs interacting with TRs as negative data for circMI dataset. Analogously, 1036 circRNAs interacting with TRs were derived to be used as negative samples for circRBP dataset. The 2172 circRNAs interacting with miRNAs or RBPs were chosen to construct the negative data of circTR dataset. In addition, we used 140 samples including 29 circRNA–miRNA interactions, 50 circRNA–RBP interactions, 40 circRNA–TR interactions, and 21 miRNA–circRNA-RBP interactions as an independent test set. This test set does not overlap the former datasets. More details can be found in **Table 1**.

Feature Extraction

Feature extraction has great influence on the predictive performance. Note that, features related to RNA circularization and circRNA regulatory information may be different. So, we separately extracted the features for circLGB and circMRT models. 188 sequence-derived features including 70 sequence composition features, 101 graph features, 12 conservation scores, and 5 ATOS features (Pan and Xiong, 2015; Chen et al., 2018) were used for circRNAs detection. Based on these features, we added three features including A-to-I, A-to-I density and IRES to train our circLGB. We extracted a 182-dimensional vector to train our circMRT for circRNAs regulatory interactions

[1]http://starbase.sysu.edu.cn/

[2]http://licpathway.net/TRCirc/view/index

TABLE 1 | Summary of circMI, circRBP, circTR datasets and the independent test set.

Model	Dataset	Positive data	Negative data
circRNA-miRNA	circMI	1046 circRNAs interacting with miRNAs	1046 circRNAs interacting with TRs
circRNA-RBP	circRBP	1036 circRNAs interacting with RBPs	1036 circRNAs interacting with TRs
circRNA-TR	circTR	2172 circRNAs interacting with TRs	2172 circRNAs interacting with miRNAs or RBPs
circMRT	Independent test set	–	–

prediction. These features were divided into four groups: sequence-based features, graph features, genome context and regulatory information. The value of each feature was normalized to the interval from 0 to 1. More details were summarized in **Supplementary Tables S1, S2**.

Features of circLGB for Classifying circRNA From Other lncRNAs

Group 1: Basic sequence features

The basic sequence features were extracted using the same processing scheme described in Pan and Xiong (2015). These features contain a wide range of possible explanatory attributes from 64 trinucleotide frequencies and other sequence component composition features, e.g., sequence length, GC content, frequencies of AG, GT, AGGT, and GTAG. GT/AG signal has an impact on forming the circRNAs, such as back-splicing and exon-junction (Kitamura-Abe et al., 2004). A detailed description can be referred to Pan and Xiong (2015).

Group 2: Graph feature

RNA structure plays an important role in gene splicing, which has an influence on back-splicing (Ding et al., 2014). Secondary structures play important role in identifying of the hypothetical interacting sites of circRNAs (Cuesta and Manrubia, 2017). In RNA graph, the nodes are nucleotides while edges represent backbone connection or bond relations between the nucleotides (Maticzka et al., 2014). RNA graph features reflect the relationships between nucleotides and represent the relations of the abstract structure annotations predicted from RNA shapes (Steffen et al., 2006). GraphProt is a machine learning-based framework considering both sequence and full secondary structure information that can find RBP sequence and structure-binding preferences from the high-throughput data (Maticzka et al., 2014). In this work, we applied GraphProt to calculate RNA secondary structures. In addition, it was adopted in previous studies (Pan and Xiong, 2015; Chen et al., 2018; Pan et al., 2018; Ilik et al., 2020). We initially extracted a 32,768-dimensional RNA graph feature vector for the candidate transcript using GraphProt 1.0.1. To improve the feature representation ability, Pan et al. employed RF to rank the extracted features by their importance scores and chose the top 101 features (Pan and Xiong, 2015). For fair comparison, we used these 101 features for analysis. The RF importance ranking list of the selected features can be downloaded from https://github.com/xypan1232/PredcircRNA/blob/master/features/all_fea_ranking.

Group 3: Conservation scores

Previous studies showed that circRNAs are significantly enriched with conserved nucleotides (Memczak et al., 2013). On the contrary, lncRNAs have a low level of sequence conservation compared with other functional transcripts (Marques and Ponting, 2009). Thus, conservation scores may help to discriminate circRNAs from lncRNAs. These scores were extracted by downloading the placent_phylop46way[3] from the UCSC database (Karolchik et al., 2003). We calculated the mean, maximum, and variance of conservation scores from per base phyloP conservation score for each transcript (Lowe et al., 2011). Furthermore, the frequencies of bases with conservation scores greater than 0.3, 0.6, 0.9 and smaller than 0.9 were also calculated.

Group 4: ALU and tandem repeat, ORF, SNP, IRES, A-to-I, and A-to-I density

ALU repeats contribute to RNA circularization by making the splice sites recognize each other (Liang and Wilusz, 2014). We downloaded the annotated ALU repeat sites from UCSC and calculated the number of ALU repeats for each transcript. Tandem duplications within a gene have a great impact on back-splicing (Ulitsky et al., 2011). Tandem repeats were detected by employing Tandem Repeat Finder (Benson, 1999). We computed the frequency of tandem repeats. The open reading frame (ORF) length information was extracted by using txCdsPredict from UCSC. The longest ORF and ORF propensity (ORF prop) defined by the length of an ORF divided by the total length of the transcript were calculated. Single nucleotide polymorphism data with coordinates in the genome was downloaded from the 1000 Genomes Project (Kuehn, 2008). Single nucleotide polymorphism density was computed for each transcript. A previous study suggested that A-to-I editing events occur frequently at intronic positions that were proximal to the splice sites of circularized exons (Ivanov et al., 2015). The annotated data of A-to-I was downloaded from the RADAR (Ramaswami and Li, 2014) dataset. A-to-I density was defined by the number of A-to-I divided by the sequence length for each transcript. Another work demonstrated that IRES provides the information of peptides or proteins from circRNA (Abe et al., 2015), implying that this feature has discriminative power for circRNA detection. IRES information of the given RNA sequence was extracted by IRESfinder (Zhao et al., 2018).

Features of circMRT for Predicting circRNA Regulatory Interactions

Group 1: Sequence-based features

The sequence features consist of 70 sequence composition features and one repeat feature. Note that these features were generated in the same way in section "Features of circLGB for Classifying circRNA From Other lncRNAs."

Group 2: Graph features

The 101-dimensional graph features were generated identically to the way described in section "Features of circLGB for Classifying circRNA From Other lncRNAs."

Group 3: Genome context features

We calculated the mean and standard deviation of conservation scores for each transcript. ALU, SNP density and A-to-I features were generated identically to the way described in section "Features of circLGB for Classifying circRNA From Other lncRNAs." A previous study showed that circRNA sequences are enriched for back-splice junctions (Jeck et al., 2013). Moreover, CIRI (Gao et al., 2015) and find_circ (Memczak et al., 2013) characterized circRNA by calculating the circular junctions. We derived the one-dimensional back-splice junction feature from the TRCirc database. It is a general phenomenon that circRNAs

[3]http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP46way/placentalMammals/

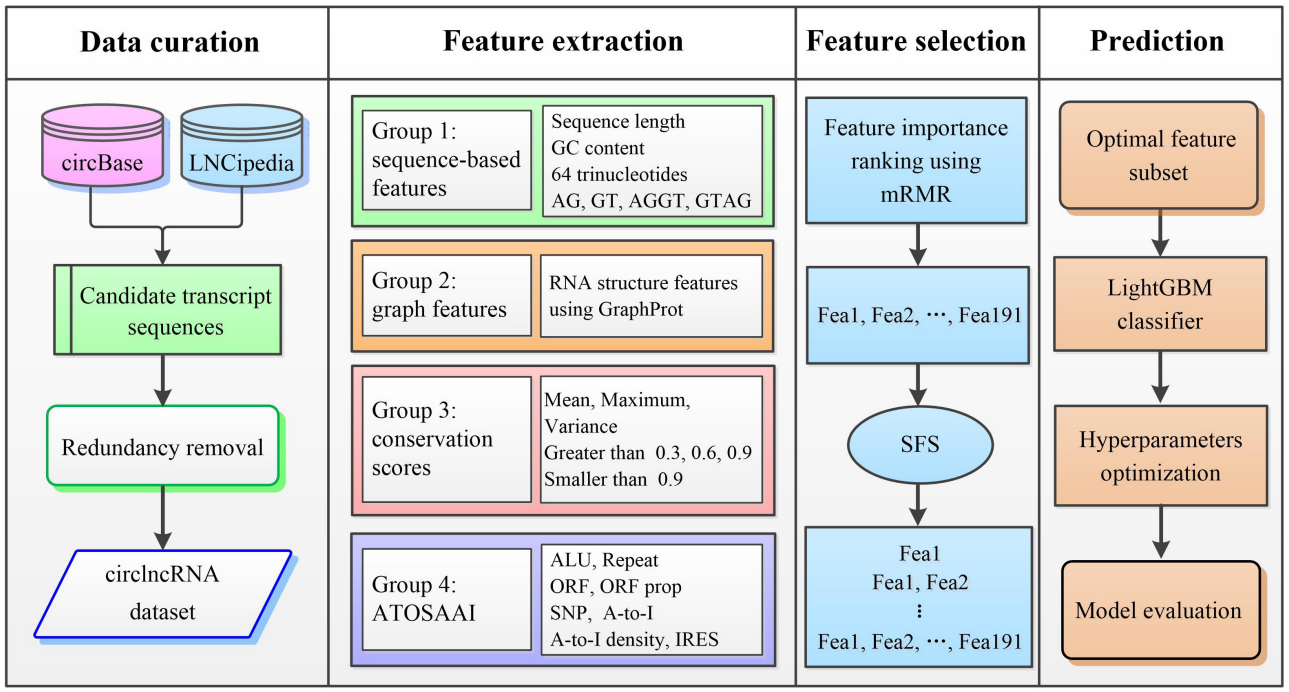

FIGURE 1 | Overview of the proposed circLGB for circRNA identification that involved the following steps: **(i)** construction of circlncRNA dataset; **(ii)** extraction of sequence-derived features including sequence composition, graph features, conservation scores and ATOSAAI for training and testing the circLGB model; **(iii)** ranking the features using mRMR algorithm according their importance and generation of the optimal feature subset using SFS; and **(iv)** construction of the final prediction by applying LightGBM classifier that separates the input into circRNAs and lncRNAs. ALU, transposable element; ATOSAAI, ATOS, A-to-I, A-to-I density and IRES; ORF, open reading frame; SNP, single nucleotide polymorphism; A-to-I, adenosine to inosine; mRMR, minimal redundancy and maximal relevance; SFS, sequential forward search. These abbreviations also apply to **Figures 2**, **3**.

compete with other RNAs for binding miRNAs. For example, ciRS-7 contains over 70 selective conserved miRNA target sites (Hansen et al., 2013). We integrated the one-dimensional miRNA binding sites as one feature.

Group 4: Regulatory information features

Transcriptional regulation involves in a complex and meticulous pattern of activities that incorporates with transcription factors (TFs; Rowell et al., 2014). A recent study indicated that TFs can selectively promote the expression of circular Cul2 rather than the host gene (Meng et al., 2018). circRNAs are regulated by TFs and other correlative information, such as H3K27ac signals. Yang et al. found N^6-methyladenosine boosts the efficient initiation of protein translation from circRNAs in human cells (Yang et al., 2017). We obtained the one-dimensional of TF feature vector, methylation feature vector, H3K27ac feature vector from TRCirc for each sequence, thereby leading to a 3-dimensional vector.

Model Training and Optimization

LightGBM

Gradient boosting decision tree (Friedman, 2001) is an iterative decision tree algorithm with various effective implementations such as XGBoost (Chen and Guestrin, 2016). However, the efficiency and scalability are still ungratified when feature dimension is high and data size is large (Ke et al., 2017). Recently, LightGBM (Ke et al., 2017) has been proposed to address this issue, which can effectively solve the time-consuming problem of conventional GBDT while retaining high classification ACC. LightGBM possesses two novel techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). Gradient-based one-side sampling excludes a significant proportion of data instances with small gradients and uses the remaining to estimate the information gain. Hence, this technique can effectively reduce the number of data at the time of calculation and further improve the efficiency. Exclusive feature bundling bundles mutually exclusive features to reduce the number of features. Features with larger gradients contribute more to the information gain and are thus more important for classification. Compared with GBDT, LightGBM speeds up the training process significantly because the number of bundled features will be much smaller than those of the original features. The speed of model training in LightGBM is 20 times faster than GBDT under the premise of achieving almost the same ACC (Ke et al., 2017). We employed the LightGBM algorithm using the lightgbm package in Python[4].

Support Vector Machine

Support vector machine is one of the most widely used machine learning algorithms for classification problems (Noble, 2006). The main idea of SVM is based on kernel functions that map

[4]https://github.com/Microsoft/LightGBM

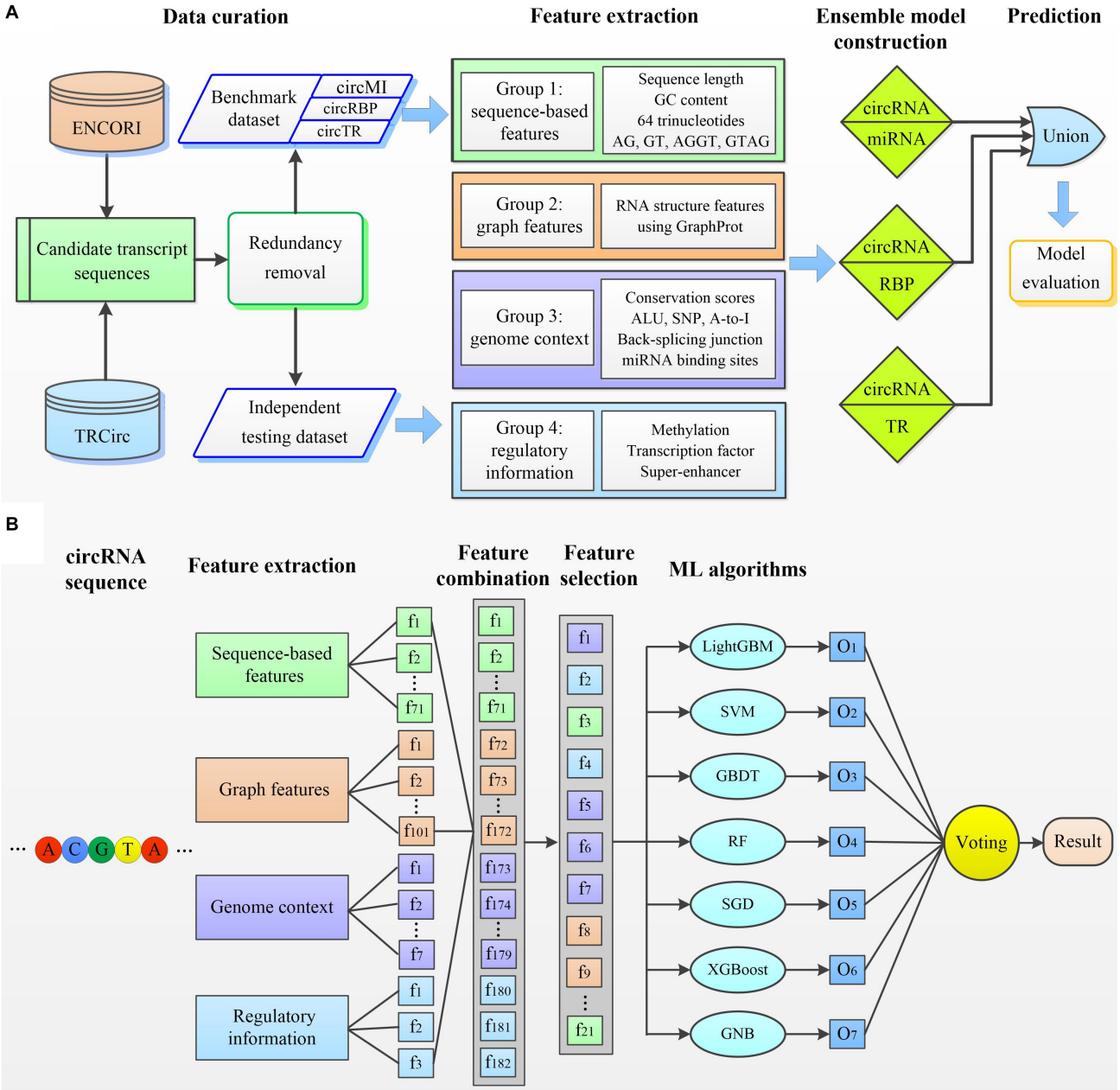

FIGURE 2 | The overall framework of circMRT. **(A)** An outline of the overall flowchart of circMRT. The development of circMRT involved four major steps: **(i)** data collection and preprocessing, **(ii)** feature extraction, **(iii)** ensemble model construction, and **(iv)** model prediction and performance assessment. **(B)** An illustration of the detailed procedures for constructing the circRNA-miRNA classifier. (i) Four groups of features including sequence composition, graph features, genome context and regulatory information are extracted for each candidate circRNA. (ii) The optimal feature set is selected by applying the proposed two-step feature selection strategy. **(iii)** Based on the optimal feature set, we train the prediction models using seven machine learning-based algorithms. **(iv)** Three individual model's outputs are integrated by using majority voting algorithm.

the input data into a high dimensional space. Support vector machine aims to search the hyperplane to maximize the margin between two support vectors. In this study, SVM with the "linear" kernel was implemented using the Scikit-learn library in Python. We optimized the parameter cost C from the choice of (1.0, 1.1, 1.2, 1.3, 1.4) by grid search. After optimization, the parameter of C was set as 1.0.

Random Forest

Random forest (Liaw and Wiener, 2002) is an ensemble learning method for regression and classification which involves multiple decision trees. Random forest assumes that there are P samples with Q features in the original training set, and it selects P samples from the training data by bootstrapping and randomly selects q features ($q \ll Q$) to train a decision tree. By repeating the step

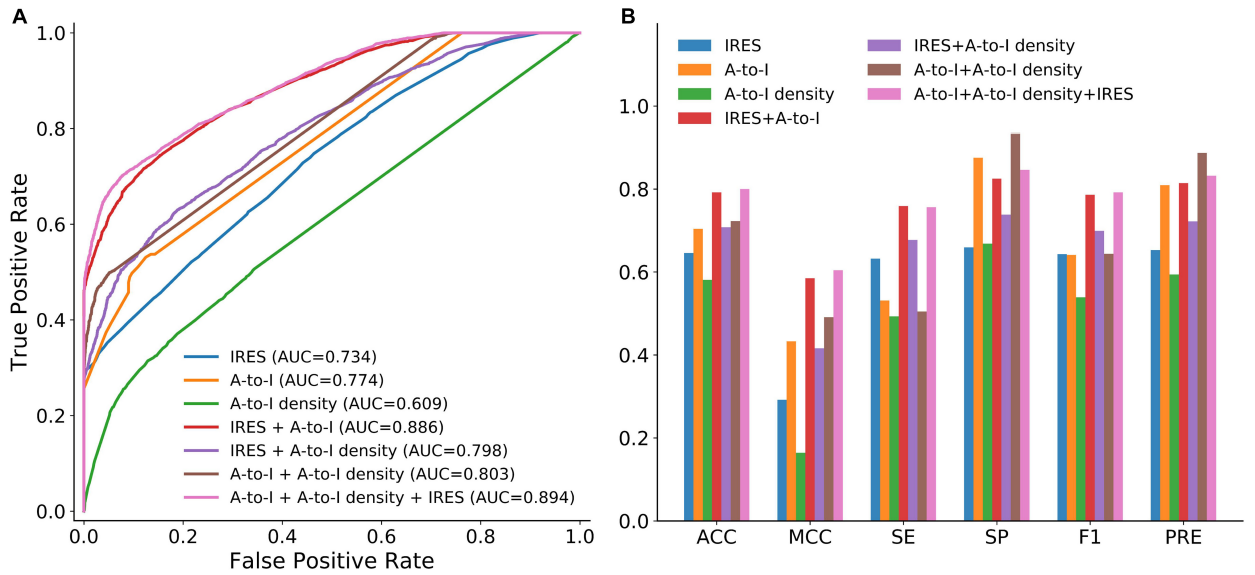

FIGURE 3 | Performance of circLGB for circRNA identification on the circlncRNA dataset by various combinations of sequence-derived features including IRES, A-to-I and A-to-I density. Panel **(A)** shows comparison of ROC curves and panel **(B)** shows comparison of ACC, MCC, SE, SP, F1,3] and PRE.

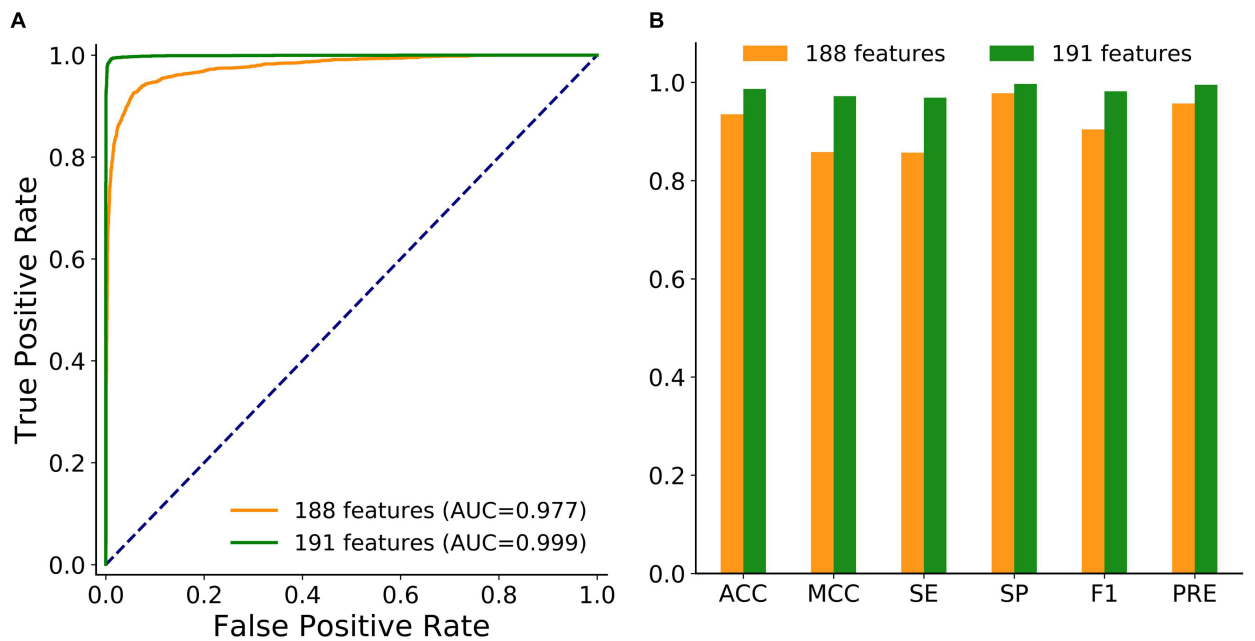

FIGURE 4 | (A) ROC curves and **(B)** histograms showing the performance of circLGB by extracting 188 and 191 sequence-derived features on the circlncRNA dataset under 10-time 5-fold cross-validation. The performance comparison in terms of ACC, MCC, SE, SP, F1, and PRE.

above, numerous decision trees are trained, and their outputs are integrated in the ensemble model to make a final prediction. We trained the RF with 20 decision trees using Scikit-learn.

Stochastic Gradient Descent

Stochastic gradient descent (SGD; Friedman, 2001) is an effective method for solving large scale supervised machine learning problems. It generally confers a significant decrease in training time without sacrificing ACC. In particular, SGD with early stopping at a fixed number of interactions approximately halves the training time. In this work, SGD was applied using Scikit-learn.

Gaussian Naive Bayes

A Naive Bayes (NB) classifier calculates the probability of a given example belonging to a certain class. When the

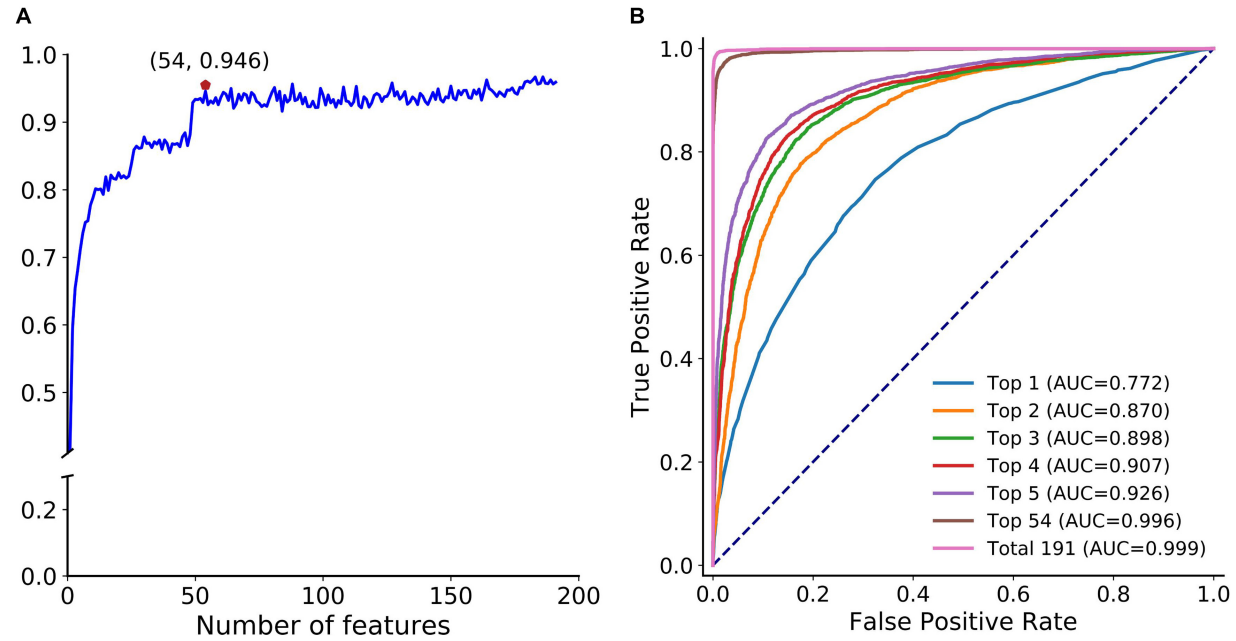

FIGURE 5 | Feature importance analyses. **(A)** SFS curve of MCC with increasing number of ranked features. The features are selected by mRMR feature importance list in descending order. X-axis represents the number of selected features. The maximum MCC (0.946) obtained by integrating the top 54 features on the curve is marked by a red pentagon. This notation also applies to **Figure 7A**. **(B)** ROC curves of circLGB for discriminating circRNAs and lncRNAs by using the top 1 to top 5, top 54 and total 191 features.

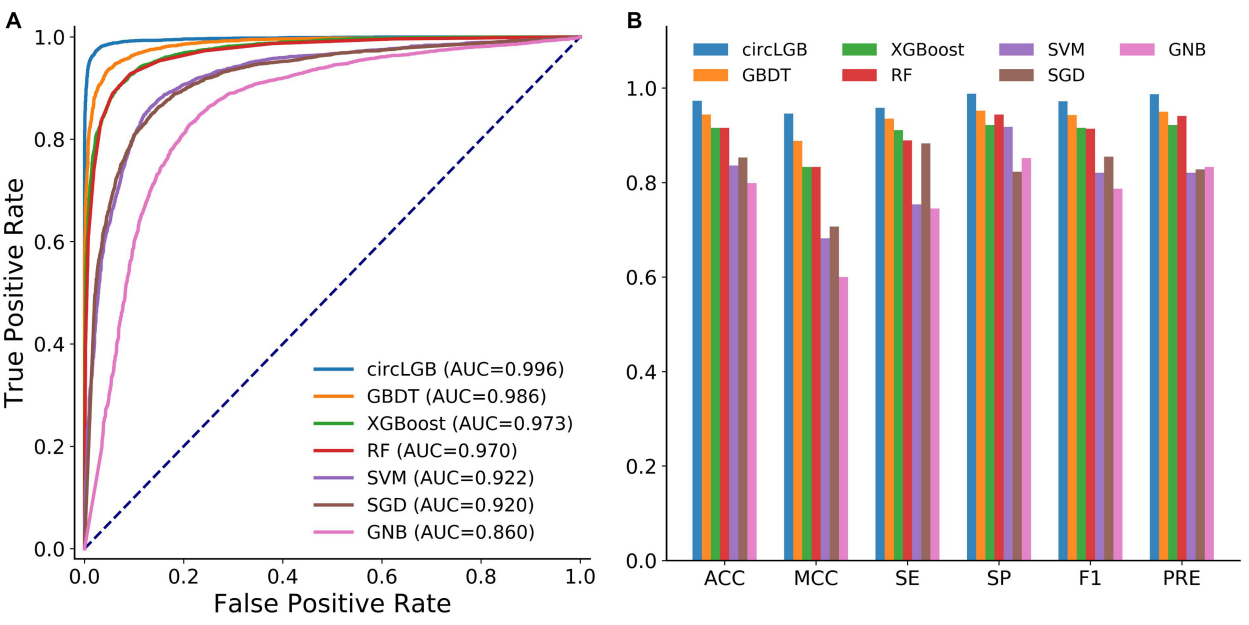

FIGURE 6 | (A) ROC curves and **(B)** histograms of evaluation metrics show the superior performance of circLGB over GBDT, XGBoost, RF, SVM, SGD, and GNB for circRNA identification on the circlncRNA dataset under 10-time 5-fold cross-validation. Evaluation metrics including ACC, MCC, SE, SP, F1, and PRE. GBDT, gradient boosting decision tree; RF, random forest; SVM, support vector machine; SGD, stochastic gradient descent; GNB, Gaussian naive Bayes.

likelihood of the features is assumed to be Gaussian, the NB classifier is called Gaussian naive Bayes (GNB; John and Langley, 2013). Gaussian naive Bayes supposes that features are independent from each other. Gaussian naive Bayes is simpler and faster than other sophisticated methods. Thus, it is usually used for prediction problems in bioinformatics (Murakami and Mizuguchi, 2010). Here, GNB was also implemented using Scikit-learn.

TABLE 2 | Performance evaluation of our circLGB and other two learning-based algorithms on the CIRCdeep dataset.

Model	ACC	MCC	F1	References
circLGB	**0.998**	**0.995**	**0.998**	–
circDeep	0.942	0.883	0.940	(Chaabane et al., 2019)
PredcircRNA	0.806	0.611	0.811	(Pan and Xiong, 2015)

*The best performance across different evaluation metrics is highlighted in bold for clarification. These highlights also apply to **Tables 3, 4**.*

TABLE 3 | Performance of three ensemble machine learning-based classifiers for circRNA regulatory interactions prediction based on different groups of sequence-derived features.

Features	ROC	SE	SP	ACC	MCC	F1	PRE
(A) circRNA-miRNA classifier							
Sequence-based	**0.995**	**1.000**	**0.991**	**0.995**	**0.990**	**0.995**	**0.990**
Graph features	0.958	0.980	0.937	0.957	0.915	0.956	0.933
Genome context	0.884	0.876	0.892	0.883	0.766	0.889	0.904
Regulation information	0.952	0.922	0.981	0.952	0.906	0.950	0.979
(B) circRNA-RBP classifier							
Sequence-based	**0.991**	**0.986**	**0.995**	**0.990**	**0.981**	**0.991**	**0.995**
Graph features	0.969	0.954	0.985	0.969	0.938	0.970	0.986
Genome context	0.895	0.833	0.956	0.894	0.794	0.888	0.951
Regulation information	0.955	0.925	0.985	0.954	0.910	0.954	0.985
(C) circRNA-TR classifier							
Sequence-based	**0.993**	**0.989**	**0.998**	**0.993**	**0.986**	**0.993**	**0.998**
Graph features	0.961	0.935	0.987	0.962	0.925	0.959	0.985
Genome context	0.891	0.954	0.829	0.891	0.788	0.896	0.846
Regulation information	0.971	0.978	0.965	0.971	0.943	0.970	0.962

circLGB

We proposed a machine-learning framework called circLGB to classify circRNA from lncRNAs. As shown in **Figure 1**, the major procedures of circLGB can be summarized as below: (i) The collected human circRNAs and lncRNAs transcripts are combined to construct the circlncRNA dataset. (ii) Four groups of sequence-derived features are extracted from various toolkits and databases. (iii) minimum redundancy-maximum relevance (mRMR; Ding and Peng, 2005) feature selection framework is applied to rank the extracted features according their importance scores. Then, sequential forward search (SFS) is utilized to determine the optimal feature subset which yields the best Matthews correlation coefficient (MCC). **Supplementary Table S3** summarizes the feature importance scores on the circlncRNA dataset. (iv) The resulting feature vector is fed into the LightGBM classifier for circRNA identification. Finally, performance metrics are calculated for model evaluation.

circMRT

We next developed circMRT to predict the regulatory information for circRNAs, including their interactions with miRNA, RBP, and TR. Note that, one interaction may exist simultaneously for a given circRNA. We first developed three binary classifiers to explore whether the given circRNA has associations with miRNA, RBP, and TR, respectively. Then, the outputs of these classifiers were fused to make a final prediction. The circMRT methodology (**Figure 2**) consists of four major steps: (i) Datasets circMI, circRBP and circTR are constructed to train the circRNA–miRNA, circRNA–RBP and circRNA–TR classifiers, respectively. Besides, independent test set is generated to evaluate the generalization of circMRT. (ii) The candidate circRNA sequence is input for feature encoding by extracting four types of features. (iii) The extracted features are fed into the abovementioned classifiers for training and testing. Each classifier is trained on its own optimal features selected by applying the proposed feature optimization strategy. (iv) The independent test set is respectively fed into three well-trained classifiers for prediction. Finally, the outputs are fused by a union operator to predict the regulatory interactions for a given circRNA.

Feature Selection

We utilized a two-step feature selection strategy to improve the feature representation ability. We first used mRMR to achieve the ranked feature list according to the importance scores of the learned features. Features with higher scores were more predictive. Second, SFS was applied to investigate the optimal combination of features that can yield the best performance. We ranked the features in a descending order from the mRMR features list. Subsequently, incremental feature selection approach was employed to select the optimal top-k features. We added the features from the ranked feature list one by one and trained the proposed model. The feature subset with the relative higher values of MCC was regarded as the most discriminative features. It is worth noting that we here used the MCC since it is a balanced measurement, even if the sizes of positive and negative samples are imbalanced. Therefore, the MCC is a better indicator to assess the performance of the models.

Hyperparameters Optimization

cricLGB and circMRT were implemented using Python 2.7. All experiments were carried out on a desktop computer with Intel (R) Core (TM) i7-7800X CPU @ 3.50GHz, Ubuntu 16.04.5 LTS and 16 GB RAM. To ensure the ACC and robustness of the proposed algorithms, we employed the grid-search parameter adjustment to achieve the optimal parameters. Specifically, we used grid-search to tune six parameters including learning rate, number of leaves, feature fraction, bagging fraction, reg_alpha, and reg_lambda for each dataset. The Grid search range of each parameter was as below: learning rate from the choice (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2), number of leaves from the choice (20, 25, 30, 35, 40, 45, 50), feature fraction from

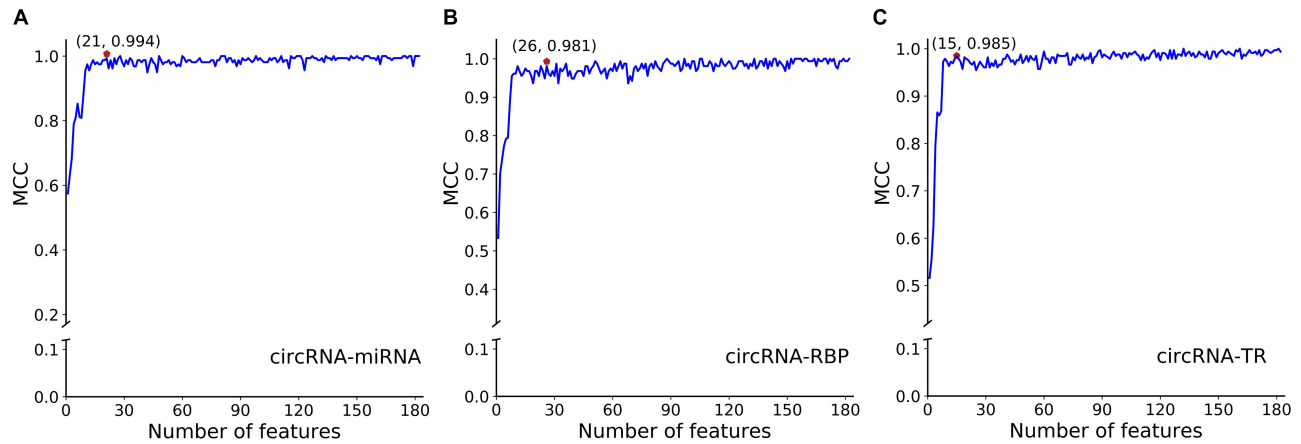

FIGURE 7 | SFS curves of MCC with increasing number of selected features for classifiers of panel **(A)** circRNA-miRNA, **(B)** circRNA-RBP, and **(C)** circRNA-TR, respectively. The features are selected by their estimated feature importance in descending order. We choose the top 21, top 26, and top 15 features for the above three classifiers, with MCC values of 0.994, 0.981, and 0.985, respectively.

TABLE 4 | Performance evaluation of three binary classifiers on datasets circMI, circRBP, and circTR, respectively.

Classifier	Dataset	ROC	SE	SP	ACC	MCC	F1	PRE
circRNA-miRNA	circMI	0.981	0.986	0.976	0.981	**0.994**	0.981	0.977
circRNA-RBP	circRBP	0.990	**0.991**	0.990	0.990	0.981	0.991	0.991
circRNA-TR	circTR	**0.992**	0.988	**0.997**	**0.992**	0.985	**0.992**	**0.997**

the choice (0.5, 0.6, 0.7, 0.8, 0.9), bagging fraction from the choice (0.5, 0.6, 0.7, 0.8, 0.9), reg_alpha from the choice (0.001, 0.01, 0.03, 0.05), and reg_lambda from the choice of (0.001, 0.01, 0.03, 0.05). The proposed methods are binary classification problems, we used "binary" of "objective" and "auc" of 'metric' with 100 times iteration and "stopping patience" of 10.

Considering that the grid-search for all the parameters requires a large computation cost, we adjusted the above parameters in batches to maximize the value of AUC under 5-fold cross-validation. We took the optimal hyperparameters for the model once the performance does not improve. The tuned optimal parameters were regarded as the input parameters to tune the next parameters. We first tuned the parameters of learning rate and number of leaves. Then, we adjusted the feature fraction and bagging fraction. Next, we tuned the regularization parameter including alpha and lambda. The combination of the optimal parameters for circLGB from the learning rate of 0.1 were, number of leaves of 60, feature faction of 0.5, bagging fraction of 0.6, reg_alpha of 0.01 and reg_lambda of 0.001. The determination of the optimal parameters of circRNA–miRNA, circRNA–RBP and circRNA-TR classifiers were as follows: learning rate of (0.1, 0.1, 0.1), number of leaves of (20, 40, 35), feature faction of (0.6, 0.6, 0.6), bagging fraction of (0.6, 0.6, 0.5), reg_alpha of (0.01, 0.001, 0.01), and reg_lambda of (0.03, 0.001, 0.01).

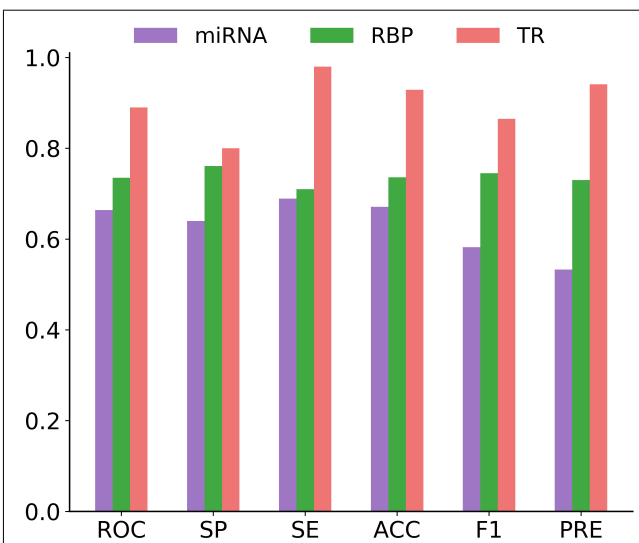

FIGURE 8 | Histograms showing the performance of circRNA-miRNA, circRNA-RBP and circRNA-TR classifiers in terms of ROC, SP, SE, ACC, F1, and PRE on the independent testing dataset.

Performance Evaluation

To evaluate the performance of our models and to compare with existing state-of-the-art methods, sensitivity (SE), specificity (SP), precision (PRE), F1 score (F1), ACC, and MCC were calculated. These indicators are widely used to measure the quality of binary classification defined as follows:

$$\mathrm{SE} = \frac{TP}{TP + FN} \tag{1}$$

$$\mathrm{SP} = \frac{TN}{TP + FP} \tag{2}$$

$$\mathrm{PRE} = \frac{TP}{TP + FP} \tag{3}$$

$$\mathrm{F_1} = 2 \times \frac{SN \times PRE}{SN + FRE} \tag{4}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$\mathrm{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{6}$$

where TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively. Receiver Operating Characteristic (ROC) curves were employed to visualize the performance between different methods together with the area under ROC curve (AUC).

RESULTS

circLGB for circRNA Identification

The Effect of Three New Sequence-Derived Features

We first examined whether A-to-I, A-to-I density or IRES could be used as effective features for circRNA identification. To this end, we trained circLGB with these features on the circlncRNA dataset under 10-time 5-fold cross-validation. As shown in **Figure 3**, two observations can be made: (i) circLGB trained using IRES achieved the highest SE value of 0.632. (ii) circLGB trained with A-to-I achieved more optimal performance than those using A-to-I density or IRES. These results indicated that no single feature contains enough useful patterns and characteristics for classifying circRNAs.

To achieve better performance, the combination of these three new features was modeled in circLGB. As depicted in **Figure 3**, IRES combined with A-to-I outperformed any other combinations of two features, reaching AUC value of 0.886. Interestingly, though A-to-I density alone showed relatively poor performance, it gained great progress by incorporating with A-to-I or IRES, reaching AUC values of 0.803 and 0.798, respectively. circLGB achieved an overall AUC of 0.894 using these three features. So, we added them with commonly used features to train our model. As expected, circLGB trained using 191 features, achieved better performance than that on 188 features, reaching AUC values of 0.999 and 0.977, respectively (**Figure 4A**). Similar results on other evaluation metrics can be found in **Figure 4B**. Together, the addition of three new features can boost the prediction ability of circLGB.

Feature Importance Analysis for circLGB

Next, we adopted the proposed optimization strategy to enhance the feature representation ability. **Figure 5A** depicts the SFS curve of MCC of circLGB on the circlncRNA dataset by adding features one by one from the ranked feature list (**Supplementary Table S3**). Apparently, it increased quickly as the features were integrated. The MCC reached a relatively high value of 0.946 when adding the top 54 features. However, the performance fluctuated when incorporating more features. This implied that the improvement of the low-ranked features is not obvious, and they even lead to a decline of the performance. Moreover, we compared the performance of circLGB using the top 1 to top 5, top 54 and all features under 10-time 5-fold cross-validation. Obviously, the performance of circLGB trained on the selected feature sets improved when gradually adding the top ranked features (**Figure 5B**). The predictive results using the optimal features showed comparable performance with those using 191 features, reaching ROC values of 0.996 and 0.999, respectively. Therefore, these 54 features were regarded as the optimal features.

Supplementary Figure S2 illustrates the feature importance distribution of the optimal features based on the importance scores. There were 30 graph features, 10 sequence-based features, 9 conservation scores, and 5 ATOSAAI features (ATOS, A-to-I, A-to-I density and IRES) amongst them. This result was consistent with a recent study that shows that graph features are the most predictive features for circRNA detection (Chen et al., 2018). We noted that A-to-I density, A-to-I, and IRES features, respectively, were ranked in the 3rd, 26th, and 49th place, which verified their superior ability in identifying circRNA.

Comparison With Learning-Based Methods

We compared the performance of circLGB with six machine learning algorithms including GBDT, XGBoost, RF, SVM, SGD, and GNB on the circlncRNA dataset using the optimal features under 10-time 5-fold cross-validation. All the machine learning methods were run under their optimal parameters for fair comparisons. As shown in **Figure 6A**, of the seven algorithms tested here, circLGB was the most predictive, with ROC of 0.996. Moreover, circLGB outperformed others with remarkable ACC, MCC, SE, SP, F1, and PRE values of 0.973, 0.946, 0.958, 0.988, 0.972, and 0.987, respectively (**Figure 6B**). Furthermore, we compared circLGB with two state-of-the-art predictors (e.g., circDeep and PredcircRNA) on the CIRCdeep dataset. For fair comparison, we randomly separated the dataset into a training dataset, a validation dataset, and an independent testing set with 75, 10, and 15%, respectively. Overall, circLGB achieved the most powerful predictive ability, with ACC, MCC, and F1 values of 0.998, 0.995, and 0.998, respectively (**Table 2**).

circMRT for Predicting circRNA Regulatory Interactions

Feature Importance Analysis for circMRT

We first compared the performance of the proposed classifiers. As indicated in **Table 3**, sequence-based features were more important than other groups of features for each classifier. The regulation information features had strong discriminating power for predicting circRNA–TR interactions. **Supplementary Tables S4–S6** present the ranked feature list of circRNA–miRNA, circRNA–RBP, and circRNA–TR classifiers on datasets circMI, circRBP, and circTR, respectively. Some interesting conclusions can be drawn: (i) The predicted results of circRNA-miRNA interactions were strongly influenced by ALU and miRNA. (ii)

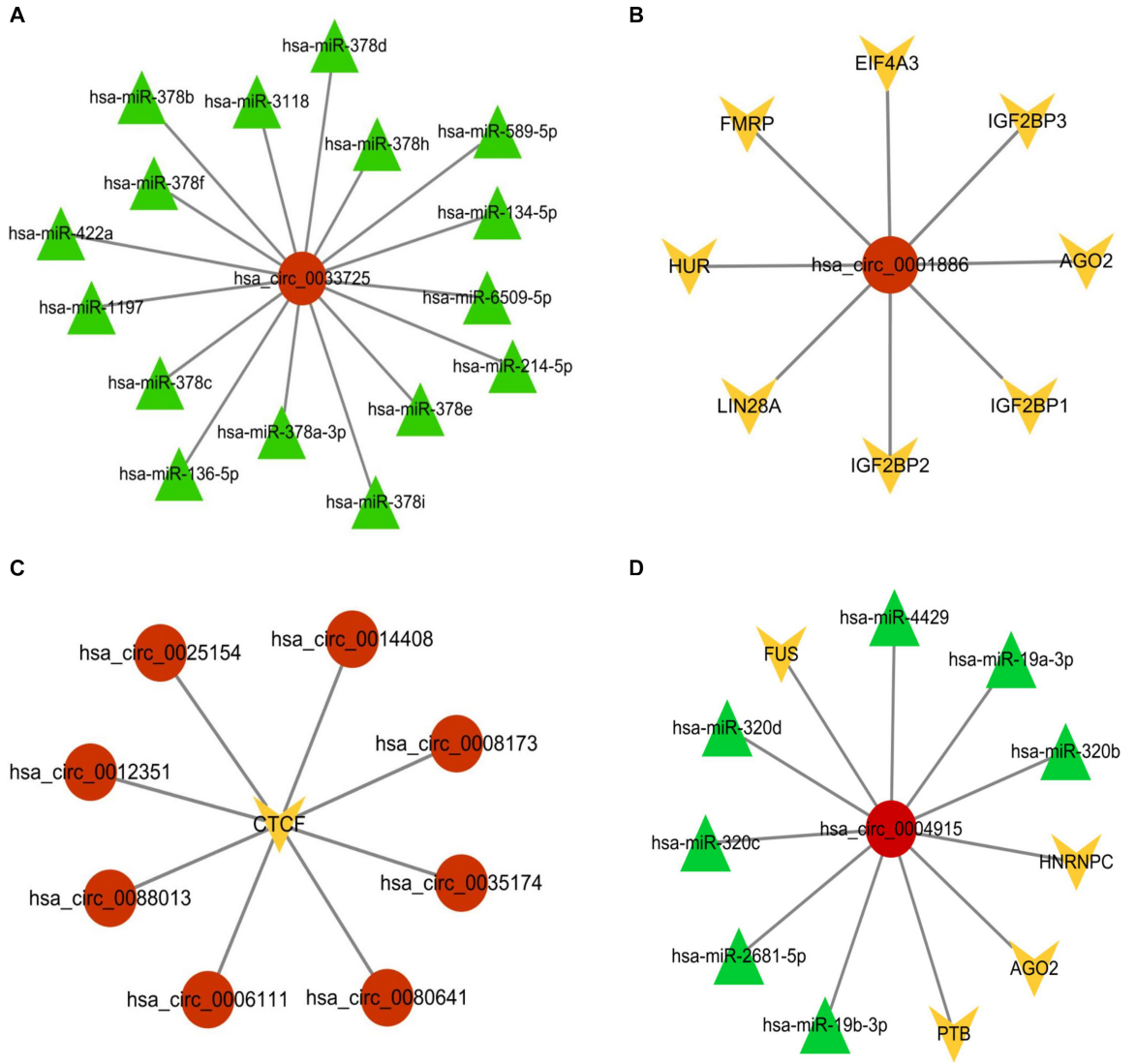

FIGURE 9 | Visualization of the circRNA-associated interactions according to databases ENCORI, TRCirc and Interactome. **(A)** Visualization of the has_circ_0033725 associated with miRNAs interactions. **(B)** Visualization of the has_circ_001886 associated with RBPs interactions. **(C)** Visualization of has_circ_0006111, has_circ_0008173, has_circ_0012351, has_circ_0014408, has_circ_0025154, has_circ_0035174, has_circ_0080641, and has_circ_0088103 associated with TR interactions. **(D)** Visualization of has_circ_0004915 associated with miRNAs and RBPs interactions.

Junction and repeat features contributed most for the circRNA–RBP classifier. (iii) Junction and methylation were the most predictive for the circRNA–TR classifier. (iv) Conservation scores ranked in the top seven features of all three classifiers. Therefore, conservation information was very predictive to distinguish circRNA regulatory interactions.

To avoid overfitting, we performed the proposed feature optimization strategy to obtain the representative features for each classifier. **Figure 7** depicts the MCC curves of these classifiers by gradually integrating features from the ranked feature list. It can be observed that the maximum MCC values of circRNA–miRNA, circRNA–RBP, and circRNA–TR classifiers were 0.994, 0.981, and 0.985 (**Table 4**) when the top 21, 26, and 15 features from their own ranked feature list were used. Therefore, circRNA associated with miRNA, RBP, and TR were predicted using the proposed classifiers with their own optimal features.

Performance Evaluation of circMRT on the Independent Test Set

In this section, we focused on measuring the generalizability of circMRT for unseen data. For this purpose, we evaluated the performance of circMRT on the constructed independent test set. This dataset was split into 60, 20, and 20% classes, subsequently being used as the training set, validation set, and testing set, respectively. As depicted in **Figure 8**, circRNA-TR classifier exhibited the best predictive power, with the maximal ROC and ACC values of 0.890 and 0.929. The circRNA-RBP classifier was the second most predictive, with ROC and ACC of

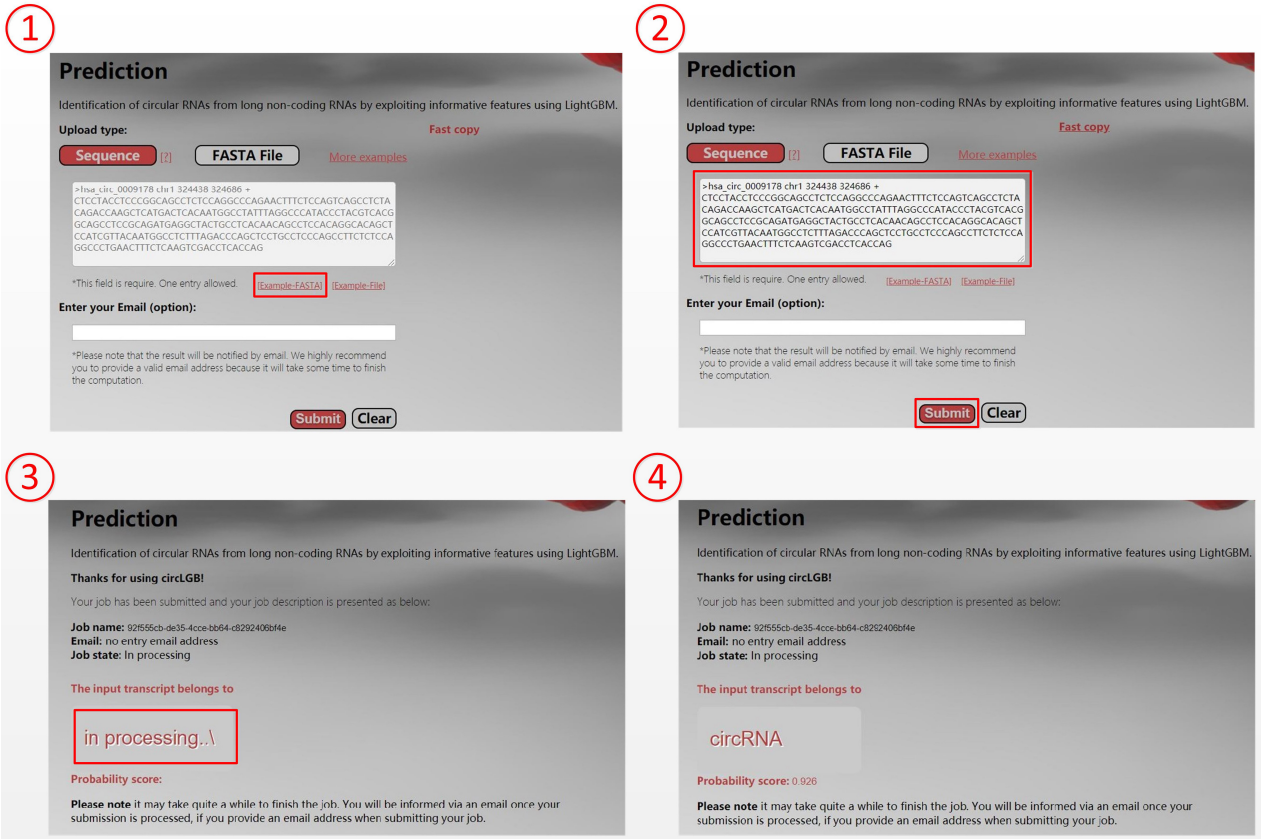

FIGURE 10 | Screenshots show the prediction of hsa_circ_0009178. Users can submit query sequence in the following four steps: **(1)** Click the START button to be taken to the prediction page. **(2)** Input the hsa_circ_0009178 sequence into the input box and then click the Submit button to make prediction. **(3)** Click the Submit button to upload hsa_circ_0009178 sequence. **(4)** The web server returns the prediction result "circRNA"ĺ with a probability score of 0.926 in the Gray box.

0.735 and 0.736, respectively. The circRNA-miRNA classifier also performed well, but with relatively lower ROC and ACC.

Taking has_circ_0033725 as an example, our circMRT predicted that it has interactions with miRNAs. According to the ENCORI database, has_circ_0033725 has interactions with 16 miRNAs (**Figure 9a**). circMRT predicted that has_circ_001886 has association with RBP. Databases ENCORI and Interactome[5] shows that has_circ_001886 has interactions with AGO2, EIF4A3, FMRP, HUR, IGF2BP1, IGF2BP2, IGF2BP3, and LIN28A (**Figure 9b**). circMRT suggested that circRNAs has_circ_0006111, has_circ_0008173, has_circ_0012351, has_circ_0014408, has_circ_0025154, has_circ_0035174, has_circ_0080641, and has_circ_0088103 have associations with TR. According to the TRCirc database, all the above circRNAs have interactions with CTCF (**Figure 9c**). Moreover, has_circ_0004915 was predicted to have interactions with miRNA and RBP. From ENCORI, has_circ_0004915 has interactions with AGO2, FUS, HNRNPC, PTB, has_miR_19b-3p, has_miR_19a-3p, has_miR_2681-5p, has_miR_320c, has_miR_320b, has_miR_320d, and has_miR_4429 (**Figure 9d**). More details can be found in the **Supplementary Material**.

[5]https://circinteractome.nia.nih.gov/

Availability of Online Webserver

For the convenience of researchers, we have developed an easy-to-use webserver that implements our circLGB, which is freely accessible through http://www.circlgb.com. The following description provides a step-by-step instruction on how to use the webserver to obtain the prediction result. First, users need to submit the query sequence into the input box or upload a FASTA sequence file to make a prediction. Note that the input sequence must only contain the following four canonical bases "A," "C," "G," and "T." The FASTA formatted sequence begins with a single line description, followed by lines of sequence data. The definition line is distinguished from the sequence data by a greater-than ">" character at the beginning. The rest of the definition line must contain five columns including sequence name, chromosome, start position, end position, and strand. Second, click the Submit button to upload the query sequence (FASTA file) for prediction. Upon submitting the sequence, the software will extract the features for the given sequence from a server. The Prediction page will show the job description including job ID, job name, email address, and job state. The web server will return the prediction result in the Gray box when the job is completed. **Figure 10** shows an example for using the web server.

DISCUSSION

Here we present two machine learning-based methods, circLGB and circMRT, to classify circRNA from other lncRNAs and to predict its regulatory interactions using diverse sources of sequence-derived features, respectively. The feature section is important, in addition to the modeling approach for predicting activity. In recent years, considerable research efforts have been made in identifying circRNA, thus generating several groups of features for RNAs representation. Inspired by these studies (Pan and Xiong, 2015; Chen et al., 2018), we integrated the commonly used sequence features to generate the feature space of circLGB. To achieve optimal performance, A-to-I and A-to-I density, and IRES features were modeled in the circLGB model. The success of circLGB lies in the enriched representative features and powerful machine learning model incorporating the feature optimization strategy. Compared to existing tools, circLGB has the following merits: (i) It successfully integrates three new features that can enhance the discrimination ability for circRNA detection. (ii) It takes advantage of the feature optimization strategy to determine the most important features, thus reducing the feature dimensions and avoiding overfitting. (iii) circLGB provides a user-friendly webserver to identify circRNA for a new query RNA sequence.

Many studies focus on the interactions between circRNAs and miRNAs (e.g., TargetScan, miRanda), RBPs (e.g., ENCORI), and TRs (e.g., TRCirc). However, there is a lack of a comprehensive human circRNA regulatory information database. circMRT is an efficient computational ensemble machine learning model for simultaneous prediction of circRNA potential interacted miRNAs, RBPs, and TRs, further facilitating interpretation and its functional mechanisms. circMRT incorporates several features from other freely available web resources and toolkits, such as UCSC, TRCirc, and GraphProt. It enables the user to find the potential regulatory interactions for an unseen circRNA sequence. Together, circMRT will accelerate our efforts to understand the roles of circRNAs in biological processes related to health and disease.

Several future improvements are expected. First, we have currently designed circLGB and circMRT only for human circRNAs. They will be expanded to include other species in the future. Second, manual design of proper RNA sequence features will definitely enhance the prediction ability of models. Here, we use the commonly used sequence-derived features as well as explore three new features for RNA sequence representations and show that feature engineering really boosts the performance. Future directions can combine feature engineering and feature selection strategies for improving the prediction performance. Third, the number of the available training sample sizes have great influence on the predictive performance. However, after removing the duplications, the sample sizes of circRNA–miRNA, circRNA–RBP, and circRNA–TR interactions are relatively small, which brings a challenge for an unseen query sequence. Consequently, appropriate data augmentation techniques await exploration. Finally, though circLGB and circMRT achieve the desired performance for circRNA identification and prediction of its regulatory interactions, both of them rely heavily on the considerable domain expertise to design the feature extractor. We believe that simple and modern deep learning models will contribute to enhancements for these issues.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/Peppags/circLGB-circMRT.

AUTHOR CONTRIBUTIONS

GZ and BY wrote the analysis source code. GZ analyzed the data and drafted the full manuscript. YD and QL collected and compiled the data from the literature and public database. ZD developed the data analysis and participated in discussion of the project. YC and XD critically revised the final manuscript. All authors contributed to the project design and read and approved the final manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (NSFC) (Grant No. 61872396), Natural Science Foundation (NSF) of Guangdong Province (Grant No. 2014A030308014), and also by 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (Project No. 2020LKSFG06C).

ACKNOWLEDGMENTS

We thank Dr. Baoxian Yu for useful discussions and proofreading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00655/full#supplementary-material

REFERENCES

Abe, N., Matsumoto, K., Nishihara, M., Nakano, Y., Shibata, A., Maruyama, H., et al. (2015). Rolling Circle Translation of Circular RNA in Living Human Cells. *Sci. Rep.* 5:16435. doi: 10.1038/srep16435

Ahmed, F., Ansari, H. R., and Raghava, G. P. (2009). Prediction of guide strand of microRNAs from its sequence and secondary structure. *BMC Bioinformatics* 10:105. doi: 10.1186/1471-2105-10-105

Ahmed, F., Kaundal, R., and Raghava, G. P. (2013). PHDcleav: a SVM based method for predicting human Dicer cleavage sites using sequence and

secondary structure of miRNA precursors. *BMC Bioinformatics* 14(Suppl. 14):S9. doi: 10.1186/1471-2105-14-s14-s9

Awasthi, R., Singh, A. K., Mishra, G., Maurya, A., Chellappan, D. K., Gupta, G., et al. (2018). An Overview of Circular RNAs. *Adv. Exp. Med. Biol.* 1087, 3–14. doi: 10.1007/978-981-13-1426-1_1

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

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Boutros, M., and Ahringer, J. (2008). The art and design of genetic screens: RNA interference. *Nat. Rev. Genet.* 9, 554–566. doi: 10.1038/nrg2364

Buratti, E., and Baralle, F. E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell. Biol.* 24, 10505–10514. doi: 10.1128/mcb.24.24.10505-10514.2004

Chaabane, M., Williams, R. M., Stephens, A. T., and Park, J. W. (2019). circDeep: deep learning approach for circular RNA classification from other long non-coding RNA. *Bioinformatics* 36, 73–80. doi: 10.1093/bioinformatics/btz537

Chen, L., Zhang, Y. H., Huang, G., Pan, X., Wang, S., Huang, T., et al. (2018). Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol. Genet. Genomics* 293, 137–149. doi: 10.1007/s00438-017-1372-1377

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, (New York, NY: ACM), 785–794.

Chen, X., Han, P., Zhou, T., Guo, X., Song, X., and Li, Y. (2016). circRNADb: a comprehensive database for human circular RNAs with protein-coding annotations. *Sci. Rep.* 6:34985. doi: 10.1038/srep34985

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

Cuesta, J. A., and Manrubia, S. (2017). Enumerating secondary structures and structural moieties for circular RNAs. *J. Theor. Biol.* 419, 375–382. doi: 10.1016/j.jtbi.2017.02.024

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/s0219720005001004

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505, 696–700. doi: 10.1038/nature12756

Dong, R., Ma, X. K., Li, G. W., and Yang, L. (2018). CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison. *Genom. Proteom. Bioinform.* 16, 226–233. doi: 10.1016/j.gpb.2018.08.001

Dudekula, D. B., Panda, A. C., Grammatikakis, I., De, S., Abdelmohsen, K., and Gorospe, M. (2016). CircInteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs. *RNA Biol.* 13, 34–42. doi: 10.1080/15476286.2015.1128065

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232.

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16:4. doi: 10.1186/s13059-014-0571-573

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

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111

Ilik, I. A., Aktas, T., Maticzka, D., Backofen, R., and Akhtar, A. (2020). FLASH: ultra-fast protocol to identify RNA-protein interactions in cells. *Nucleic Acids Res.* 48:e15. doi: 10.1093/nar/gkz1141

Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H. T., Orejuela, M. R., et al. (2015). Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.* 10, 170–177. doi: 10.1016/j.celrep.2014.12.019

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19, 141–157. doi: 10.1261/rna.035667.112

John, G. H., and Langley, P. (2013). Estimating continuous distributions in bayesian classifiers. *arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1302.4964 (accessed February 20, 2013).

Ju, Y., Yuan, L., Yang, Y., and Zhao, H. (2019). CircSLNN: identifying RBP-binding sites on circRNAs via sequence labeling neural networks. *Front. Genet.* 10:1184. doi: 10.3389/fgene.2019.01184

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., et al. (2003). The UCSC genome browser database. *Nucleic Acids Res.* 31, 51–54. doi: 10.1093/nar/gkg129

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "Lightgbm: a highly efficient gradient boosting decision tree," in *Proceedings of the Advances in Neural Information Processing Systems*, (Cambridge, MA: The MIT Press), 3146–3154.

Kitamura-Abe, S., Itoh, H., Washio, T., Tsutsumi, A., and Tomita, M. (2004). Characterization of the splice sites in GT-AG and GC-AG introns in higher eukaryotes using full-length cDNAs. *J. Bioinform. Comput. Biol.* 2, 309–331. doi: 10.1142/s0219720004000570

Kuehn, B. M. (2008). 1000 Genomes Project promises closer look at variation in human genome. *JAMA* 300:2715. doi: 10.1001/jama.2008.823

Lasda, E., and Parker, R. (2014). Circular RNAs: diversity of form and function. *RNA* 20, 1829–1842. doi: 10.1261/rna.047126.114

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

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Li, X., Yang, L., and Chen, L. L. (2018). The biogenesis, functions, and challenges of circular RNAs. *Mol. Cell.* 71, 428–442. doi: 10.1016/j.molcel.2018.06.034

Liang, D., and Wilusz, J. E. (2014). Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.* 28, 2233–2247. doi: 10.1101/gad.251926.114

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R. News* 2, 18–22.

Liu, Y. C., Li, J. R., Sun, C. H., Andrews, E., Chao, R. F., Lin, F. M., et al. (2016). CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.* 44, D209–D215. doi: 10.1093/nar/gkv940

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., et al. (2011). Three periods of regulatory innovation during vertebrate evolution. *Science* 333, 1019–1024. doi: 10.1126/science.1202702

Marques, A. C., and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* 10:R124. doi: 10.1186/gb-2009-10-11-r124

Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* 15:R17. doi: 10.1186/gb-2014-15-1-r17

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

Meng, J., Chen, S., Han, J. X., Qian, B., Wang, X. R., Zhong, W. L., et al. (2018). Twist1 Regulates Vimentin through Cul2 Circular RNA to Promote EMT in hepatocellular carcinoma. *Cancer Res.* 78, 4150–4162. doi: 10.1158/0008-5472.CAN-17-3009

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 12:489. doi: 10.1186/1471-2105-12-489

Murakami, Y., and Mizuguchi, K. (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 26, 1841–1848. doi: 10.1093/bioinformatics/btq302

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi: 10.1038/nbt1206-1565

Pan, X., and Xiong, K. (2015). PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. *Mol. Biosyst.* 11, 2219–2226. doi: 10.1039/c5mb00214a

Pan, X., Xiong, K., Anthon, C., Hyttel, P., Freude, K., Jensen, L., et al. (2018). WebCircRNA: classifying the Circular RNA Potential of Coding and Noncoding RNA. *Genes* 9:536. doi: 10.3390/genes9110536

Panda, A. C. (2018). Circular RNAs Act as miRNA Sponges. *Adv. Exp. Med. Biol.* 1087, 67–79. doi: 10.1007/978-981-13-1426-1_6

Pervouchine, D. D. (2019). Circular exonic RNAs: when RNA structure meets topology. *Biochim. Biophys. Acta Gene Regul. Mech.* 1862:194384. doi: 10.1016/j.bbagrm.2019.05.002

Ramaswami, G., and Li, J. B. (2014). RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* 42, D109–D113. doi: 10.1093/nar/gkt996

Rowell, J., Koitabashi, N., Kass, D. A., and Barth, A. S. (2014). Dynamic gene expression patterns in animal models of early and late heart failure reveal biphasic-bidirectional transcriptional activation of signaling pathways. *Physiol. Genomics* 46, 779–787. doi: 10.1152/physiolgenomics.00054.2014

Shalem, O., Sanjana, N. E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* 16, 299–311. doi: 10.1038/nrg3899

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* 22, 500–503. doi: 10.1093/bioinformatics/btk010

Tang, Z., Li, X., Zhao, J., Qian, F., Feng, C., Li, Y., et al. (2018). TRCirc: a resource for transcriptional regulation information of circRNAs. *Brief Bioinform.* 20, 2327–2333. doi: 10.1093/bib/bby083

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147, 1537–1550. doi: 10.1016/j.cell.2011.11.055

Volders, P. J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41, D246–D251. doi: 10.1093/nar/gks915

Wang, L., You, Z. H., Huang, Y. A., Huang, D. S., and Chan, K. C. C. (2019). An Efficient Approach based on Multi-sources Information to Predict CircRNA-disease Associations Using Deep Convoltional Neural Network. *Bioinformatics* 36, 4038–4046. doi: 10.1093/bioinformatics/btz825

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

Wang, Y., and Wang, Z. (2015). Efficient backsplicing produces translatable circular mRNAs. *RNA* 21, 172–179. doi: 10.1261/rna.048272.114

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806. doi: 10.1126/science.1254806

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, K., Pan, X., Yang, Y., and Shen, H. B. (2019). CRIP: predicting circRNA-RBP interaction sites using a codon-based encoding and hybrid deep neural networks. *RNA* 25, 1604–1615. doi: 10.1261/rna.070565.119

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

Zhang, Z., Yang, T., and Xiao, J. (2018). Circular RNAs: promising Biomarkers for Human Diseases. *EBioMedicine* 34, 267–274. doi: 10.1016/j.ebiom.2018.07.036

Zhao, J., Wu, J., Xu, T., Yang, Q., He, J., and Song, X. (2018). IRESfinder: identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features. *J. Genet. Genom.* 45:403. doi: 10.1016/j.jgg.2018.07.006

Zheng, L. L., Li, J. H., Wu, J., Sun, W. J., Liu, S., Wang, Z. L., et al. (2016). deepBase v2.0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Res.* 44, D196–D202. doi: 10.1093/nar/gkv1273

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: 22 July 2020
doi: 10.3389/fgene.2020.00784

Nearest-Neighbor Projected Distance Regression for Epistasis Detection in GWAS With Population Structure Correction

*Marziyeh Arabnejad[1], Courtney G. Montgomery[2], Patrick M. Gaffney[2] and Brett A. McKinney[1,3]**

[1] Tandy School of Computer Science, University of Tulsa, Tulsa, OK, United States, [2] Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, United States, [3] Department of Mathematics, University of Tulsa, Tulsa, OK, United States

Nearest-neighbor Projected-Distance Regression (NPDR) is a feature selection technique that uses nearest-neighbors in high dimensional data to detect complex multivariate effects including epistasis. NPDR uses a regression formalism that allows statistical significance testing and efficient control for multiple testing. In addition, the regression formalism provides a mechanism for NPDR to adjust for population structure, which we apply to a GWAS of systemic lupus erythematosus (SLE). We also test NPDR on benchmark simulated genetic variant data with epistatic effects, main effects, imbalanced data for case-control design and continuous outcomes. NPDR identifies potential interactions in an epistasis network that influences the SLE disorder.

Keywords: epistasis, feature selection, GWAS, machine learning, nearest-neighbors

OPEN ACCESS

Edited by:
Christian Darabos, Dartmouth College, United States

Reviewed by:
Kira Vyatkina, Saint Petersburg Academic University (RAS), Russia
Marta E. Alarcon-Riquelme, Junta de Andalucía de Genómica e Investigación Oncológica (GENYO), Spain

***Correspondence:**
Brett A. McKinney
brett-mckinnney@utulsa.edu; brett.mckinney@gmail.com

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

Received: *29 January 2020*
Accepted: *01 July 2020*
Published: *22 July 2020*

Citation:
Arabnejad M, Montgomery CG, Gaffney PM and McKinney BA (2020) Nearest-Neighbor Projected Distance Regression for Epistasis Detection in GWAS With Population Structure Correction. Front. Genet. 11:784. doi: 10.3389/fgene.2020.00784

INTRODUCTION

An important challenge for machine learning in GWAS is to perform computationally efficient screening for variants involved in complex genetic models, including epistatic effects. The identification of interactions in GWAS may lead to an increased understanding of pathogenic mechanisms and potential therapeutic targets, but low minor allele frequencies and the curse of dimensionality make interaction detection difficult. Machine learning methods also face the challenge of identifying statistical thresholds that limit false discoveries and handling the intricacies of biomedical studies such as covariates and population structure.

Recently we developed a flexible nearest-neighbor-based machine learning feature selection method called Nearest-neighbor Projected Distance Regression (NPDR) to address these challenges (Le et al., 2020). NPDR integrates a regression formalism to allow statistical significance testing with projected nearest-neighbor machine learning to enable detection of complex multivariate models in high dimensional data. The projection of nearest neighbors from high dimensions onto single feature dimensions allows NPDR to detect features involved in complex patterns with other features in high-dimensional data that influence phenotypic variance. The regression formalism of NPDR maintains the ability to detect interactions while providing a statistical basis for feature selection thresholding and control of false discoveries due to multiple hypothesis testing.

In the current study, we demonstrate the capabilities of the NPDR framework to detect variants involved in complex genetic models and to adjust for population structure. We compare the

performance of NPDR with random forest and univariate analysis on a panel of benchmark simulated genetic variant data described by Urbanowicz et al. (2018). We analyze data with multivariate main effects and multiple epistatic effects and outcomes with balanced and imbalanced cases-control ratios as well as continuous variation. Consistent with our previous studies (McKinney et al., 2009; Le et al., 2020), we show that random forest is able to detect interactions when the number of predictors is small but its power diminishes with the dimensionality of the data. NPDR is less susceptible to the curse of dimensionality as we show it is able to detect interactions with statistical significance in both low and high dimensional contexts.

In addition to adjustment for multiple testing, NPDR enables the adjustment for covariates such as sex, age, or population structure – due to population stratification or cryptic relatedness. Population structure leads to linkage disequilibrium (LD) and this deviation from independence may increase false associations (McCarthy et al., 2008; Chen et al., 2016). The confounding effect of population structure may be exacerbated for complex models involving interactions between variants. Covariate adjustment is challenging for many machine learning methods that have the flexibility of being model free (Le et al., 2020). NPDR is model free in its use of nearest neighbors for detecting interactions, but it includes a statistical model for the projected distance for each feature. This generalized linear model (GLM) of projected distances then allows for the inclusion of projected distance covariates such as principal components (PCs).

Systemic lupus erythematosus (SLE) is an autoimmune inflammatory disease characterized by antinuclear autoanti bodies, complement and interferon activation, and tissue destruction. It predominantly affects women. Numerous immune-related genes and genes with other functions have been shown to predispose to SLE (Harley et al., 2008; Gregersen and Olsson, 2009), but there is a need to identify other genomic factors that may be interacting with each other as pairs or in a higher-order network to influence the development of this complex disease (Davis et al., 2013; Tyler et al., 2019). We use NPDR to enrich for interactions in the systemic lupus erythematosus genetics (SLEGEN) GWAS, which consists of females of European ancestry (720 SLE and the 2,337 controls) (Harley et al., 2008). Although the SLEGEN data is a homogeneous sample, we demonstrate the ability of NPDR to adjust for possible cryptic relatedness by including PCs as covariates. Identifying additional interacting variants may lead to a better understanding of the pathways affecting SLE.

MATERIALS AND METHODS

Nearest-Neighbor Projected-Distance Regression

Relief-based methods are known for their ability to identify interactions with computational efficiency but generally do not account for statistical significance of the attributes that may lead to high misclassification rate. In order to control false discoveries and adjust for covariates, we developed NPDR to use the GLM to perform regression between nearest-neighbor pair distances projected onto each predictor dimension (Le et al., 2020). We define the NPDR neighborhood set $\mathcal{N}$ of ordered pair indices of subjects as follows.

In NPDR, instance (e.g., subject) i is a point in p attribute (e.g., variant) dimensions, and the topological neighborhood of i is labeled by N_i. This neighborhood is a set of other instances trained on the dataset $X^{m\times p}$ of m instances and p attributes and depends on the type of Relief neighborhood method (e.g., fixed-k or adaptive radius) and the type of metric (e.g., Manhattan or Euclidean). If instance j is in the neighborhood of i ($j \in N_i$), then the ordered pair is in the overall neighborhood ($(i, j) \in \mathcal{N}$) for the projected-distance regression analysis. The ordered pairs constituting the overall neighborhood can then be represented as nested sets:

$$\mathcal{N} = \{\{(i, j)\}_{i=1}^{m}\}_{\{j\neq i: j\in N_i\}}.$$

The cardinality of the set $\{j \neq i : \ j \in N_i\}$ is k_i, the number of nearest neighbors for subject i. In the analyses in the current study, we use an adaptive k for hits and misses, $k = 0.154$ (m-1), that has shown good balance between detecting main effects and interaction effects (Le et al., 2019, 2020).

We compute the distance between two instances i and j in the space of the set A of all attributes with an L_q metric

$$D_{ij}^{(q)} = \left(\sum_{a\in A} \left|d_{ij}(a)\right|^q\right)^{1/q},$$

where $|A| = p$ is the number of attributes in the dataset. We use q = 1 (Manhattan) in this study. The projected difference or diff function $[d_{ij}(a)]$ between two instances i and j onto a SNP is of critical importance to the NDPR algorithm and can be computed by various difference functions. The standard difference used by Relief-based algorithms for categorical variables is a binary mismatch. For SNPs, this genotype mismatch (GM) is a 0 or 1 difference between two individuals (R_i, R_j) for a SNP, a, based on the individuals' genotypes for this SNP. Specifically, the diff function is

$$\begin{aligned} d_{ij}^{GM}(a) \ &= \text{diff}_{GM}(a, R_i, R_j) \\ &= \begin{cases} 0, & \text{genotype}\,(a,\ R_i) = \text{genotype}\,(a, R_j) \\ 1, & \text{otherwise} \end{cases} \end{aligned}$$

where genotype $(a,\ R_i)$ is the genotype for individual R_i for SNP a. In other words, two individuals have zero diff if they have identical genotypes and they have unit diff if they have different genotypes.

A potential drawback of GM is that it is not sensitive to heterozygous genotype differences when computing the diff. The following allele mismatch (AM) diff accounts for the difference in the number of alleles for a SNP when computing the distance between two individuals (Arabnejad et al., 2018). The AM difference of two individuals can be calculated by the following formula:

$$\begin{aligned} d_{ij}^{AM}(a) \ &= \text{diff}_{AM}\ (g_v, R_i, R_j) \\ &= \frac{1}{2} \times |\text{genotype}\,(a,\ R_i) - \text{genotype}\,(a,\ R_j)| \end{aligned}$$

TABLE 1 | Properties of simulated data from epistasis benchmarking repository (https://github.com/EpistasisLab/rebate-benchmark/).

Dataset	Predictive features (influence ratio)	Total features	Heritability	MAF	Instances (case/ctl)
Main effect pair	2 (50:50)	20	0.4	0.2	800/800
4-main effect	4 (25:25:25:25)	20	0.4	0.2	800/800
4 interactions	4 (2 interacting pairs)	20	0.4	0.2	800/800
Interacting pair	2	20	0.05	0.2	800/800
Continuous outcome interactions	2	20	0.4	0.2	1,600
Imbalanced outcome interactions	2	20	0.4	0.2	960/640
10,000 variants	2	10,000	0.4	0.2	800/800

All scenarios are case-control except one continuous outcome dataset. All scenarios have 20 features except for one with 10,000 features.

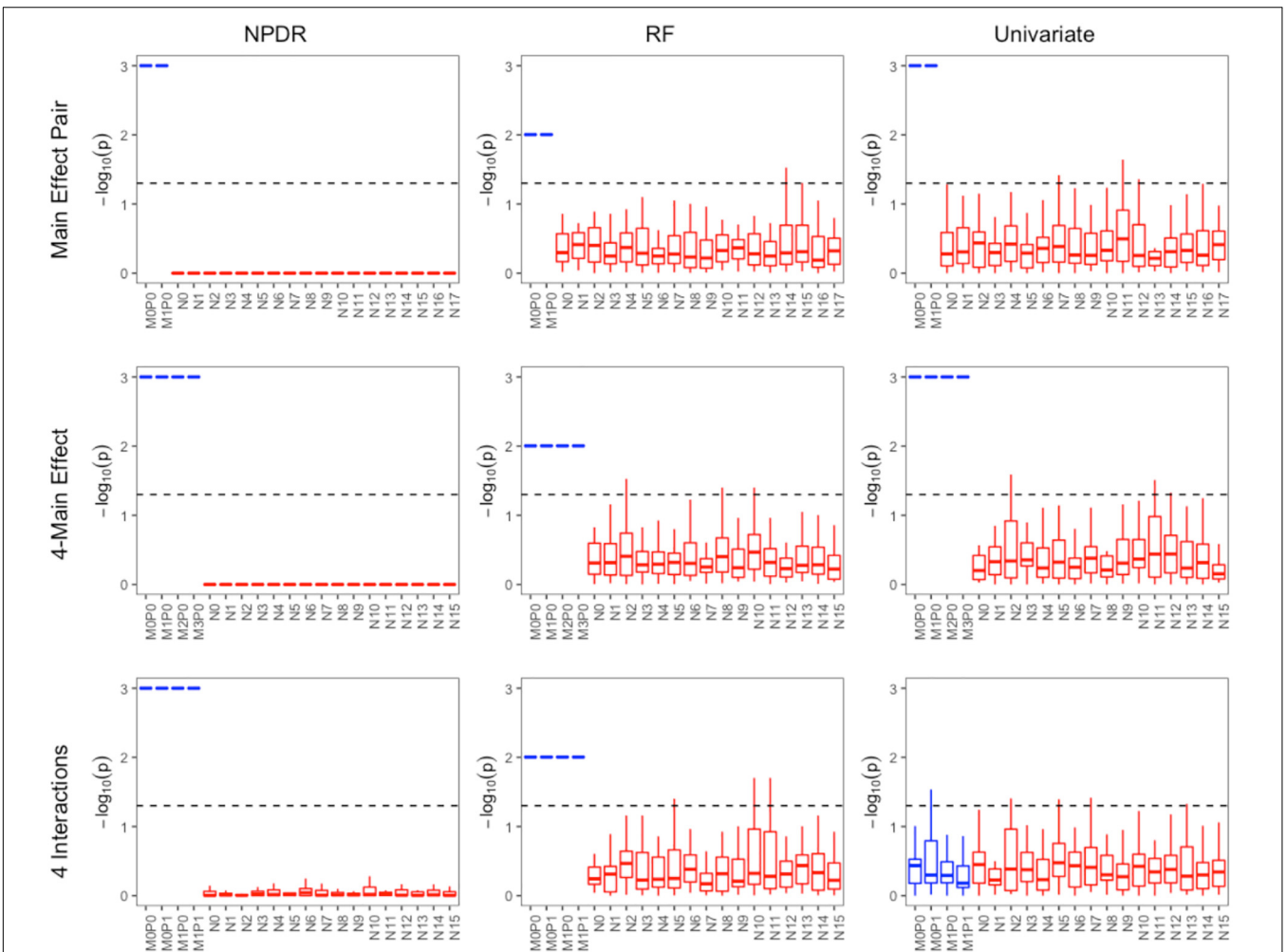

FIGURE 1 | Performance of three feature selection algorithms (from left to right): NPDR, random forest (RF) using permutation *P*-value, and univariate regression for three simulation models from https://github.com/EpistasisLab/rebate-benchmark/. Simulated models include (from top to bottom) main effect of two variants, main effect of four variants, and heterogeneous interaction of two interacting pairs. The –log(adjusted *P*-value) is plotted for the 20 variants in each dataset for 30 replicate simulations. The functional variable names (blue) begin with letter M, and the background variable names (red) start with letter N. Datasets have 1,600 samples (800 cases and 800 controls). Additional dataset details are given in **Table 1**. The dashed line represents Bonferroni adjusted *P*-value of 0.05.

where genotype $(a,\ R_i)$ is the genotype encoding for number of copies of the reference allele for SNP a for individual i. In other words, the value of genotype $(a,\ R_i)$ is the number of minor alleles in the genotype: 0, 1, or 2. Then the return value of $d_{ij}^{AM}(a)$ is 0, 0.5, or 1 when the two individual have 2, 1, or 0 alleles in common, respectively. Other projected differences and metrics are described by Arabnejad et al. (2018).

Nearest-neighbor Projected-Distance Regression uses the GLM to perform regression between nearest neighbors. For each attribute, the NPDR model fits a GLM to the attribute's

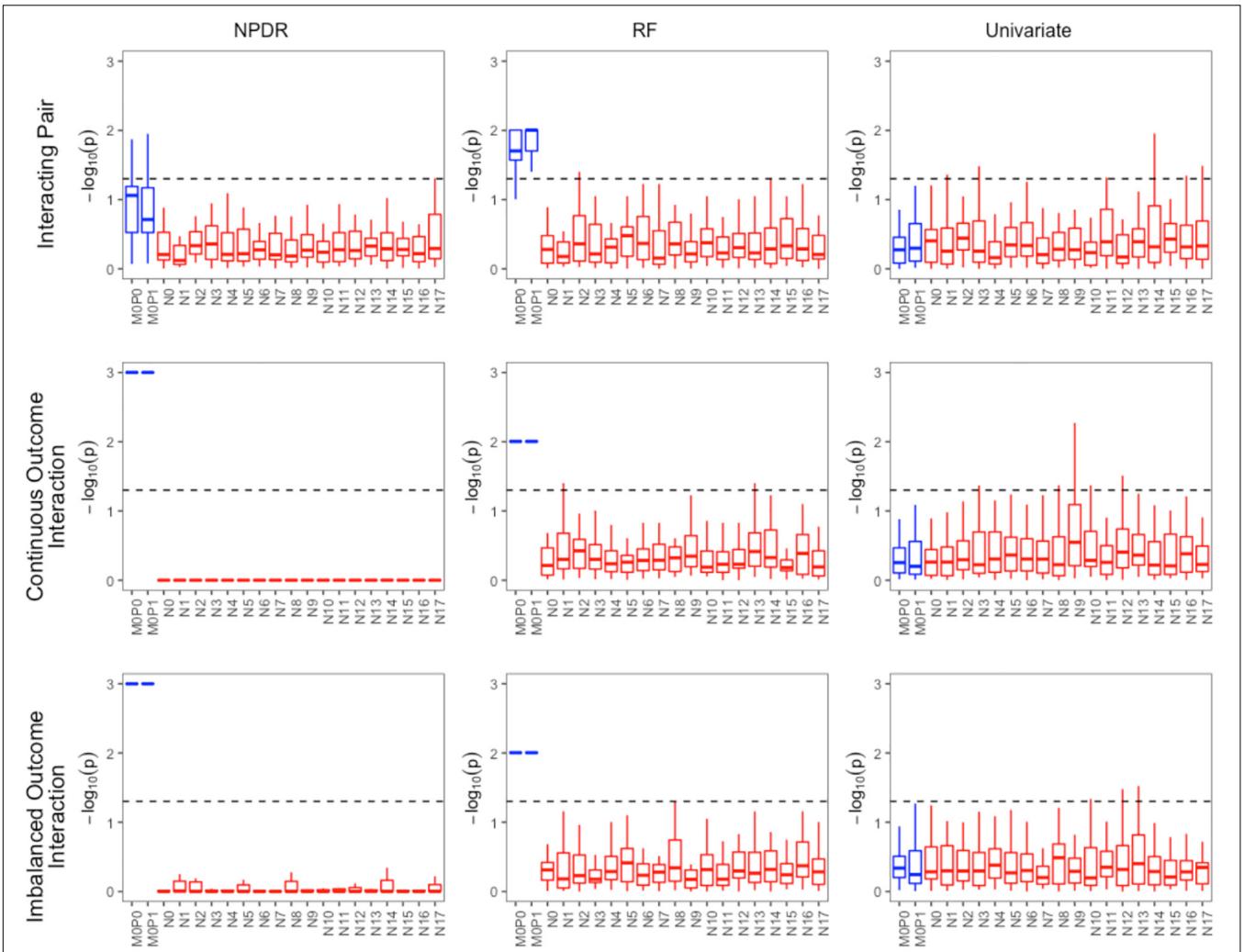

FIGURE 2 | Performance of three feature selection algorithms (from left to right): NPDR, random forest (RF) using permutation *P*-value, and univariate regression for three simulation models from https://github.com/EpistasisLab/rebate-benchmark/. Simulated models include (from top to bottom) an interacting pair, an interacting pair with a continuous outcome, and imbalanced case-control (60%). The –log(adjusted *P*-value) is plotted for the 20 variants in each dataset for 30 replicate simulations. The functional variable names (blue) begin with letter M, and the background variable names (red) start with letter N. Datasets have 1,600 samples (800 cases and 800 controls). Additional dataset details are given in **Table 1**. The dashed line represents Bonferroni adjusted *P*-value of 0.05.

projected distances between all pairs of nearest neighbors. The regression coefficients are calculated to minimize the least-squares error. For case-control phenotypes, p_{ij}^{miss} is the probability that subjects i and j are in the opposite phenotype class (misses) versus the same class (hits). We model this probability from the projected differences for SNP a with logit function:

$$\text{logit}\left(p_{ij}^{miss}\right) = \beta_0 + \beta_a d_{ij}(a) + \varepsilon_{ij}.$$

The NPDR test statistic for attribute a is the β_a estimate with one-sided hypotheses:

$$H_0: \quad \beta_a < 0.$$
$$H_1: \quad \beta_a \geq 0.$$

The quantity e^{β_a} is the predicted change in odds of neighbors being in opposite classes when the difference of the attribute a changes by one unit. For a continuous outcome (quantitative trait), NPDR uses linear regression of the numerical difference $d_{ij}^{\text{num}}(y)$ of the outcome y between neighbors:

$$d_{ij}^{\text{num}}(y) = \beta_0 + \beta_a d_{ij}(a) + \varepsilon_{ij}$$

and feature importance and significance are again determined from the coefficient β_a.

False-positive associations can arise in GWAS due to population stratification or cryptic relatedness. A standard approach to correct for population structure is to include PCs in the regression model to account for the genetic background. Many machine learning feature selection algorithms have limited ability to adjust for population structure or other potentially confounding covariates. However, the NPDR formalism can adjust for multiple covariates by including projected differences $d_{ij}^{\text{num}}(\vec{y}_{\text{covs}})$ for each covariate attribute in the regression model.

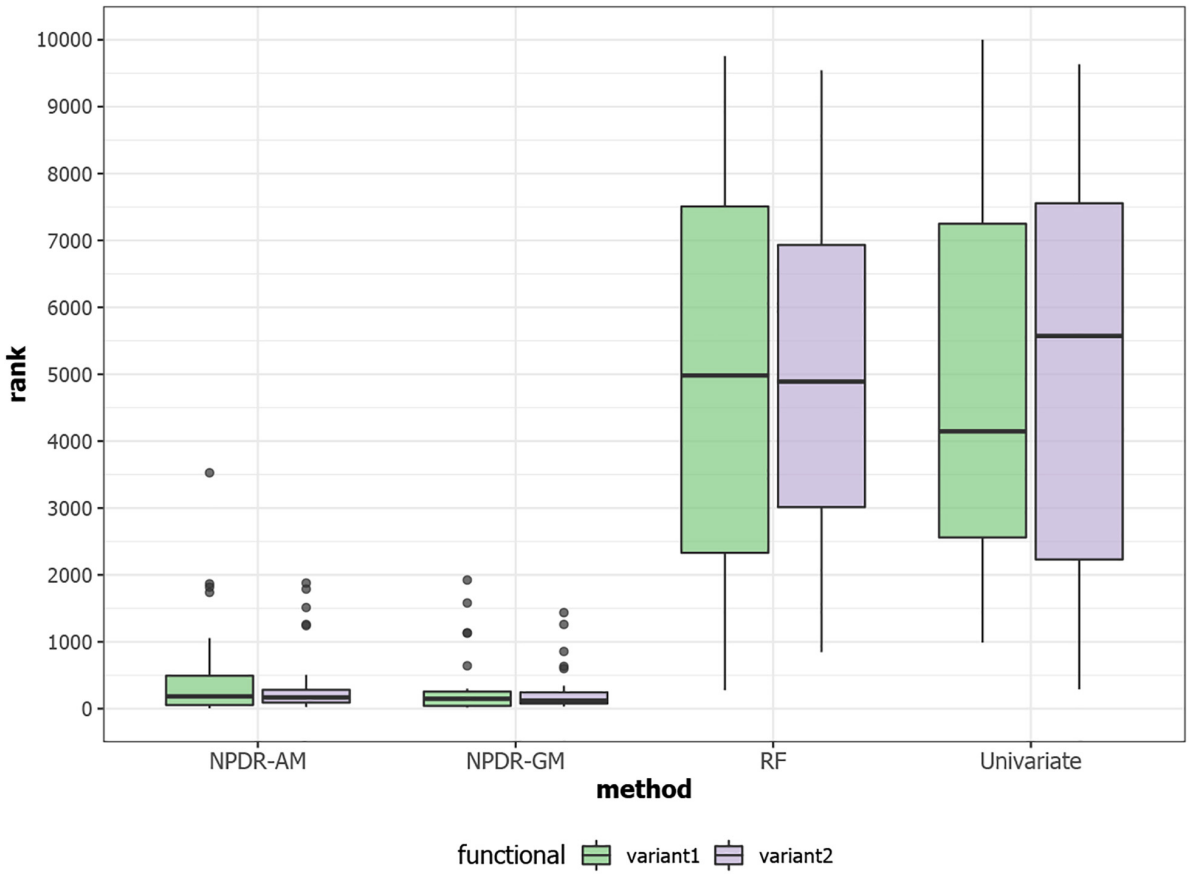

FIGURE 3 | Rank (lower is better) of the two functional interacting variants [variant 1 (green) and variant 2 (purple)] in simulated datasets with 10,000 total variants from https://github.com/EpistasisLab/rebate-benchmark/. The ranks of the two functional variables are averaged over 30 replicate simulations. NPDR *P*-value ranking is performed for allele mismatch (AM) and genotype mismatch (GM) projected distance metrics. Random forest ranking is performed by permutation importance score (RF) and univariate uses the regression coefficient *P*-value. The 10,000 simulated variants have average minor allele frequency 0.2 and heritability 0.4, and datasets have 800 cases and 800 controls.

The covariate adjusted model then becomes,

$$\text{logit}\left(p_{ij}^{\text{miss}}\right) = \beta_0 + \beta_a d_{ij}(a) + \vec{\beta}_{\text{covs}}^{T} d_{ij}(\vec{y}_{\text{covs}}) + \varepsilon_{ij}$$

where the covariate coefficient vector for 10 PCs is,

$$\vec{\beta}_{\text{covs}}^{T} = (\beta_{PC_1}, \beta_{PC_2}, \ldots, \beta_{PC_{10}})$$

and again p_{ij}^{miss} is the probability of neighbors having a different phenotype. Neighbors are still determined in the attribute (variant) space, but we add additional covariate diffs to the NDPR regression model.

Simulated Data

We compare methods using existing simulated data from the epistasis benchmark described by Urbanowicz et al. (2018) and available at https://github.com/EpistasisLab/rebate-benchmark/. For simplicity, many of the benchmark simulations use 20 total variants, but we also compare performance for multiple replicate simulations with 10,000 total variants to demonstrate computational feasibility and the effect of higher dimensionality (**Table 1** summary of datasets). For case-control data, we use data with 1,600 balanced instances (800 cases and 800 controls) and one imbalanced scenario with (60% cases). Datasets have a heritability effect size of 0.4, minor allele frequency of 0.2 and include models with 2–4 functional variants and models with additive main effects and epistatic effects. We also use a dataset with a pair of interacting variants that influence a continuous endpoint.

Real GWAS

We apply NPDR to a study of females with European ancestry with genotyping data for 317,503 SNPs for 720 SLE subjects and 2,337 controls from the SLEGEN consortium (Harley et al., 2008). All women with SLE satisfied the revised criteria for classification of SLE from the American College of Rheumatology. The study sample consisted of 730 unrelated women with SLE and 475 controls from SLEGEN. Additional "out-of-study" European ancestry female controls were added from Illumina's iControlDB. The majority of iControlDB are from the Robert S. Boas Center for Genomics and Human Genetics at the

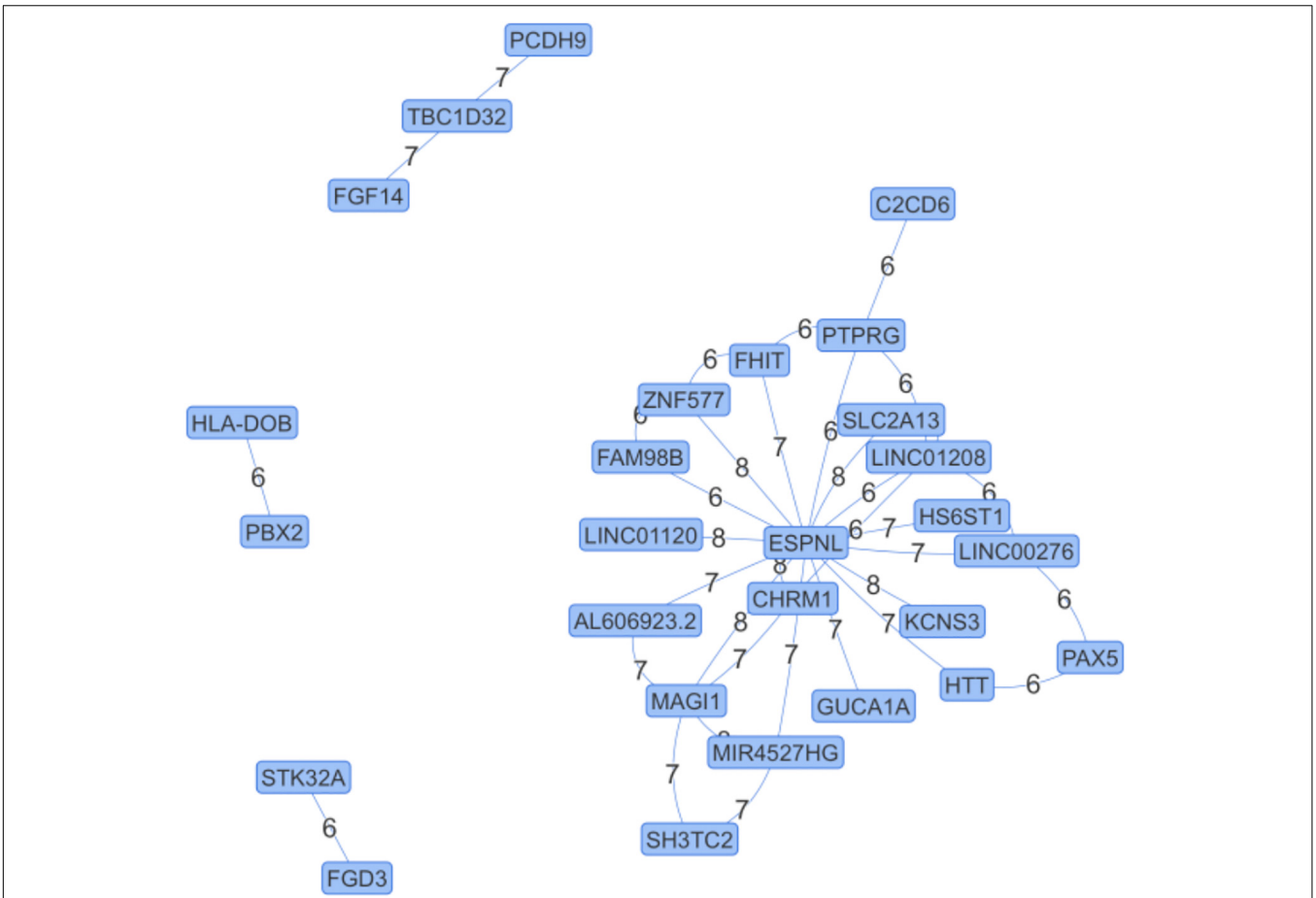

FIGURE 4 | Epistasis Network for Lupus GWAS. Edges are significant pairwise interactions with adjusted *P*-value $< 10^{-6}$ between variants in genes computed after filtering. The edge weights are the magnitude of the statistical interaction between SNPs calculated by $-\log_{10}$ (adjusted *P*-value). The espin-like (ESPNL) gene is an epistasis hub with 15 edges. The interaction on the left (HLA-DOB and PBX2) is between genes in the MHC II region.

Feinstein Institute for Medical Research. We reduced the data dimensionality using LD pruning with a correlation threshold of 0.5, minor allele frequency threshold of 0.01, Hardy-Weinberg Expectations (HWE) in controls $P > 0.01$, and HWE in cases $P > 0.0001$. LD pruning helps remove redundant features, where a SNP from a pair in high LD is removed from the data (Calus and Vandenplas, 2018). Initial filtering reduced the number of SNPs to 184,170. Due to computer memory constraints of the current implementation of NPDR, we further filtered to 10,000 SNPs based on univariate association. This filtering risks removing some interaction effects but should capture a considerable amount of variation in the data. Future implementations of NPDR will improve memory efficiency and incorporate additional variants. An advantage of NPDR is the ability to incorporate covariates into feature selection. We used the top 10 PCs from the variance-standardized relationship matrix. We mapped the top SNPs found by NPDR to Ensembl gene IDs based on proximity. If the SNP is not in an intron or exon of a gene, the algorithm computes the distance of the variant to the nearest two genes and the SNP is mapped to the gene symbol of the closest gene.

RESULTS

Simulation Analysis

For low-dimensional simulations with only 20 total attributes (**Figures 1**, **2**), both NPDR with AM metric and random forest rank the functionally interacting attributes at the top for all simulation scenarios. All methods detect the main effects (**Figure 1**), but as expected the univariate analysis cannot detect interaction effects. For low dimensional datasets (20 attributes), random forest is able to exhaustively sample all attributes and find a tree with the interacting attributes. When the total number of attributes increases from 20 to 10,000 (**Figure 3**), random forest is unable to detect the functional interactions with average ranks near random (5,000). In this case of 10,000 attributes, the random forest ranking is very similar to a univariate ranking, while NPDR has good rankings for interacting attributes using either the GM or AM projected difference metrics (**Figure 3**). This is consistent with our previous results that random forest is unable to detect interactions beyond random chance in high dimensional data, whereas Relief-based methods are less affected by the curse of dimensionality (McKinney et al., 2009). NPDR also performs

TABLE 2 | Logistic regression interaction statistics for pairs of variants in genes in the epistasis network (**Figure 4**) for SLE GWAS with adjusted *P*-values $< 10^{-6}$.

SNP 1	Gene 1	SNP 2	Gene 2	*P*-value	Adjusted *P*-value
rs10210979	ESPNL	rs2067477	CHRM1	5.149E-12	3.51E-09
rs10210979	ESPNL	rs11564281	SLC2A13	1.055E-11	3.51E-09
rs10210979	ESPNL	rs4832401	KCNS3	1.243E-11	3.51E-09
rs10210979	ESPNL	rs7573771	LINC01120	4.106E-11	8.70E-09
rs10210979	ESPNL	rs9814172	MAGI1	1.585E-10	2.69E-08
rs10210979	ESPNL	rs9807842	ZNF577	2.431E-10	3.08E-08
rs9814172	MAGI1	rs6507759	MIR4527HG	2.539E-10	3.08E-08
rs10210979	ESPNL	rs7694687	HTT	4.235E-10	4.16E-08
rs9814172	MAGI1	rs2067477	CHRM1	4.415E-10	4.16E-08
rs10210979	ESPNL	rs1446540	LINC00276	9.587E-10	8.13E-08
rs10210979	ESPNL	rs7762152	GUCA1A	1.189E-09	9.17E-08
rs9814172	MAGI1	rs17653341	SH3TC2	1.568E-09	1.05E-07
rs10210979	ESPNL	rs9311738	FHIT	1.606E-09	1.05E-07
rs10210979	ESPNL	rs6507759	MIR4527HG	2.278E-09	1.38E-07
rs11920836	LINC01208	rs11564281	SLC2A13	2.638E-09	1.44E-07
rs17083190	TBC1D32	rs9317652	PCDH9	2.717E-09	1.44E-07
rs17653341	SH3TC2	rs6507759	MIR4527HG	3.133E-09	1.56E-07
rs17083190	TBC1D32	rs978268	FGF14	3.502E-09	1.61E-07
rs10210979	ESPNL	rs12477083	HS6ST1	3.608E-09	1.61E-07
rs10210979	ESPNL	rs6936115	AL606923.2	4.387E-09	1.86E-07
rs9814172	MAGI1	rs6936115	AL606923.2	6.287E-09	2.54E-07
rs11920836	LINC01208	rs6445245	PTPRG	8.55E-09	3.30E-07
rs9311738	FHIT	rs6445245	PTPRG	1.216E-08	4.48E-07
rs10210979	ESPNL	rs11920836	LINC01208	1.283E-08	4.53E-07
rs10210979	ESPNL	rs11073328	FAM98B	1.449E-08	4.85E-07
rs9311738	FHIT	rs9807842	ZNF577	1.486E-08	4.85E-07
rs10210979	ESPNL	rs6445245	PTPRG	1.566E-08	4.92E-07
rs11920836	LINC01208	rs1446540	LINC00276	1.935E-08	5.86E-07
rs12464623	C2CD6	rs6445245	PTPRG	2.036E-08	5.95E-07
rs2067477	CHRM1	rs11920836	LINC01208	2.463E-08	6.96E-07
rs7694687	HTT	rs4507859	PAX5	3.037E-08	8.31E-07
rs204995	PBX2	rs11244	HLA-DOB	3.244E-08	8.45E-07
rs1446540	LINC00276	rs4507859	PAX5	3.288E-08	8.45E-07
rs10992568	FGD3	rs4705038	STK32A	3.749E-08	9.28E-07
rs11073328	FAM98B	rs9807842	ZNF577	3.832E-08	9.28E-07

well for multiple pairwise interactions (**Figure 1**), interactions for a continuous outcome (**Figure 2**, middle row) and imbalanced case-control data (**Figure 2**, bottom row).

SLE GWAS

We apply NPDR to the SLEGEN data, which is a real GWAS composed of females with SLE and healthy controls. Although the study is composed of European ancestry individuals, we include 10 PCs as covariates in NPDR to account for possible cryptic relatedness. Following filtering we create a network from significant (adjusted *P*-value < 1e-6) pairwise interactions (**Figure 4**). This edge significance threshold results in 35 edges (**Table 2**) in the epistasis network. The espin-like (ESPNL) gene is a hub of the network, involved in 15 of the 35 significant interactions. The particular interacting SNP (rs10210979) in ESPNL is an intron variant on chromosome 2. Although ESPNL is involved in hearing, it is ubiquitously expressed and some of its interactions may indicate novel function for the immune system. In addition to this hub gene, there is an interesting interaction between HLA-DOB and PBX2 (Pre-B-cell leukemia homeobox 2). Both of these genes are located within the major histocompatibility complex (MHC) class II region on chromosome 6, and HLA-DOB is a beta chain of the MHC class II molecule.

CONCLUSION

Machine learning feature selection methods are needed to enrich for attributes involved in complex interaction network effects in high dimensional data, such as GWAS and gene expression, for case-control and quantitative trait studies. In addition to interactions, machine learning methods need to handle complicated modeling scenarios, such as controlling for

potential confounders from demographic data or population structure, which is a perennial challenge for GWAS data.

In the current study of GWAS data, we applied a new feature selection technique called NPDR that uses the GLM to perform regression between nearest-neighbor pair distances projected onto predictor dimensions. NPDR detects interaction structure using local nearest-neighbor information in the full space of predictors, which may be SNPs or expression levels. Using simulated GWAS, we showed that NPDR has good power to detect functional variants in a variety of simulation scenarios including case-control data with and without imbalance, quantitative trait outcomes, main effects, and multiple pairwise epistatic effects. Similar to our previous findings (McKinney et al., 2009; Le et al., 2020), NPDR is less susceptible to the curse of dimensionality than random forest because when the total number of variants increases to 10,000, the ranking of interacting variants by random forest is consistent with a random ranking, while NPDR consistently ranks the functional variants near the top.

We demonstrated NPDR's ability to handle covariates by including the first 10 PCs in the NPDR models for a GWAS of SLE. Previously we showed that using the covariate term in NPDR can remove genes from nearest-neighbor feature selection that are confounded by sex (Le et al., 2020). In the current study, we constructed a candidate epistasis network for SLE from the filtered data, and found the ESPNL gene is a hub in the network with 15 of the 35 statistically significant interactions. There is no prior evidence for the role of ESPNL in autoimmunity, but it is ubiquitously expressed. Replication and functional investigation of these interactions are needed to identify mechanisms in the pathogenesis of autoimmunity. The lupus epistasis network also contains a significant interaction between HLA-DOB and PBX2, which are both located within the MHC class II region on chromosome 6. A limitation of this discovery analysis was the lack of a replication dataset. Another limitation is the need for SNP filtering in the current NPDR implementation in R for GWAS. Future implementations will take advantage of binary GWAS data formats for improved memory management.

DATA AVAILABILITY STATEMENT

Software available at: https://github.com/insilico/npdr.

AUTHOR CONTRIBUTIONS

MA and BM conceived of the project and wrote the initial draft. MA performed the analyses. CM and PG provided the statistical and biological expertise. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Institute of Health GM121312 and GM103456, and the William K. Warren Jr. Foundation.

REFERENCES

Arabnejad, M., Dawkins, B. A., Bush, W. S., White, B. C., Harkness, A. R., and Mckinney, B. A. (2018). Transition-transversion encoding and genetic relationship metric in ReliefF feature selection improves pathway enrichment in GWAS. *BioData Min.* 11:23.

Calus, M. P. L., and Vandenplas, J. (2018). SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genet Sel Evol.* 50:34. doi: 10.1186/s12711-018-0404-z

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet.* 98, 653–666. doi: 10.1016/j.ajhg.2016.02.012

Davis, N. A., Lareau, C. A., White, B. C., Pandey, A., Wiley, G., Montgomery, C. G., et al. (2013). Encore: genetic association interaction network centrality pipeline and application to SLE exome data. *Genet. Epidemiol.* 37, 614–621. doi: 10.1002/gepi.21739

Gregersen, P. K., and Olsson, L. M. (2009). Recent advances in the genetics of autoimmune disease. *Annu. Rev. Immunol.* 27, 363–391. doi: 10.1146/annurev.immunol.021908.132653

Harley, J. B., Alarcón-Riquelme, M. E., Criswell, L. A., Jacob, C. O., Kimberly, R. P., Moser, K. L., et al. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM. PXK, KIAA1542 and other loci. *Nat. Genet.* 40, 204–210. doi: 10.1038/ng.81

Le, T. T., Dawkins, B. A., and McKinney, B. A. (2020). Nearest-neighbor Projected-Distance Regression (n.d.) for detecting network interactions with adjustments for multiple tests and confounding. 36, 2770–2777. *Bioinformatics* doi: 10.1093/bioinformatics/btaa024

Le, T. T., Urbanowicz, R. J., Moore, J. H., and McKinney, B. A. (2019). STatistical Inference Relief (STIR) feature selection. *Bioinformatics* 35, 1358–1365. doi: 10.1093/bioinformatics/bty788

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369. doi: 10.1038/nrg2344

McKinney, B. A., Crowe, J. E., Guo, J., and Tian, D. (2009). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet.* 5:e1000432. doi: 10.1371/journal.pgen.1000432

Tyler, A., Mahoney, J. M., and Carter, G. W. (2019). Genetic interactions affect lung function in patients with systemic sclerosis. *G3* 10, 151–163. doi: 10.1534/g3.119.400775

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., and Moore, J. H. (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. *J Biomed Inform.* 85, 168–188. doi: 10.1016/j.jbi.2018.07.015

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: info@frontiersin.org | +41 21 510 17 00

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