Fundamental Research 4 (2024) 761–769

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

GWASTool: A web pipeline for detecting SNP-phenotype associations

Xin Wang [a,b], Beibei Xin [c], Maozu Guo [d], Guoxian Yu [a,b,*], Jun Wang [b,*]

[a] School of Software, Shandong University, Jinan 250101, China
[b] Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan 250101, China
[c] College of Agronomy & Biotechnology, China Agricultural University, Beijing 100193, China
[d] College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

ARTICLE INFO

Article history:
Received 25 December 2023
Received in revised form 19 February 2024
Accepted 11 March 2024
Available online 22 March 2024

Keywords:
Genome-wide association studies
SNP interactions
Associated loci detection
SNP visualization
Web server

ABSTRACT

The genome-wide association study (GWAS) aims to detect associations between individual single nucleotide polymorphisms (SNPs) or SNP interactions and phenotypes to decipher the genetic mechanism. Existing GWAS analysis tools have different focuses and advantages, but suffer a series of tedious and heterogeneous configurations for computation. It is inconvenient for researchers to simply choose and apply these tools, statistically and biologically analyze their results for different usages. To address these issues, we develop a user friendly web pipeline GWASTool for detecting associations, which includes simulation data generation, associated loci detection, result visualization, analysis and comparison. GWASTool provides a unified and plugin-able framework to encapsulate the heterogeneity of GWAS algorithms, simplifies the analysis steps and energizes GWAS tasks. GWASTool is implemented in Java and is freely available for public use at http://www.sdu-idea.cn/GWASTool. The website hosts a comprehensive collection of resources, including a user manual, description of integrated algorithms, data examples and standalone version for download.

1. Introduction

The influx of genome-wide data has accelerated genome-wide association study (GWAS). The aim of GWAS is to explore the genetic associations between small variations, such as single nucleotide polymorphisms (SNPs), and complex traits or diseases. SNPs are the most common genetic variation in the DNA sequences [1]. Association analysis results can be valuable in numerous scenarios. Studying variants associated with diseases can provide guidance for early prevention and targeted treatments. Detection of SNPs associated with plant traits can be utilized to select high-yield and high-quality plant lines based on their genomes, eliminating the need for field planting. It reduces the breeding cycle and experimental costs. Single marker analysis methods and multi-locus methods have been proposed to dissect the genetic foundation of traits. Single marker methods test the association between a single SNP and phenotype each time [2]. In contrast, multi-locus methods examine the associations between multiple loci and phenotype simultaneously [3]. They are more in accordance with biological rules. Besides, it has been recognized that many complex traits and diseases are caused by interactions between loci [4]. SNP interactions (a.k.a. **epistasis**) can further uncover the unknown heritability of complex traits [5].

Available epistasis detection methods can be categorized into three groups: exhaustive [6], stepwise [7,8] and machine learning-based search methods [9,10]. Exhaustive methods often have the highest coverage but take a long time to run and face the challenge of "curse of dimensionality". Stepwise methods gradually reduce the candidate set and have better efficiency, but may miss SNP interactions associated with phenotypes. Machine learning-based approaches do not rely on specific models but lack interpretability and accuracy.

Existing algorithms have been developed on diverse running environments, implemented using heterogeneous programming languages, and utilized various input file formats. Their heterogeneity prevents researchers to select the most suitable algorithms or compare multiple algorithms for more reliable and comprehensive results [11]. Besides, conducting biological analysis based on these results needs solid knowledge in biology and a complex, comprehensive search process. Given those, we develop GWASTool, a user-friendly and plugin-able web pipeline for detecting SNPs or SNP combinations associated with diseases or phenotypes. GWASTool is a complete GWAS pipeline from data generation and processing, associated loci detection, to result analysis, without many technical barriers.

To our best knowledge, GWASTool provides the first online platform that assembles multi-type algorithms and the whole pipeline for detect-

* Corresponding authors.
E-mail addresses: gxyu@sdu.edu.cn (G. Yu), kingjun@sdu.edu.cn (J. Wang).

https://doi.org/10.1016/j.fmre.2024.03.005

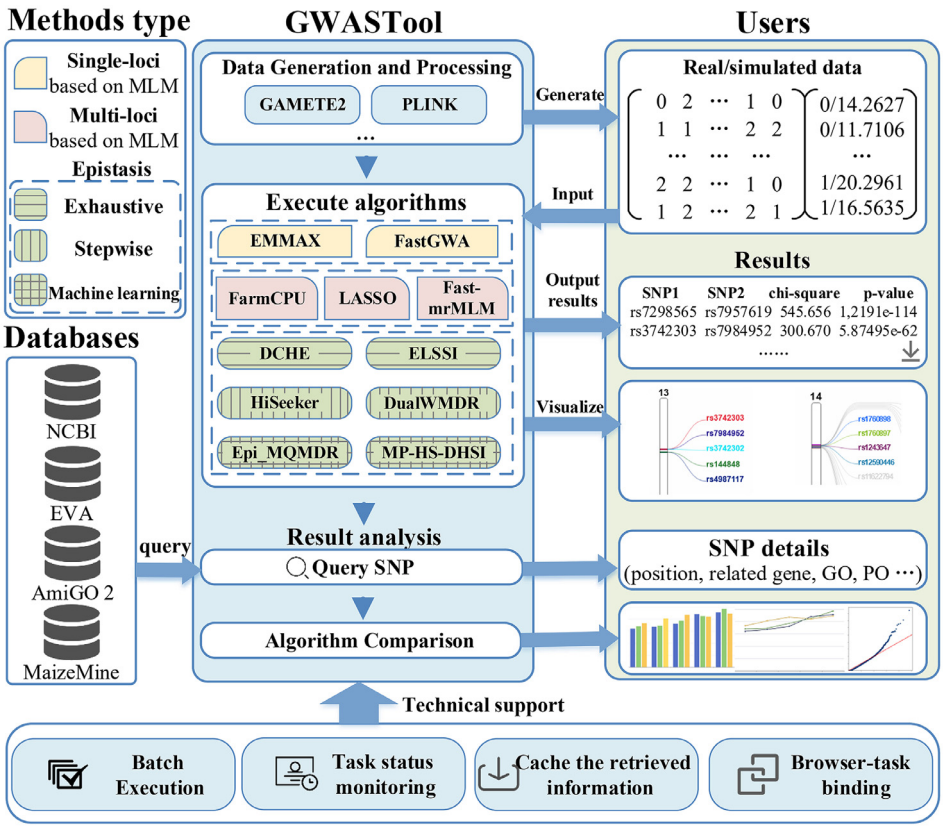

Fig. 1. The overall architecture of GWASTool mainly consists of four parts: simulated data generation and real data processing, multiple detection algorithms execution, result analysis and performance comparison. The "Data Generation and Processing" module offers tools to generate qualitative or quantitative simulated datasets and preprocess existing datasets. The "Execute algorithms" module provides single-loci, multi-loci and epistasis detection algorithms to detect associations. The "Result analysis" module facilitates the analysis and interpretation of detected SNPs. The "Algorithm comparison" module enables users to evaluate and compare the performance of different algorithms.

ing associated SNPs and epistasis. GWASTool is a valuable resource for association detection.

2. Methods

GWASTool mainly contains four modules: data generation and processing, algorithm execution, result analysis and performance comparison, as illustrated in Fig. 1. Each module serves a specific purpose to facilitate efficient analysis and interpretation of genetic data. The data generation and processing module in GWASTool offers a range of tools to generate simulated data sets or preprocess existing data sets. This enables users to prepare their data for analysis. The algorithm execution module is a crucial component of GWASTool. It provides users with a selection of 11 diverse algorithms. These algorithms are designed to detect phenotype-associated SNPs or SNP interactions. Users can choose the most suitable algorithm based on their specific requirements. GWASTool also offers result visualization and query module to facilitate the analysis and interpretation of detected SNPs and epistasis. Additionally, GWASTool provides tool-kits for easy performance evaluation and comparison. These toolkits simplify the process of assessing the performance of different algorithms with different parameters or comparing the results obtained from various analyses. It is important to note that all modules can be used independently, allowing users to tailor their analysis according to their specific needs and research requirements. GWASTool is designed as parallel, batch-able and task-based to improve platform performance and stability. Researchers can conveniently detect associations by choosing different algorithms and easily add new algorithms as needed.

2.1. Data generation and processing

GWASTool offers convenient tools to simplify simulated data generation and data preprocessing, as shown in Fig. 2b. Three types of tools are integrated. The quantitative trait values generation tool [12] is introduced to simulate trait values based on different genetic models. Those models include those with no polygenic variances, an additive polygenic background or an epistatic background. The second tool integrated is GAMETES [13]. As a canonical tool, GAMETES is supported to generate complex biallelic SNP-disease models in simulation studies. Besides, GWASTool uses PLINK [14] to execute SNP filtering and data processing. Users also can download these tools from the corresponding column of GWASTool, enabling them to access more functions.

2.2. Algorithm execution

GWASTool integrates a variety of current representative algorithms, including two single-locus algorithms based on mixed linear model (MLM) framework, EMMAX [15] and FastGWA [16]; three multi-locus methods based on MLM framework, LASSO [17], FarmCPU [18] and FastmrMLM [12]; and six epistasis detection algorithms, exhaustive methods DCHE [19] and ELSSI [20], stepwise methods HiSeeker [7] and DualWMDR [8], machine learning based methods MP-HS-DHSI [9] and Epi-MQMDR [21] in the latest version. The detailed procedures of these algorithms are given in the Supplementary file.

As shown in Fig. 2c, GWASTool provides a user-friendly interface for each algorithm, allowing users to input the required genotype and phenotype data files and parameters. The genotype data files contain M SNPs of N individuals, with SNP genotypes encoded as 0, 1, or 2,

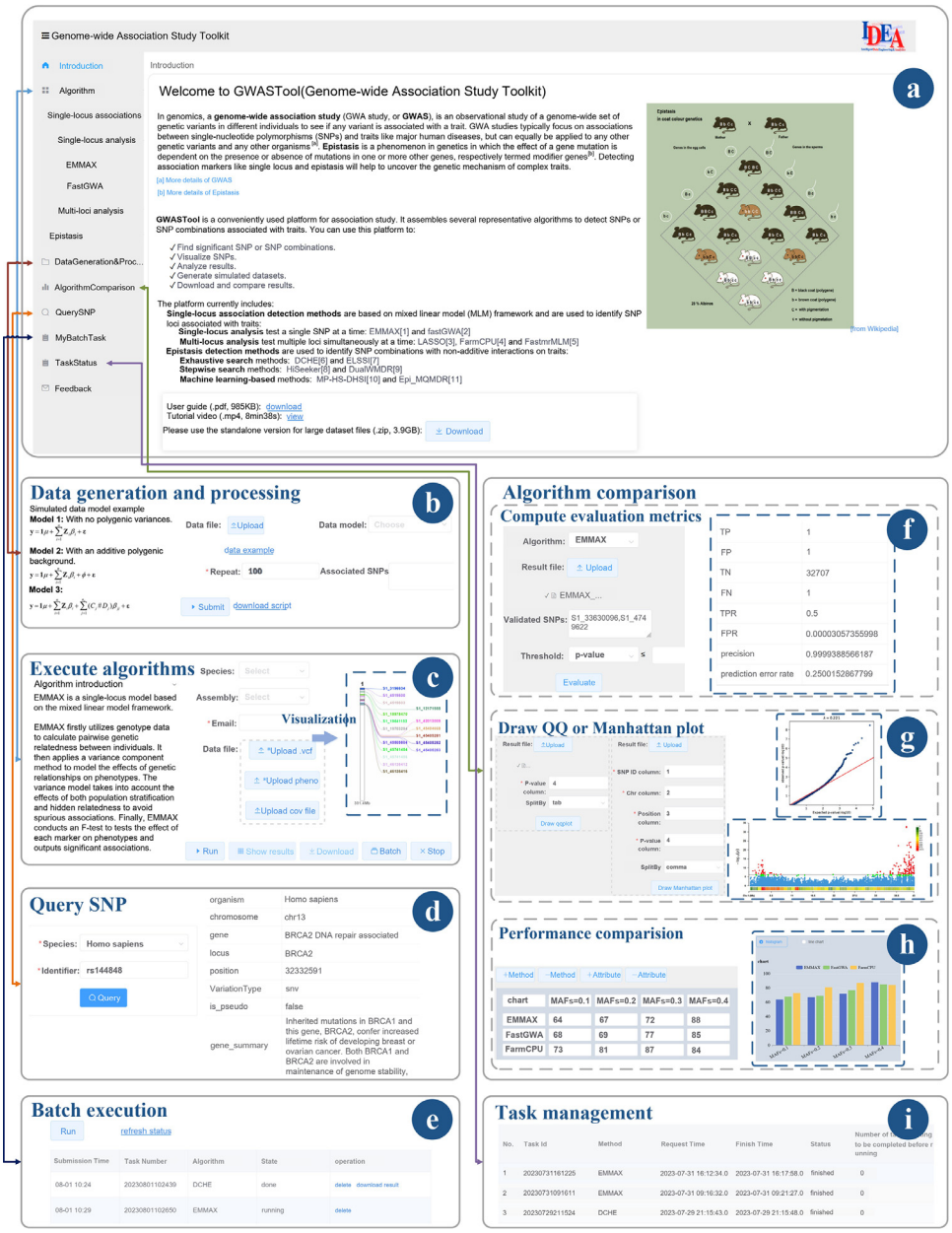

Fig. 2. The usage of GWASTool. (a) The homepage highlights the features of GWASTool, and introduces the concept of GWAS and epistasis. (b) The "Data generation and processing" page offers tools for generating simulated datasets and preprocessing available datasets. (c) The "Algorithm" page provides multiple association detection algorithms for analyzing the data, along with result visualization. (d) The "Query SNP" page allows users to retrieve basic information of SNPs, such as position, related genes, and other relevant details. (e) The "MyBatchTask" page supports the execution of multiple algorithms in a batch. Panel (f-h) separately compute typical evaluation metrics using result file, generate QQ or Manhattan plots of association results, compare the performance of different methods. (i) The "TaskStatus" page enables users to check the status of submitted tasks.

according to the number of minor alleles present at each locus. The phenotype data files contain quantitative trait values or disease status (1 for case and 0 for control) of N samples. To ensure data security, GWASTool uses a task-based computing architecture. In the current version, data files and their results are stored for two days, allowing users to download them. After this expiration period, we guarantee that data files are automatically deleted and will not be used for any other purpose. Our commitment to privacy and data protection is clearly outlined in the website's privacy policy (http://www.sdu-idea.cn/GWASTool/privacyPolicy). On the input page, all parameters for each algorithm are clearly explained and preset with default values. These default parameter values have been extensively tested in corresponding literature and are suitable for most cases. The user can also update the parameters within the provided value range as needed. Additionally, GWASTool offers performance comparison tools to assist users in selecting optimal parameters.

A significant advantage of GWASTool is its support for batch execution of multiple algorithms and efficient task management. This feature greatly improves computational efficiency and resource utilization. Users can add tasks to the task queue and run them in batches (Fig. 2e). The "TaskStatus" page provides comprehensive information about all submitted tasks, including the submitted time, execution status, order in the queue, and completion time, as shown in Fig. 2i. Besides, GWASTool has good scalability. It is convenient to plugin new detection algorithms

for personal needs. The details for new algorithms adding are given in the online document.

2.3. Result analysis

GWASTool provides the visualization, download and annotation analysis for the obtained results. Users can visualize the positions of detected significant SNPs on chromosomes, their located genes and other relevant information if the SNP names are given. Users can also conveniently query basic information of SNPs (Fig. 2d), including their chromosome, position, related gene, Gene Ontology (GO), Plant Ontology (PO) and the nearby genes. It provides convenience for users to evaluate the validity of the results and do better analysis. The execution and analysis results of the detected associated SNPs can be downloaded from the algorithm page or batchtask page. They will also be emailed to the user-specified address, which can ensure the stability of submitted tasks, even in cases where the tool or page is closed or interrupted.

2.4. Performance comparison

GWASTool provides a comprehensive set of tools for results evaluation, performance analysis and visualization. Canonical evaluation metrics, such as true positive rate (TPR), false positive rate (FPR), precision and accuracy, can be automatically calculated for a specific algorithm using the user-defined thresholds(Fig. 2f). These metrics can be used to quantitatively evaluate the algorithm performance. As shown in Fig. 2g, the QQ plot and Manhattan plot are also supported to visualize the results of association analysis. In addition, a real-time chart generation tool is embedded in GWASTool, which enables researchers to visually compare the performance trends of multiple methods under different conditions.

2.5. Implementation

GWASTool takes genotype and phenotype data files as input to detect associated loci. It is developed in Java and operates on a server equipped with an Intel Xeon 6248R processor, 512GB RAM and an Ubuntu 18.04 system. The software stack includes R-4.0.0, Node 16.15.0, Npm 8.5.5, tomcat 8.5.78, redis 6.2.6, MySQL 8.0.32 and elasticsearch 7.15.2. To detect associations, GWASTool directly executes executable files of the algorithms. Basic SNP information is stored in Redis, while task attributes are stored in MySQL. Elasticsearch is utilized to search for nearby genes of SNPs. For querying SNP corresponding information, GWASTool leverages multiple open databases, including NCBI, European Variation Archive (EVA) [22], AmiGO2 [23,24], and MaizeMine. The NCBI dbSNP database has amassed more than 900 million distinct variants from over 200,000 subjects for homo sapiens up to the latest version in August 3, 2023. AmiGO 2 contains 879,963 biological process items and 44,997 GO terms up to 2020. EVA hosts more than 3 billion genomic variants of over 130 species. MaizeMine integrates the Zea mays Zm-B73-REFERENCE-NAM-5.0 genome assembly and genome assemblies of 25 other NAM founder lines with annotation data sets. These databases are widely used and support a comprehensive analysis of the detected associated SNPs and their effect mechanisms. Besides, we will also keep up with the latest research progress and update the resources used in GWASTool in future. We have created a docker container for easy deployment of the running environment.

3. Results

3.1. Comparison with existing detection tools

The functional differences between GWASTool and other related tools are outlined in Table 1. Compared with existing tools, GWASTool provides a whole and multi-functional pipeline for detecting associated SNPs or SNP interactions and result analysis. It applies batch execution and caches data retrieved from public databases to enhance efficiency and convenience, which facilitates users to conduct comprehensive association detection and result analysis. In contrast, other tools often only execute a single algorithm at a time, lack algorithm comparison, in-depth result inspection and low coverage of existing detection algorithms.

3.2. Detailed information of the test datasets

We conducted extensive tests of GWASTool using both simulated and real datasets. The real datasets we collected include a Breast Cancer (BC) dataset with dichotomous qualitative traits and a Maize dataset with quantitative traits. The BC dataset comprises 5,607 SNPs from 1,045 affected individuals and 3,893 controls. The Maize dataset contains 127,669 SNPs from 6,957 samples after quality control. For the simulation studies, we generated two-loci epistasis datasets based on disease model DME-2 (with marginal effects disease model) and DNME-2 (with no marginal effects disease model), and three-loci epistasis datasets based on DME-3 (with marginal effects) and DNME-3 (without marginal effects) using GAMETES. Their parameters and the penetrance values are shown in the Supplementary file. Additionally, to enable a comprehensive performance comparison, we simulated five effective SNP combinations in DNME-2 and DNME-3 datasets. For further performance analysis on quantitative traits, we sampled 10,000 SNPs and 200 samples from the collected Maize dataset. We simulated the phenotype values using five genetic models: M-a (with no polygenic variances), M-b (with an additive polygenic), M-c_AxA (with Additive × Additive epistasis background), M-c_AxD (with Additive × Dominance epistasis background) and M-c_DxD (with Dominance × Dominance epistasis background). The simulated dataset for quantitative trait testing includes six effective SNPs, their effects and positions can be found in the Supplementary file.

3.3. Performance of GWASTool

To provide users with a performance reference for algorithm selection, we evaluated the performance of integrated algorithm in GWASTool on simulated data and real data. The runtime of epistasis detection algorithms is closely related to the number of loci and samples. Thus, we conducted runtime tests on simulated data files generated by GAMETES with different numbers of SNPs and samples. The results are shown in Tables 4 and 5, where N and M are the number of SNPs and samples, respectively. The default parameters were used for the tests and the runtime varies with different input parameters. Once the algorithm finishes, the runtime is displayed in the prompt. As the number of samples and SNPs increases, the algorithm runs slower. Detecting higher-order epistasis takes more runtime, due to the increased number of SNP combinations that need to be evaluated. MP-HS-DHSI, HiSeeker, DCHE generally perform faster than DualWMDR, Epi-MQMDR and ELSSI. Since HiSeeker can simultaneously output two-loci and three-loci association results, its runtimes remain the same in Tables 4 and 5. Epi-MQMDR doesn't support three-loci association detection. Thus, its runtimes for three-loci model are not reported in Table 5.

We also test the runtime of assembled algorithms on Breast Cancer (BC) dataset and Maize dataset, as shown in Tables 6 and 7, respectively. For quantitative traits such as Maize dataset, the algorithms EMMAX, FastGWA, LASSO, FarmCPU, FastmrMLM and Epi-MQMDR can be used, while DCHE, ELSSI, HiSeeker, DualWMDR and MP-HS-DHSI target for dichotomous qualitative traits such as BC dataset. Among these detection methods, LASSO and MP-HS-DHSI demonstrated the fastest runtimes.

Besides, we evaluate epistasis detection performance of integrated disease-associated methods on DME-2, DNME-2, DME-3 and DNME-3. To measure the detection capability, we adopt power, precision, recall and F1-score as evaluation metrics. As shown in Table 8, HiSeeker

Table 1
Differences between ViSEN [25], QTLNetwork [26], GWASpro [27], CASMAP [28], GBOOST [29], mrMLM [12] and GWASTool.

	Features	Online service	Multiple methods	Parallel execution	Visualization	Result inspection
ViSEN	A software that reads main effects and interactions, quantifies the effects of SNP attributes with information-theoretic quantities, and visualizes them in a network.	×	×	×	✓	×
QTLNetwork	A package that can dissect the genetic architecture of complex traits into single-locus effects, epistasis, and QTL-environment interactions, and visualize the analysis results by graphs.	✓	×	×	✓	×
GWASpro	A high-performance web server can provide data analyses and build complex design matrices to account for replicated phenotypic observations.	✓	×	✓	✓	×
CASMAP	A package that can detect region-based association studies and allow the correction of categorical covariates.	×	✓	×	×	×
GBOOST	A GPU-based tool implements the Boolean operation-based screening and testing (BOOST), and a gene-gene interaction analysis method.	×	×	✓	✓	×
mrMLM	An R package integrates several multi-locus GWAS methods.	×	✓	✓	✓	×
GWASTool	A complete pipeline for detecting SNP loci associated with complex traits, including simulated data generation, multi-type algorithms execution, result analysis and performance comparison.	✓	✓	✓	✓	✓

Table 2
Significant SNP interactions identified by DCHE, ELSSI, HiSeeker, DualWMDR and MP-HS-DHSI on Breast cancer dataset.

Method	Chromosome	SNP-SNP interaction	Related genes	p-value[a]
DCHE	(chr1, chr1)	(rs3820011, rs2278107)	(CFAP74, EPHA7)	$< 10^{-100}$
	(chr1, chr5)	(rs3820011, rs13360277)	(CFAP74, UIMC1)	$< 10^{-100}$
	(chr1, chr7)	(rs3820011, rs5763)	(CFAP74, TBXAS1)	$< 10^{-100}$
	(chr23, chr23)	(rs5969783, rs1802288)	(TXLNG, TSPAN6)	$< 10^{-100}$
	(chr22, chr19, chr20)	(rs1001587, rs5969783, rs912002)	(TCF20, TXLNG, ADGRG4)	$< 10^{-100}$
ELSSI	(chr17, chr20)	(rs434473, rs2903808)	(ALOX12, ZSWIM3)	$< 10^{-100}$
	(chr7, chr17)	(rs4987667, rs434473)	(TRPV6, ALOX12)	$< 10^{-100}$
	(chr12, chr17)	(rs2242653, rs4968318)	(LY6G6F, EFCAB13)	$< 10^{-100}$
	(chr12, chr17)	(rs13110318, rs4968318)	(TBC1D1, EFCAB13)	$< 10^{-100}$
	(chr5, chr16)	(rs1974777, rs9652589)	(GEMIN5, PDILT)	$< 10^{-100}$
HiSeeker	(chr3, chr3)	(rs1108842, rs4687657)	(GNL3, ITIH4)	$< 10^{-100}$
	(chr16, chr16)	(rs4408545, rs3785181)	(AFG3L1P, GAS8)	5.74×10^{-56}
	(chr6, chr6)	(rs2523608, rs805262)	(HLA-B, C6orf47)	1.19×10^{-41}
	(chr20, chr17, chr20)	(rs2272955, rs3827040, rs2903808)	(WFDC8, ALOX12, ZSWIM3)	$< 10^{-100}$
	(chr20, chr17, chr20)	(rs2272955, rs3827040, rs4638862)	(WFDC8, ALOX12, SNX21)	$< 10^{-100}$
DualWMDR	(chr3, chr6)	(rs2289247, rs757256)	(GNL3, LINC02829)	/
	(chr18, chr21)	(rs3809970, rs2070417)	(ALPK2, TIAM1)	/
	(chr6, chr23)	(rs757256, rs1129980)	(LINC02829, GPC4)	/
	(chr6, chr17)	(rs757256, rs12449313)	(LINC02829, SMCR8)	/
MP-HS-DHSI	(chr14, chr20)	(rs976272, rs3827040)	(TRMT5, SPATA25)	/
	(chr15, chr20)	(rs2242047, rs3827040)	(SLC28A1, SPATA25)	/
	(chr16, chr20)	(rs7192210, rs3827040)	(ACSM5, SPATA25)	/

a '/' means that p-value is not included in the result of the method.

achieves the highest F1-score on DME-2. ELSSI displays the highest power and recall on DNME-2 and the highest F1-score on DNME-3. MP-HS-DHSI has the highest precision on DME-2 and DNME-3, while DCHE has the highest precision on DME-3. DCHE, ELSSI and HiSeeker successfully detected associated SNP interactions in all simulated data files in DNME-3. Both DCHE and ELSSI can be applied to models without marginal effects. However, ELSSI ignores the main effect in its base methods when detecting epistasis. As a result, ELSSI tends to have lower power than DCHE for models with marginal effect. Besides, ELSSI performs better when most of its base methods accurately identify epistasis, and vice versa. DualWMDR calculates partial mutual information during its dual screening process to exclude SNPs. However, its accuracy can be influenced by the presence of multiple effective SNP combinations with similar genetic models in our simulated datasets. We want to remark that DualWMDR can handle datasets with diverse genetic models and thus we also integrate it into our pipeline.

In the simulation studies on quantitative traits, we adopt power, mean square error (MSE) and false positive rate (FPR) as the evaluation metrics [30]. Power focuses on the identification capability of specific effective loci. FPR assesses the algorithms' capability to avoid false positives, which refers to the erroneous identification of loci as associated ones when they are actually irrelevant to phenotypes. While MSE measures the variance and bias of effect estimates. These three metrics evaluate algorithms from different perspectives. The results are shown in Tables 9–11. In most cases, FastGWA has the highest power. However, FastmrMLM performs better in detecting the second and third associated markers and FarmCPU performs better in detecting the fifth marker in models M-c_AxA and M-c_AxD. EMMAX exhibits the lowest MSE in the detection of simulated associated markers in model M-c_DxD. FastmrMLM has comparable or lower MSE in model M-a. LASSO has the lowest FPR. Tables 8–11 can serve as a reference for users to select algorithms based on their specific requirements. More detailed evaluations and analysis of algorithms can be found in the corresponding literature. Users can execute multiple algorithms for more accurate and comprehensive results.

We also evaluate integrated algorithms with Breast Cancer dataset in Table 2 and Maize dataset in Table 3. ELSSI detects locus rs144848 in gene $BRCA2$ [31] on Chromosome 13 and locus rs434473 in gene

Table 3
Significant SNPs associated with leaf length trait identified by EMMAX, FastGWA, LASSO, FarmCPU, FastmrMLM and SNP-SNP interactions identified by Epi-MQMDR on Maize dataset.

Method	SNP or SNP-SNP interaction	Related genes	*p*-value[a]
EMMAX	S1_51041338	Zm00001eb126390	3.36×10^{-4}
	S5_150626025	Zm00001eb239430	1.57×10^{-4}
	S6_161803077	Zm00001eb294350	2.15×10^{-4}
FastGWA	S4_5012512	Zm00001eb166550	1.88×10^{-7}
	S3_211765068	Zm00001eb166550	2.17×10^{-7}
	S4_2513671	Zm00001eb165270	4.25×10^{-7}
	S4_2513673	Zm00001eb157700	4.25×10^{-7}
	S4_2513683	Zm00001eb165270	4.25×10^{-7}
LASSO	S1_41428002	Zm00001eb012560	/
	S2_203085168	Zm00001eb105950	/
	S3_28921363	Zm00001eb126390	/
	S6_161803077	Zm00001eb294350	/
FarmCPU	S4_120643323	Zm00001eb182500	3.44×10^{-11}
	S6_154173295	Zm00001eb290690	1.93×10^{-14}
	S1_202299516	Zm00001eb038710	9.43×10^{-11}
	S2_83451063	Zm00001eb086680	2.62×10^{-7}
	S1_27861046	Zm00001eb086680	1.61×10^{-6}
FastmrMLM	S7_172583812	Zm00001eb330380	$< 10^{-4}$
	S10_39954466	Zm00001eb411880	$< 10^{-3}$
	S2_4662352	Zm00001eb067980	$< 10^{-3}$
	S10_15221316	Zm00001eb409000	$< 10^{-3}$
Epi-MQMDR	(S1_75691971, S1_267907289)	(Zm00001eb020490, Zm00001eb055010)	/
	(S1_39038251, S1_215761002)	(Zm00001eb011970, unknown)	/
	(S2_43203235, S2_236352049)	(Zm00001eb081090, unknown)	/
	(S3_37359861, S3_176356543)	(Zm00001eb127900, unknown)	/
	(S4_153048434, S4_192205670)	(Zm00001eb186760, unknown)	/
	(S5_212596204, S5_215261920)	(unknown, Zm00001eb258590)	/
	(S6_141077911, S6_166820527)	(unknown, Zm00001eb297120)	/
	(S6_62741234, S6_166820527)	(Zm00001eb288140, Zm00001eb297120)	/
	(S7_83882682, S7_173802089)	(unknown, Zm00001eb330990)	/
	(S8_142012041, S8_162756199)	(Zm00001eb357860, Zm00001eb364930)	/
	(S8_142012041, S8_162756193)	(Zm00001eb357860, Zm00001eb364930)	/
	(S10_121750464, S10_138051720)	(Zm00001eb423960, Zm00001eb429220)	/
	(S10_4750083, S10_9214870)	(Zm00001eb406480, Zm00001eb407760)	/
	(S10_115518504, S10_137828247)	(Zm00001eb422190, Zm00001eb429130)	/

a '/' means that *p*-value is not included in the result of the method.

Table 4
The runtime with different numbers of SNPs (N) and samples (M) on two-locus epistasis model.

	M	*N*			
		1,000	2,000	5,000	10,000
DCHE	1,000	9 s	25 s	145 s	685 s
	3,000	12 s	35 s	219 s	913 s
	5,000	15 s	45 s	301 s	1311 s
ELSSI	1,000	69 s	228 s	1,396 s	5,791 s
	3,000	82 s	331 s	2,620 s	10,736 s
	5,000	120s	405 s	3,661 s	16,657 s
HiSeeker	1,000	12 s	24 s	91 s	334 s
	3,000	25 s	52 s	180 s	772 s
	5,000	34 s	76 s	321 s	1,147 s
DualWMDR	1,000	103 s	57 s	472 s	1,533 s
	3,000	175 s	160 s	798 s	2,926 s
	5,000	42 s	282 s	1,313 s	4,725 s
MP-HS-DHSI	1,000	45 s	45 s	67 s	66 s
	3,000	100 s	64 s	74 s	91 s
	5,000	78 s	79 s	100 s	127 s
Epi-MQMDR	1,000	37 s	95 s	529 s	1,779 s
	3,000	188 s	307 s	1,568 s	4,217 s
	5,000	589 s	806 s	2,772 s	8,889 s

Table 5
The runtime of DCHE, ELSSI, DualWMDR and MP-HS-DHSI with different numbers of SNPs and samples on three-locus epistasis model.

	M	*N*			
		1000	2,000	5,000	10,000
DCHE	1,000	25 s	67 s	313 s	1,064 s
	3,000	34 s	169 s	448 s	1,272 s
	5,000	37 s	127 s	532 s	1,360 s
ELSSI	1,000	88 s	925 s	2,499 s	7,818 s
	3,000	163 s	1,360 s	4,256 s	15,219 s
	5,000	193 s	3,412 s	6,285 s	15,033 s
HiSeeker	1,000	12 s	24 s	91 s	334 s
	3,000	25 s	52 s	180 s	772 s
	5,000	34 s	76 s	321 s	1,147 s
DualWMDR	1,000	145 s	130 s	313 s	1,457 s
	3,000	40 s	106 s	941 s	2,653 s
	5,000	43 s	160 s	1,487 s	3,552 s
MP-HS-DHSI	1,000	109 s	490 s	166 s	199 s
	3,000	172 s	208 s	226 s	291 s
	5,000	235 s	247 s	426 s	339 s

Table 6
The runtime of detecting associated SNP combinations with DCHE, ELSSI, HiSeeker, DualWMDR and MP-HS-DHSI on Breast cancer dataset.

DCHE	ELSSI	HiSeeker	DualWMDR	MP-HS-DHSI
533 s	3,342 s	715 s	3,531 s	94 s

Table 7
The runtime of detecting associated SNPs by EMMAX, FastGWA, LASSO, FarmCPU, FastmrMLM and Epi-MQMDR on Maize dataset.

EMMAX	FastGWA	LASSO	FarmCPU	FastmrMLM	Epi-MQMDR
559 s	4,231 s	493 s	1,472 s	72,956 s	251,892 s

Table 8
***Power, precision, recall* and *F1-score* of DCHE, ELSSI, HiSeeker, DualWMDR and MP-HS-DHSI on simulated datasets.** Each dataset contains **N = 1000** SNPs, 800 cases and 800 controls.

		DCHE	ELSSI	HiSeeker	DualWMDR	MP-HS-DHSI
Power	DME-2	0.7300	0.5900	0.0100	0.0100	0.5000
	DNME-2	0.4400	0.4600	0.3300	0.0000	0.0000
	DME-3	0.5000	0.4900	0.3800	0.0100	0.0900
	DNME-3	1.0000	1.0000	1.0000	0.1700	0.5500
Precision	DME-2	0.0365	0.0295	0.0042	0.0005	0.2135
	DNME-2	0.0635	0.0315	0.0278	0.0000	0.0000
	DME-3	0.6729	0.0417	0.0191	0.0005	0.2290
	DNME-3	0.2400	0.2690	0.1489	0.0110	0.4150
Recall	DME-2	0.7300	0.5900	0.0100	0.0100	0.5000
	DNME-2	0.1140	0.1260	0.0860	0.0000	0.0000
	DME-3	0.5000	0.4900	0.3800	0.0100	0.0900
	DNME-3	0.7880	0.7600	0.5900	0.0440	0.1560
F1-score	DME-2	0.0471	0.0476	0.2500	0.0476	0.2238
	DNME-2	0.0848	0.05478	0.0632	0.0000	0.0000
	DME-3	0.4054	0.0700	0.0478	0.0476	0.3268
	DNME-3	0.1802	0.1933	0.1189	0.0518	0.1225

$ALOX12$ [32], they are confirmed as the risk loci/genes of breast cancer. SNP rs4987667 is located on gene $TRPV6$, which encodes the TRPV6 protein, an endothelial calcium entry channel that has a large influence on breast cancer cell proliferation [33]. It also detects locus rs1974777 in gene $GEMIN5$. Dysregulation of this gene may play a role in tumor cell motility [34]. SNP rs9652589 is located on gene $PDILT$ of the PDI family, the overexpression of PDI is closely associated with breast cancer cell proliferation [35]. DCHE successfully detects rs2278107 in gene $EPHA7$, increased expression of which is associated with carcinoma [36], and rs13360277 in gene $UIMC1$, which encodes a nuclear protein that interacts with BRCA1 [37]. It also detects rs1802288 on gene $TSPAN6$, which controls the migration and recruitment of B cells to breast cancer tissues, B lymphocytes play an important role in anticancer immunity [38]. 4 loci are detected by both ELSSI and DCHE. MP-HS-DHSI detects several significant combinations including (rs2242047, rs3827040). All loci detected by MP-HS-DHSI are also considered by DCHE to be related to BC. HiSeeker detects rs11088402 in gene $GNL3$ and rs3785181 in gene $GAS11$. The protein encoded by $GNL3$ may interact with p53 and be involved in tumorigenesis [39]. $GAS11$ is associated with breast cancer [40]. DualWMDR detects rs2289247 in gene $GNL3$. It also detects rs2070417 in gene $TIAM1$, which plays an important role in cell invasion, metastasis, and carcinogenesis [41]. 3 loci are detected by both HiSeeker and DualWMDR. It is worth noting that $TRPV6$ is one of store-operated Ca^{2+} channels, Ca^{2+} entry through CRAC(Ca^{2+} releaseactivated Ca^{2+}) channels stimulates arachidonic acid release [42], and the $ALOX12$ gene encodes arachidonic acid 12-lipoxygenase [43]. Both of them are associated with breast cancer but the interaction between $TRPV6$ and $ALOX12$ has not been exten-

Table 9
Comparison of powers for EMMAX, FastGWA, LASSO, FarmCPU, FastmrMLM.

	SNP	EMMAX	FastGWA	LASSO	FarmCPU	FastmrMLM
M-a	1	0.04	0.35	0.02	0.02	0.00
	2	0.04	0.10	0.04	0.02	0.01
	3	0.08	0.72	0.02	0.10	0.00
	4	0.04	0.28	0.03	0.06	0.00
	5	0.00	0.48	0.00	0.01	0.00
	6	0.00	0.02	0.00	0.00	0.00
M-b	1	0.09	0.50	0.02	0.07	0.10
	2	0.01	0.10	0.02	0.04	0.11
	3	0.09	0.78	0.04	0.10	0.03
	4	0.07	0.44	0.04	0.05	0.14
	5	0.00	0.52	0.00	0.01	0.00
	6	0.01	0.01	0.01	0.00	0.06
M-c_AxA	1	0.02	0.07	0.00	0.02	0.03
	2	0.06	0.11	0.00	0.05	0.26
	3	1.00	0.92	0.73	0.00	1.00
	4	0.29	0.19	0.05	0.00	0.14
	5	0.00	0.00	0.00	0.01	0.00
	6	0.00	0.00	0.00	0.00	0.00
M-c_AxD	1	0.02	0.07	0.00	0.02	0.03
	2	0.06	0.11	0.00	0.05	0.26
	3	1.00	0.92	0.73	0.00	1.00
	4	0.29	0.19	0.05	0.00	0.14
	5	0.00	0.00	0.00	0.01	0.00
	6	0.00	0.00	0.00	0.00	0.00
M-c_DxD	1	0.01	0.13	0.00	0.01	0.02
	2	0.37	0.64	0.22	0.05	0.19
	3	0.17	0.48	0.05	0.00	0.05
	4	0.74	0.83	0.60	0.32	0.76
	5	0.00	0.37	0.00	0.02	0.00
	6	0.07	0.22	0.03	0.02	0.00

Table 10
Comparison of MSE for EMMAX, FastGWA, FastmrMLM. Only estimated effect of significant associated markers are shown in LASSO and FarmCPU. Thus, we ignore the MSE of LASSO and FarmCPU.

	SNP	EMMAX	FastGWA	FastmrMLM
M-a	1	0.9833	1.2660	0.0217
	2	0.0059	0.0269	0.0119
	3	0.2483	0.4805	0.0018
	4	1.0432	1.3887	0.0060
	5	0.2631	0.5089	0.0109
	6	0.0047	0.0185	0.0108
M-b	1	1.2338	1.5713	0.9714
	2	0.0107	0.0437	0.7232
	3	0.3275	0.6090	0.3035
	4	1.3421	1.7628	0.6897
	5	0.3435	0.6405	1.9419
	6	0.0046	0.0230	1.0260
M-c_AxA	1	0.8345	0.4699	1.1403
	2	0.2039	2.3659	4.6603
	3	0.0816	0.0280	13.0624
	4	0.5830	0.0660	2.0840
	5	0.0548	0.0558	1.8992
	6	0.1918	3.1734	1.3036
M-c_AxD	1	0.8345	0.4699	1.1403
	2	0.2039	2.3659	4.6603
	3	0.0816	0.0280	13.0624
	4	0.5830	0.0660	2.0842
	5	0.0548	0.0558	1.8992
	6	0.1918	3.1734	1.3036
M-c_DxD	1	1.1110	1.3012	1.2527
	2	0.0275	0.3586	2.4165
	3	0.1901	0.4166	0.5125
	4	1.0785	1.2307	3.8275
	5	0.18858	0.4581	1.6196
	6	0.0293	0.5197	2.2124

Table 11
Comparison of FPR for EMMAX, FastGWA, LASSO, FarmCPU and FastmrMLM.

	EMMAX	FastGWA	LASSO	FarmCPU	FastmrMLM
M-a	0.0008	0.0943	0.0006	0.0027	0.0020
M-b	0.0010	0.0969	0.0007	0.0021	0.0018
M-c_AxA	0.0052	0.0041	0.0008	0.0017	0.0027
M-c_AxD	0.0051	0.0041	0.0008	0.0017	0.0027
M-c_DxD	0.0019	0.0755	0.0009	0.0021	0.0014

sively studied to date. Therefore, these algorithms can recommend potential directions for further research.

EMMAX detects S5_150626025 and S6_161803077, FastGWA detects S3_211765068, S4_5012512, S4_2513671, S4_2513673 and S4_2513683. LASSO detects S1_41428002 and S2_203085168. FarmCPU detects S1_202299516, S1_27861046, S4_120643323 and S6_154173295. FastmrMLM detects S5_211179420, S7_172583812 and S10_39954466. Their corresponding genes are related to leaf tip and leaf base. 10 loci are considered to be related to trait by both FarmCPU and FastmrMLM. All parameters used in the experiments are the default ones.

In summary, the assembled algorithms can detect SNPs or SNP combinations that are significantly associated with the traits. Our GWASTool greatly simplifies the GWAS workflow and empowers GWAS tasks.

4. Conclusion

GWASTool offers a pipeline that integrates single/multiple SNPs and epistasis detection methods, practical simulated data generation tools, data processing and result analysis tools. It encapsulates the tedious data processing, running environment configurations and result analysis of diverse detection methods, and offers a user friendly web interface to researchers. Besides, it is easy to plugin new methods to meet the users' specific requirement. Users can also download and run GWASTool on their own machines when processing large or private datasets. In the future, we will pay attention to the latest research progress and promptly update GWASTool's functions and data resources to meet the evolving needs of users. Additionally, we will mine SNP interactions in higher dimensions, aiming to further unravel the genetic mechanisms underlying complex traits.

Availability

GWASTool is freely available for public use at http://www.sdu-idea.cn/GWASTool.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62031003 and 62072380) and Shandong Provincial Key Research and Development Program (2021CXGC010506).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.fmre.2024.03.005.

References

[1] J.-S. Milanese, C. Tibiche, N. Zaman, et al., Etumormetastasis: A network-based algorithm predicts clinical outcomes using whole-exome sequencing data of cancer patients, Genomics, Proteomics Bioinf. 19 (6) (2021) 973–985.
[2] H.J. Cordell, Detecting gene–gene interactions that underlie human diseases, Nat. Revi. Genet. 10 (6) (2009) 392–404.
[3] V. Segura, B.J. Vilhjálmsson, A. Platt, et al., An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations, Nat. Genet. 44 (7) (2012) 825–830.
[4] T.F. Mackay, J.H. Moore, Why epistasis is important for tackling complex human disease genetics, Genome Med. 6 (6) (2014) 1–3.
[5] W.-H. Wei, G. Hemani, C.S. Haley, Detecting epistasis in human complex traits, Nat. Rev. Gen. 15 (11) (2014) 722–733.
[6] X. Wan, C. Yang, Q. Yang, et al., BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies, Am. J. Hum. Genet. 87 (3) (2010) 325–340.
[7] J. Liu, G. Yu, Y. Jiang, et al., HiSeeker: Detecting high-order SNP interactions based on pairwise SNP combinations, Genes 8 (6) (2017) 153.
[8] X. Cao, G. Yu, W. Ren, et al., DualWMDR: Detecting epistatic interaction with dual screening and multifactor dimensionality reduction, Hum. Mutat. 41 (3) (2020) 719–734.
[9] S. Tuo, H. Liu, H. Chen, Multipopulation harmony search algorithm for the detection of high-order SNP interactions, Bioinformatics 36 (16) (2020) 4389–4398.
[10] A. Aghazadeh, H. Nisonoff, O. Ocal, et al., Epistatic net allows the sparse spectral regularization of deep neural networks for inferring fitness functions, Nat. Com. 12 (1) (2021) 1–10.
[11] E. Ip, G. Chapman, D. Winlaw, et al., VPOT: A customizable variant prioritization ordering tool for annotated variants, Genomics Proteomics Bioinf. 17 (5) (2019) 540–545.
[12] Y.-W. Zhang, C.L. Tamba, Y.-J. Wen, et al., mrMLM v4. 0.2: An R platform for multi--locus genome-wide association studies, Genomics Proteomics Bioinf. 18 (4) (2020) 481–487.
[13] R.J. Urbanowicz, J. Kiralis, N.A. Sinnott-Armstrong, et al., GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures, BioData Min. 5 (1) (2012) 1–14.
[14] C.C. Chang, C.C. Chow, L.C. Tellier, et al., Second-generation plink: Rising to the challenge of larger and richer datasets, GigaScience 4 (1) (2015) s13742–015.
[15] H.M. Kang, J.H. Sul, S.K. Service, et al., Variance component model to account for sample structure in genome-wide association studies, Nat. Genet. 42 (4) (2010) 348–354.
[16] L. Jiang, Z. Zheng, T. Qi, et al., A resource-efficient tool for mixed model association analysis of large-scale data, Nat. Genet. 51 (12) (2019) 1749–1755.
[17] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodological) 58 (1) (1996) 267–288.
[18] X. Liu, M. Huang, B. Fan, et al., Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies, PLoS Genet. 12 (2) (2016) e1005767.
[19] X. Guo, Y. Meng, N. Yu, et al., Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering, BMC Bioinf. 15 (1) (2014) 1–16.
[20] X. Wang, X. Cao, Y. Feng, et al., ELSSI: Parallel SNP–SNP interactions detection by ensemble multi-type detectors, Brief. Bioinfo. 23 (4) (2022) bbac213.
[21] X. Wang, J. Wang, G. Yu, et al., Maize epistasis detection by multi-class quantitative multifactor dimensionality reduction, in: IEEE Inter. Conf. on Bioinf. and Biomed., 2021, pp. 314–319.
[22] T. Cezard, F. Cunningham, S.E. Hunt, et al., The european variation archive: A fair resource of genomic variation for all species, Nucl. Acids Res. 50 (D1) (2022) D1216–D1220.
[23] G.O. Consortium, The gene ontology resource: Enriching a gold mine, Nucl. Acids Res. 49 (D1) (2021) D325–D334.
[24] S. Carbon, A. Ireland, C.J. Mungall, et al., AmiGO: Online access to ontology and annotation data, Bioinformatics 25 (2) (2009) 288–289.
[25] T. Hu, Y. Chen, J.W. Kiralis, et al., ViSEN: Methodology and software for visualization of statistical epistasis networks, Genet. Epidem. 37 (3) (2013) 283–285.
[26] J. Yang, C. Hu, H. Hu, et al., QTLNetwork: Mapping and visualizing genetic architecture of complex traits in experimental populations, Bioinformatics 24 (5) (2008) 721–723.
[27] B. Kim, X. Dai, W. Zhang, et al., GWASpro: A high-performance genome-wide association analysis server, Bioinformatics 35 (14) (2019) 2512–2514.
[28] F. Llinares-López, L. Papaxanthos, D. Roqueiro, et al., CASMAP: Detection of statistically significant combinations of SNPs in association mapping, Bioinformatics 35 (15) (2019) 2680–2682.
[29] L.S. Yung, C. Yang, X. Wan, et al., GBOOST: A GPU-based tool for detecting gene—gene interactions in genome–wide case control studies, Bioinformatics 27 (9) (2011) 1309–1310.
[30] W.-L. Ren, Y.-J. Wen, J.M. Dunwell, et al., pKWmEB: Integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study, Heredity 120 (3) (2018) 208–218.
[31] R. Krupa, T. Sliwinski, Z. Morawiec, et al., Association between polymorphisms of the BRCA2 gene and clinical parameters in breast cancer, Exp. Oncol. 31 (4) (2009) 250–251.
[32] A.E. Connor, R.N. Baumgartner, K.B. Baumgartner, et al., Associations between ALOX, COX, and CRP polymorphisms and breast cancer among hispanic and non-hispanic white women: The breast cancer health disparities study, Mol. Carcinog. 54 (12) (2015) 1541–1553.
[33] K.A. Bolanz, M.A. Hediger, C.P. Landowski, The role of TRPV6 in breast carcinogenesis, Mol. Cancer Ther. 7 (2) (2008) 271–279.
[34] F.M. Spinelli, D.L. Vitale, A. Icardi, et al., Hyaluronan preconditioning of monocytes/macrophages affects their angiogenic behavior and regulation of TSG-6 expression in a tumor type-specific manner, FEBS J. 286 (17) (2019) 3433–3449.

[35] S. Yang, C. Jackson, E. Karapetyan, et al., Roles of protein disulfide isomerase in breast cancer, Cancers 14 (3) (2022) 745.
[36] I. Nikas, C. Giaginis, K. Petrouska, et al., EPHA2, EPHA4, and EPHA7 expression in triple-negative breast cancer, Diagnostics 12 (2) (2022) 366.
[37] J. Wu, M.S. Huen, L.-Y. Lu, et al., Histone ubiquitination associates with BRCA1-dependent dna damage response, Mol. Cell. Biol. 29 (3) (2009) 849–860.
[38] G. Molostvov, M. Gachechiladze, A.M. Shaaban, et al., Tspan6 stimulates the chemoattractive potential of breast cancer cells for b cells in an EV-and LXR-dependent manner, Cell Rep. 42 (3) (2023) 112207.
[39] R. Krishnan, M. Murugiah, N.P. Lakshmi, et al., Guanine nucleotide binding protein like-1 (GNL1) promotes cancer cell proliferation and survival through AKT/p21 CIP1 signaling cascade, Mol. Biol. Cell 31 (26) (2020) 2904–2919.
[40] S.A. Whitmore, C. Settasatian, J. Crawford, et al., Characterization and screening for mutations of the growth arrest-specific 11 (GAS11) and C16orf3 genes at 16q24. 3 in breast cancer, Genomics 52 (3) (1998) 325–331.
[41] M.E. Minard, L.-S. Kim, J.E. Price, et al., The role of the guanine nucleotide exchange factor Tiam1 in cellular migration, invasion, adhesion and tumor progression, Breast Cancer Res. Treat. 84 (2004) 21–32.
[42] W.-C. Chang, A.B. Parekh, Close functional coupling between Ca^{2+} release-activated Ca^{2+} channels, arachidonic acid release, and leukotriene C4 secretion, J. Biol. Chem. 279 (29) (2004) 29994–29999.
[43] Z. Zheng, Y. Li, G. Jin, et al., The biological role of arachidonic acid 12-lipoxygenase (ALOX12) in various human diseases, Biomed. Pharmacother. 129 (2020) 110354.

Xin Wang received the BEng degree from the School of Software, Shandong University, Jinan, China, in July 2021, where she is currently pursuing the MPhil degree with the School of Software. Her research interests include data mining and bioinformatics.

Guoxian Yu (BRID: 03183.00.78117) is a professor at the School of Software, Shandong University, Jinan, China. He received the PhD in computer science from South China University of Technology, Guangzhou, China in 2013. His current research interests include artificial intelligence, data mining and bioinformatics. He has served as associate editor for *Neurocomputing, Interdisciplinary Sciences: Computational Life Sciences, Frontier in Genetics*.

Jun Wang (BRID: 09837.00.70895) is a professor with at the Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan, China. She received BSc degree in computer science, MEng degree in computer science and PhD in artificial intelligence from Harbin Institute of Technology, Harbin, China in 2004, 2006 and 2010, respectively. Her current research interests include machine learning, data mining and their applications in bioinformatics.