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

immediate

July 15, 2020

Abstract

Summary: The Genome-Wide Association Studies (GWAS) are essential to determine the association between genetic variants across individuals. One way to support the results is by using different tools to validate the reproducibility of the associations. Currently, software for GWAS in diploids is well-established but for polyploids species is scarce. Each GWAS software has its characteristics, which can cost time and effort to use them successfully. Here, we present MultiGWAS, a tool to do GWAS analysis in tetraploid organisms by executing in parallel and integrating the results from four existing GWAS software: two available for polyploids (GWASpoly and SHEsis) and two frequently used for diploids (PLINK and TASSEL). The tool deals with all the elements of the GWAS process in the four software, including (1) the use of different control quality filters for the genomic data, (2) the execution of two GWAS models, the full model with control for population structure and individual relatedness and the Naive model without any control. The summary report generated by MultiGWAS provides the user with tables and plots describing intuitively the significant association found by both each one and across four software, which helps users to check for false-positive or false-negative results.

MultiGWAS generates five summary results integrating the four tools. (1) Score tables with detailed information on the associations for each tool. (2) Venn diagrams of shared SNPs among the four tools. (3) Heatmaps of significative SNP profiles among the four tools. (4) Manhattan and QQ plots for the association found by each tool. And (5) Chord diagrams for the chromosomes vs. SNP by each tool. **Contact:** phreyes@agrosavia.co

Keywords: GWAS, tetraploids, SNPs, polyploids, software

Comentatios Ivania

Comentatios Luis

1 Introduction

The Genome-wide association studies (GWAS) are used to identify which variants through the whole genome of a large number of individuals are associated with

a specific trait ((**?**; **?**)). This methodology started with humans and several model plants, such as rice, maize, and *Arabidopsis* ((**?**; **?**; **?**; **?**; **?**)). 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 ((**?**; **?**)). Thus, the GWAS is becoming the standard tool to understand the genetic bases of either ecological or economic phenotypic variation for both model and non-model organisms. This increment in GWAS includes complex species such as polyploids (Fig. **??**) ((**?**; **?**)).

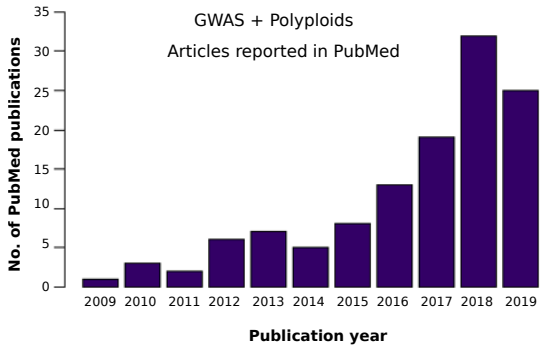


Figure 1: Timeline for articles reported for GWAS studies on polyploid species in PubMed. We present data for completed years.

The GWAS for polyploid species has fourth related challenges. First, as all GWAS, we should replicate the study as a reliable method to validate the results and recognize real associations. This replication involves finding the same associations either in several replicates from the study population using the same software or testing different GWAS tools among the same study population. This approach involved the use of different parameters, models, or conditions, to test how consistent the results are ((**?**; **?**)). However, the performance of different GWAS software could affect the results. For example, the threshold $pvalue$ for SNP significance change through four GWAS software (i.e., PLINK, TASSEL, GAPIT, and FaST-LMM) when sample size varies ((**?**)). It means that well-ranked SNPs from one package can be ranked differently in another.

Second, although there are many GWAS software available to repeat the analysis under different conditions ((**?**)), most of them are designed exclusively for the diploid data matrix ((**?**)). Therefore, it is often necessary to "diploidizing" the polyploid genomic data in order to replicate the analysis.

Third, 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 know, there is only two software with this service in mind, such as iPAT and easyGWAS.

The iPAT allows running in a graphic interface three well-known command-line GWAS software such as GAPIT, PLINK, and FarmCPU ((**?**)) . However, the output from each package is separated. On the other hand, the easyGWAS allows running a

Me confunde al hablar de software y toois

"The latter approach..."

Software or tools ??

GWAS analysis on the web using different algorithms. This analysis could run independently of both the computer capacity and operating system. However, it needs either several datasets available or a dataset with a large number of individuals to make replicates in order to compare among algorithms. Moreover, the output from different algorithms is separated ((**?**)). Thus, for both software iPAT and easyGWAS, the integrative and comparative outputs among software or algorithms are missing.

Fourth, the GWAS on polyploids generates a new level of complexity to understand how allele dosage affect the phenotype expression on quantitative traits. Therefore, any tool that compares among software but also models with different allele dosage will contribute to gain a better understanding in how redundancy or complex interaction among alleles affect the phenotype expression and the evolution of new phenotypes among polyploid species .

Aqui falta la cita del paper de blueberry y el paper de Rosyara 2016

To contribute to sort out all the above fourth challenges, we developed the MultiGWAS tool that performs GWAS analyses for tetraploid species using four software in parallel. Our tool include GWASpoly ((**?**)) and the SHEsis tool ((**?**)) that accept polyploid genomic data, and PLINK ((**?**)) and TASSEL ((**?**)) with the use of a "diploidized" genomic matrix. The tool deals with input file formats, data preprocessing, search for associations by running four GWAS tools in parallel, and creation of comparative reports from the output of each software to help the user to decide more intuitively the true or false associations.

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 and phenotype, and 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 along with the parameters that will be used by the four tools. Once the configuration file is read and processed, the genomic data files (genotype and phenotype) are preprocessed by cleaning, filtering, and checking 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:

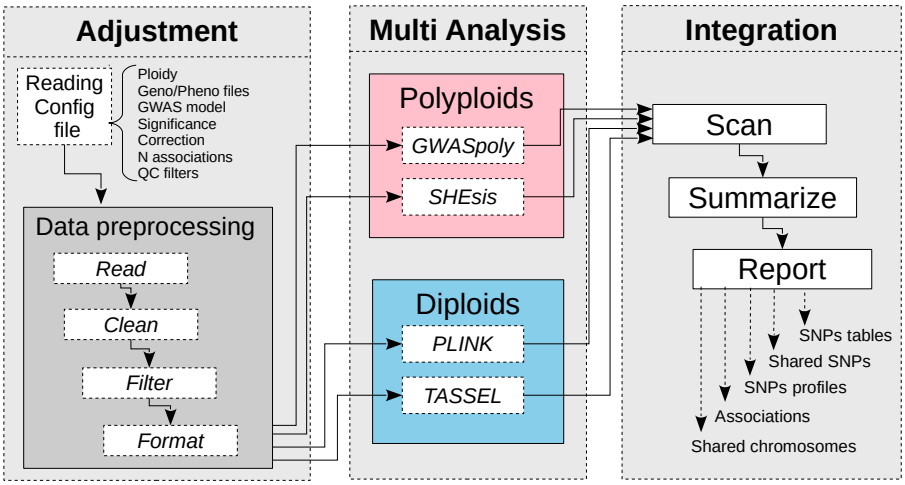


Figure 2: MultiGWAS flowchart has three steps: adjustment, multi analysis, and integration. The first step deals with input data management, reading the configuration file, reading and preprocessing the input genomic data (genotype and phenotype). The second step deals with GWAS analysis, configuring and running the four packages in parallel. And the third step deals with summarizing and reporting results using different tabular and graphical visualizations.

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 genotype and one for phenotype. Genotypes files can be either in GWASpoly format ((**?**)) using SNP markers in rows and samples in columns (Fig. **??**.a) or Variant Call Format (VCF) (Fig.**??**.b) which is transformed into GWASpoly format using NGSEP 4.0.2 ((**?**)). The phenotype file contains only one trait and uses a matrix format with the first column for the sample names and the second column for the trait values (Fig. **??**.c).

a.
```
Marker, Chrom, Pos, sample01, sample02, sample03,...
c2_41437, 0, 805179,  AAAG, AAGG, AAGG, ...
c2_24258, 0, 1252430, AAGG, AGGG, GGGG, ...
c2_21332, 0, 3499519, TTCC, TTCC, TTCC, ...
```

c.
```
Individual,Trait
sample01, 3.59
sample02, 4.07
sample03, 1.05
```

b.
```
##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM POS ID  REF ALT QUAL FILTER INFO FORMAT sample01 sample02 sample03
0 805179  c2_41437  A   G   .   .   PR  GT     0/1/1/0  0/1/1/0  0/1/0/0
0 1252430 c2_24258  G   A   .   .   PR  GT     0/1/0/0  0/1/1/0  0/0/1/0
0 3499519 c2_21332  T   C   .   .   PR  GT     0/1/1/0  0/1/1/1  0/1/1/0
```

Figure 3: **Examples of MultiGWAS input file formats**. Figures a and b show genotype files in GWASpoly and VCF formats, respectively, while figure c shows a phenotype file in matrix format. a. Genotype file in GWASpoly format containing column headers and with the first three columns for markers names, chromosomes and positions. The following columns correspond to the marker data of the samples in "ACGT" format (e.g. AAGG, CCTT for tetraploids, AG, CT for diploids). b. 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 ((**?**)): ACN, ADP or BSDP. c. Phenotype file in 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 know in the literature as the Q+K model (**?**) and includes control for structure (Q) and relatedness between samples (K), whereas the *naive model* does not include any type of correction. Both models are based on linear regression approaches and variations of them are implemented by the four GWAS packages used by MultiGWAS. 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) (**?**), which is expressed with the following equation:

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

where y is the vector of observed phenotypes; β is a vector of fixed effects other than SNP or population group effects; α is a vector of SNP effects (Quantitative Trait Nucleotides); ν is a vector of population effects; μ is a vector of polygene background effects; e is a vector of residual effects; 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 (**?**; **?**), and MultiGWAS uses an α of 0.05 for the four GWAS packages. But the threshold is adjusted according to each package, as

some packages as GWASpoly and TASSEL calculates the SNP effect for each genotypic class using different gene action models (see "Multi analysis stage"). So, the number of tested markers may be different in each model (see below) that results in different *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 common 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 stringent 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 in considering only statistically significant associations as the only possible correct associations (**?**; **?**). Many of low *p-value* associations, closer to being significant, are discarded due to the stringent significance levels, and consequently increasing the number of false negatives. To help to analyze both significant and non-significant associations, 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 we can observe replicability in the results for different programs. Nevertheless, we present each association with the corresponding *p-value*.

Quality control filters: A control step is necessary to check the input data for 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:*** filters out SNPs which have Hardy-Weinberg equilibrium exact test *p-value* below the x threshold.

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

GWAS tools: List of names of GWAS packages to run and integrate into MultiGWAS analysis. Currently four packages, two for tetraploid organisms: GWASpoly and SHEsis, and two for diploids: PLINK and TASSEL.

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, individuals and SNPs with poor quality are removed by considering 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, the genomic data, genotype and phenotype, are converted 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 coeficients across individuals ((**?**)). 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 ((**?**)).

2.2.1 GWASpoly

GWASpoly ((**?**)) is an R package designed for GWAS in polyploid species used in several studies in plants ((**?**; **?**; **?**; **?**)). 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. As a consequence, the number of statistical test 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 ((**?**; **?**)).

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 ((**?**)). 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 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 ((**?**; **?**)). 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, and so the significance threshold depends on each action model.

MultiGWAS is using 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 both 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 real associations are those their *p-value* drops below a predefined significance threshold. However, the four GWAS packages compute differently *p-values* with the consequence to compute them too high or too low. If *p-values* is too high, it would lead to false negatives or SNPs with real associations with the phenotype, but that does not reach the significance threshold. Conversely, if *p-values* are too low, then it would lead to false positives or SNPs with false associations with the phenotype, but that reaches the significance threshold.

To overcome these difficulties, in the case of too high *p-values*, MultiGWAS identifies and reports both significant and best-ranked associations (the ones closest to being statistically significant). Whereas, in the case of too low *p-values*, 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 (defined by the user in the configuration file), and m (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.

2.3.2 Selection of significant and best-ranked associations

MultiGWAS selects two groups of associations from the results of each GWAS package: statistically significant and best-ranked. The latter equally important to the former as they are associations with lowest *p-values* not reaching the significance threshold but representing interesting associations for further analysis (possible false negatives).

The significant associations are selected from the ones with *p-values* falling below a significant threshold, calculated for each GWAS package; and the best-ranked associations are selected as the closest N to being statistically significant, with N defined by the user in the configuration file.

The selection of these groups takes into account whether the GWAS package uses only one gene action model, as PLINK and SHEsis do, or uses several ones, as GWASpoly and TASSEL do. In the first case, there is only one resulting set of associations and the selection is straighforward, as described above. However, in the second case, there are several resulting sets of associations, one for each model, and MultiGWAS selects both groups of associations by choosing the gene action model with the highest number of shared SNPs and with the inflation factor closest to one, according to the following equation:

$$score(M_i) = \frac{\sum_{j=1}^{k} sharedSNPs(M_i, M_j)}{k * N} + 1 - |1 - \lambda(M_i)|$$

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; $sharedSNPs(M_i, M_j)$ is the number of shared SNPs between models M_i and M_j; N is the number of closest SNPs to being statistically significant, as it was described above; and $\lambda(M_i)$ is the inflation factor for the model M_i.

The score is high when a model M_i both identifies a high number of shared SNPs and has an inflation factor λ close to 1. Conversely, the score is low when the model M_i both identifies a small number of shared SNPs and has an inflation factor λ either low (close to 0) or high ($\lambda > 2$). In any other case, the score is balanced between the number of shared SNPs and the inflation factor.

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:

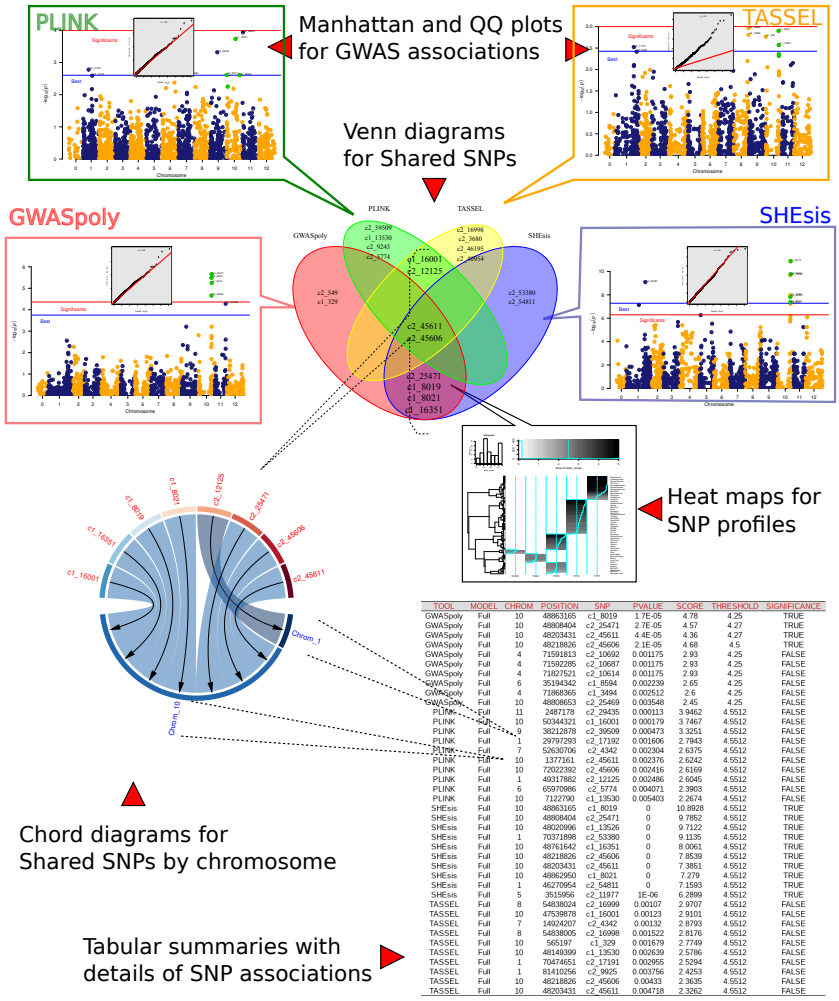
- A Quantile-Quantile (QQ) plots for the resultant *p-values* of each tool and the corresponding inflation factor λ to assess the degree of the test statistic inflation.
- A Manhattan plot of each tool with two lower thresholds, one for the best-ranked SNPs, and another for the significant SNPs.

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 shows SNPs predicted exclusively by one tool and intersections that help to identify the SNPs predicted by one, two, three, or all the tools. Also, we provide detailed tables for the two sets.

For each SNP identified more than once, we provide what we call the 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 phenotype 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.

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



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	0.004071	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	48148399	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 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, respectively, is the lower limit for the best ranked and significative SNPs. 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.

needs one. In the supplementary information, we include a report and a description of the report content (Supplementary Material 1)

In the following section, we present the results of the functionality of the tool applied on a open dataset of a diversity panel of a tetraploid potato, genotyped and phenotyped as part of the USDA-NIFA Solanaceae Coordinated Agricultural Project (SolCAP) ((?)).

AquÃ deberÃamos decir especÃficamente quÃ© datos son y de donde salieron

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, in the diversity panel of tetraploid potato described above are shown in the supplementary information at `https://github.com/agrosavia-bioinformatics/multiGWAS-Supplementary`.

3.1 Manhattan and QQ plots for GWAS associations

MultiGWAS uses classical Manhattan and Quantile–Quantile plots (QQ plots) to visualize the results of each package. 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. While the QQ plot compares the expected distribution of *p-values* (y-axis) with the observed distribution (x-axis)..

Corregido: "among packages" y cita de polygenic traits

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

Revisar si esto del color verde es correcto "shared among all packages". TambiÃ©n revisar si estÃ¡ bien dicho lo de la comparación entre pocos genes y muchos genes. No se si sea necesario cuantificar la desviación. Originalmente se cuantificaba pero era confuso

best-ranked SNPs are above a blue line. In addition, 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. But, it is also expected that few SNPs deviate from the diagonal for a truly polygenetic trait ((**?**)).

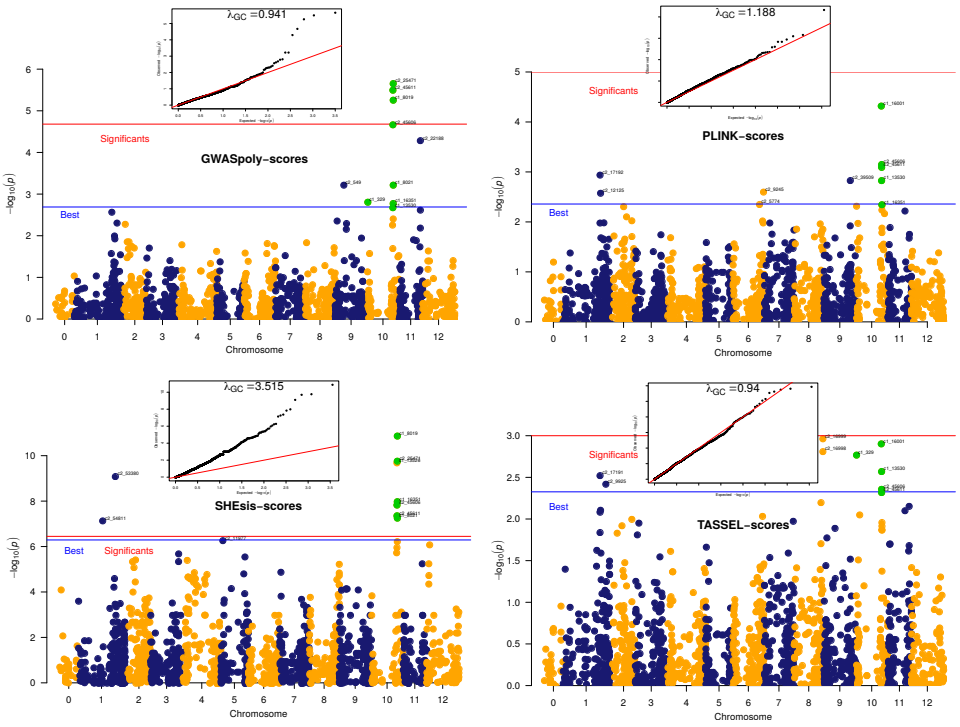


Figure 5: Associations in the tetraploid potato dataset. MultiGWAS shows the associations identified by the four GWAS packages using Manhattan and QQ plots. In the case of the tetraploid potato, several SNPs are observed to be shared between the four packages (green dots). The best-ranked SNPs are above the blue line, but only GWASpoly and SHEsis identified significant associations (SNPs above the red line). 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 $-log_{10}(p\text{-}value)$ 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	MODEL	GC	SNP	CHR	POS	PVALUE	SCR	THR	SGN
GWASpoly	additive	0.96	c2_25471	10	48808	0.000002	5.67	4.50	T
GWASpoly	additive	0.96	c2_45611	10	48203	0.000003	5.51	4.50	T
GWASpoly	additive	0.96	c1_8019	10	48863	0.000005	5.27	4.50	T
GWASpoly	additive	0.96	c2_45606	10	48218	0.000021	4.68	4.50	T
GWASpoly	additive	0.96	c2_22188	11	40777	0.000050	4.30	4.50	F
GWASpoly	additive	0.96	c2_549	9	16527	0.000589	3.23	4.50	F
GWASpoly	additive	0.96	c1_8021	10	48862	0.000589	3.23	4.50	F
GWASpoly	additive	0.96	c1_329	10	56519	0.001514	2.82	4.50	F
GWASpoly	additive	0.96	c1_16351	10	48761	0.001622	2.79	4.50	F
PLINK	additive	1.19	c1_16001	10	47539	0.000047	4.33	4.55	F
PLINK	additive	1.19	c2_45606	10	48218	0.000688	3.16	4.55	F
PLINK	additive	1.19	c2_45611	10	48203	0.000786	3.10	4.55	F
PLINK	additive	1.19	c2_17192	1	70472	0.001123	2.95	4.55	F
PLINK	additive	1.19	c2_39509	9	50174	0.001440	2.84	4.55	F
PLINK	additive	1.19	c1_13530	10	48149	0.001443	2.84	4.55	F
PLINK	additive	1.19	c2_9245	6	57953	0.002455	2.61	4.55	F
PLINK	additive	1.19	c2_12125	1	71450	0.002593	2.59	4.55	F
PLINK	additive	1.19	c2_5774	6	50345	0.004336	2.36	4.55	F
SHEsis	general	1.47	c1_8019	10	48863	0.000000	7.64	4.55	T
SHEsis	general	1.47	c1_13526	10	48020	0.000000	6.94	4.55	T
SHEsis	general	1.47	c2_25471	10	48808	0.000000	6.94	4.55	T
SHEsis	general	1.47	c2_53380	1	70371	0.000000	6.46	4.55	T
SHEsis	general	1.47	c1_16351	10	48761	0.000004	5.45	4.55	T
SHEsis	general	1.47	c2_45606	10	48218	0.000004	5.38	4.55	T
SHEsis	general	1.47	c2_45611	10	48203	0.000010	4.98	4.55	T
SHEsis	general	1.47	c1_8021	10	48862	0.000012	4.93	4.55	T
SHEsis	general	1.47	c2_54811	1	46270	0.000014	4.86	4.55	T
TASSEL	additive	0.86	c2_16999	8	54838	0.000247	3.61	3.89	F
TASSEL	additive	0.86	c2_16998	8	54838	0.000329	3.48	3.89	F
TASSEL	additive	0.86	c2_12125	1	71450	0.003287	2.48	3.89	F
TASSEL	additive	0.86	c1_16001	10	47539	0.006105	2.21	3.89	F
TASSEL	additive	0.86	c2_3680	11	39908	0.006701	2.17	3.89	F
TASSEL	additive	0.86	c2_46195	1	64259	0.007116	2.15	3.89	F
TASSEL	additive	0.86	c2_40954	1	63756	0.011097	1.95	3.89	F
TASSEL	additive	0.86	c2_45606	10	48218	0.011369	1.94	3.89	F
TASSEL	additive	0.86	c2_45611	10	48203	0.012091	1.92	3.89	F

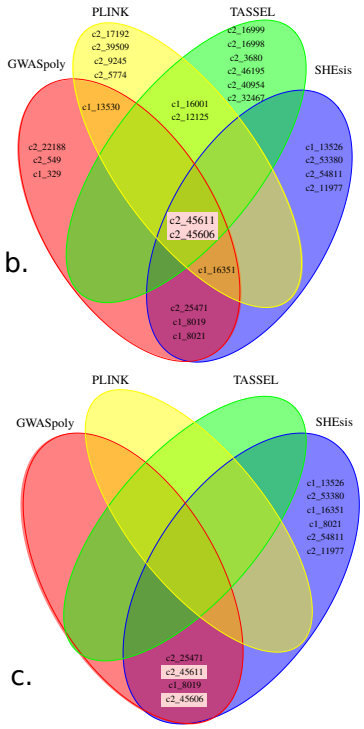
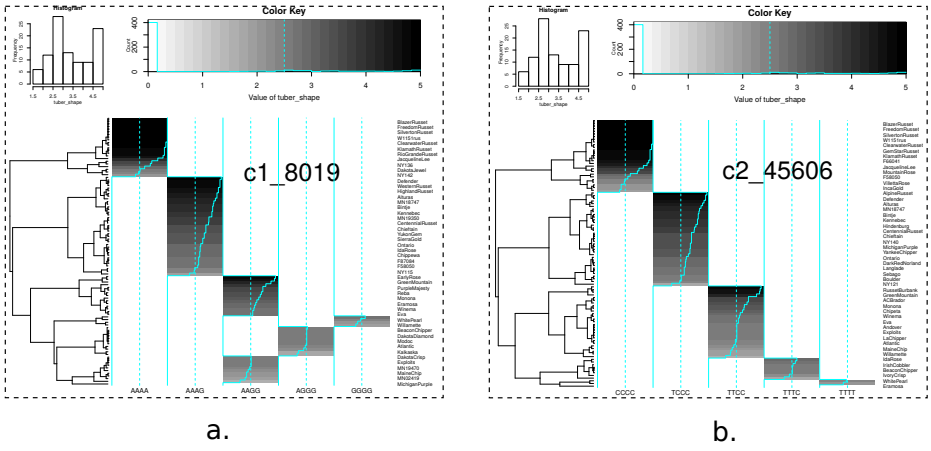


Figure 6: Shared SNPs Views. Tabular and graphical views of SNP associations identified by one or more GWAS packages (shared SNPs). SNPs identified by all packages are marker with red background in all figures. **(a)** Table with details of the N=9 best-ranked SNPs from each GWAS package. Each row corresponds to a single SNP, and the nine columns are tool name, the model used by the tool, genomic control factor (inflation factor), SNP name, chromosome, position in the genome, *p-value*, score as $-log_{10}(p\text{-}value)$, significance threshold as $-log_{10}(\alpha/m)$ where α is the significance level, and m is the number of tested markers, and significance as true (T) or false (F) whether score > threshold or not. **(b)** Venn diagram of the N=9 best-ranked SNPs. SNPs identified by all packages are in the central intersection. Other SNPs identified by more than one packages are in both upper central and lower central intersections. **(c)** Venn diagram of the significant SNPs (score > threshold).

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 by 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 chromosomes 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).

En esta sección fui bastante radical y terminÃ© eliminando dos pÃ¡rrafos que no me parecen relevantes. Sin embargo los dejo seÃ±alados por si ustedes consideran que es importante dejarlo. Si se vuelen a incorporar al texto sugiero reorganizarlos mejor porque el mensaje es confuso en mi opinión. PÃ¡rrafo 1. Generally, in a typical GWAS analysis the strongest associations are signaled by several nearby-correlated SNPs located in the same chromosome, as in manhattan plots, where these associations form neat peaks with several SNPs showing the same signal. Conversely, no peaks are shown when few SNPs correlate with a trait. Leyendo

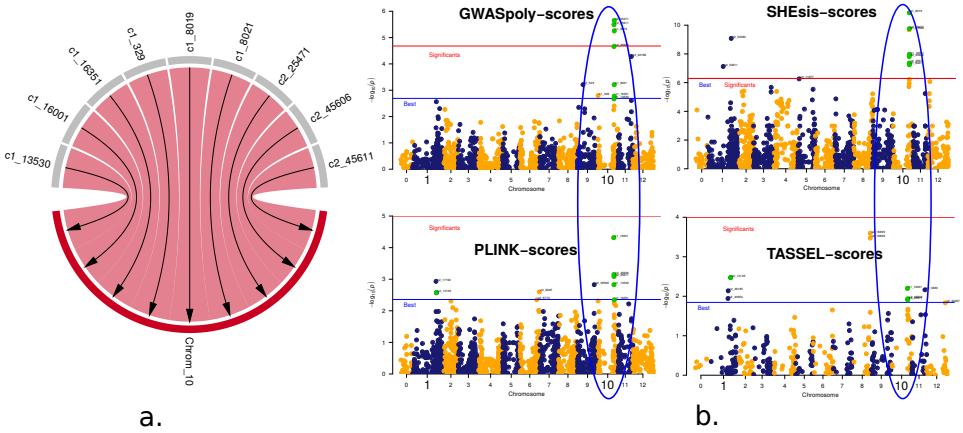


Figure 8: SNPs by chromosome. The position of best-ranked SNPs across chromosomes using two different visualizations. **(a)** Chord diagram showing that best-ranked SNPs located in chromosome 10. The SNPs are at the top and the chromosomes at the bottom. The arrows connect the best-ranked SNPs with their 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 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`.

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 shown in the Figure(Figure **??**.a.)

```
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 ploidy level of the organism (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 simple. It only needs to open a Linux console, change to the folder where the configuration file was created, and type the name of the executable tool followed by the 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 the subsection **??**. The interface can be executed by calling the following command from a linux console:

```
jmultiGWAS
```

Arreglado, no hay folder de salida en este archivo. El orden si coincide

Yo no veo en la figura 9 el nombre del folder de salida que dice en la leyenda que aparece en el archivo de configuración. TambiÃ©n aquÃ hay que revisar el orden de cada descripción, que coincida con la figura. Yo no puede revisarlo porque si agrando la imagen, se me pierde

CambiÃ© "origina graphics" por "source graphics"

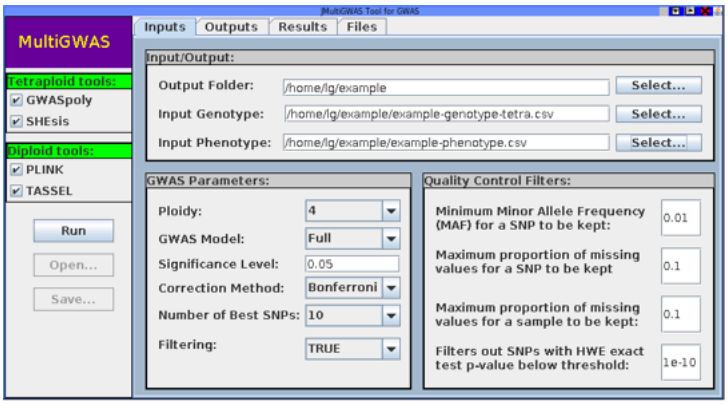


Figure 10: Main view of the MultiGWAS graphical user interface. The interface shows a main view on the center, a toolbar on the left and four tabs on top. From the main view, users can specify the input parameters for the analysis. From the toolbar, users can select the GWAS packages to be used in the analysis (two for tetraploids and two for diploids), and start the analysis with the current parameters (or load a previously saved configuration). And, from the tabs, in addition to specifying the input parameters, users can view the outputs of the process, the results of the analysis as an html report, and browse the source files that support the report.

5 Discussion

XXXXXXXXXXXXXXXXXXXXXXX

Challenges studying polyploid organisms are related to the complexity of (1) the data, and (2) inheritance mechanisms that are under study ((**?**)). The difficulties regarding the data complexity are the uncertainty in the allele dosage and null alleles, but these problems are opportunities to improve software at the variant calling stage. Moreover, there is an on-going understanding of the inheritance mechanisms for polyploids. For autopolyploids, most loci have a polysomic inheritance. However, sections of the genome that did not duplicate lead to disomic inheritance for some loci ((**?**; **?**; **?**)).

MultiGWAS does not face the problem of the allele dosage uncertainty. It is necessary to measure the impact of the allele dosage uncertainty at the GWAS stage to understand the effects arisen from this problem at the association stage. However, MultiGWAS addresses the second challenge variation of inheritance mechanisms by using different for existing software: two designed for polysomic inheritance ((**?**; **?**)) together with two for disomic inheritance ((**?**; **?**)). Thus it is a useful tool for researchers because it looks for significative associations that involve both types of inheritance.

Moreover, GWASpoly ((**?**)) offers models for different types of polyploid gene action additive, diploidized additive, duplex dominant, simplex dominant, and general. On the other hand, TASSEL ((**?**)) also models different types of gene action for diploids general, additive and dominant. We propose an automatic selection of

Luis esta parte revisala por fa.

the gene action model for both tools based on a balance between the factor of inflation and the replicability of the identified SNPs. We inform the user of the selected model based on the automatic strategy; we consider this information helps to understand the gene action model for the trait of interest. Even though the main focus is on the resultant SNPs, the model has assumptions that reflect the gene actions for a specific phenotype.

Replicability is a strategy to assess results derived from different methods. MultiGWAS integrates results to check for replicability in the results from different software. By combining results, we can compare the outputs and look for coincidences. Some software are more sensible, and others are more specific, but the integration allows MultiGWAS to balance specificity and sensitivity.

We provide the user with different graphic outputs to help to interpret the results. We report the SNP profile for the SNPs identified by more than one software. The SNP profile gives the researcher visual feedback from the SNP. We previously check for significative SNPs based on the p-value; however, 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, we designed the SNP profile.

6 Acknowledgements

This research was possible thanks to AGROSAVIA project *Investigación en conservación, caracterización y uso de los recursos genéticos vegetales*. COLCIENCIAS, today Minister of Science, Technology and Innovation of the republic of Colombia, for supporting the postdoctoral researcher L. Garreta at AGROSAVIA during 2019-2020 under the supervision of ICS and PHRH, (Grant number 811-2019). Colombian Corporation for Agricultural Research's editorial fund thanks for financing this publication. We thank Andres J. Cortes for helpful discussion.

Este agradecimiento es mÃis actualizado

Luis revisa si tenemos que agrecer a colciencias en otro modo