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

L. Garreta[1], I. Cerón-Souza[1], M.R. Palacio[2], and P.H. Reyes-Herrera[1]

[1]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI Tibaitatá, Kilómetro 14, Vía a Mosquera, 250047, Colombia
[2]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI El Mira, Kilómetro 38, Vía Tumaco Pasto, Colombia

August 14, 2020

Abstract

Summary: The Genome-Wide Association Studies (GWAS) are essential to determine the genetic bases of either ecological or economic phenotypic variation across individuals within populations of model and non-model organisms. For this research question, current practice is the replication of the GWAS testing different parameters and models to validate the reproducibility of results. 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, two for polyploid data (GWASpoly and SHEsis) and two for diploids data (PLINK and TASSEL). MultiGWAS has several advantages. It runs either in the command line or in an interface. It manages different genotype formats, including VCF. It executes both the full and naÃŕve models using several quality filters. Besides, it calculates a score to choose the best gene action model across GWASPoly and TASSEL. Finally, it generates several reports that facilitate the identification of false associations from both the significant and the best-ranked association SNP among the four software. We tested MultiGWAS with tetraploid potato data. The execution demonstrated that the Venn diagram and the other companion reports (i.e., Manhattan and QQ plots, heatmaps for associated SNP profiles, and chord diagrams to trace associated SNP by chromosomes) were useful to identify associated SNP shared among different models and

parameters. Therefore, we confirmed that MultiGWAS is a suitable wrapping tool that successfully handles GWAS replication in both diploid and tetraploid organisms.
Contact: phreyes@agrosavia.co
Keywords: GWAS on polyploids, GWASPoly, PLINK, SNP, SHEsis, software, TASSEL

1 Introduction

The Genome-Wide Association Studies (GWAS) comprise statistical tests that identify which variants through the whole genome of a large number of individuals are associated with a specific trait (**cantor2010prioritizing**; **begum2012comprehensive**). This methodology started with humans and several model plants, such as rice, maize, and *Arabidopsis* (**lauc2010genomics**; **tian2011genome**; **cao2011whole**; **korte2013advantages**; **han2013sequencing**). Because of the advances in the next-gen 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 (**ekblom2011applications**; **ellegren2014genome**). 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 non-model organisms. This increment includes complex species such as polyploids (Fig. **??**) (**ekblom2011applications**; **santure2018wild**).

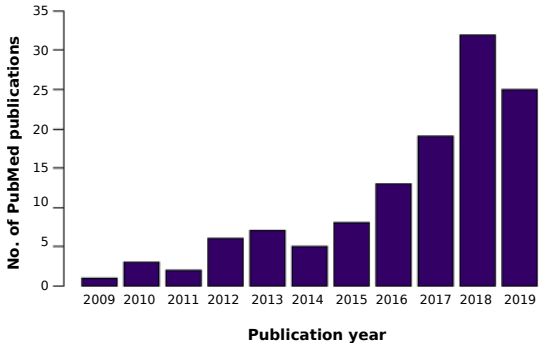


Figure 1: The number of peer-reviewed papers that contains the keywords "GWAS" and "polyploid" in the PubMed database between 2009 and 2019.

The GWAS for polyploid species has fourth related challenges. First, replication is critical to validate GWAS results and capture real associations.This approach involved using different parameters, models, or conditions to test how consistent the results are in the same software or different GWAS tools (**De2014**; **Pearson2008**). However, the performance of different GWAS software could affect the results. For example, the significance threshold for *pvalue* changes through four GWAS software (i.e., PLINK, TASSEL, GAPIT, and FaST-LMM) when the sample size varies (**Yan2019**). It means that well-ranked SNPs from one package can be ranked differently in another.

Second, there are very few tools focused on the integration of several GWAS software, to make comparisons under different parameters and conditions across them. As far as we are aware, there is only two software with this service in mind, which are iPAT and easyGWAS.

The iPAT allows running in a graphic interface three well-known command-line GWAS software such as GAPIT, PLINK, and FarmCPU (**Zhang2018**). 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 (**Grimm2017**). Thus, although both software iPAT and easyGWAS integrate with different programs or algorithms, an output that allows them to compare similitudes and differences in the association is missing.

Third, although there are different GWAS software available to repeat the analysis under different conditions (**Gumpinger2018**), most of them are designed exclusively for the diploid data matrix (**Bourke2018**). Therefore, it is often necessary to "diploidizing" the polyploid genomic data in order to replicate the analysis. The main consequence of this process is missing the complexity of polyploid data to understand how allele dosage affects the phenotype expression (**Ferrao2018**).

Finally, for polyploid species, any tool that integrates and compares different gene action among software is key to understanding how redundancy or complex interaction among alleles affects the phenotype expression and the evolution of new phenotypes (**Bourke2018**; **Rosyara2016**; **Ferrao2018**).

To overcome these challenges, we developed the MultiGWAS tool that performs GWAS analyses for both diploid and tetraploid species using four software in parallel. Our tool includes GWASpoly (**Rosyara2016**) and the SHEsis tool (**Shen2016**) that accept polyploid genomic data, and PLINK (**Purcell2007**) and TASSEL (**Bradbury2007**), designed exclusively for diploids, but that in the case of tetraploid data, their use require "diploidizing" genomic matrix. This wrapping tool deals with different input file formats, including VCF. Besides, manage data preprocessing, search for associations by running four GWAS software in parallel, and create a score to choose between gene action models in GWASPoly and TASSEL. Moreover, create comparative reports from the output of each software to help the user distinguish genuine associations from false positives.

2 Method

The MultiGWAS tool has three main consecutive steps: the adjustment, the multi analysis, and the integration (Fig. **??**). In the adjustment step, MultiGWAS processes the configuration file. Then it cleans and filters the genotype

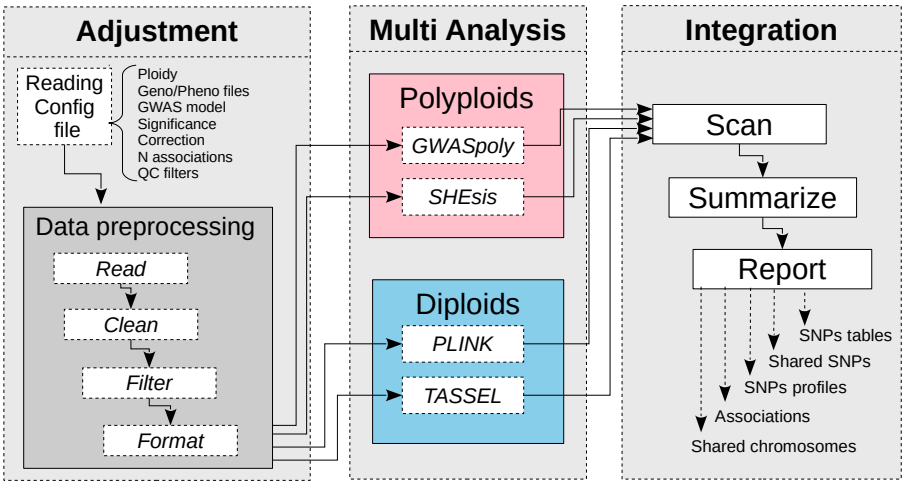


Figure 2: MultiGWAS flowchart has three steps: adjustment, multi analysis, and integration. In the first step, after the input data management upload, MultiGWAS read the configuration file, and preprocessing the input data (genotype and phenotype dataset). The second step is the GWAS analysis, where MultiGWAS configure and run the four packages in parallel. Finally, in the third step, MultiGWAS summarize the results and generate a report using different tabular and graphical visualizations.

and phenotype datasets, and in case of tetraploids, MultiGWAS "diploidize" the genomic data. Next, during the multi analysis, each GWAS tool runs in parallel. Subsequently, in the integration step, the MultiGWAS tool scans the output files from the four packages (i.e., GWASPoly, SHEsis, PLINK, and TASSEL). Finally, it generates a summary of all results that contains score tables, Venn diagrams, SNP profiles, and Manhattan plots.

2.1 Adjustment stage

MultiGWAS takes as input a configuration file where the user specifies the genomics data and the parameters used by 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: Numerical value for the ploidy level of the genotype, currently MultiGWAS supports diploids and tetraploids genotypes (2: for diploids, 4: for tetraploids).

Genotype and phenotype input files: MultiGWAS uses two input files, one for the genotype and one for the phenotype. Genotype data can be input in three different formats, including a matrix format (Fig. **??**.a), a GWASpoly format (**Rosyara2016**) (Fig. **??**.b), and Variant Call Format (VCF) (Fig.**??**.c) which is transformed into GWASpoly format using NGSEP 4.0.2 (**Duitama2019**). The phenotype file contains only one trait with the first column for the sample names and the second column for the trait values (Fig. **??**.d).

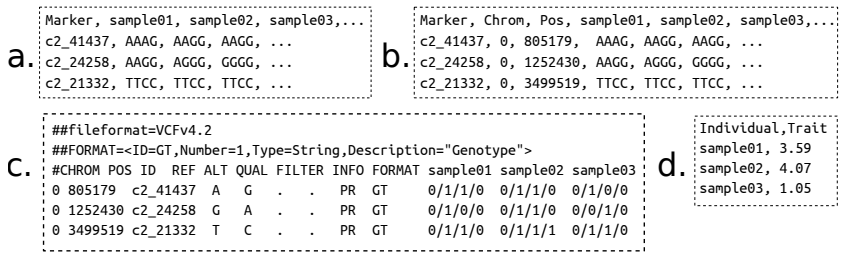


Figure 3: Examples of MultiGWAS input file formats. Figures a, b and c show examples of genotypes, while figure d shows an example of a phenotype. **a**. Genotype file in matrix format containing in the first column the marker names and in the following columns the marker data of the samples coded in "ACGT" format (e.g. AAGG, CCTT for tetraploids, AG, CT for diploids). **b**. Genotype file in GWASpoly format adding the chromosome and marker position to the matrix format.**c**. Genotype file in VCF format with metadata (first two lines) and header line. The following lines contain genotype information of the samples for each position. VCF marker data can be encoded as simple genotype calls (GT format field, e.g., 0/0/1/1 for tetraploids or 0/1 for diploids) or using the NGSEP custom format fields (**Duitama2019**): ACN, ADP or BSDP. **d**. Phenotype file in a matrix format with column headers and sample names followed by their trait values. Both GWASpoly genotype and phenotype files are in CSV (Comma Separated Values format).

GWAS model: MultiGWAS is designed to work with quantitative phenotypes and can run GWAS analysis using two types of statistical models that we have called *full* and *naive* models. The *full model* is known in the literature as the Q+K model (**Yu2006**) and includes a control for structure (Q) and relatedness between samples (K). In contrast, the *naive model* does not include any type of correction. Both models are linear regression approaches, and each one of the four GWAS packages used in MultiGWAS has 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) (**Yu2006**), following the equation:

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

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 ν 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 ν, and X, S and Z are incidence matrices of 1s and 0s relating y to β, α and μ, respectively.

Genome-wide significance: GWAS searches 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 (**Gumpinger2018**; **Rosyara2016**), 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 each of those 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 (FDR) 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 adjust 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: Criticism has arisen, considering only statistically significant associations as possible correct associations (**Thomson2011**; **Kaler2019**). Many low *p-value* associations, closer to being significant, are discarded due to the stringent significance levels, which consequently increases the number of false negatives. 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 poor quality that can lead 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):

- **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:*** (for diploids) filters out SNPs with a *p-value* below the x threshold in the Hardy-Weinberg equilibrium exact test.

MultiGWAS does the MAF filtering, and uses the PLINK package (**Gumpinger2018**) for the other three filters: MIND, GENO, and HWE.

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

2.1.2 Data preprocessing

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

At this point, the format "ACGT" suitable for the polyploid software GWASpoly and SHEsis, is "diploidized" for PLINK and TASSEL. The homozygous tetraploid genotypes are converted to diploid thus: AAAA→AA, CCCC→CC, GGGG→GG, 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

MultiGWAS runs in parallel using two types of statistical models specified in the parameters file, the Full model (Q+K) and Naive (i.e., without any control) where Q refers to population structure, and K refers to relatedness, calculated by kinship coefficients across individuals (**Sharma2018**). The Full model (Q+K) controls for both population structure and individual relatedness. For population structure, MultiGWAS uses the Principal Component Analysis (PCA) and takes the top five PC as covariates. For relatedness, MultiGWAS uses kinship matrices that TASSEL and GWASpoly calculated separately, and for PLINK and SHEsis, relatedness depends on kinship coefficients calculated with the PLINK 2.0 built-in algorithm (**Chang2015**).

2.2.1 GWASpoly

GWASpoly (**Rosyara2016**) is an R package designed for GWAS in polyploid species used in several studies in plants (**Berdugo2017**; **Ferrao2018**; **Sharma2018**; **Yuan2019**). 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. Consequently, the number of statistical tests performed can be different in each action model and so thresholds below which the *p-values* are considered significant.

MultiGWAS is using 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 is a program based on a linear regression model that includes single-locus association analysis, among others. The software design includes polyploid species. However, their use is mainly in diploids animals and humans (**Qiao2015**; **Meng2019**).

MultiGWAS is using version 1.0, which does not take account for population structure or relatedness. Despite, MultiGWAS externally estimates relatedness for SHEsis by excluding individuals with cryptic first-degree relatedness using the algorithm implemented in PLINK 2.0 (see below).

2.2.3 PLINK

PLINK is one of the most extensively used programs for GWAS in humans and any diploid species (**Power2016**). PLINK includes a range of analyses, including univariate GWAS using two-sample tests and linear regression models.

MultiGWAS is using two versions of PLINK: 1.9 and 2.0. Linear regression from PLINK 1.9 performs both naive and full model. For the full model, the software calculates the population structure using the first five principal components calculated with a built-in algorithm integrated into version 1.9. Moreover, version 2.0 calculates the kinship coefficients across individuals using a built-in algorithm that removes the close individuals with the first-degree relatedness.

2.2.4 TASSEL

TASSEL is another standard GWAS program based on the Java software developed initially for maize but currently used in several species (**Alvarez2017**;

Zhang2018). For the association analysis, TASSEL includes the general linear model (GLM) and mixed linear model (MLM) that accounts for population structure and relatedness. Moreover, as GWASPoly, TASSEL provides three-gene action models to calculate the SNP effect of each genotypic class: general, additive, and dominant. Hence, the significance threshold depends on each action model.

MultiGWAS uses TASSEL 5.0, with all gene action models used to find the N best-ranked associations and reporting the top N best-ranked associations (SNPs with lowest *p-values*). Naive GWAS uses the GLM, and full GWAS uses the MLM with two parameters: population structure that uses the first five principal components, and relatedness that uses the kinship matrix with centered IBS method, both calculated with the TASSEL built-in algorithms.

2.3 Integration stage.

The outputs resulting from the four GWAS packages are scanned and processed to identify significant and best-ranked associations with *p-values* lower than and close to a significance threshold, respectively.

2.3.1 Calculation of *p-values* and significance thresholds

GWAS packages compute *p-value* as a measure of association between each SNP and the trait of interest. The statistically significant associations are those their *p-value* drops below a predefined significance threshold. Since a GWAS analysis performs a large number of tests to look for possible associations, one for each SNP, then some correction in the *p-values* is needed to reduce the possibility of identifying false positives, or SNPs with false associations with the phenotype, but that reach the significance threshold.

MultiGWAS provides two methods for adjusting *p-values* and significance threshold: the false discovery rate (FDR) that adjust *p-values,* and the Bonferroni correction, that adjusts the threshold. By default, MultiGWAS uses the Bonferroni correction that uses the significance level α/m, with α defined by the user in the configuration file, and m is the number of tested markers to adjust the significance threshold in the GWAS study.

However, the significance threshold can be different for each GWAS package as some of them use several action models to calculate the SNP effect of each genotypic class. For both PLINK and SHEsis packages, which use only one model, m is equal to the total number of SNPs. However, for both GWASpoly and TASSEL packages, which use eight and three gene action models, respectively, m is equal to the number of tests performed in each model, which is different between models.

Furthermore, most GWAS packages compute both *p-values* and thresholds differently, with the consequence that significant associations identified by one package do not reach the threshold of significance in the others. Thus, it could result in the loss of real associations, the so-called false negatives. To overcome

these difficulties, MultiGWAS reports two sets of associations: significant and best-ranked (those closest to being statistically significant), as described below.

2.3.2 Selection of significant and best-ranked associations

MultiGWAS reports two groups of associations from the results of 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).

PLINK and SHEsis have a unique gene action model (see section 2.2.2 and 2.2.3). However, in the case of GWASpoly and TASSEL, which have eight and three models respectively, MultiGWAS automatically selects the "best gene action model" from each package and takes the associations from it. This selection within GWASPoly and TASSEL has three criteria: the inflation factor (I), the shared SNPs (R), and the significant SNPs (S).

Each gene action model is scored using the following equation:

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

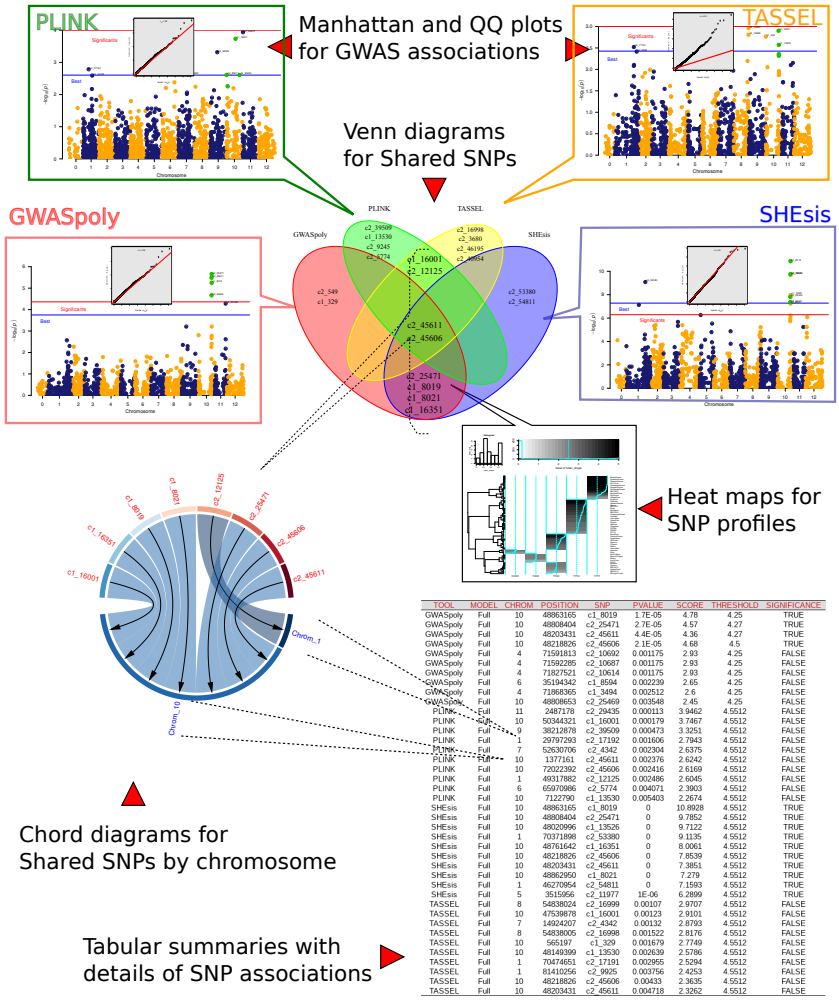
where $score(M_i)$ is the score for the gene action model M_i, with i from $1..k$, for a GWAS package with k gene action models. I_i is the score for the inflation factor defined as $I_i = 1 - |1 - \lambda(M_i)|$, where $\lambda(M_i)$ is the inflation factor for the M_i model. R_i is the score of the 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.

2.3.3 Integration of results

At this stage, MultiGWAS integrates the results to evaluate reproducible results among tools (Fig **??**). However, it still reports a summary of the results of each tool:

- A Quantile-Quantile (QQ) plots for the resultant *p-values* of each tool and the corresponding inflation factor λ to assess the degree of the test statistic inflation.



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
PLINK	Full	11	2487178	c2_29435	0.000113	3.9462	4.5512	FALSE
PLINK	Full	10	50344321	c1_16001	0.000179	3.7467	4.5512	FALSE
PLINK	Full	9	38212878	c2_39509	0.000473	3.3251	4.5512	FALSE
PLINK	Full	1	29797293	c2_17192	0.001606	2.7943	4.5512	FALSE
PLINK	Full	7	52630706	c2_4342	0.002304	2.6375	4.5512	FALSE
PLINK	Full	10	1377161	c2_45611	0.002376	2.6242	4.5512	FALSE
PLINK	Full	10	72022392	c2_45606	0.002416	2.6169	4.5512	FALSE
PLINK	Full	1	49317882	c2_12125	0.002486	2.6045	4.5512	FALSE
PLINK	Full	6	65970986	c2_5774	0.034071	2.3903	4.5512	FALSE
PLINK	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	0.002965	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 4: Reports presented by MultiGWAS. For each tool, first, a QQ plot that assesses the resultant *p-values*. Second, a Manhattan plot for each tool with two lines, blue and red, represents the lower limit for the best ranked and significative SNPs, respectively. We present two Venn diagrams, one for the significative SNPs and one for N best-ranked SNPs of each tool. We show the results for GWAspoly, PLINK, TASSEL, and SHEsis in red, green, yellow, and blue. For each SNP that is in the intersection, thus, that is predicted by more than one tool, we provide an SNP profile. SNPs by chromosome chord diagrams show that the strongest associations are limited to few chromosomes. Furthermore, we present tabular summaries with details of significant and best-ranked associations.

- A Manhattan plot of each tool with two lower thresholds, one for the best-ranked SNPs, and another for the significant SNPs.

To present the replicability, we use two sets: (1) the set of all the significative SNPs provided by each tool and (2) the set of all the best-ranked SNPs. For each set, we present a Venn diagram that displays all SNPs predicted exclusively by one tool and intersections that help identify the SNPs predicted by one, two, three, or all the tools. Also, this information is present on the tables for the two sets.

For each SNP identified more than once, MultiGWAS provides its SNP profile. That is a heat diagram for a specific SNP, where each column is a genotype state AAAA, AAAB, AABB, ABBB, and BBBB. Moreover, each row corresponds to a sample. Samples with close genotypes form together clusters. Thus to generate the clusters, we do not use the phenotype information. However, we present the phenotypic information in the figure as the color. This figure visually provides information regarding genotype and phenotype information simultaneously for the whole population. We present colors as tones between white and black for color blind people.

MultiGWAS generates a report, one document with the content previously described. Besides, there is a folder with the individual figures just in case the user needs one (Supplementary Material 1).

In the following section, we present the results of the functionality of the tool, configured with a Full GWAS model using quality filters, and applied on an open dataset of a diversity panel of a tetraploid potato, genotyped and phenotyped as part of the USDA-NIFA Solanaceae Coordinated Agricultural Project (SolCAP) **Hirsch2013**. The complete report of this analysis and the report of a second analysis using a naive GWAS model without quality filters are presented in the supplementary materials S1 and S2, respectively.

3 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 maybe not significant for the others. Alternatively, well-ranked SNPs in one package may be ranked differently in another.

To highlight these differences in the results across the four packages, MultiGWAS produces five types of results combining graphics and tables to compare, select, and interpret the set of possible SNPs associated with a trait of interest. The outputs include:

- Manhattan and Q-Q plots to show GWAS associations.
- Venn diagrams to show associations identified by single or several tools.
- Heat diagrams to show the genotypic structure of shared SNPs.

- Chord diagrams to show shared SNPs by chromosomes.
- Score tables to show detailed information of associations for both summary results from MultiGWAS and particular results from each GWAS package

The complete reports generated by MultiGWAS for both types of analysis, full and naive, applied to the diversity panel of tetraploid potato, are supplementary information at https://github.com/agrosavia-bioinformatics/MultiGWAS/tree/master/docs/supplements.

3.1 Manhattan and QQ plots for GWAS associations

MultiGWAS uses classical Manhattan and Quantile–Quantile plots (QQ 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 Fig. **??**). 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. At the same time, the QQ plot compares the expected distribution of *p-values* (y-axis) with the observed distribution (x-axis).

MultiGWAS adds distinctive marks to both plots to identify different types of SNPs: (a) In the Manhattan plots, the significant SNPs are above a red line, and the best-ranked SNPs are above a blue line. Also, SNPs shared between packages are colored green (See Fig. **??**.b). (b) In the QQ plots, a red diagonal line indicates the expected distribution under the null hypothesis of no association of SNPs with the phenotype. Both distributions should coincide, and most SNPs should lie on the diagonal line. Deviations for a large number of 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 (**Power2016**).

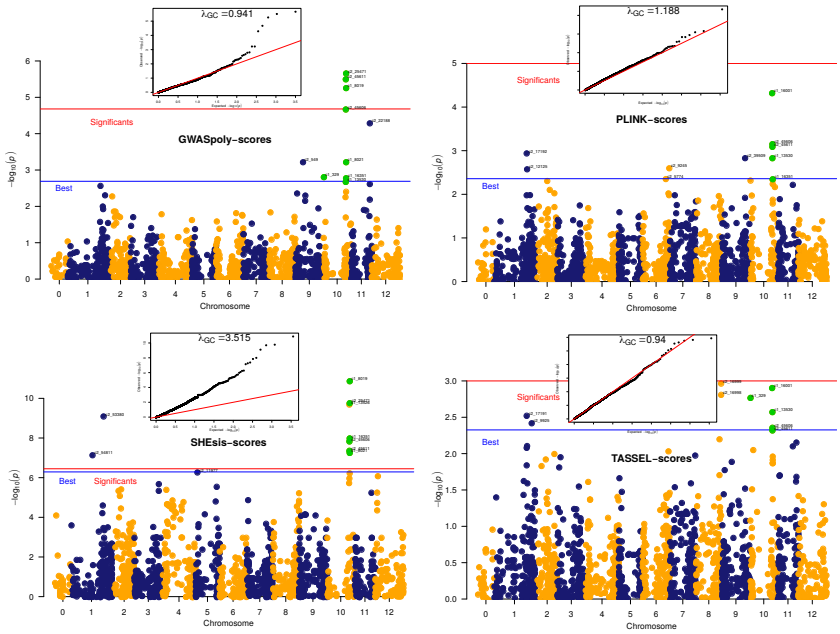


Figure 5: Associations in the tetraploid potato dataset. MultiGWAS shows the associations identified by the four GWAS packages using Manhattan and QQ plots. The tetraploid potato data showed several SNPs shared between the four software (green dots). The best-ranked SNPs are above the blue line, but 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 ($\lambda = 3.5$, at the top of the QQ plot), which is observed by the high number of SNPs deviating from the red diagonal of the QQ plot.

3.2 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 (see Figure **??**). 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 appears in both a SNPs table (Figure **??**.a), and in a Venn diagram (Figure **??**.b). The table lists them by package and sorts by decreasing score, whereas the Venn diagram emphasizes if 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 appear in a Venn diagram (Figure **??**.c), and in the SNPs table, marked with significance TRUE (T) in the table of the Figure **??**.a.

a.

TOOL	MDL	IF	SNP	CHR	POS	PVAL	SCR	THR	SGN
GWASpoly	additive	0.99	c2_25471	10	48808404	0.000	5.28	4.48	T
GWASpoly	additive	0.99	c2_45611	10	48203431	0.000	5.07	4.48	T
GWASpoly	additive	0.99	c1_8019	10	48863165	0.000	4.93	4.48	T
GWASpoly	additive	0.99	c2_45606	10	48218826	0.000	4.32	4.48	F
GWASpoly	additive	0.99	c2_549	9	16527499	0.001	3.25	4.48	F
GWASpoly	additive	0.99	c2_9925	1	81410256	0.002	2.77	4.48	F
GWASpoly	additive	0.99	c1_8021	10	48862950	0.002	2.66	4.48	F
GWASpoly	additive	0.99	c2_12125	1	71450400	0.002	2.64	4.48	F
PLINK	additive	1.28	c1_16001	10	47539878	0.000	3.94	4.52	F
PLINK	additive	1.28	c2_17192	1	70472766	0.001	2.86	4.52	F
PLINK	additive	1.28	c2_12125	1	71450400	0.002	2.75	4.52	F
PLINK	additive	1.28	c2_45606	10	48218826	0.002	2.72	4.52	F
PLINK	additive	1.28	c2_45611	10	48203431	0.002	2.64	4.52	F
PLINK	additive	1.28	c2_14903	1	87322718	0.003	2.50	4.52	F
PLINK	additive	1.28	c1_13530	10	48149399	0.003	2.50	4.52	F
PLINK	additive	1.28	c2_2201	1	77738822	0.003	2.49	4.52	F
SHEsis	general	3.56	c1_8019	10	48863165	0.000	10.99	4.52	T
SHEsis	general	3.56	c1_13526	10	48020996	0.000	10.05	4.52	T
SHEsis	general	3.56	c2_45603	10	48073593	0.000	9.89	4.52	T
SHEsis	general	3.56	c2_25471	10	48808404	0.000	9.65	4.52	T
SHEsis	general	3.56	c2_53380	1	70371898	0.000	8.97	4.52	T
SHEsis	general	3.56	c2_45606	10	48218826	0.000	8.17	4.52	T
SHEsis	general	3.56	c1_16351	10	48761642	0.000	8.00	4.52	T
SHEsis	general	3.56	c2_45611	10	48203431	0.000	7.73	4.52	T
TASSEL	general	1.00	c2_16999	8	54838024	0.001	2.96	4.52	F
TASSEL	general	1.00	c2_4342	7	14924207	0.001	2.92	4.52	F
TASSEL	general	1.00	c2_16998	8	54838005	0.001	2.86	4.52	F
TASSEL	general	1.00	c2_17191	1	70474651	0.002	2.67	4.52	F
TASSEL	general	1.00	c2_9925	1	81410256	0.002	2.65	4.52	F
TASSEL	general	1.00	c1_16001	10	47539878	0.002	2.63	4.52	F
TASSEL	general	1.00	c2_45606	10	48218826	0.005	2.34	4.52	F
TASSEL	general	1.00	c2_45611	10	48203431	0.005	2.31	4.52	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.

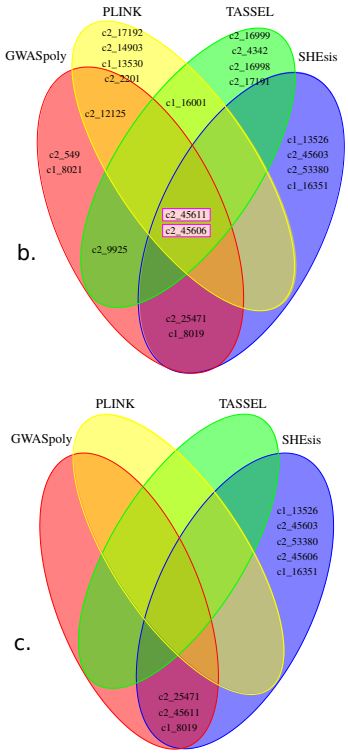


Figure 6: Shared SNPs Views. Tabular and graphical views of SNP associations identified by one or more GWAS packages (shared SNPs). SNPs identified by all packages are marker with red background in all figures. **a** Table with details of the N=8 best-ranked SNPs from each GWAS package. Each row corresponds 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).

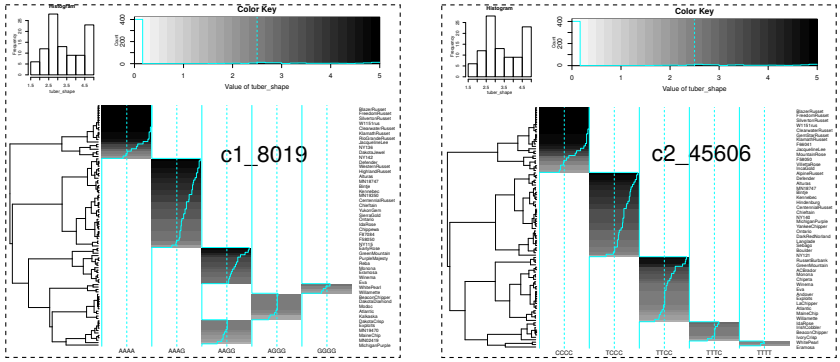
In this analysis, both the polyploid packages GWASpoly and SHEsis identified the SNPs c2_25471, c2_45611, and c1_8019. Of these SNPs, c1_8019 has been reported in previous studies to be associated with tuber shape and depth of eye traits (**Rosyara2016**; **Sharma2018**). Furthermore, in another analysis of MultiGWAS using a naive model without filters (Supplemental Material S2), the SNP c1_8019 was co-identified by three packages: GWASpoly, SHEsis, and the diploid PLINK package.

3.3 Heat diagrams for the structure of shared SNPs

MultiGWAS creates a two-dimensional representation, called the SNP profile, to visualize each trait by individuals and genotypes as rows and columns, respectively (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 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 grayscale (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 the mean of each column and a broken cyan line that indicates how far the cell deviates from the mean.

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.



a. b.

Figure 7: SNP profiles. SNP profiles for two of the best-ranked significant SNPs shown in the figure ??.b. **a.** SNP c2_45606 best-ranked by the four packages (central intersection of the Venn diagram Figure ??.b) **b.** SNP c1_8019 best-ranked by the two tetraploid packages (Figure ??.b), and also identified as significant by the same packages (at the bottom of the Figure ??.a).

3.4 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 table ??.a. Thus, in the case of the tetraploid potato, we found that they are located mostly in chromosome 10 (Figure ??.a). This visualization complements the manhattan plots from each GWAS package (Figure ??.b).

4 Availability and Implementation

The core of the MultiGWAS tool runs under R and users can interact with the tool by either a command-line interface (CLI) developed in R or a graphical user interface (GUI) developed in Java (Figure ??). Source code, examples, documentation, and installation instructions are available at https://github.com/agrosavia-bioinformatics/multiGWAS.

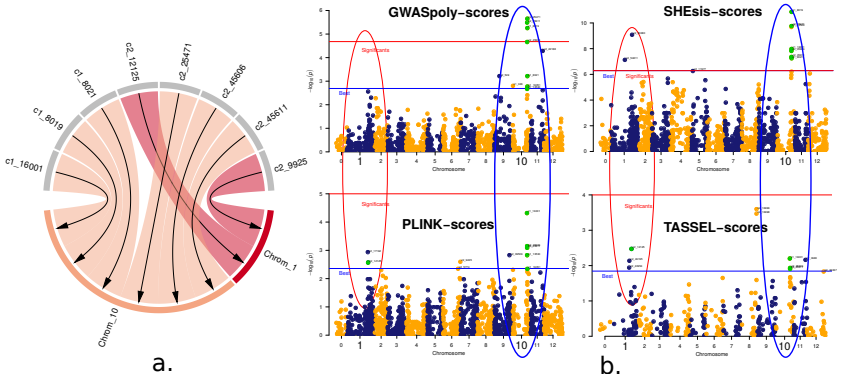


Figure 8: SNPs by chromosome. The position of best-ranked SNPs across chromosomes using two different visualizations. **a.** Chord diagram showing that best-ranked SNPs located in chromosome 10. The SNPs are at the top and the chromosomes at the bottom. The arrows connect the best-ranked SNPs with their position in the chromosomes. **b.** Manhattan plots from each GWAS packages showing two important locations of associations, chromosome 1 and chromosome 10, marked with blue and red ellipsis, respectively.

4.1 Input parameters

MutiGWAS uses as the only input a simple configuration text file with the values for the main parameters that drive the analysis. To create the configuration text file, users can choose either a text editor or the MultiGWAS GUI application. If users prefer a text file, it must have the parameter names and values separated by a colon, filenames enclosed in quotation marks, and TRUE or FALSE values to indicate if filters are applied. If the users prefer the GUI applications, they can create the configuration file using the input parameter view. In any case, this file must have the structure showed in Figure **??**.

```
default:
  ploidy            : 4
  genotypeFile      : "example-genotype-tetra.csv"
  phenotypeFile     : "example-phenotype.csv"
  significanceLevel : 0.05
  correctionMethod  : "Bonferroni"
  gwasModel         : "Full"
  nBest             : 10
  filtering         : TRUE
  MAF               : 0.01
  MIND              : 0.1
  GENO              : 0.1
  HWE               : 1e-10
  tools             : "GWASpoly SHEsis PLINK TASSEL"
```

Figure 9: Configuration file for MultiGWAS. The input parameters include the organism's ploidy level (2: for diploids, 4: for tetraploids). The input genotype/phenotype filenames. The genome-wide significance threshold. The method for multiple testing correction. The GWAS model. The number of associations to report. The quality control filters choosing TRUE or FALSE. The filters are minor allele frequency, individual missing rate, SNP missing rate, and Hardy-Weinberg threshold. Finally, the GWAS packages selected for the analysis.

4.2 Using the command line interface

The execution of the CLI tool is easy. It only needs to open a Linux console, change 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*. The results include a complete HTML report containing the different views described in the results 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.

4.3 Using the graphical user interface

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

```
jmultiGWAS
```

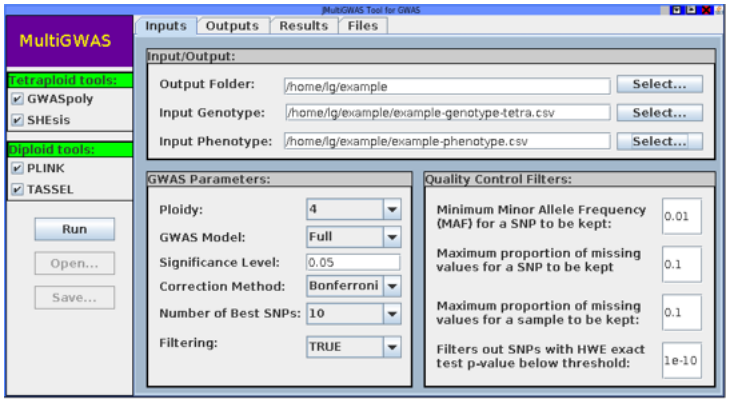


Figure 10: Main view of the MultiGWAS graphical user interface. The interface has a toolbar at the left side and four tabs at the top. In the toolbar, users can select the GWAS packages (Two for tetraploids and two for diploids). The analysis starts with the current parameters or loading a previously saved configuration. In the Input tab, users can set the 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.

5 Discussion

The reanalysis of potato data with MultiGWAS showed that this wrapping tool is handy to improve the GWAS in both diploid and tetraploid species. Through

MultiGWAS performance, we could test its effectiveness to answer some of the challenges analyzing polyploid organisms. They include the integration and replication among parameters and software, the diploidization of polyploid data, and the incorporation of different inheritance mechanisms (**dufresne2014**).

The main advantage of MultiGWAS is that it replicates the GWAS analysis among four software and integrates the results obtained across software, models, and parameters. Therefore, in MultiGWAS, users do not have to choose between specificity or 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 structures for the genomic input data. Currently, the most common format for next-generation sequencing variant data is the VCF (Variant Call Format) (**Danecek2011**; **Ebbert2014**). 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 levels of ploidy. MultiGWAS simplifies the use of the GWAS software available because it allows the VCF files as an input (but see VarStats tool in VTC).

Moreover, the MultiGWAS is the unique wrapping tool we are aware of that facilitates understanding the effect of diploidizing the tetraploid data in the performance of the analysis 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 if the SNP is a real 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 (**Rosyara2016**) provides models of different types of polyploid gene action, including additive, diploidized additive, duplex dominant, simplex dominant, and general. On the other hand, TASSEL (**Bradbury2007**) also models different types of gene action 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 the gene actions for a specific phenotype.

Finally, MultiGWAS, through the active comparison among models, addresses the search of the inheritance mechanisms by comparing among two designed for polysomic inheritance software (**Rosyara2016**; **Shen2016**) with two software for disomic inheritance (**Purcell2007**; **Bradbury2007**). 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 (**ohno1970**; **lynch2000**; **dufresne2014**). Thus it is a useful tool for re-

searchers because it looks for significative associations that involve both types of inheritance.

5.1 Future remarks

The evolution and population genomics of polyploids is an exciting novel area of research. The advancement of next-gen sequencing techniques is producing more empirical polyploid data in different model and non-model organisms (**ekblom2011applications**; **ellegren2014genome**).

Many assumptions developed for diploids in the GWAS analysis do not apply entirely for polyploids (**dufresne2014**). Those include Hardy-Weinberg equilibrium, among others. Fortunately, in the last five years, different models to calculate several parameters for population genomics on polyploids are testing and developing in both simulated and empirical data (**meirmans2018**; **hardy2016population**; **blischak2016accounting**).

For MutiGWAS, we started with the most simple ploidy, such as tetraploids. Moreover, we did not filter data yet for HWE in tetraploids before GWAS as MultiGWAs do it for diploid data. Nevertheless, future MultiGWAS versions should include more complex ploidies further than tetraploids, as well as the explicit calculation of parameters either for filtering polyploid data before GWAS analysis or complementing other population genomics' parameters of the data analyzed.

6 Acknowledgements

This research was possible thanks to AGROSAVIA five-years macroproject entitled *InvestigaciÃşn en conservaciÃşn, caracterizaciÃşn y uso de los recursos genÃľticos vegetales*.

We thanks to 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 for finatial supporting for this publication. Finally to Andres J. CortÃľs for his valuable comments to improve this manuscript.

7 Author Contributions

LG, ICS, and PHRH conceived the idea. LG developed MultiGWAS. MP tested MultiGWAS. All authors wrote and approved the final version of the manuscript.