Received: 20 November 2020 | Revised: 22 March 2021 | Accepted: 23 March 2021

DOI: 10.1002/ece3.7572

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

MultiGWAS: An integrative tool for Genome Wide Association Studies in tetraploid organisms

Luis Garreta[1] | Ivania Cerón-Souza[1] | Manfred Ricardo Palacio[2] | Paula H. Reyes-Herrera[1]

[1]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI Tibaitatá, Bogota, Colombia

[2]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI El Mira, Tumaco, Colombia

Correspondence
Paula H. Reyes-Herrera, Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI Tibaitatá, Kilómetro 14, Vía a Mosquera, 250047 Bogota, Colombia.
Email: phreyes@agrosavia.co

Funding information
AGROSAVIA; Departamento Administrativo de Ciencia, Tecnología e Innovación, Grant/Award Number: 811-2019

Abstract

The genome-wide association studies (GWASs) are essential to determine the genetic bases of either ecological or economic phenotypic variation across individuals within populations of the model and nonmodel organisms. For this research question, the GWAS replication testing different parameters and models to validate the results' reproducibility is common. However, straightforward methodologies that manage both replication and tetraploid data are still missing. To solve this problem, we designed the MultiGWAS, a tool that does GWAS for diploid and tetraploid organisms by executing in parallel four software packages, two designed for polyploid data (GWASpoly and SHEsis) and two designed for diploid data (GAPIT and TASSEL). MultiGWAS has several advantages. It runs either in the command line or in a graphical interface; it manages different genotype formats, including VCF. Moreover, it allows control for population structure, relatedness, and several quality control checks on genotype data. Besides, MultiGWAS can test for additive and dominant gene action models, and, through a proprietary scoring function, select the best model to report its associations. Finally, it generates several reports that facilitate identifying false associations from both the significant and the best-ranked association Single Nucleotide Polymorphisms (SNPs) among the four software packages. We tested MultiGWAS with public tetraploid potato data for tuber shape and several simulated data under both additive and dominant models. These tests demonstrated that MultiGWAS is better at detecting reliable associations than using each of the four software packages individually. Moreover, the parallel analysis of polyploid and diploid software that only offers MultiGWAS demonstrates its utility in understanding the best genetic model behind the SNP association in tetraploid organisms. Therefore, MultiGWAS probed to be an excellent alternative for wrapping GWAS replication in diploid and tetraploid organisms in a single analysis environment.

KEYWORDS

GAPIT, GWAS on polyploids, GWASpoly, SHEsis, SNP, software, TASSEL

1 | INTRODUCTION

The genome-wide association studies (GWASs) comprise statistical tests that identify which variants through the whole genome of a large number of individuals are associated with a specific trait (Begum et al., 2012; Cantor et al., 2010). This methodology started with humans and several model plants, such as rice, maize, and *Arabidopsis* (Cao et al., 2011; Han & Huang, 2013; Korte & Farlow, 2013; Lauc et al., 2010; Tian et al., 2011). Because of the advances in the high-throughput sequencing technology and the decline of the sequencing cost in recent years, there is an increase in the availability of genome sequences of different organisms at a faster rate (Ekblom & Galindo, 2011; Ellegren, 2014). Thus, the GWAS is becoming the standard tool to understand the genetic bases of either ecologically or economically relevant phenotypic variation for both model and nonmodel organisms. This increment includes complex species such as polyploids (Figure 1) (Ekblom & Galindo, 2011; Santure & Garant, 2018).

The GWAS for polyploid species has four related challenges. First, replication among tools is critical to validate GWAS results and capture positive associations (Chanock et al., 2007; De et al., 2014), because each tool has its assumptions (i.e., data quality control, false-positive control, and model optimizations), leading to different results such as *p*-values, significance thresholds, and genomic control inflation factors. Consequently, each tool can be considered an independent environment to replicate a GWAS analysis. For example, the significance threshold for *p*-value changes through four GWAS software (i.e., PLINK, TASSEL, GAPIT, and FaST-LMM) when the sample size varies (Yan et al., 2019). It means that well-ranked SNPs from one package can be ranked differently in another.

Second, there are very few tools focused on integrating several GWAS software to compare different parameters and conditions across them. As far as we know, there are only two software packages with this service in mind: iPAT and easyGWAS. The iPAT allows running in a graphic interface three well-known command-line GWAS software such as GAPIT, PLINK, and FarmCPU (Zhang et al., 2018). However, the output from each package is separated. On the other hand, the easyGWAS allows running a GWAS analysis on the Web using different algorithms and combining several GWAS results. This analysis runs independently of both the computer capacity and the operating system. Nevertheless, it needs either several datasets to obtain the different GWAS results to make replicates or GWAS results already computed. In either case, the results from different algorithms are also separated (Grimm et al., 2017). Thus, although both software iPAT and easyGWAS integrate with different programs or algorithms, an output that allows to compare similitude and differences in the association is missing.

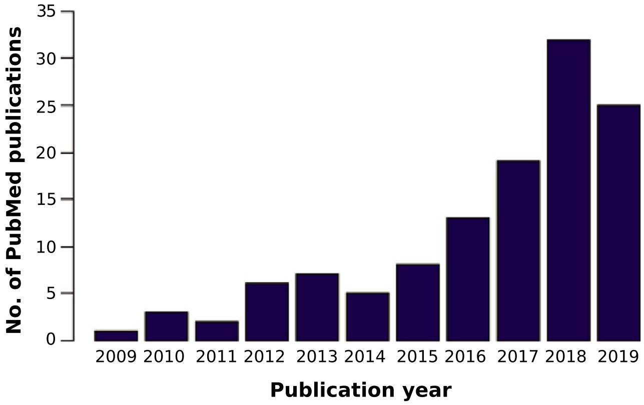


FIGURE 1 The number of peer-reviewed papers that contains the keywords “GWAS” and “polyploid” in the PubMed database between 2009 and 2019

Third, although there are different GWAS software packages available to repeat the analysis under different conditions (Gumpinger et al., 2018), most of them are designed exclusively for the diploid data matrix (Bourke et al., 2018). Therefore, it is often necessary to “diploidizing” the polyploid genomic data to replicate the analysis. This process could withdraw how allele dosage affects the phenotype expression in polyploid species (Ferrão et al., 2018). However, some genome sections of autopolyploid species did not duplicate, leading to loci's disomic inheritance (Dufresne et al., 2014; Lynch & Conery, 2000; Ohno, 1970). Moreover, the inheritance mechanism of most of the polyploid species is still unknown. Therefore, software that accounts for both polyploid and diploid data facilitates analyzing both inheritance types in polyploids.

Finally, for polyploid species, any tool that integrates and compares different gene action models among software is key to understanding how redundancy or complex interaction among alleles affects the phenotype expression and the evolution of new phenotypes (Bourke et al., 2018; Ferrão et al., 2018; Rosyara et al., 2016).

This study developed the MultiGWAS tool that performs GWAS analyses for both diploid and tetraploid species using four software packages in parallel to overcome these challenges. The tool includes GWASpoly (Rosyara et al., 2016) and SHEsis (Shen et al., 2016) that accept polyploid genomic data. Also, it includes GAPIT (Tang et al., 2016) and TASSEL (Bradbury et al., 2017), designed for GWAS in plants, but that in the case of tetraploid data, their use requires “diploidizing” genomic matrix. This wrapping tool deals with different input file formats that come from several polyploid genotypes calling software, including VCF. Besides, MultiGWAS manages data preprocessing, searches associations by running four GWAS software packages in parallel, and creates a score to choose between gene action models in GWASpoly and TASSEL. This study describes the utilities of MultiGWAS and its evaluation through simulation studies and one public GWAS dataset, demonstrating its advantages.

2 | METHOD

The MultiGWAS tool has three main steps: the adjustment, the multi-analysis, and the integration (Figure 2). In the adjustment step, MultiGWAS processes the configuration file. Then, it cleans and filters the genotype and phenotype datasets, and in the case of tetraploids, MultiGWAS “diploidizes” the genomic data. Next, during the multi-analysis, each GWAS tool runs in parallel. Subsequently,

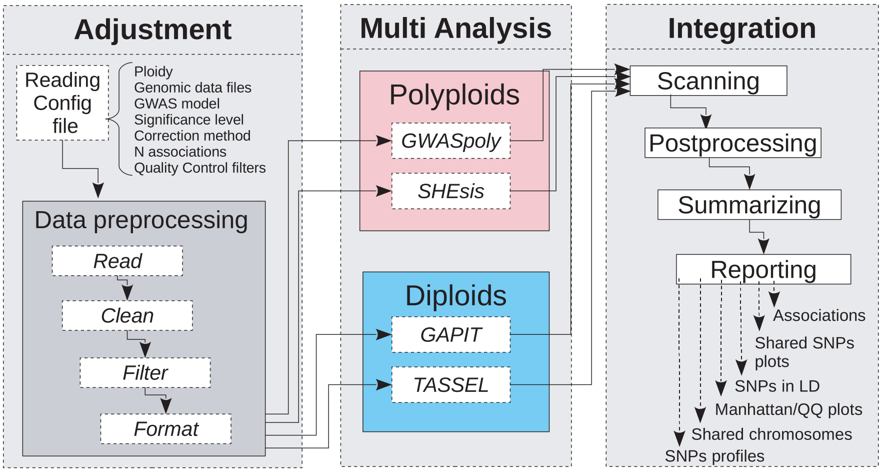


FIGURE 2 MultiGWAS flowchart has three steps: adjustment, multi-analysis, and integration. In the first step, after the input data management upload, MultiGWAS read the configuration file and preprocess the input data (genotype and phenotype dataset). The second step is the GWAS analysis, where MultiGWAS configures and runs the four packages in parallel. Finally, in the third step, MultiGWAS summarizes the results and generates a report using different tabular and graphical visualizations

in the integration step, the MultiGWAS tool scans the output data files from the four packages (i.e., GWASpoly, SHEsis, GAPIT, and TASSEL); postprocesses the data; and finally, it generates a summary of all results that contain associations tables, Venn diagrams showing associated SNPs shared among tools, SNPs in linkage disequilibrium (LD), Manhattan and quantile–quantile (Q–Q) plots, chord diagrams showing the position in the chromosome of the associated SNPs, and finally SNP profiles (see Section 2.3.3).

2.1 | Adjustment stage

MultiGWAS takes as input a configuration file where the user specifies the genomic data and the parameters for the four tools. Once the configuration file is read and processed, the genomic data files (genotype and phenotype) are then cleaned, filtered, and checked for data quality. The output of this stage corresponds to the inputs for the four programs at the multi-analysis stage.

2.1.1 | Reading configuration file

The configuration file includes the following settings that we briefly describe:

Ploidy

Currently, MultiGWAS supports diploid and tetraploid genotypes, where two for diploids and four for tetraploids.

Genomic input files

MultiGWAS mainly uses two input files for the genotype and the phenotype files, and depending on the genotype format (see below) could be needed a map file with marker information (chromosome name, genomic position, reference allele, and alternate allele).

For genotypes aligned with a reference genome, the N chromosomes/contigs displayed in the plots were specified. Chromosomes are sorted in decreasing order by size. Chromosome/contig size is approximated to the highest variant position. And, when chromosome/contig names are numerical or are too large, they are changed with the string prefix "contig" and a sequential number from 1 to *N*.

MultiGWAS allows genotype data in five different formats: "gwaspoly" (Rosyara et al., 2016), "vcf" (Parra-Salazar et al., 2020; Team, 2015), "matrix," "fitpoly" (Zych et al., 2018), and "updog" (Gerard & Ferrão, 2020). The former two formats already include marker information, but the last three formats do not, and they need the additional map file. VCF files are transformed into GWASpoly format using NGSEP 4.0.2 (Tello et al., 2019). A detailed information on these files and formats is available in the GitHub tool (https://github.com/agrosavia-bioinfo/MultiGWAS).

Test model

One of the main factors in detecting real trait–marker associations depends on the gene action models. A unique feature offered by MultiGWAS is to test the different gene action models supported by the tools (see Section 2.2). The additive model is the default model supported by all the tools. However, GWASpoly supports eight, TASSEL three, GAPIT two, and SHEsis only supports one. To integrate the different models in one wrapping tool, MultiGWAS offers three testing options: "additive" (supported by all the tools), "dominant" (supported by all tools except SHEsis), and "all" (for testing all effects supported by the tools, including both additive and dominant effects). In any of the three tests, MultiGWAS reports its top N associations with low *p*-value (with N defined by the user; see below). Taking these associations is straightforward for the former two tests, but not for the last one, as tools report different associations for each gene action model. For this last test, we have created a method that automatically selects the "best" gene action model described in Section 2.3.1.

GWAS model

MultiGWAS works with quantitative phenotypes and runs two types of GWAS, either with control for population structure and relatedness between samples or without any control. The first is known in the literature as the Q + K or *full model*, where Q refers to population structure and K to relatedness, and the second is known as the *naive model* (Sharma et al., 2018).

Both models are linear regression approaches, and GWAS tools used by MultiGWAS implements some variations of those models. The *naive* is modeled with generalized linear models (GLMs, Phenotype + Genotype), and the *full* is modeled with mixed linear models (MLMs, Phenotype + Genotype + Structure + Kinship). The default model used by MultiGWAS is the *full model* (Q + K) (Yu et al., 2006), and the equation is as follows:

$$y = X\beta + S\alpha + Qv + Z\mu + e$$

In this equation, the y is the vector of observed phenotypes. Moreover, the β is a vector of fixed effects other than SNP or population group effects, the α is a vector of SNP effects (quantitative trait nucleotides), the v is a vector of population effects, the μ is a vector of polygene background effects, and the e is a vector of residual effects. Besides, Q, modeled as a fixed effect, refers to the incidence matrix for subpopulation covariates relating y to v, and X, S, and Z are incidence matrices of 1s and 0s relating y to β, α, and μ, respectively.

Genome-wide significance

Genome-wide association studies search SNPs associated with the phenotype in a statistically significant manner. A threshold or significance level α is specified and compared with the p-value derived for each association score. Standard significance levels are 0.01 or 0.05 (Gumpinger et al., 2018; Rosyara et al., 2016), and MultiGWAS uses an α of 0.05 for the four GWAS packages. However, in GWASpoly and TASSEL, which calculates the SNP effect for each genotypic class using different gene action models (see “Multi-analysis stage”), the threshold is adjusted according to these two packages. Therefore, the number of tested markers may be different in each model (see below), impacting the p-value thresholds.

Multiple testing correction

Due to the massive number of statistical tests performed by GWAS, it is necessary to perform a correction method for multiple hypothesis testing and adjusting the p-value threshold accordingly. Two standard methods for multiple hypothesis testing are the false discovery rate and the Bonferroni correction. The latter is the default method used by MultiGWAS, which is one of the most rigorous methods. However, instead of adjusting the p-values, MultiGWAS adjusts the threshold below which a *p-value* is considered significant, that is, α/m, where α is the significance level and m is the number of tested markers from the genotype matrix.

Number of reported associations

The use of stringent significance levels could discard many p-value associations closer to significant threshold, generating a high number of false negatives (Kaler & Purcell, 2019; Thompson et al., 2011). To avoid this problem, MultiGWAS provides the option to specify the number of best-ranked associations (lower p-values), adding the corresponding p-value to each association found. In this way, it is possible to enlarge the number of results and their replicability across the different programs. Nevertheless, the report displays each association with its corresponding p-value.

Quality control filters

A control step is necessary to check the input data for the genotype or phenotype errors or low quality, leading to spurious GWAS results. MultiGWAS provides the option to select and define thresholds for the following filters that control the data quality: minor allele frequency (MAF), individual missing rate (MIND), SNP missing rate (GENO), and Hardy–Weinberg threshold (HWE). All of these filters are built-in implementations of MultiGWAS, except the HWE for tetraploids:

- *MAF of x:* filters out SNPs with minor allele frequency below x (default 0.01);
- *MIND of x:* filters out all individuals with missing genotypes exceeding x*100% (default 0.1);
- *GENO of x:* filters out SNPs with missing values exceeding x*100% (default 0.1);
- *HWE of x:* filters out SNPs with a p-value below the x threshold in the Hardy–Weinberg equilibrium exact test. In the case of tetraploid genotypes, this calculation is taken from SHEsis (Shen et al., 2016).

GWAS tools

List of the four GWAS software names to run and integrate into MultiGWAS analysis are as follows: GWASpoly and SHEsis (designed for polyploid data), and GAPIT and TASSEL (designed for diploid data).

Linkage Disequilibrium threshold (R^2)

It is defined as user-defined squared correlation threshold (R) above which a pair of SNPs is considered to be in LD (see Section 2.3.3 for details).

2.1.2 | Data preprocessing

Once the configuration file is processed, the genomic data are read and cleaned by selecting individuals present in both genotype and phenotype. Then, MultiGWAS removes individuals and SNPs with low quality following the previously selected quality control filters and their thresholds.

At this point, the format "ACGT" suitable for the polyploid software GWASpoly and SHEsis is "diploidized" for GAPIT and TASSEL. The homozygous tetraploid genotypes are converted to diploid: AAAA→AA, CCCC→CC, GGGG→GG, and TTTT→TT. Moreover, for tetraploid heterozygous genotypes, the conversion depends on the reference and alternate alleles calculated for each position (e.g., AAAT→AT, …,CCCG→CG).

After this process, MultiGWAS converts the genomic data, genotype, and phenotype datasets to the specific formats required for each of the four GWAS packages.

2.2 | Multi-analysis stage

As described in Section 2.1.1, MultiGWAS can run two types of GWAS: *naive* without any genotype data control and *full* with control for population structure and relatedness. GWASpoly, GAPIT, and TASSEL support both models. However, SHEsis supports only the *naive* model. To control population structure and relatedness in the *full* model, MultiGWAS uses built-in algorithms to calculate both principal components as covariates and kinship among pairs of individuals. A more detailed description of each of the GWAS tools is given below.

2.2.1 | GWASpoly

GWASpoly (Rosyara et al., 2016) is an R package designed for GWAS in polyploid species used in several studies in plants (Berdugo-Cely et al., 2017; Ferrão et al., 2018; Sharma et al., 2018; Yuan et al., 2020). GWASpoly uses a Q + K linear mixed model with biallelic SNPs that account for population structure and relatedness. Also, to calculate the SNP effect for each genotypic class, GWASpoly provides eight gene action models: general, additive, simplex dominant alternative, simplex dominant reference, duplex dominant alternative, duplex dominant, diplo-general, and diplo-additive. Therefore, each gene action model calculates p-values differently. GWASpoly considers the number of obtained p-values that vary among gene action models and thus varies their respective significance threshold (Bonferroni), resulting in different thresholds depending on the model.

MultiGWAS uses GWASpoly version 1.3 with all gene action models available to find associations. The MultiGWAS reports the top *N* best-ranked (the SNPs with lowest *p*-values) that the user specified in the *N* input configuration file. The *full* model used by GWASpoly includes the population structure and relatedness, which are estimated using the first five principal components and the kinship matrix, respectively, both calculated with the GWASpoly built-in algorithms.

2.2.2 | SHEsis

SHEsis (Shen et al., 2016) is a program based on a linear regression model that includes single-locus association analysis for polyploids, among other analyses. However, it has been used mainly by animals and humans, both diploids (Meng et al., 2019; Qiao et al., 2015).

MultiGWAS uses version 1.0, which does not take into account of population structure or relatedness. However, MultiGWAS externally estimates relatedness for SHEsis by excluding individuals with cryptic first-degree relatedness using the kinship matrix calculated by GWASpoly built-in algorithm.

2.2.3 | GAPIT

GAPIT is an R-based program designed for plants. This tool implements the classical MLM for the *full* model correcting by population structure and relatedness. Also, it uses the GLM approach for the *naive* model without any correction (Tang et al., 2016).

GAPIT offers two models of gene action: additive and dominant. For both models, the genotype must be in numerical format. For the additive model, the genotype is implicitly transformed by GAPIT, using 0 for homozygous genotypes with recessive allele combinations, 2 for homozygous genotypes with dominant allele combinations, and 1 for heterozygous genotypes. For the dominant model, MultiGWAS transforms the genotype, using 0 for the two types of homozygous genotypes and 1 for heterozygous genotypes, as indicated by the authors (Tang et al., 2016). MultiGWAS uses the latest version 3, which also implements several state-of-the-art methods developed for statistical genomics (Wang & Zhang, 2020).

2.2.4 | TASSEL

TASSEL is another standard GWAS program developed initially for maize but currently used in several species (Álvarez et al., 2017; Zhang et al., 2018). TASSEL is a java package that runs either using a graphic user interface developed in JAVA or a command-line interface through a Perl pipeline. In MultiGWAS is implemented the Perl pipeline.

For the association analysis, TASSEL includes the general linear model (GLM) for a naive analysis. Moreover, it uses the MLM for a full analysis controlling for population structure, a principal component analysis, and controlling relatedness using a kinship matrix with a centered IBS method with TASSEL built-in algorithms. Moreover, as GWASpoly, TASSEL provides three-gene action models to calculate each genotypic class SNP effect: general, additive, and dominant. Hence, the significance threshold depends on each action model.

2.3 | Integration stage

The outputs resulting from the four GWAS packages are post-processed to identify SNP with either significative *p*-value association or best-ranked association (i.e., with *p*-values close to a significance threshold).

2.3.1 | Selection of best gene action model

MultiGWAS offers three testing options: “additive,” “dominant,” and “all.” Taking the best associations from “additive” and “dominant” tests is straightforward. However, for the option "all," MultiGWAS has a method to select within each tool the "best" gene action model and takes the top associations.

The method works by scoring each gene action model using three criteria: inflation factor (I), shared SNPs (R), and significant SNPs (S), using the following equation:

$$score(M_i) = I_i + R_i + S_i$$

where $score(M_i)$ is the score for the gene action model M_i, with i from $1 \ldots k$, for a GWAS package with k gene action models. I_i is the score for the inflation factor defined as $I_i = 1 - |1 - \lambda(M_i)|$, where $\lambda(M_i)$ is the inflation factor for the M_i model. R_i is the score of the shared SNPs defined as $R_i = \sum_{j=1}^{k} |M_i \sim M_j|$, where $|M_i \sim M_j|$ is the number of SNPs shared between M_i and M_j models, normalized by the maximum number of SNPs shared between all models. And, S_i is the number of significant SNPs of model M_i normalized by the total number SNPs shared among all models.

The score is high when an M_i model has an inflation factor λ close to 1, identifies a high number of shared SNPs, and contains one or more significant SNPs. Conversely, the score is low when the M_i model has an inflation factor λ either low (close to 0) or high ($\lambda > 2$), which identifies a small number of shared SNPs, and contains 0 or few significant SNPs. In any other case, the score results from the balance among the inflation factor, the number of shared SNPs, and the number of significant SNPs.

2.3.2 | Selection of significant and best-ranked associations

MultiGWAS reports two groups of associations from the four GWAS packages: the statistically significant associations with *p*-values below a threshold of significance, and the best-ranked associations with the lowest *p*-values, but not reaching the limit to be statistically significant. However, they are representing interesting associations for further analysis (possible false negatives).

2.3.3 | Integration of results

All four GWAS packages adopted by MultiGWAS use linear regression approaches. However, they often produce different association results for the same input. Computed *p*-values for the same set of SNPs are different between packages. Therefore, SNPs with significant *p*-values for one package may be not significant for the others. Alternatively, well-ranked SNPs in one package may be ranked differently in another.

MultiGWAS integrates the results of the four tools generating six types of outputs that combine graphics and tables to compare, select, and interpret the set of possible SNPs associated with a trait of interest (Figure 3). The unified output is one HTML document that contains the tables and figures to cover all user's needs to present results and includes the following:

Q–Q plots for GWAS associations

The Q–Q plot shows how well most SNPs fit the null hypothesis of no association with the phenotype. Both distributions should coincide, and most SNPs should lie on the red diagonal line. Deviations for many SNPs may reflect inflated *p*-values due to population structure or cryptic relatedness. Nevertheless, few SNPs deviate from the diagonal for a truly polygenetic trait (Power et al., 2016). MultiGWAS adds the top of each Q–Q plot the corresponding inflation factor λ to assess the test statistic inflation degree.

Manhattan plot for GWAS associations

MultiGWAS uses classical Manhattan plots to visualize each package's results. In both plots, the points are the SNPs and their *p*-values are transformed into scores like $-\log_{10}(p\text{ - values})$ (see Figure 3). The Manhattan plot shows the strength of association of the SNPs (y-axis) distributed at their genomic location (*x*-axis), so the higher the score, the stronger the association. MultiGWAS adds distinctive marks to the plot; significant SNPs are above a red line, best-ranked SNPs are above a blue line, and SNPs shared between packages are colored green.

Tables and Venn diagrams for single and shared SNPs

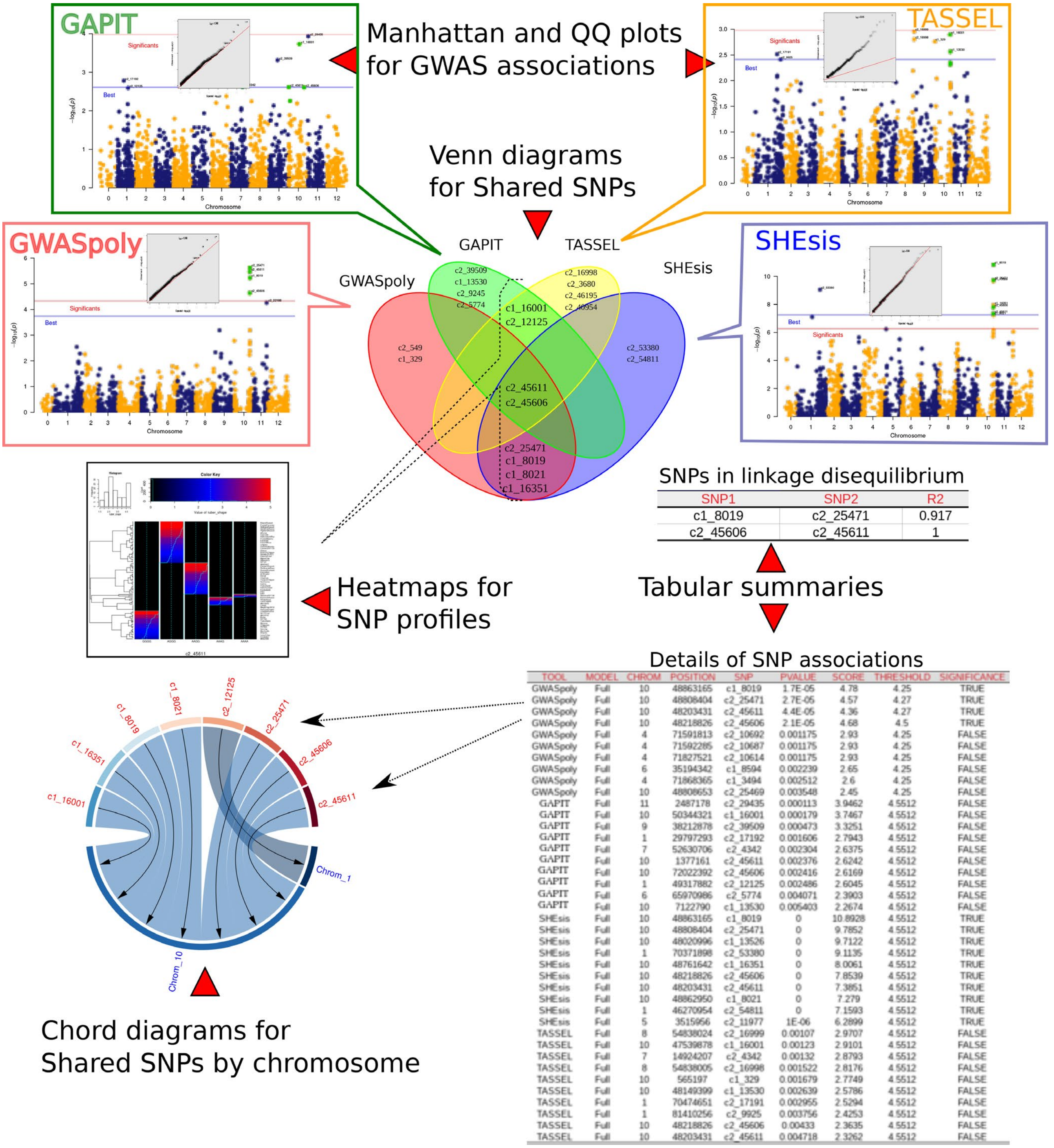
MultiGWAS provides tabular and graphic views to report the best-ranked and significant SNPs identified by the four GWAS packages in an integrative way (Figure 3). Both *p*-values and significance levels have been scaled as $-\log_{10}(p\text{ - values})$ to give high scores to the best statistically evaluated SNPs.

First, best-ranked SNPs correspond to the top-scored *N* SNPs, whether they were assessed significant or not by its package, and with *N* defined by the user in the configuration file. These SNPs appear in both an SNP table and in a Venn diagram. The table lists them by package and sorts them by decreasing score, whereas the Venn diagram emphasizes whether they were best-ranked either in a single package or in several at once (shared). Second, the significant SNPs correspond to the ones valued statistically significant by each package. They also appear in a Venn diagram and the SNP table, marked with significance TRUE (T).

Views of SNPs in LD

MultiGWAS reports a Venn diagram and a table (Figure 7a, b, respectively) for pairs of SNPs with squared correlation equal to or greater than the threshold *R*, where *R* is defined by the user in the configuration file (see Section 2.1.1). MultiGWAS joins the N best associations found for each GWAS packages (SNPs with the lowest *p*-value), and calculates for each pair of SNPs the *R* using the R ldsep library for LD in polyploids (Gerard, 2021). Finally, it summarizes the results in a table with pairs of SNPs per row along with their calculated *R*.

Pairs of SNPs in LD are assigned a new ID (LD_SNP) and reported in a Venn diagram, highlighting the shared SNPs in LD detected between the GWAS software packages. This view allows for quick identification of related SNPs with different names instead of a plain table, as most GWAS packages report their results.



SNPs in linkage disequilibrium

SNP1	SNP2	R2
c1_8019	c2_25471	0.917
c2_45606	c2_45611	1

Details of SNP associations

TOOL	MODEL	CHROM	POSITION	SNP	PVALUE	SCORE	THRESHOLD	SIGNIFICANCE
GWASpoly	Full	10	48863165	c1_8019	1.7E-05	4.78	4.25	TRUE
GWASpoly	Full	10	48808404	c2_25471	2.7E-05	4.57	4.27	TRUE
GWASpoly	Full	10	48203431	c2_45611	4.4E-05	4.36	4.27	TRUE
GWASpoly	Full	10	48218826	c2_45606	2.1E-05	4.68	4.5	TRUE
GWASpoly	Full	4	71591813	c2_10692	0.001175	2.93	4.25	FALSE
GWASpoly	Full	4	71592285	c2_10687	0.001175	2.93	4.25	FALSE
GWASpoly	Full	4	71827521	c2_10614	0.001175	2.93	4.25	FALSE
GWASpoly	Full	6	35194342	c1_8594	0.002239	2.65	4.25	FALSE
GWASpoly	Full	4	71868365	c1_3494	0.002512	2.6	4.25	FALSE
GWASpoly	Full	10	48808653	c2_25469	0.003548	2.45	4.25	FALSE
GAPIT	Full	11	2487178	c2_29435	0.000113	3.9462	4.5512	FALSE
GAPIT	Full	10	50344321	c1_16001	0.000179	3.7467	4.5512	FALSE
GAPIT	Full	9	38212878	c2_39509	0.000473	3.3251	4.5512	FALSE
GAPIT	Full	1	29797293	c2_17192	0.001606	2.7943	4.5512	FALSE
GAPIT	Full	7	52630706	c2_4342	0.002304	2.6375	4.5512	FALSE
GAPIT	Full	10	1377161	c2_45611	0.002376	2.6242	4.5512	FALSE
GAPIT	Full	10	72022392	c2_45606	0.002416	2.6169	4.5512	FALSE
GAPIT	Full	1	49317882	c2_12125	0.002486	2.6045	4.5512	FALSE
GAPIT	Full	6	65970986	c2_5774	0.004071	2.3903	4.5512	FALSE
GAPIT	Full	10	7122790	c1_13530	0.005403	2.2674	4.5512	FALSE
SHEsis	Full	10	48863165	c1_8019	0	10.8928	4.5512	TRUE
SHEsis	Full	10	48808404	c2_25471	0	9.7852	4.5512	TRUE
SHEsis	Full	10	48020996	c1_13526	0	9.7122	4.5512	TRUE
SHEsis	Full	1	70371898	c2_53380	0	9.1135	4.5512	TRUE
SHEsis	Full	10	48761642	c1_16351	0	8.0061	4.5512	TRUE
SHEsis	Full	10	48218826	c2_45606	0	7.8539	4.5512	TRUE
SHEsis	Full	10	48203431	c2_45611	0	7.3851	4.5512	TRUE
SHEsis	Full	10	48862950	c1_8021	0	7.279	4.5512	TRUE
SHEsis	Full	1	46270954	c2_54811	0	7.1593	4.5512	TRUE
SHEsis	Full	5	3515956	c2_11977	1E-06	6.2899	4.5512	TRUE
TASSEL	Full	8	54838024	c2_16999	0.00107	2.9707	4.5512	FALSE
TASSEL	Full	10	47539878	c1_16001	0.00123	2.9101	4.5512	FALSE
TASSEL	Full	7	14924207	c2_4342	0.00132	2.8793	4.5512	FALSE
TASSEL	Full	8	54838005	c2_16998	0.001522	2.8176	4.5512	FALSE
TASSEL	Full	10	565197	c1_329	0.001679	2.7749	4.5512	FALSE
TASSEL	Full	10	48149399	c1_13530	0.002639	2.5786	4.5512	FALSE
TASSEL	Full	1	70474651	c2_17191	0.002955	2.5294	4.5512	FALSE
TASSEL	Full	1	81410256	c2_9925	0.003756	2.4253	4.5512	FALSE
TASSEL	Full	10	48218826	c2_45606	0.00433	2.3635	4.5512	FALSE
TASSEL	Full	10	48203431	c2_45611	0.004718	2.3262	4.5512	FALSE

FIGURE 3 Reports presented by MultiGWAS. Results for GWASpoly, GAPIT, TASSEL, and SHEsis. For each tool, we present, first, a Q–Q plot that assesses the resultant *p*-values, and second, a Manhattan plot with two lines, blue and red, representing the lower limit for the best-ranked and significant SNPs, respectively. Also, we present Venn diagrams that visualize the reproducibility of results; SNPs profiles that visualize SNPs by a two-dimensional representation; chord diagrams that show how the strongest associations are limited to a few chromosomes. Furthermore, we present tabular summaries with details of SNPs in linkage disequilibrium and significant and best-ranked associations

Heat maps for the structure of shared SNPs

For each SNP identified more than once, MultiGWAS provides its SNP profile. It is a two-dimensional heat map representing the SNP that visualizes each trait by individuals and genotypes as rows and columns, respectively. Within the figure, at the left, the individuals are grouped in a dendrogram by their genotype. At the right, there is the name or ID of each individual. At the bottom, the genotypes are ordered from left to right, starting from the major allele to the minor allele (i.e., AAAA, AAAB, AABB, ABBB, BBBB). At the top, there is a description of the trait based on a histogram of frequency (top left) and an assigned color for each numerical phenotype value using a color scale (top right). Thus, each individual appears as a colored line by its phenotype value on its genotype column. For each column, there is a solid cyan line with each column's mean and a broken cyan line that indicates how far the cell deviates from the mean (Figure 3).

Because each MultiGWAS report shows one specific trait at a time, the histogram and color key will remain the same for all the best-ranked SNPs.

Chord diagrams for SNPs by chromosome

The chord diagrams visualize the location across the genome of the best-ranked associated SNPs shared among the four packages and described in the tables. This visualization complements the Manhattan plots from each GWAS package (Figure 3).

3 | AVAILABILITY AND IMPLEMENTATION

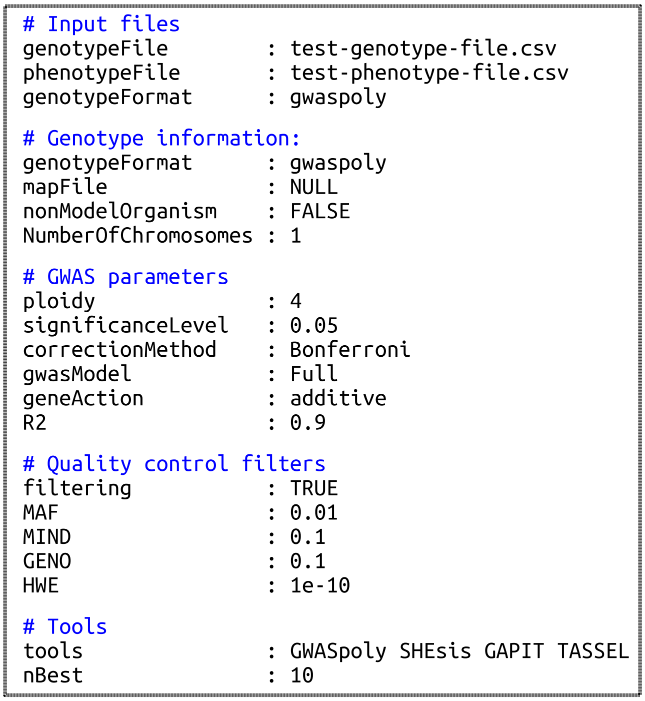
MultiGWAS is a wrapping tool developed in R (R>=3.6). However, it is not an R package or run in the R interface. Instead, it runs on Linux environments because it integrates four external GWAS software packages implemented in different languages. GWASpoly and GAPIT are R packages; SHEsis is a binary program developed in C++, and TASSEL is a Java package that runs through a pipeline implemented in Perl. Consequently, users can run MultiGWAS either by a command-line interface (an R script) or a graphic user interface (a Java application). For detailed instructions and usage examples, refer to https://github.com/agrosavia-bioinfo/MultiGWAS#running-the-examples.

3.1 | Input parameters

MultiGWAS uses a single configuration text file with the values for the main parameters that drive the analysis. If users prefer a text file, it must have the parameter names and values separated by a colon, filenames without blank spaces, TRUE or FALSE values to indicate whether filters are applied, and NULL value to indicate that there is no value for the parameter. This file must have the structure shown in Figure 4. In contrast, if users prefer the GUI application, they can create the configuration file using the GUI described in Section 3.2.2. The input files (genotype/phenotype/map) do not need to be in the working directory, but if this is the case, MultiGWAS needs the absolute path.

3.2 | Installing and using MultiGWAS

MultiGWAS offers different installations from scratch, precompiled versions, a virtual machine, and a docker image. Specific instructions for the different installation types, including a ready-to-use Linux virtual machine (VM) for running MultiGWAS on other platforms (Windows, OS X), are available in the GitHub tool (https://github.com/agrosavia-bioinfo/MultiGWAS).



```
# Input files
genotypeFile        : test-genotype-file.csv
phenotypeFile       : test-phenotype-file.csv
genotypeFormat      : gwaspoly

# Genotype information:
genotypeFormat      : gwaspoly
mapFile             : NULL
nonModelOrganism    : FALSE
NumberOfChromosomes : 1

# GWAS parameters
ploidy              : 4
significanceLevel   : 0.05
correctionMethod    : Bonferroni
gwasModel           : Full
geneAction          : additive
R2                  : 0.9

# Quality control filters
filtering           : TRUE
MAF                 : 0.01
MIND                : 0.1
GENO                : 0.1
HWE                 : 1e-10

# Tools
tools               : GWASpoly SHEsis GAPIT TASSEL
nBest               : 10
```

FIGURE 4 Example of the configuration text file for running MultiGWAS. The input parameters have five sections: (1) input files for genotype and phenotype file paths, (2) genotype information for additional information of the genotype, (3) GWAS parameters for setting the main parameter driving the GWAS analysis, (4) quality control filters to enable/disable and set values for quality control filters on the data, and (5) tools for setting the software to include in the analysis and the number of associations to be reported. The details of each parameter are in Section 2.1.1

3.2.1 | Using the command-line interface

The execution of the CLI tool is simple. In a Linux console, move to the folder where is the configuration file, and type the executable tool's name, followed by the filename of the configuration file, like this: multiGWAS Test01.config.

Then, the tool starts the execution, showing information on the process in the console window. When it finishes, the results are in a new subfolder called *"out-Test01,"* containing a subfolder for each trait in the phenotype file. The results in each trait subfolder include a complete HTML report containing the different views described in the Methods section, the source graphics and tables supporting the report, and the preprocessed tables from the results generated by the four GWAS packages used by MultiGWAS.

3.2.2 | Using the graphical user interface

The interface allows users to save, load, or specify the different input parameters for MultiGWAS in a friendly way (Figure 5). The input parameters correspond to the settings included in the configuration file described in subsection 2.1.1. It executes by calling the following command from a Linux console: jmultiGWAS.

4 | TESTING MULTIGWAS

4.1 | Testing MultiGWAS in real data

We tested MultiGWAS in real data using an open dataset of a diversity panel of phenotype and genotype information for tetraploid potato. These data are part of the USDA-NIFA SolCAP Project (Hirsch et al., 2013). We limited the experiment only for the tuber shape trait, testing both the full model and the naive GWAS model.

4.2 | Testing MultiGWAS in simulated data

We created two different simulated genotyping–phenotyping datasets as an experiment to determine the advantages of running a wrapping tool as MultiGWAS compared with an individual analysis of each of the four GWAS software packages that integrate MultiGWAS (i.e., GWASpoly, SHEsis, GAPIT, and TASSEL). The first simulated dataset had an additive inheritance model, and the second one had a dominant inheritance model.

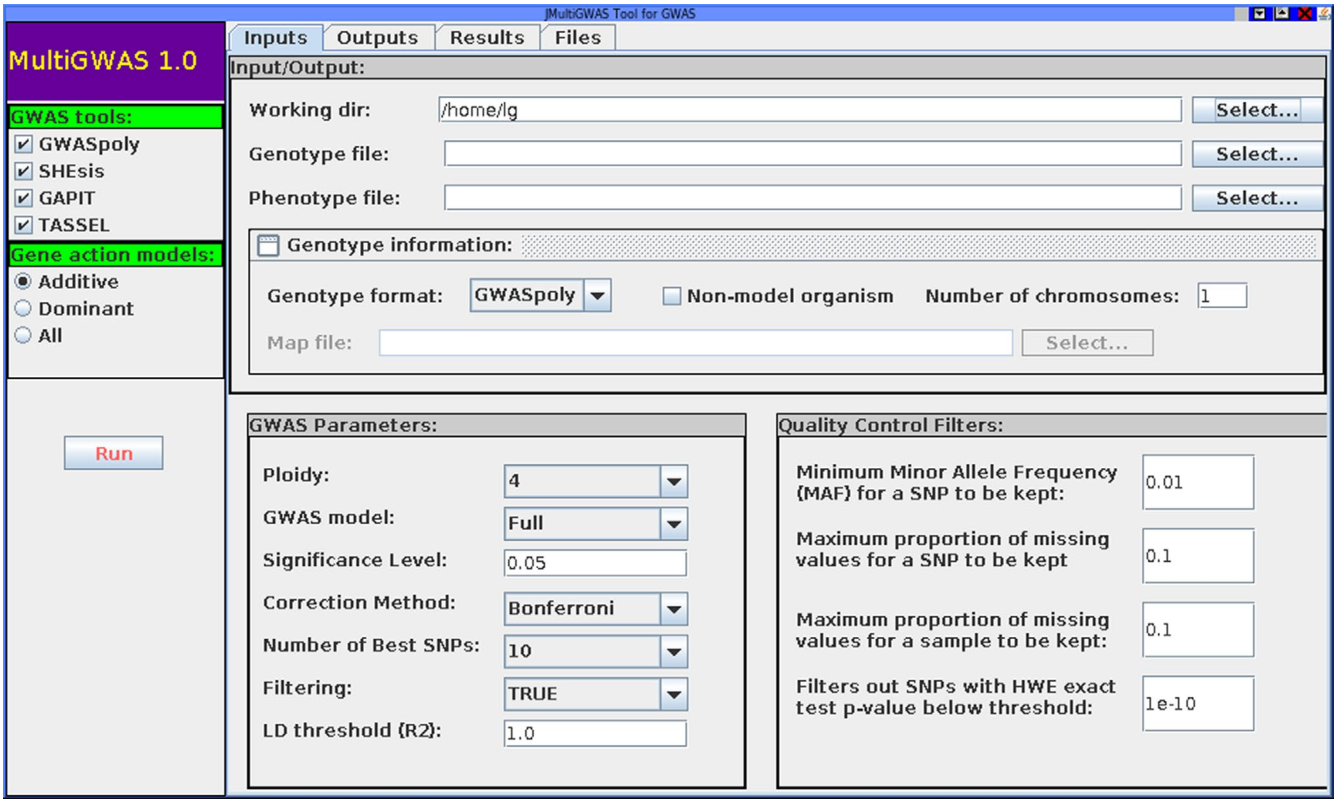
In both simulations, we used as a founder population a subset of 400 SNP and 150 individuals from tetraploid potato data described by Enciso-Rodriguez et al. (2018). To create both simulations of phenotypes, we sampled either additive or dominant effects from a gamma distribution Γ (shape = 0.2 and scale = 5) and specified 10 SNPs as causal SNPs along with their effects under the Phyton 3 SeqBreed software (Pérez-Enciso et al., 2020), inspired in the pSBVB software created to generate polyploid data (Zingaretti et al., 2019).

Both simulated datasets were analyzed in MultiGWAS using the following parameters: gene action, either additive or dominant, false filtering, Bonferroni correction method, and naive GWAS model using the founder genotype and either the additive or dominant phenotype. After MultiGWAS analysis, we summarized the top SNPs (i.e., the N best-ranked SNPs found by the tool) and significant SNPs found by each GWAS tool. Then, we calculated two metrics: true-positive rates (TPRs) and true- negative rates (TNRs) expressed in the following equations:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

where TP is the number of SNPs correctly identified as causal SNPs, and FN is the number of SNPs incorrectly identified as noncausal SNPs.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$



GWAS tools:
☑ GWASpoly
☑ SHEsis
☑ GAPIT
☑ TASSEL

Gene action models:
● Additive
○ Dominant
○ All

Input/Output:
Working dir: /home/lg
Genotype file:
Phenotype file:

Genotype information:
Genotype format: GWASpoly
□ Non-model organism
Number of chromosomes: 1
Map file:

GWAS Parameters:
Ploidy: 4
GWAS model: Full
Significance Level: 0.05
Correction Method: Bonferroni
Number of Best SNPs: 10
Filtering: TRUE
LD threshold (R2): 1.0

Quality Control Filters:
Minimum Minor Allele Frequency (MAF) for a SNP to be kept: 0.01
Maximum proportion of missing values for a SNP to be kept: 0.1
Maximum proportion of missing values for a sample to be kept: 0.1
Filters out SNPs with HWE exact test p-value below threshold: 1e-10

FIGURE 5 Main view of the MultiGWAS graphical user interface. The interface has a toolbar at the left side and four tabs at the top. In the toolbar, users can select the GWAS tools (GWASpoly, SHEsis, TASSEL, and GAPIT). In the Input tab, at top, users can specify the working directory where outputs will be saved, along with genotype, phenotype, and additional information of the genotype. And at bottom, users can set the GWAS parameters and quality control filters. The Output tab shows the execution of each process. In the Results tab, users can browse the HTML report of the current analysis generated by the tool. Finally, in the Files tab, users can browse the source files of each software and access the produced data across the analysis. The details of each parameter are described in Section 2.1.1

where TN is the number of SNPs correctly identified as noncausal SNPs, and FP is the number of SNPs incorrectly identified as causal SNPs.

5 | RESULTS

5.1 | MultiGWAS performance in real data

We run MultiGWAS for the tuber shape of the tetraploid potato dataset using a full GWAS model controlling the population structure and relatedness (Hirsch et al., 2013).

The full GWAS analysis found several associated SNPs (table of Figure 6a). From them, three SNPs named as c2_25471, c2_45606, and c2_45611 were detected from top SNPs across the four GWAS packages (central intersection in Figure 6a). Two SNPs, named as c1_8019 and c2_25471, were identified as significant by the polyploid packages GWASpoly and SHEsis (Figure 6b). Previous association studies also reported these SNPs where the SNP c1_8019 is associated with potato tuber shape and eye depth traits (Rosyara et al., 2016; Sharma et al., 2018), while the SNPs c2_45606 and c2_45611 are associated with eye depth (Totsky et al., 2020).

MultiGWAS strengthened the replicability of these associated SNPs by the four GWAS packages. Also, the LD analyzed confirmed this replicability. Furthermore, when the naive GWAS model was used to analyze the same dataset, MultiGWAS showed all four tools simultaneously detected the SNP c1_8019 as a significant associated SNP, highlighting it as a reliable association (see Supplemental Materials S1 and S2 at https://github.com/agrosavia-bioinfo/multiGWAS/tree/master/docs.)

Two pairs of SNPs resulted in LD, c2_8019 with c2_25471 and c2_45606 with c2_45611, named by MultiGWAS as LD_SNP1 and LD_SNP2, respectively (table of Figure 7a). The Venn diagram (Figure 7b) shows that almost one SNP of the pairwise SNPs in LD

(a)

TOOL	MDL	GC	SNP	CHR	POS	PVAL	SCR	THR	SGN
GWASpoly	additive	1.02	c1_8019	10	48863165	0.00	5.18	4.65	T
GWASpoly	additive	1.02	c2_25471	10	48808404	0.00	4.81	4.65	T
GWASpoly	additive	1.02	c2_45611	10	48203431	0.00	3.88	4.65	F
GWASpoly	additive	1.02	c2_549	9	16527499	0.00	3.82	4.65	F
GWASpoly	additive	1.02	c2_45606	10	48218826	0.00	3.76	4.65	F
GWASpoly	additive	1.02	c2_9925	1	81410256	0.00	3.23	4.65	F
GWASpoly	additive	1.02	c1_8021	10	48862950	0.00	3.18	4.65	F
GWASpoly	additive	1.02	c2_26504	9	21610960	0.00	2.86	4.65	F
SHEsis	additive	3.93	c1_8019	10	48863165	0.00	11.46	5.39	T
SHEsis	additive	3.93	c1_13526	10	48020996	0.00	10.80	5.39	T
SHEsis	additive	3.93	c2_53380	1	70371898	0.00	10.61	5.39	T
SHEsis	additive	3.93	c2_25471	10	48808404	0.00	9.63	5.39	T
SHEsis	additive	3.93	c2_54811	1	46270954	0.00	9.55	5.39	T
SHEsis	additive	3.93	c2_45606	10	48218826	0.00	8.20	5.39	T
SHEsis	additive	3.93	c2_11977	5	3515956	0.00	8.17	5.39	T
SHEsis	additive	3.93	c2_45611	10	48203431	0.00	7.79	5.39	T
GAPIT	additive	0.91	c2_25471	10	48808404	0.00	3.47	4.69	F
GAPIT	additive	0.91	c1_16001	10	47539878	0.00	3.40	4.69	F
GAPIT	additive	0.91	c2_45611	10	48203431	0.00	3.39	4.69	F
GAPIT	additive	0.91	c2_45606	10	48218826	0.00	3.39	4.69	F
GAPIT	additive	0.91	c1_8019	10	48863165	0.00	3.02	4.69	F
GAPIT	additive	0.91	c1_8021	10	48862950	0.00	2.93	4.69	F
GAPIT	additive	0.91	c1_16351	10	48761642	0.00	2.93	4.69	F
GAPIT	additive	0.91	c1_329	10	565197	0.00	2.74	4.69	F
TASSEL	additive	1.01	c1_14512	1	43409359	0.00	2.53	4.27	F
TASSEL	additive	1.01	c2_46195	1	64259758	0.01	2.28	4.27	F
TASSEL	additive	1.01	c1_10202	4	70396887	0.01	2.28	4.27	F
TASSEL	additive	1.01	c1_16001	10	47539878	0.01	2.27	4.27	F
TASSEL	additive	1.01	c2_45606	10	48218826	0.01	2.21	4.27	F
TASSEL	additive	1.01	c2_45611	10	48203431	0.01	2.18	4.27	F
TASSEL	additive	1.01	c2_3680	11	39908878	0.01	2.15	4.27	F
TASSEL	additive	1.01	c2_25471	10	48808404	0.01	2.06	4.27	F

Column headers: MDL: Model, **IF**: Inflation factor, **SNP:** marker name, **CHR**: Chromosome, **PVAL:** *p-value,* **SCR:** score as -log10 (*p-value*), **THR:** significance threshold as -log10 (α / *m*), where α is the significance level, and *m* is the number of tested markers, and **SGN:** significance threshold as true (T) or false (F) whether score > threshold or not.

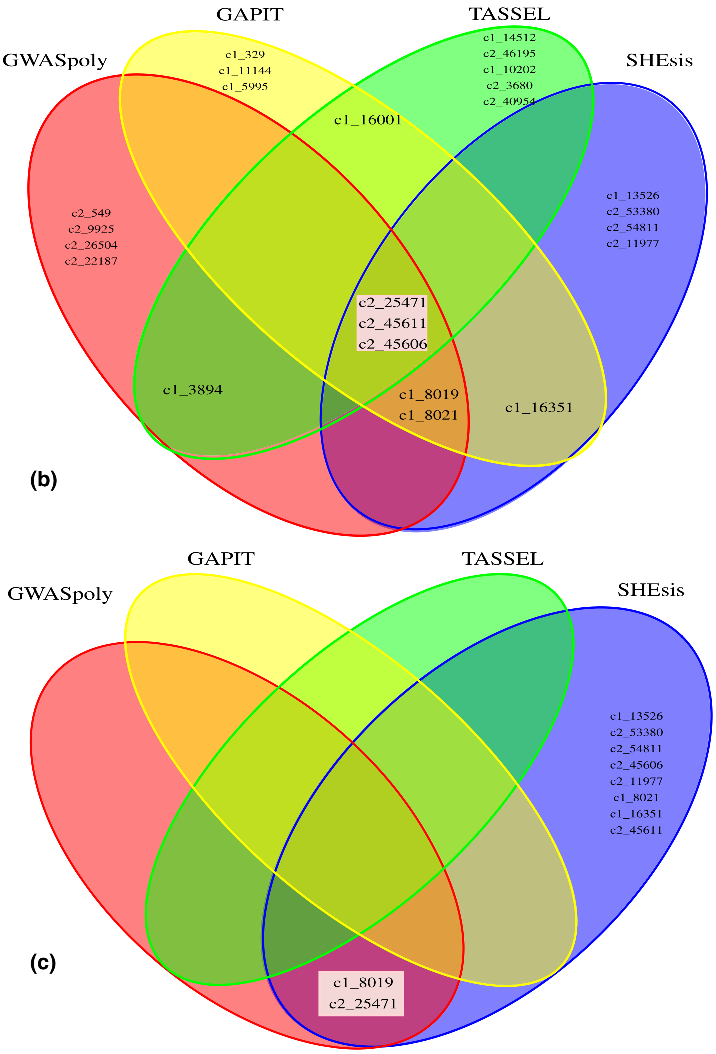
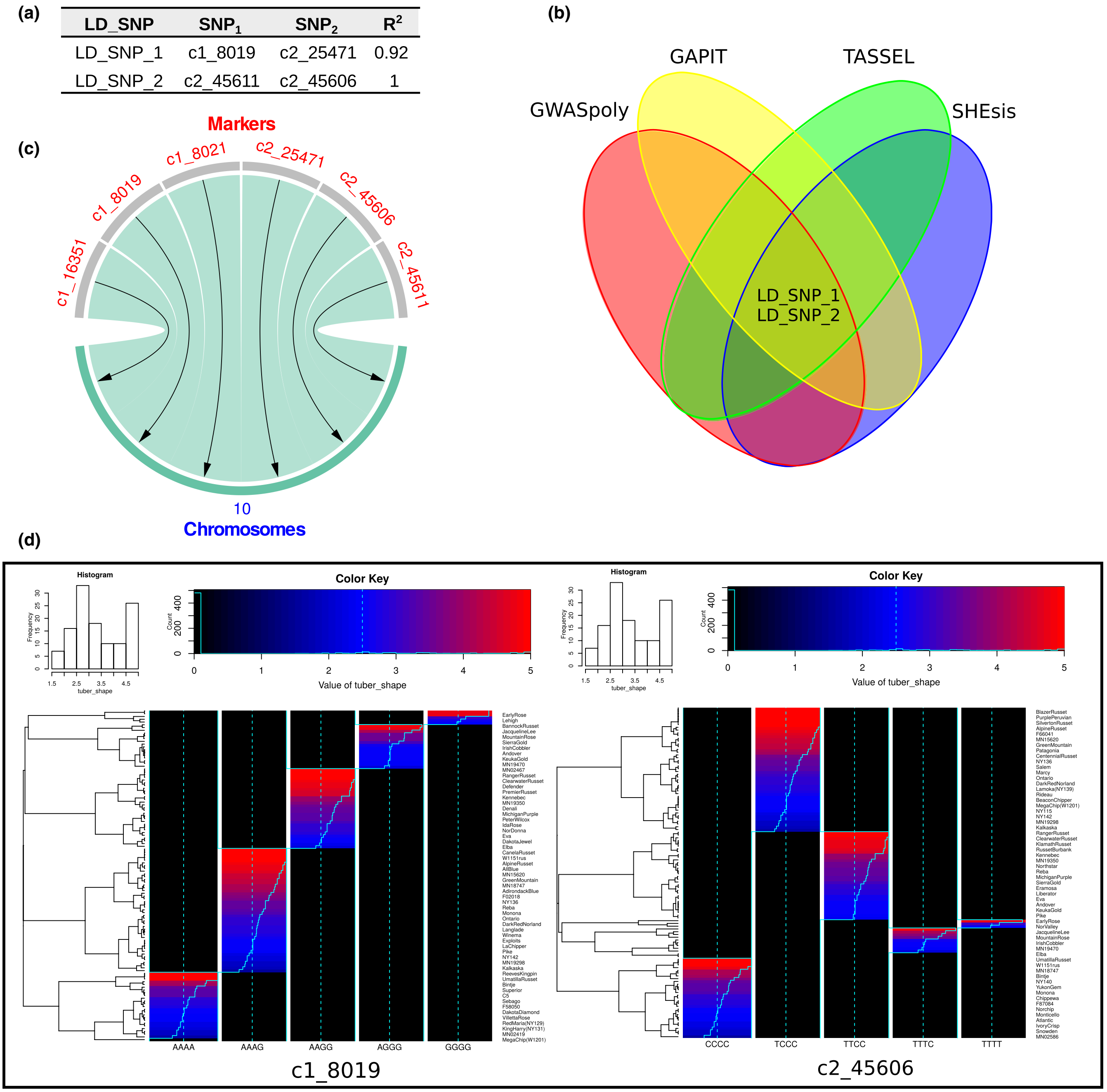


FIGURE 6 Shared SNP views. Tabular and graphical views of SNP associations identified by one or more GWAS packages (shared SNPs). SNPs identified by all packages are marker with red background in all figures. (a) Table with details of the $N = 8$ best-ranked SNPs from each GWAS package. Each row corresponds to a single SNP. (b) Venn diagram of the best-ranked SNPs. SNPs identified by all packages are in the central intersection. Other shared SNPs are in both upper central and lower central intersections. (c) Venn diagram of the significant SNPs (score threshold)



LD_SNP	SNP_1	SNP_2	R^2
LD_SNP_1	c1_8019	c2_25471	0.92
LD_SNP_2	c2_45611	c2_45606	1

FIGURE 7 SNPs in LD, SNPs by chromosome, and SNP structure. MultiGWAS brings different views to figure out SNP relationships and structure. (a) Table of pairs of SNPs in LD with columns LD_SNP for an ID given by MultiGWAS to the pair of SNPs; SNP1 and SNP2 for the names of SNPs; and R^2 for the squared correlation value between SNPs. (b) Venn diagram of pairwise SNPs in LD. LD_SNPs in the center show that the four packages simultaneously detected at least one SNP from the pairwise SNPs shown in the table. (c) Chord diagram showing that best-ranked SNPs are located in chromosome 10. (d) SNP profiles showing the structure of one of the pairwise SNPs in LD

was detected by the four GWAS packages, showing the replicability of the SNPs in the four packages. Moreover, the chord diagrams show that most of the best-ranked SNPs were in chromosome 10 (Figure 7c). Finally, the best-ranked SNP's heat maps show visible differences that related the association of the genotype with the phenotype for tuber shape (Figure 7d).

The Manhattan plot for each GWAS package showed that four packages found that the associated SNP location (i.e., SNP above the blue line) was chromosome 10 (Figure 8). GWASpoly and SHEsis find significant SNPs (above the red line). Both SNP groups, the strong associated and the significant, are present in both the shared table and Venn diagram (Figure 6).

Additionally, for most GWAS packages, except for SHEsis, the majority of observed *p*-values corresponded to the expected *p*-values, as it is shown in the Q–Q plots generated from the associations found for each package (Q–Q plots above Manhattan plots in Figure 8). For SHEsis, its genomic inflation factor λ was far above 1.0, meaning that its calculated scores were inflated, and explaining because SHEsis does not control for population structure and relatedness.

5.2 | MultiGWAS performance in simulated data

For MultiGWAS, we present the results using different sets to evidence the effect of replicability in the performance (MultiGWAS_1: predicted by one software package, MultiGWAS_2: predicted by two software packages, MultiGWAS_3: predicted by three software packages, and MultiGWAS_4: predicted by four software packages).

For the additive effect simulation, GWASpoly (green) and SHEsis (blue) had the best performance based on true-positive rate (TPR) and true-negative rate (TNR) in the detection of the best-ranked SNP. The two diploid software packages GAPIT and TASSEL have similar results but lower performance in both statistics. In parallel, MultiGWAS performance changes depending on the number of software involved in the predicted SNP intersection; the TPR was progressively lower, and TNR was progressively higher. Consequently, in the two more restrictive cases (i.e., the intersection of predicted SNP by three and all the four software packages, MultiGWAS_3 and MultiGWAS_4, respectively), the TPR was similar to that obtained by TASSEL and GAPIT. However, the TNR was higher than even GWASpoly and SHEsis (Figure 9a).

For the dominant effect simulation, GAPIT (cyan) tends to have a higher TPR than the other three software packages. Moreover, SHEsis had a lower value of both TPR and TNR since it was designed only to detect associations with additive effects. Comparing these four software's performance with a wrapping tool as MultiGWAs, it had a similar performance to the additive effect simulation. As the more restrictive the intersection is, the TPR was progressively lower and TNR was progressively higher. However, in the two more restrictive cases (i.e., the intersection of predicted SNP by three and all the four software packages, MultiGWAS_3 and MultiGWAS_4, respectively), the TNR was higher than all the four software packages (Figure 9b).

In the case of significant SNPs, for additive effects, SHEsis (blue) has the highest TPR but has the lowest TNR, suggesting that SHEsis probably is overestimating the significative p-value association. Therefore, true and false associations are reaching the

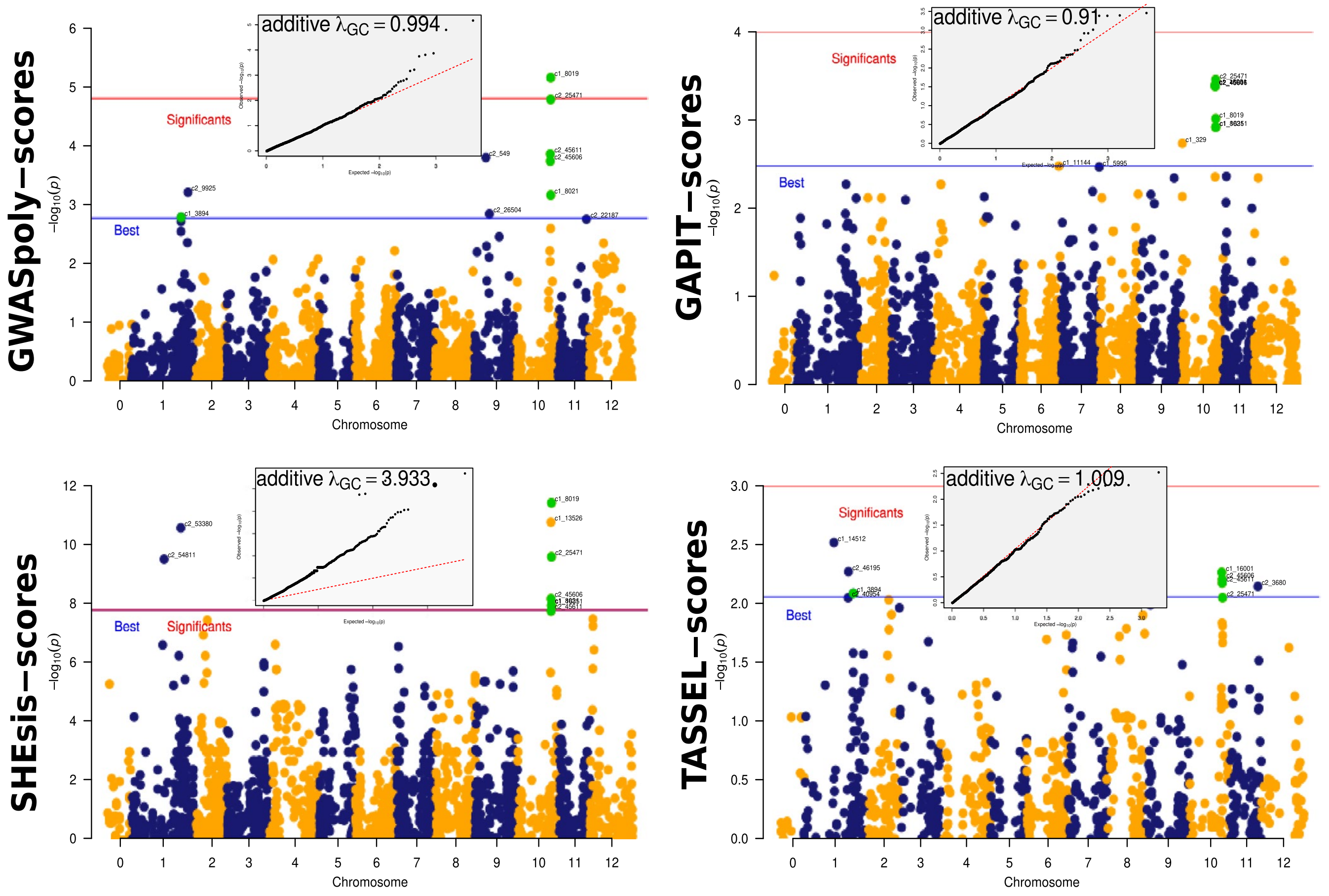


FIGURE 8 Associations for each GWAS package. MultiGWAS shows the associations identified by the four GWAS packages using Manhattan and Q–Q plots. The tetraploid potato data showed several SNPs shared between the four GWAS packages (green dots). The best-ranked SNPs are above the blue line, but only GWASpoly and SHEsis identified significant associations (SNPs above the red line) for this dataset. However, the inflation factor given by SHEsis is too high ($\lambda = 3.9$, at the top of the Q–Q plot), which is observed by the high number of SNPs deviating from the red diagonal of the Q–Q plot

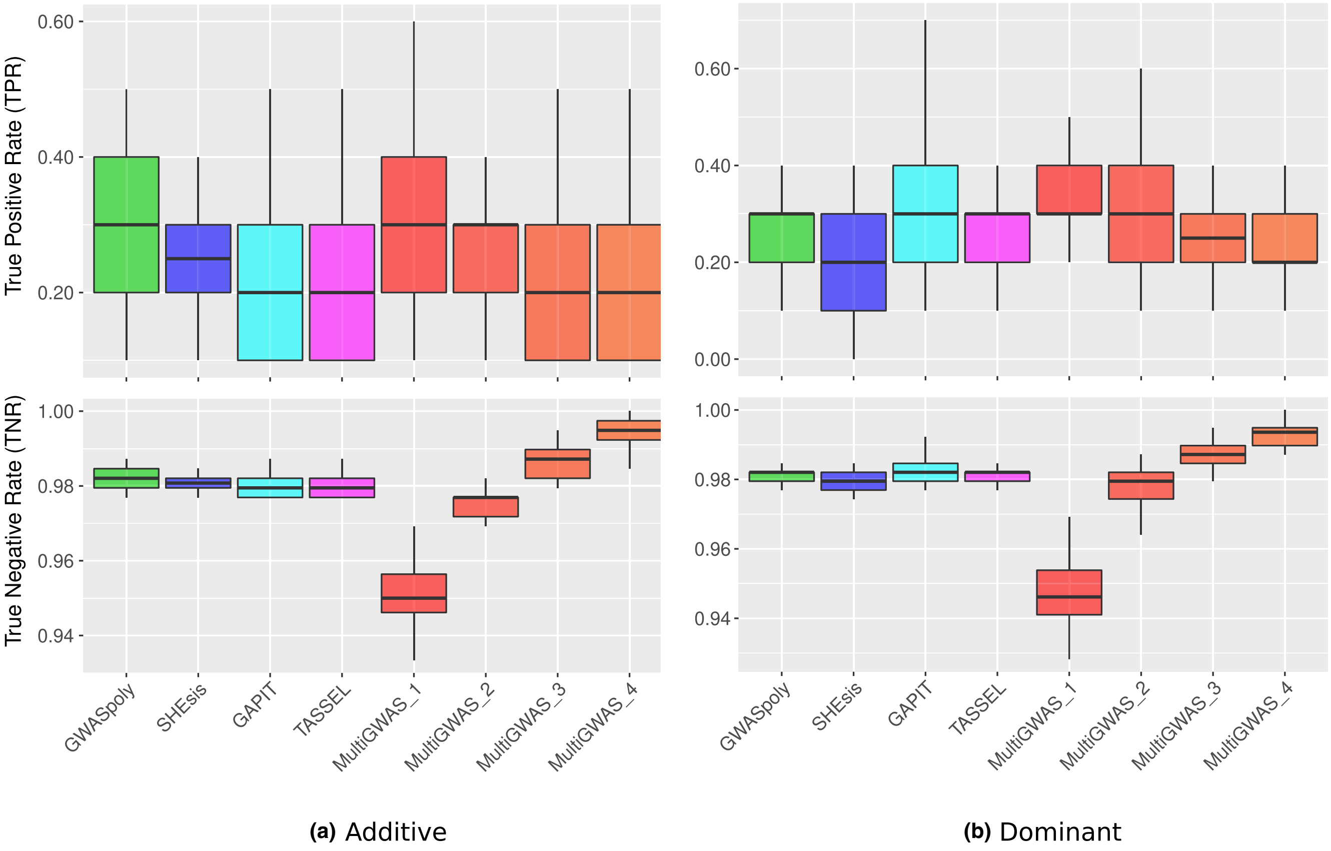


FIGURE 9 Boxplots for the true-positive rates (TPRs) and true-negative rates (TNRs) for the top SNPs (i.e., the N best-ranked) identified for each GWAS software using simulated datasets under either (a) additive or (b) dominant inheritance model after 50 replicates. Each panel compares MultiGWAS with each of the four software packages that integrate it (i.e., GWASpoly, SHEsis, GAPIT, and TASSEL). The MultiGWAS results are separated into four groups: MultiGWAS_1, predicted by one tool; MultiGWAS_2, predicted by two tools; MultiGWAS_3: predicted by three tools; and MultiGWAS_4: predicted by four tools

significance threshold. In comparison, MultiGWAS_4, GWASpoly, and GAPIT are more conservative, with closer TPR but high TNR (Figure 10a).

For dominant effect simulation, GWASpoly had the higher TPR with a lower TNR. Thus, this software was the most sensitive detecting significative associations, but at the same time, it was one of the least specific. In comparison, MultiGWAS_4 and GAPIT had TPR slightly lower than GWASpoly, but with the highest TNR. This pattern suggests that both are less sensitive in detecting significative association but more specific than GWASpoly. Therefore, MultiGWAS_4 provides very accurate associations (Figure 10b).

6 | DISCUSSION

The reanalysis of both potato and simulated data with MultiGWAS showed that this wrapping tool is handy to improve the GWAS in both diploid and tetraploid species for additive and dominant gene action effects. Through MultiGWAS performance, we could test its effectiveness to answer some of the challenges of analyzing polyploid organisms. They include integrating and replicating among parameters and software, the diploidization of polyploid data, and the incorporation of different inheritance mechanisms (Dufresne et al., 2014).

The main advantage of MultiGWAS is that it replicates the GWAS analysis among four software packages and integrates the results obtained across software, models, and parameters. Therefore, in MultiGWAS, users do not have to choose between specificity and sensitivity because they can observe their effect in the analysis within the same wrapping environment.

Another difficulty for replication among software is the variability of formats for the genomic input data. MultiGWAS receives the genotype data in five different formats, including two software outputs used to call polyploid allele dosage. Currently, the most common format for next-generation sequencing variant data is the VCF (variant call format) (Danecek et al., 2011; Ebbert et al., 2014). One of the advantages of VCF is its versatility in summarizing important genome information for hundreds or thousands of individuals and SNP, including information about ploidy levels. MultiGWAS facilitates users; it receives VCF files as input (but see VarStats tool in VTC) and runs the four GWAS software mentioned by adjusting internally to specific formats.

Moreover, the MultiGWAS is the unique wrapping tool we are aware of that facilitates understanding the effect of diploidizing the

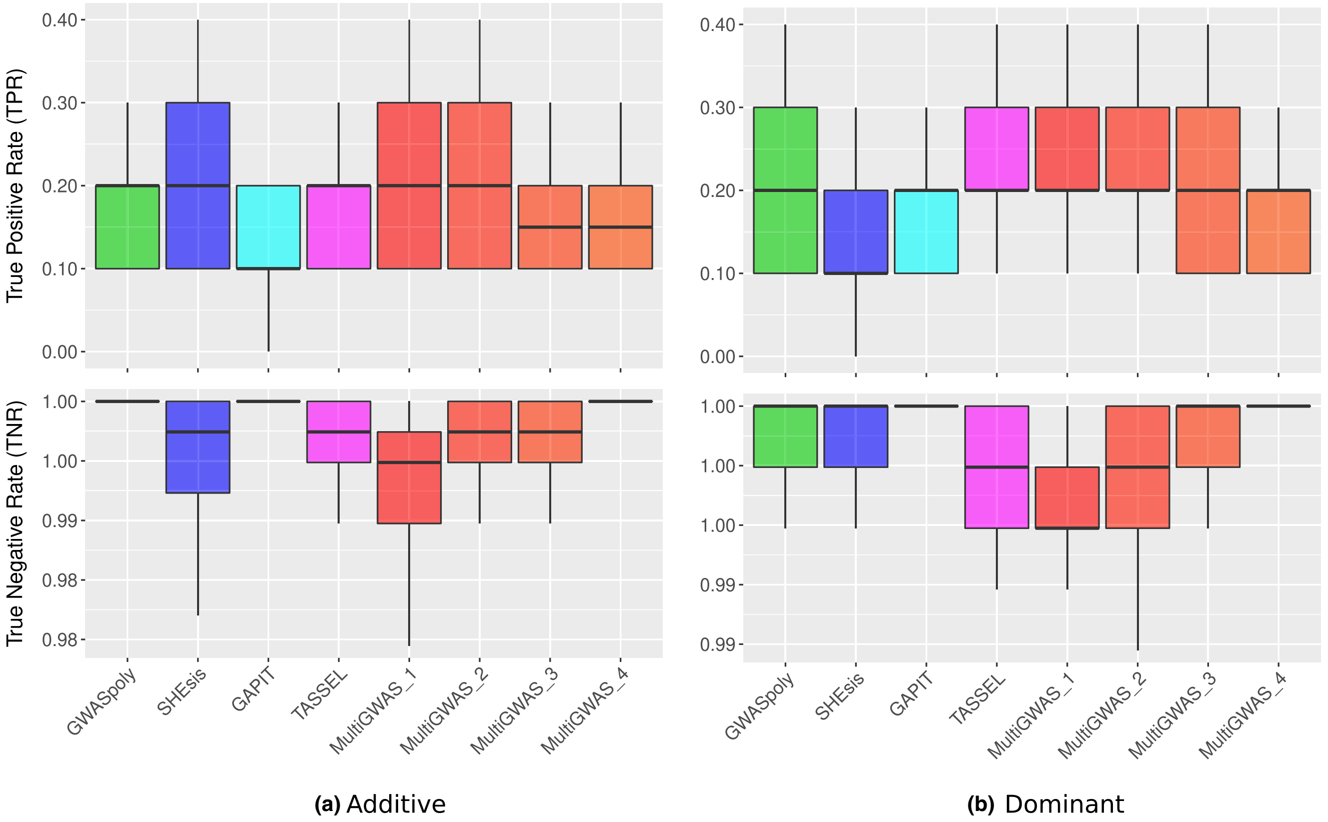


FIGURE 10 Boxplots for the true-positive rates (TPRs) and true-negative rates (TNRs) for the identification of significant SNP identified for each GWAS software using simulated datasets under either (a) additive or (b) dominant inheritance model after 50 replicates. Each panel compares MultiGWAS with each of the four software packages that integrate it (i.e., GWASpoly, SHEsis, GAPIT, and TASSEL). The MultiGWAS results are separated into four groups: MultiGWAS_1, predicted by one tool; MultiGWAS_2, predicted by two tools; MultiGWAS_3: predicted by three tools; and MultiGWAS_4: predicted by four tools

tetraploid data in the analysis performance directly. The SNP profile allows identifying what the significant associations detected by more than one software are. Furthermore, although MultiGWAS checks for significative SNPs based on the *p*-value, it is essential to go back to the data and check whether the SNP is a true association between the genotype and phenotype. For this purpose, the SNP profile gives visual feedback for the accuracy of the association.

Furthermore, the MultiGWAS allows comparing among the gene action models that offer GWASpoly and TASSEL. GWASpoly (Rosyara et al., 2016) provides models of polyploid gene action models, including additive, diploidized additive, duplex dominant, simplex, and general. On the other hand, TASSEL (Bradbury et al., 2007) also models different gene action types for general, additive, and dominant diploids. To choose among models, we propose an automatic selection of the gene action model for both tools based on a balance between three criteria: the inflation factor, the replicability of identified SNPs, and the significance of identified SNPs. This inflation index is a new tool for comparison that does not offer either GWASpoly or TASSEL. This automatic strategy will help to understand the gene action model for the trait of interest. Although the main focus is on the resultant SNPs, the model has assumptions that reflect a specific phenotype's gene actions.

Finally, MultiGWAS, through the active comparison among models, addresses the search of the inheritance mechanisms by comparing among two software packages designed for polysomic inheritance (Rosyara et al., 2016; Shen et al., 2016) with two software packages designed for disomic inheritance (Bradbury et al., 2007; Purcell et al., 2007.). Understanding the inheritance mechanisms for polyploid organisms is an open question. For autopolyploids, most loci have a polysomic heritage. However, sections of the genome that did not duplicate lead to disomic inheritance for some loci (Dufresne et al., 2014; Lynch & Conery, 2000; Ohno, 1970). Thus, it is a valuable tool for researchers because it looks for significant associations that involve both types of inheritance.

6.1 | Future remarks

The evolution and population genomics of polyploids are an exciting novel area of research. The advancement of next-generation sequencing techniques produces more empirical polyploid data in different model and nonmodel organisms (Ekblom & Galindo, 2011; Ellegren, 2014).

Many assumptions developed for diploids in the GWAS analysis do not apply entirely for polyploids (Dufresne et al., 2014). Fortunately, in the last five years, different models to calculate several parameters for population genomics on polyploids are testing and developing in simulated and empirical data (Blischak et al., 2016; Hardy, 2016; Meirmans et al., 2018).

For MultiGWAS, we started with the most simple ploidy, such as tetraploids. Nevertheless, future MultiGWAS versions should include more complex ploidies than tetraploids and the explicit calculation of parameters either for filtering polyploid data before GWAS analysis or for complementing other population genomics' parameters of the data analyzed.

ACKNOWLEDGMENTS

This research possibly thanks AGROSAVIA, a five-year macroproject entitled *Investigación en conservación, caracterización y uso de los recursos genéticis vegetales*. We thank the Minister of Science, Technology and Innovation of the Republic of Colombia (previously COLCIENCIAS), for supporting the postdoctoral researcher L. Garreta at AGROSAVIA during 2019–2020 under the supervision of ICS and PHRH (Grant Number 811-2019). The editorial of AGROSAVIA gave financial support for this publication. Finally, we thank Andres J. Cortés for his valuable comments to improve this manuscript.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Luis Garreta: Conceptualization (equal); investigation (equal); methodology (equal); software (lead); validation (lead); writing—original draft (equal). **Ivania Cerón-Souza:** Conceptualization (equal); funding acquisition (equal); investigation (equal); methodology (equal); supervision (equal); writing—original draft (equal). **Manfred Ricardo Palacio:** Software (equal); validation (equal). **Paula Reyes-Herrera:** Conceptualization (equal); funding acquisition (equal); investigation (equal); methodology (equal); software (equal); supervision (equal); writing—original draft (equal).

DATA AVAILABILITY STATEMENT

We used two datasets: (1) real data are an open dataset (Hirsch et al., 2013) and (2) simulated data in this case the scripts used to generate this dataset are available at https://github.com/agrosavia-bioinfo/multiGWAS-sim

ORCID

Paula H. Reyes-Herrera https://orcid.org/0000-0003-0502-4266

REFERENCES

Álvarez, M. F., Angarita, M., Delgado, M. C., García, C., Jiménez-Gomez, J., Gebhardt, C., & Mosquera, T. (2017). Identification of novel associations of candidate genes with resistance to late blight in *Solanum tuberosum* group Phureja. *Frontiers in Plant Science*, *8*, 1040.

Begum, F., Ghosh, D., Tseng, G. C., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research*, *40*(9), 3777–3784.

Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., & Yockteng, R. (2017). Genetic diversity and association mapping in the colombian central collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. *PLoS One*, *12*(3), e0173039.

Blischak, P. D., Kubatko, L. S., & Wolfe, A. D. (2016). Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Molecular Ecology Resources*, *16*(3), 742–754.

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., & Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Frontiers in Plant Science*, *9*, 513.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635.

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, *86*(1), 6–22.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., & Wang, X. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, *43*(10), 956.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni, J. F., Freimer, N. B., Gerhard, D. S., Gunter, C., & Guttmacher, A. E., … Group, N.-N. W (2007). Replicating genotype-phenotype associations. *Nature*, *447*(7145), 655–660.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R., & Group, G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

De, R., Bush, W. S., & Moore, J. H. (2014). *Bioinformatics challenges in Genome-Wide Association Studies (GWAS)* (pp. 63–81). Springer.

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*(1), 40–69.

Ebbert, M. T., Wadsworth, M. E., Boehme, K. L., Hoyt, K. L., Sharp, A. R., DO'Fallon, B., Kauwe, J. S., & Ridge, P. G. (2014). Variant tool chest: An improved tool to analyze and manipulate variant call format (vcf) files. *BMC Bioinformatics*, *15*(S7), S12.

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*(1), 1–15.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63.

Enciso-Rodriguez, F., Douches, D., Lopez-Cruz, M., Coombs, J., & de los Campos, G.(2018). Genomic selection for late blight and common scab resistance in tetraploid potato (*Solanum tuberosum*). *G3: Genes, Genomes. Genetics*, *8*(7), 2471–2481.

Ferrão, L. F. V., Benevenuto, J., Oliveira, I. D. B., Cellon, C., Olmstead, J., Kirst, M., Resende, M. F. R., & Munoz, P. (2018). Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. *Frontiers in Ecology and Evolution*, *6*, 107.

Gerard, D. (2021). Pairwise linkage disequilibrium estimation for polyploids. *Molecular Ecology Resources*, *21*(4), 1230–1242. https://doi.org/10.1111/1755-0998.13349

Gerard, D., & Ferrão, L. F. V. (2020). Priors for genotyping polyploids. *Bioinformatics*, *36*(6), 1795–1800.

Grimm, D. G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Schölkopf, B., Weigel, D., & Borgwardt, K. M. (2017). easyGWAS: A cloud-based platform for comparing the results of Genome-Wide Association Studies. *The Plant Cell*, *29*(1), 5–19.

Gumpinger, A. C., Roqueiro, D., Grimm, D. G., & Borgwardt, K. M. (2018). Methods and tools in genome-wide association studies. In *Computational Cell Biology* (pp. 93-136). Humana Press.

Han, B., & Huang, X. (2013). Sequencing-based genome-wide association study in rice. *Current Opinion in Plant Biology*, *16*(2), 133–138.

Hardy, O. J. (2016). Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources*, *16*(1), 103–117.

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R. E., Jansky, S., Bethke, P., Douches, D. S., & Buell, C. R. (2013). Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. *G3: Genes, Genomes, Genetics*, *3*(6), 1003–1013.

Kaler, A. S., & Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC Genomics*, *20*(1), 618.

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, *9*(1), 29.

Lauc, G., Essafi, A., Huffman, J. E., Hayward, C., Knežević, A., Kattla, J. J., Polašek, O., Gornik, O., Vitart, V., Abrahams, J. L., Pučić, M., Novokmet, M., Redžić, I., Campbell, S., Wild, S. H., Borovečki, F., Wang, W., Kolčić, I., Zgaga, L., … Rudan, I. (2010). Genomics meets glycomics—the first GWAS study of human n-glycome identifies hnf1α as a master regulator of plasma protein fucosylation. *PLoS Genetics*, *6*(12), e1001256. https://doi.org/10.1371/journal.pgen.1001256

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–1155.

Meirmans, P. G., Liu, S., & van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *Journal of Heredity*, *109*(3), 283–296.

Meng, J., Song, K., Li, C., Liu, S., Shi, R., Li, B., Wang, T., Li, A., Que, H., Li, L., & Zhang, G. (2019). Genome-wide association analysis of nutrient traits in the oyster *Crassostrea gigas*: Genetic effect and interaction network. *BMC Genomics*, *20*(1), 1–14.

Ohno, S. (1970). *Evolution by gene duplication*. Springer.

Parra-Salazar, A., Gomez, J., Lozano-Arce, D., Reyes-Herrera, P. H., & Duitama, J. (2020). Robust and efficient software for reference-free genomic diversity analysis of GBS data on diploid and polyploid species. *bioRxiv*. https://doi.org/10.1101/2020.11.28.402131

Pérez-Enciso, M., Ramírez-Ayala, L. C., & Zingaretti, L. M. (2020). SeqBreed: A python tool to evaluate genomic prediction in complex scenarios. *Genetics Selection Evolution*, *52*(1), 1–9.

Power, R. A., Parkhill, J., & De Oliveira, T. (2016). Microbial genome-wide association studies: Lessons from human GWAS. *Nature Reviews Genetics*, *18*(1), 41–50.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Qiao, H. P., Zhang, C. Y., Yu, Z. L., Li, Q. M., Jiao, Y., & Cao, J. P. (2015). Genetic variants identified by GWAS was associated with colorectal cancer in the Han Chinese population. *Journal of Cancer Research and Therapeutics*, *11*(2), 468–470.

Rosyara, U. R., De Jong, W. S., Douches, D. S., & Endelman, J. B. (2016). Software for genome-wide association studies in autopolyploids and its application to potato. *The Plant Genome*, *9*(2), 1–10.

Santure, A. W., & Garant, D. (2018). Wild gwas—association mapping in natural populations. *Molecular Ecology Resources*, *18*(4), 729–738.

Sharma, S. K., MacKenzie, K., McLean, K., Dale, F., Daniels, S., & Bryan, G. J. (2018). Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, *8*(10), 3185–3202.

Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., & Shi, Y. (2016). SHEsisPlus, a toolset for genetic studies on polyploid species. *Scientific Reports*, *6*, 1–10.

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., Buckler, E. S., & Zhang, Z. (2016). GAPIT Version 2: An enhanced integrated tool for genomic association and prediction. *The Plant Genome*, *9*(2):plantgenome2015.11.0120.

Team, F. F. T. (2015). The Variant Call Format (VCF) Version 4.2 specification. http://github.com/samtools/hts-specs

Tello, D., Gil, J., Loaiza, C. D., Riascos, J. J., Cardozo, N., & Duitama, J. (2019). NGSEP3: Accurate variant calling across species and sequencing protocols. *Bioinformatics*, *35*(22), 4716–4723.

Thompson, J. R., Attia, J., & Minelli, C. (2011). The meta-analysis of genome-wide association studies. *Briefings in Bioinformatics*, *12*(3), 259–269.

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T. R., McMullen, M. D., Holland, J. B., & Buckler, E. S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics*, *43*(2), 159–162.

Totsky, I. V., Rozanova, I. V., Safonova, A. D., Batov, A. S., Gureeva, Y. A., Kochetov, A. V., & Khlestkina, E. K. (2020). Genomic regions of *Solanum tuberosum* L. associated with the tuber eye depth. *Vavilovskii Zhurnal Genetiki i Selektsii*, *24*(5), 465–473.

Wang, J., & Zhang, Z. (2020). GAPIT Version 3: Boosting power and accuracy for genomic association and prediction. *bioRxiv*. https://doi.org/10.1101/2020.11.29.403170

Yan, Y. Y., Burbridge, C., Shi, J., Liu, J., & Kusalik, A. (2019). Effects of input data quantity on genome-wide association studies (GWAS). *International Journal of Data Mining and Bioinformatics*, *22*(1), 19–43.

Yu, J., Pressoir, G., Briggs, W. H., Vroh, I. B., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., & Kresovich, S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208.

Yuan, J., Bizimungu, B., De Koeyer, D., Rosyara, U., Wen, Z., & Lague, M. (2020). Genome-Wide Association Study of resistance to potato common scab. *Potato Research*, *63*(2), 253–266. https://doi.org/10.1007/s11540-019-09437-w

Zhang, S., Chen, X., Lu, C., Ye, J., Zou, M., Lu, K., Feng, S., Pei, J., Liu, C., Zhou, X., Ma, P., Li, Z., Liu, C., Liao, Q., Xia, Z., & Wang, W. (2018). Genome-wide association studies of 11 agronomic traits in cassava (Manihot esculenta crantz). *Frontiers in Plant Science*, *9*(April), 1–15.

Zingaretti, M. L., Monfort, A., & Pérez-Enciso, M. (2019). pSBVB: A versatile simulation tool to evaluate genomic selection in polyploid species. *G3: Genes, Genomes, Genetics*, *9*(2), 327–334.

Zych, K., Gort, G., Maliepaard, C. A., Jansen, R. C., Voorrips, R. E. (2019). FitTetra 2.0 – improved genotype calling for tetraploids with multiple population and parental data support. *BMC Bioinformatics*, *20*(1). http://dx.doi.org/10.1186/s12859-019-2703-y

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Garreta L, Cerón-Souza I, Palacio MR, Reyes-Herrera PH. MultiGWAS: An integrative tool for Genome Wide Association Studies in tetraploid organisms. *Ecol Evol*. 2021;11:7411–7426. https://doi.org/10.1002/ece3.7572