

1 MultiGWAS: A tool for GWAS analysis on
2 tetraploid organisms by integrating the results
3 of four GWAS software

4 L. Garreta¹, I. Cerón-Souza¹, M.R. Palacio², and P.H. Reyes-Herrera¹

5 ¹Corporación Colombiana de Investigación Agropecuaria
6 (AGROSAVIA), CI Tibaitatá, Kilómetro 14, Vía a Mosquera, 250047,
7 Colombia

8 ²Corporación Colombiana de Investigación Agropecuaria
9 (AGROSAVIA), CI El Mira, Kilómetro 38, Vía Tumaco Pasto,
10 Colombia

11 May 15, 2020

12 Abstract

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

30 **Contact:** phreyes@agrosavia.co

31
32 **Keywords:** GWAS, tetraploids, SNPs,XXX

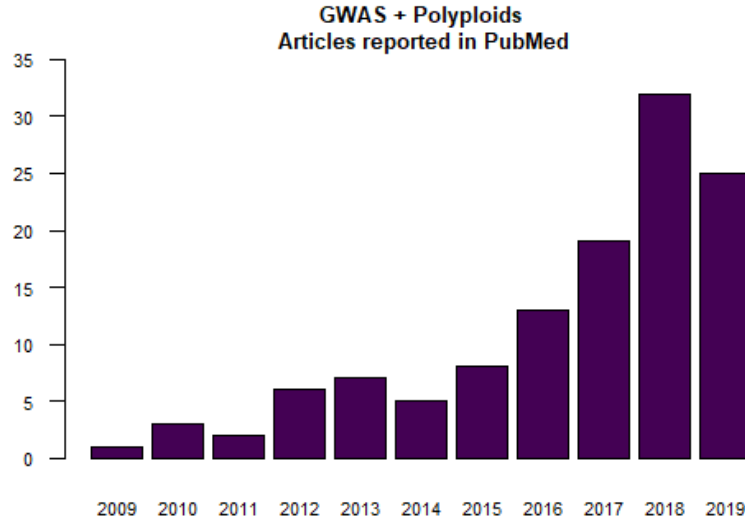


Figure 1: Timeline for articles reported for GWAS studies on polyploid species in PubMed. We present data for completed years.

1 Introduction

The GWAS (Genome-Wide Association Study) is used to identify which variants through the whole genome of a large number of individuals are associated with a specific trait [6, 2]. This methodology started with humans and several model plants, such as rice, maize, and *Arabidopsis* [20, 31, 7, 19, 15]. 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 [10, 11]. 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 1) [10, 26].

The GWAS for polyploid species has three 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 [9, 17]. 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

53 sample size varies [32]. It means that well-ranked SNPs from one package can be
54 ranked differently in another.

55 Second, although there are many GWAS software available to repeat the anal-
56 ysis under different conditions [14], most of them are designed exclusively for the
57 diploid data matrix [4]. Therefore, it is often necessary to "diploidizing" the poly-
58 ploid genomic data in order to replicate the analysis.

59 Third, there are very few tools focused on the integration of several GWAS soft-
60 ware, to make comparisons under different parameters and conditions across them.
61 As far as we know, there is only two software with this service in mind, such as iPAT
62 and easyGWAS.

63 The iPAT allows running in a graphic interface three well-known command-
64 line GWAS software such as GAPIT, PLINK, and FarmCPU (Chen and Zhang, 2018).
65 However, the output from each package is separated. On the other hand, the easyG-
66 WAS allows running a GWAS analysis on the web using different algorithms. This
67 analysis could run independently of both the computer capacity and operating sys-
68 tem. However, it needs either several datasets available or a dataset with a large
69 number of individuals to make replicates in order to compare among algorithms.
70 Moreover, the output from different algorithms is separated [13]. Thus, for both
71 software iPAT and easyGWAS, the integrative and comparative outputs among soft-
72 ware or algorithms are missing.

73 To solve all the three challenges above, we developed the MultiGWAS tool that
74 performs GWAS analyses for tetraploid species using four software in parallel. Our
75 tool include GWASpoly [25] and the SHEsis tool [28] that accept polyploid genomic
76 data, and PLINK [23] and TASSEL [5] with the use of a "diploidized" genomic ma-
77 trix. The tool deals with preprocessing data, running four GWAS tools in parallel,
78 and create comparative reports from the output of each software to help the user
79 to decide more intuitively the true or false associations.

80 2 Method

81 The MultiGWAS tool has three main steps: the adjustment stage, the multi analysis
82 stage, and the integration step (Fig. 4). In the first stage, MultiGWAS processes
83 the configuration file. Then it cleans and filters the genotype and phenotype, and
84 MultiGWAS "diploidize" the genomic data. In the second stage, each GWAS tool runs
85 in parallel. In the last stage, after the output files scanning, a summary of results is
86 generated in a report that contains score tables, Venn diagrams, SNP profiles, and
87 Manhattan plots.

88 2.1 Adjustment stage

89 MultiGWAS takes as input a configuration file where the user specifies the genomics
90 data along with the parameters that will be used by the four tools. Once the config-
91 uration file is processed, MultiGWAS preprocess the data that is cleaning, filtering,
92 and checking data quality. The output of this stage corresponds to the inputs for
93 the four programs at the Multi Analysis stage.

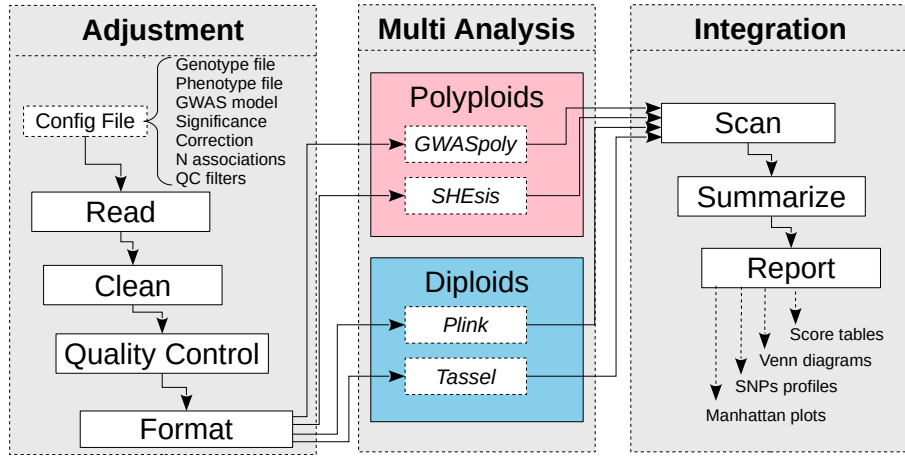


Figure 2: MultiGWAS flowchart has three stages: adjustment, multi analysis, and integration.

2.1.1 Configuration file

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

Input genotype and phenotype files: Currently, MultiGWAS uses two input files, one for genotype and the other for the phenotype. Both data correspond to data matrices with column and row names (Figure 3). The genotype file uses SNP markers in rows and samples in columns (Figure 3a). The phenotype file uses samples in rows and traits in columns (Figure 3b) with the first column corresponding to the sample name and the second column to trait value.

Marker,Chrom,Pos,Indiv01,Indiv02,Indiv03,...	Individual,Traitname
c2_41437,0,805179,AAAG,AAGG,AAGG,...	Indiv01, 3.59
c2_24258,0,1252430,AAGG,AGGG,GGGG,...	Indiv02, 4.07
c2_21332,0,3499519,TTCC,TTCC,TTCC,...	Indiv03, 1.05
...	...

Figure 3: **MultiGWAS genotype and phenotype formats.** Both files are in CSV format (Comma Separated Values) and contain as first row the header labels of the columns. Although the header labels are arbitrary, the column order is obligatory. **a.** Genotype file format, where “Marker”, “Chrom”, and “Pos”, correspond to the names for marker name, chromosome, and position in the first three columns respectively. The next columns correspond to the columns for the samples content. **b.** Phenotype file format, where “Individual” and “Traitname” are the column names for the individual and trait names, respectively.

GWAS model: MultiGWAS implements two types of GWAS analysis: (1) *naive* and (2) *full*. The *naive* model without any control for structure or relatedness between samples (without any additional parameters). Whereas the *full model*, known as the Q+K model, takes into account population structure (Q) and relatedness (K) to prevent false associations. We estimate Q and K inside MultiGWAS.

Both models use linear regression. The *naive* is modeled with Generalised 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) [33], which is described with the following equation:

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

where y is the vector of observed phenotypes; β is a vector of fixed effects; α is a vector of SNP effects (Quantitative Trait Nucleotides); γ 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 γ ; and X , S and Z are incidence matrices of ones and zeros relating y to β , α and μ , respectively.

Genome-wide significance and multiple testing correction: 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 [14, 25]. However, due to the massive number of statistical tests performed by GWAS, the *p-values* must be adjusted appropriately, performing a correction method for multiple hypothesis testing (MHT).

MultiGWAS allows two MHT methods for adjusting *p-values*: the Bonferroni correction and the false discovery rate (FDR). In the case of the Bonferroni correction, the threshold is α/m , where m is the number of valid markers from the genotype matrix (not missing or null values). (default $\alpha=0.05$, Correction=Bonferroni)

Quality Control filters: A control step is necessary to check the input data for genotype or phenotype errors that can lead to spurious GWAS results. MultiGWAS provides the filtering option (TRUE or FALSE) to check the input data by applying four quality control filters: Minor Allele Frequency (MAF), individual missing rate (MIND), SNP missing rate (GENO), and HardyWeinberg 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 {x} filters out SNPs with missing values exceeding {x}*100% (default 0.1); and HWE filters out SNPs which have Hardy-Weinberg equilibrium exact test *p-value* below the {x} threshold.

MultiGWAS does the MAF filtering. We use the PLINK package [14] for the other three filters: MIND, GENO, and HWE. Once we have the samples and SNPs that pass the filters, then we generate an input file for each package used.

Number of best ranked associations: Criticism has arisen in considering only statistically significant associations as the only possible correct associations [30, 18]. 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, 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*.

2.1.2 Data preprocessing

Once the configuration file is processed, then the genomic data is preprocessed by selecting individuals present in both genotype and phenotype and excluding individuals and SNPs that have poor quality. Moreover, 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: (e.g., 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 transform the genomic data into the formats required for each software.

2.2 Multi analysis stage

We have selected four GWAS software tools to be integrated into our multiGWAS tool. Two designed specifically for polyploid species: GWASpoly [25] and SHEsis [29]. Another two designed for diploids species and extensively used in humans and plants: PLINK [23, 8] and TASSEL [5], respectively. We briefly describe each tool as follows.

As MultiGWAS implements two types of GWAS analysis, *naive* and *full*, each tool is called in two different ways.

2.2.1 GWASpoly

GWASpoly [25] is an R package designed for GWAS in polyploid species used in several studies in plants [3, 12, 27, 34]. 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 a general gene action model along with four additional models: additive, simplex dominant, and duplex dominant. We use all the models to find associations.

MultiGWAS is using GWASpoly version 1.3. The population structure and relatedness, used in the *full* model, are estimated using the first five principal components and the kinship matrix, respectively, both calculated with the algorithms built-in GWASpoly.

181 2.2.2 SHEsis

182 SHEsis [28] is a program written in C designed for polyploid species. It includes
183 single locus association analysis, among others. It is based on a linear regression
184 model, and it has been used in some studies of animals and humans [24, 21].

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

189 2.2.3 PLINK

190 PLINK [23] is a program written in C frequently used in diploids species in particular
191 humans [22]. PLINK includes univariate GWAS using two-sample tests and linear
192 regression models.

193 MultiGWAS is using two versions of PLINK: 1.9 and 2.0. We use the linear
194 regression from PLINK 1.9 to achieve both types of analysis, *naive* and *full*. We use
195 the first five principal components calculated with the PLINK 1.9 built-in algorithm
196 to estimate the population structure. We estimate the relatedness from the kinship
197 coefficients calculated with the PLINK 2.0 built-in algorithm, removing the close
198 relatives or individuals with the first-degree relatedness.

199 2.2.4 TASSEL

200 TASSEL [5] is a GWAS program written in Java, frequently used for diploid plant
201 studies. TASSEL was developed for maize data [1, 35]. For association analysis,
202 TASSEL includes the general linear model (GLM) and mixed linear model (MLM)
203 that accounts for population structure and relatedness.

204 MultiGWAS is using TASSEL 5.0. The *naive* GWAS uses GLM. The *full* GWAS
205 uses MLM with two parameters: one for population structure, using the first five
206 principal components, and another for relatedness, using the kinship matrix with
207 centered IBS method, both calculated with the TASSEL built-in algorithms.

208 2.3 Integration stage.

209 The outputs resulting from the four software are scanned and processed to iden-
210 tify both significant and best-ranked associations. MultiGWAS corrects (correction
211 method defined in the configuration file) the p-values and calculates the threshold
212 value for each associated marker. The calculation uses the number of valid genotype
213 calls (i.e., non-missing genotypes, phenotypes, and covariates).

214 At this stage, We integrate the results to evaluate reproducible results among
215 tools (Fig 4). But, We still report a summary for the results of each tool:

- 216 • A Quantile-Quantile (QQ) plots for the resultant p-values of each tool and
217 the corresponding degree of inflation λ to asses the degree of the test statistic
218 inflation.

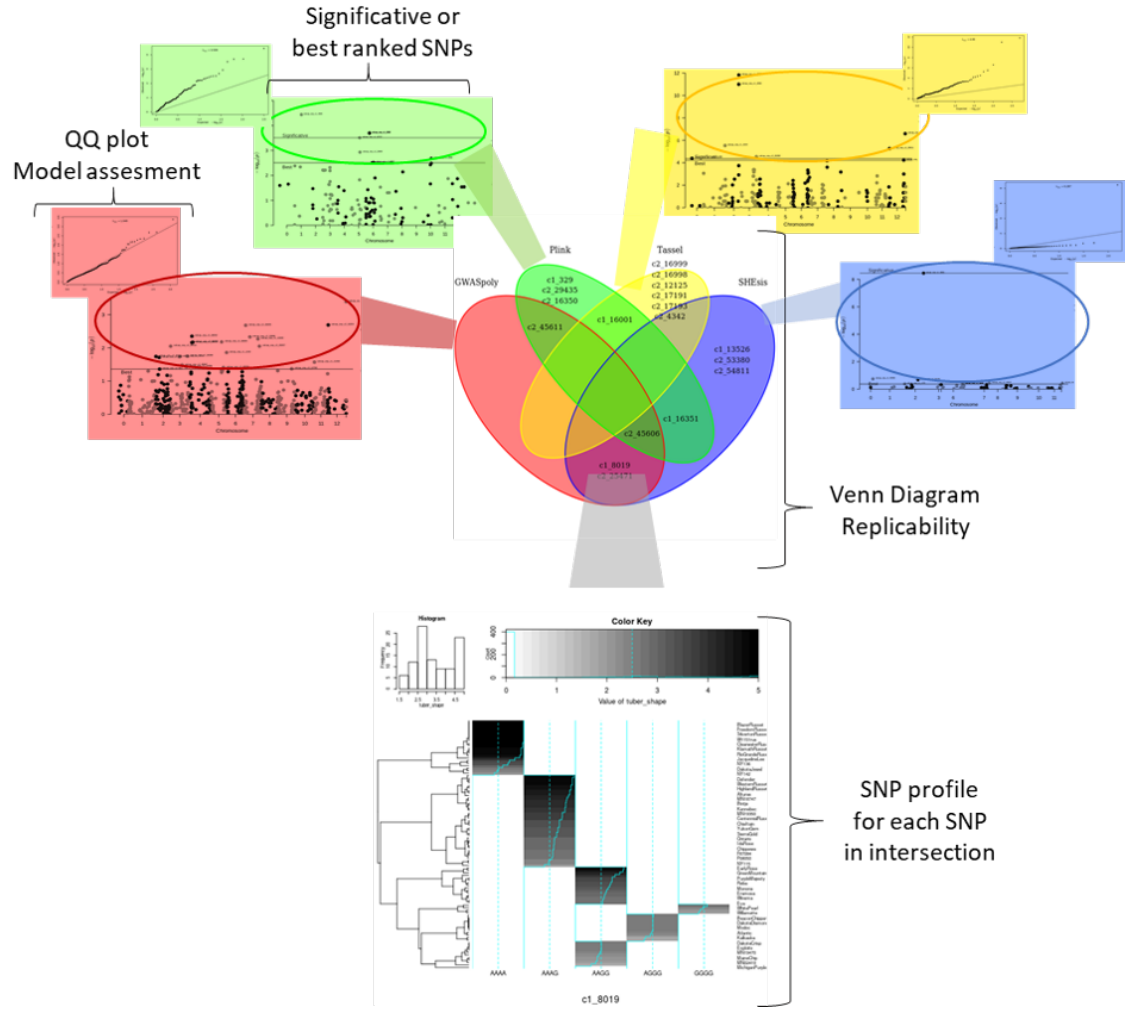


Figure 4: We present several reports. For each tool, first a QQ plot that asses the resultant p-values. Second, a Manhattan plot for each tool with two lines, blue and red, respectively, is the lower limit for the best ranked and significant SNPs. We present two Venn diagrams, one for the significant 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, respectively. Moreover, for each SNP that is in the intersection; thus, that is predicted by more than one tool we provide SNP profile.

- AA 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 significant 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. In addition, we provide detailed tables for the two sets.

For each SNP predicted 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, AAB, ABBB, BBBB. And 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. In the supplementary information, we include a report and a description of the report content (supplementary information XXX)

In the following section, we present the results applied to a public dataset.

3 Results

Although most of the GWAS packages used by MultiGWAS use linear regression approaches, they often produce different association results for the same input. For example, computed *p-values* for the same set of SNPs are different between packages; SNPs with significant *p-values* for one package maybe not significant for the others, or well-ranked SNPs in one package may be ranked differently in another. To alleviate these difficulties, MultiGWAS produces four types of outputs using different graphics and tabular views, including score tables, Venn diagrams, Manhattan and Q-Q plots, and SNP profiles. We designed these outputs to help users visually to compare, select, and interpret the set of possible SNPs associated with a trait of interest.

As an example of the functionality of the tool, here we show the results of running MultiGWAS tool in the genomic data from a tetraploid potato diversity panel, genotyped and phenotyped as part of the USDA-NIFA Solanaceae Coordinated Agricultural Project (SolCAP) [16]. The reports include: significant SNPs, best-ranked SNPs, profile SNPs, and visualization of associations. First, the best-ranked SNPs (Figure 6.b), where the SNP c2_45606 was evaluated with a high score by the four packages, but other four SNPs were also ranked with high scores by two packages simultaneously. Second, the significant SNPs (Figure 6.c), where the two polyploid software, GWASpoly and SHEsis, found as significant three SNPs, c1_8019, c2_25471, and c2_45606. In particular, the c1_8019 was also the most significant association found in the same potato dataset analyzed by Rosyara et al. (2016).

For each SNPs identified by more than one package, we provide the SNP profile (Figure 7), where for each significant association, a heat map figure is generated to summarize the genotype associated with a trait for each individual. Here we present, the SNP profile for the SNPs ranked among the best for the four tools c2_45606. We also include the SNP profile for the c1_8019 the most significant SNP for both polyploid tools. The SNP profiles for the markers present in the inter-sections are in the (supplementary information XXX)

And fourth, the visualization of associations (Figure 5), where for each package, a Manhattan and QQ plots are generated using special marks to help to identify significative, best-ranked, and shared SNPs (found by more than one tool).

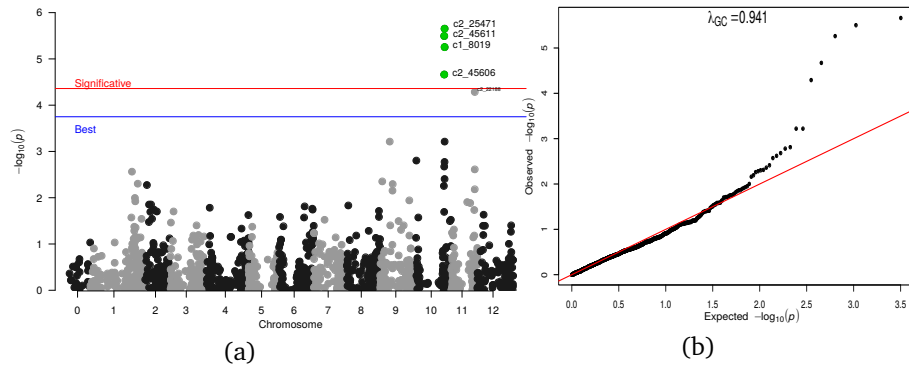


Figure 5: MultiGWAS visualization of associations. MultiGWAS adds special marks to the Manhattan and QQ plots to help identify different types of SNPs: (a) In Manhattan plots, significant SNPs are above a red line, best-ranked SNPs are above a blue line, and shared SNPs (See Figure 6.b) are colored in green (b) In QQ plots, a red diagonal line indicates the expectation, so potential associations can be observed when the number of SNPs deviating from the diagonal is small, as in the case of monogenic traits, or when this number is somewhat higher, as in the case of truly polygenic traits. However, deviations for a high number of SNPs could reflect inflated p -values owing to population structure or cryptic relatedness.

The complete report from MultiGWAS for the naive and full model is in the Supplementary information (<https://github.com/agrosavia-bioinformatics/multiGWAS>)

3.1 Visualization of shared SNPs

GWAS packages rely on p -value as a measure of association between each individual SNP and the trait of interest. The SNPs are considered statistically significant, and so possible true associations, when their p -value falls below a predefined significance level, usually 0.01 or 0.05. But, most GWAS packages compute differently both p -values and significance levels, it could result in non-significant SNPs. Consequently, it is important to know the significant SNPs. It is equally important to know the best-ranked SNPs closer to being statistically significant, as they may represent important associations to consider for posterior analysis (e.g. false negatives).

MultiGWAS provides tabular and graphic views to report in an integrated way both the best-ranked and significant SNPs identified by the four GWAS packages (see Figure 6). 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, the best-ranked SNPs correspond to the top-scored N SNPs that are the ones corresponding to the N lower p-values. 6.a) and in a Venn diagram (Figure 6.b). The table lists them by package and sorts by decreasing score, whereas the Venn diagram shows them emphasizing if these ones were best-ranked either in a single package or in several at once (shared). And second, the significant SNPs correspond to the ones assessed statistically significant by each package (depending on the threshold), they are shown in a Venn diagram (Figure 6.b), and they are also shown in the SNPs table, marked with significance TRUE and score greater than threshold, columns SGN, SCR, and THR, respectively in the table of the Figure 6.a.

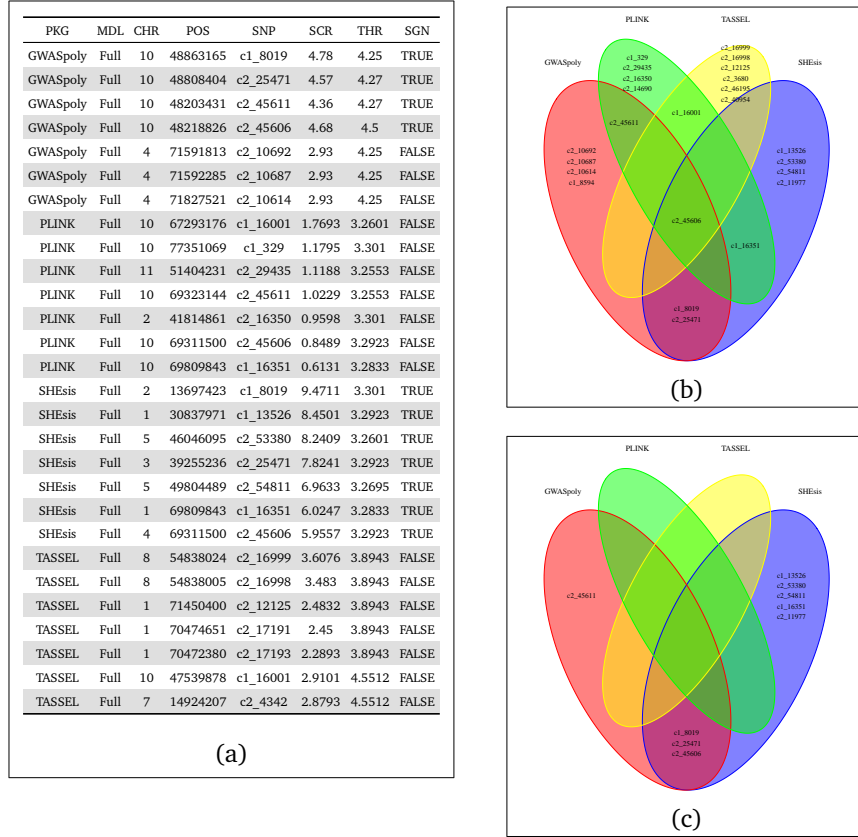


Figure 6: Visualization of shared SNPs. Tabular and graphical views of the best-ranked and significant SNPs identified by the four packages. (a) Tabular view with detailed information of each SNPs, including: package name (PKG), GWAS model used (MDL), chromosome (CHR), position in the genome (POS), ID (SNP), score (SCR), threshold (THR), and significance flag (SGN), whether the SNP was evaluated statistically significant or not (score > threshold). (b) Venn diagram with the best-ranked SNPs, showing that one SNP was shared by the four packages (c2_45606), other two only by the two polyploid packages GWASpoly and SHESis (c1_8019 and c2_25471), and other one only by the two diploid packages PLINK and TASSEL (c1_16001). (c) Venn diagram with the significant SNPs, showing that only three SNPs (c1_8019, c2_25471, and c2_45606) were evaluated as significant by the two polyploid packages GWASpoly and SHESis.

3.2 Visualization of shared SNPs

MultiGWAS creates a two-dimensional representation, what we called SNP profile, to visualize each trait by individuals and genotypes as rows and columns, respectively (Figure 7). 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.

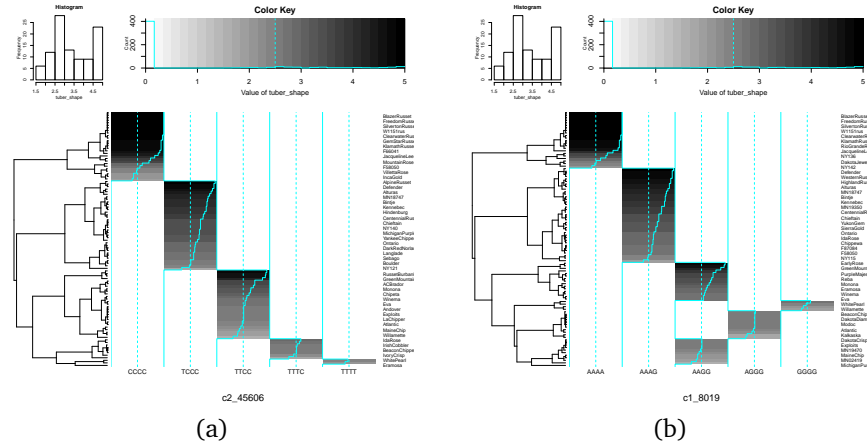


Figure 7: SNP profiles. SNP profiles for two of the best-ranked significant SNPs shown in the figure 6.b. (a) SNP c2_45606 best-ranked by the four packages (central intersection of the Venn diagram Figure 6.b) (b) SNP c1_8019 best-ranked by the two tetraploid packages (Figure 6.b), and also identified as significant by the same packages (at the bottom of the Figure 6.a).

4 Availability and implementation:

The core of the MultiGWAS tool was developed in 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 8). Source code, examples, documentation and installation instructions are available at <https://github.com/agrosavia-bioinformatics/multiGWAS>.

315 4.1 Input parameters

316 MultiGWAS uses as the only input a simple configuration text file where users set the
317 values for the main parameters that drives the GWAS process. The input parameters
318 include: the output folder where results will be written, input genotype/phenotype
319 filenames, genome-wide significance threshold, method for multiple testing cor-
320 rection, GWAS model, number of associations to be reported, and TRUE or FALSE
321 whether to use quality control filters or not. The filters are: minor allele frequency,
322 individual missing rate, SNP missing rate, and Hardy-Weinberg threshold.

323 The configuration file can be created either using a general text editor or using
324 the GUI application. In the first case, the file must have the structure shown in the
325 Figure 8.a, where parameter names and values are separated by colon, filenames
326 are enclosed in quotation marks, and TRUE or FALSE indicates wheter filters are ap-
327 plied or not. Moreover examples for the config file [https://github.com/agrosavia-](https://github.com/agrosavia-bioinformatics/MultiGWAS/tree/master/examples)
328 [bioinformatics/MultiGWAS/tree/master/examples](https://github.com/agrosavia-bioinformatics/MultiGWAS/tree/master/examples)

329 In the second case, the user creates the config file in a simple and straightforward
330 way using the input parameter view from the GUI application (Figure 8.b) and
331 clicking the “Save” button.

332 4.2 Using the command line interface

333 The execution of the tool in command line is simple, it only needs to open a linux
334 console, change to the folder where the configuration file was created, and type the
335 name of the executable tool followed by the filename of the configuration file, like
336 this:

```
337 multiGWAS full.config
```

338 Then, the tool starts the execution, showing information of the process in the
339 console window, and when it finishes the results are saved to a new subfolder called
340 “outgwas/reports. Results include a full html report containing the different views
341 described in the results section, along with the original graphics and summary tables
342 created by MultiGWAS and used to create the html report. Additionally, results
343 include the preprocessed tables of the main outputs generated by the four GWAS
344 packages used by MultiGWAS.

345 4.3 Using the graphical user interface

346 The MultiGWAS GUI application can be executed either by running from a Linux
347 console the *jmultiGWAS* command or by clicking on the Java application file *JMulti-*
348 *GWAS.jar* located in the “multiGWAS/sources” subfolder. After it opens, it shows a
349 main frame with four tabs at the top (Figure 8b): “Inputs”, “Outputs”, “Results”, and
350 “Files”. The “Inputs” tab shows the form to create the configuration file and run the
351 application. The “Outputs” tab shows the messages from the running process after
352 it starts the execution. The “Results” tab shows the full html report described above.
353 And the “Files” tab shows an embedded file browser pointing to the subfolder that
354 contains the original files used in the html report and described above.

```

Test01.config
default:
  genotypeFile      : "example-genotype.tbl"
  phenotypeFile     : "example-phenotype.tbl"
  significanceLevel : 0.05
  correctionMethod  : "Bonferroni"
  gwasModel         : Full
  nBest             : 5
  filtering         : TRUE
  MAF               : 0.01
  MIND              : 0.1
  GENO              : 0.1
  HWE               : 1e-10

```

(a)

(b)

Figure 8: MultiGWAS inputs and interaction. MultiGWAS uses as input a simple configuration text file and can be executed using either a command line interface script in R (CLI) or a graphic user interface application in Java (GUI). (a) An example of a configuration text file named “*Test01.config*” including the parameters that drive the GWAS process. It can be created using a general text editor or using the GUI application (see below) (b) Main view of the MultiGWAS GUI application (“Inputs” view) where users can create the configuration file by setting values for input parameters. The GUI contains other three views: “Outputs” view shows the logs of the running process. “Results” view shows a report in html format with the tabular and graphics described in the results section. And, the “Files” view shows an embedded file manager pointing to the subfolder that contains the files created by MultiGWAS and used to create the report.

5 Discussion

XXXXXXXXXXXXXXXXXXXXXXXXXXXX

References

- [1] María F. Álvarez et al. “Identification of Novel Associations of Candidate Genes with Resistance to Late Blight in *Solanum tuberosum* Group Phureja”. In: *Frontiers in Plant Science* 8 (2017), p. 1040. ISSN: 1664-462X. DOI: 10.3389/fpls.2017.01040. URL: <http://journal.frontiersin.org/article/10.3389/fpls.2017.01040/full>.

- [2] Ferdouse Begum et al. “Comprehensive literature review and statistical considerations for GWAS meta-analysis”. In: *Nucleic acids research* 40.9 (2012), pp. 3777–3784.
- [3] Jhon Berdugo-Cely et al. “Genetic diversity and association mapping in the colombian central collection of solanum tuberosum L. Andigenum group using SNPs markers”. In: *PLoS ONE* 12.3 (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0173039.
- [4] Peter M. Bourke et al. “Tools for Genetic Studies in Experimental Populations of Polyploids”. In: *Frontiers in Plant Science* 9 (2018), p. 513. ISSN: 1664-462X. DOI: 10.3389/fpls.2018.00513. URL: <http://journal.frontiersin.org/article/10.3389/fpls.2018.00513/full>.
- [5] Peter J Bradbury et al. “TASSEL: software for association mapping of complex traits in diverse samples”. In: *Bioinformatics* 23.19 (2007), pp. 2633–2635. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm308. URL: <https://doi.org/10.1093/bioinformatics/btm308>.
- [6] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. “Prioritizing GWAS results: a review of statistical methods and recommendations for their application”. In: *The American Journal of Human Genetics* 86.1 (2010), pp. 6–22.
- [7] Jun Cao et al. “Whole-genome sequencing of multiple Arabidopsis thaliana populations”. In: *Nature genetics* 43.10 (2011), p. 956.
- [8] Christopher C. Chang et al. “Second-generation PLINK: Rising to the challenge of larger and richer datasets”. In: *GigaScience* 4.1 (2015), pp. 1–16. ISSN: 2047217X. DOI: 10.1186/s13742-015-0047-8. arXiv: 1410.4803.
- [9] Rishika De, William S Bush, and Jason H Moore. “Bioinformatics Challenges in Genome-Wide Association Studies (GWAS)”. In: *Clinical Bioinformatics*. Ed. by Ronald Trent. New York, NY: Springer New York, 2014, pp. 63–81. ISBN: 978-1-4939-0847-9. DOI: 10.1007/978-1-4939-0847-9_5. URL: https://doi.org/10.1007/978-1-4939-0847-9_5.
- [10] Robert Ekblom and Juan Galindo. “Applications of next generation sequencing in molecular ecology of non-model organisms”. In: *Heredity* 107.1 (2011), pp. 1–15.
- [11] Hans Ellegren. “Genome sequencing and population genomics in non-model organisms”. In: *Trends in ecology & evolution* 29.1 (2014), pp. 51–63.
- [12] Luís Felipe V. Ferrão et al. “Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context”. In: *Frontiers in Ecology and Evolution* 6 (2018), p. 107. ISSN: 2296-701X. DOI: 10.3389/fevo.2018.00107. URL: <https://www.frontiersin.org/articles/10.3389/fevo.2018.00107/full>.

- [13] Dominik G Grimm et al. “easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies”. In: *The Plant Cell* 29.1 (2017), pp. 5–19. ISSN: 1040-4651. DOI: 10.1105/tpc.16.00551. URL: <http://www.plantcell.org/content/29/1/5>.
- [14] Anja C Gumpinger et al. *Methods and Tools in Genome-wide Association Studies*. Vol. 1819. 2018. ISBN: 9781493986187.
- [15] Bin Han and Xuehui Huang. “Sequencing-based genome-wide association study in rice”. In: *Current opinion in plant biology* 16.2 (2013), pp. 133–138.
- [16] Candice N. Hirsch et al. “Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries”. In: *G3: Genes, Genomes, Genetics* 3.6 (2013), pp. 1003–1013. ISSN: 21601836. DOI: 10.1534/g3.113.005595.
- [17] “How to interpret a genome-wide association study”. In: *JAMA - Journal of the American Medical Association* 299.11 (2008), pp. 1335–1344. ISSN: 00987484. DOI: 10.1001/jama.299.11.1335.
- [18] Avjinder S Kaler and Larry C Purcell. “Estimation of a significance threshold for genome-wide association studies”. In: *BMC Genomics* 20.1 (2019), p. 618. ISSN: 1471-2164. DOI: 10.1186/s12864-019-5992-7. URL: <https://doi.org/10.1186/s12864-019-5992-7>.
- [19] Arthur Korte and Ashley Farlow. “The advantages and limitations of trait analysis with GWAS: a review”. In: *Plant methods* 9.1 (2013), p. 29.
- [20] Gordan Lauc et al. “Genomics meets glycomics—the first GWAS study of human N-glycome identifies HNF1 α as a master regulator of plasma protein fucosylation”. In: *PLoS genetics* 6.12 (2010).
- [21] Jie Meng et al. “Genome-wide association analysis of nutrient traits in the oyster *Crassostrea gigas*: Genetic effect and interaction network”. In: *BMC Genomics* 20.1 (2019), pp. 1–14. ISSN: 14712164. DOI: 10.1186/s12864-019-5971-z.
- [22] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. “Microbial genome-wide association studies: lessons from human GWAS”. In: *Nature Reviews Genetics* 18.1 (2016), pp. 41–50. ISSN: 14710064. DOI: 10.1038/nrg.2016.132.
- [23] Shaun Purcell et al. “PLINK: A tool set for whole-genome association and population-based linkage analyses”. In: *American Journal of Human Genetics* 81.3 (2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795.
- [24] Hui Ping Qiao et al. “Genetic variants identified by GWAS was associated with colorectal cancer in the Han Chinese population”. In: *Journal of Cancer Research and Therapeutics* 11.2 (2015), pp. 468–470. ISSN: 19984138. DOI: 10.4103/0973-1482.150346.

- 442 [25] Umesh R. Rosyara et al. "Software for Genome-Wide Association Studies
443 in Autopolyploids and Its Application to Potato". In: *The Plant Genome* 9.2
444 (2016), pp. 1–10. ISSN: 1940-3372. DOI: 10.3835/plantgenome2015.
445 08.0073. URL: [https://dl.sciencesocieties.org/publications/
446 tpg/abstracts/9/2/plantgenome2015.08.0073](https://dl.sciencesocieties.org/publications/tpg/abstracts/9/2/plantgenome2015.08.0073).
- 447 [26] Anna W Santure and Dany Garant. "Wild GWAS—association mapping in
448 natural populations". In: *Molecular ecology resources* 18.4 (2018), pp. 729–
449 738.
- 450 [27] Sanjeev Kumar Sharma et al. "Linkage disequilibrium and evaluation of genome-
451 wide association mapping models in tetraploid potato". In: *G3: Genes, Genomes,
452 Genetics* 8.10 (2018), pp. 3185–3202. ISSN: 21601836. DOI: 10.1534/g3.
453 118.200377.
- 454 [28] Jiawei Shen et al. "SHEsisPlus, a toolset for genetic studies on polyploid
455 species". In: *Scientific Reports* 6 (2016), pp. 1–10. ISSN: 20452322. DOI: 10.
456 1038/srep24095. URL: <http://dx.doi.org/10.1038/srep24095>.
- 457 [29] Yong Yong Shi and Lin He. "SHEsis, a powerful software platform for analyses
458 of linkage disequilibrium, haplotype construction, and genetic association at
459 polymorphism loci (Cell Research (2005) 15, (97-98) DOI: 10.1038/sj.cr.7290272)".
460 In: *Cell Research* 16.10 (2006), p. 851. ISSN: 10010602. DOI: 10.1038/sj.
461 cr.7310101.
- 462 [30] John R Thompson, John Attia, and Cosetta Minelli. "The meta-analysis of
463 genome-wide association studies". In: *Briefings in Bioinformatics* 12.3 (2011),
464 pp. 259–269. ISSN: 1467-5463. DOI: 10.1093/bib/bbr020. URL: [https:
465 //doi.org/10.1093/bib/bbr020](https://doi.org/10.1093/bib/bbr020).
- 466 [31] Feng Tian et al. "Genome-wide association study of leaf architecture in the
467 maize nested association mapping population". In: *Nature genetics* 43.2 (2011),
468 pp. 159–162.
- 469 [32] Yan Y. Yan et al. "Effects of input data quantity on genome-wide association
470 studies (GWAS)". In: *International Journal of Data Mining and Bioinformatics*
471 22.1 (2019), pp. 19–43. ISSN: 17485681. DOI: 10.1504/IJDMB.2019.
472 099286.
- 473 [33] J Yu et al. "A unified mixed-model method for association mapping that ac-
474 counts for multiple levels of relatedness." In: *Nature genetics* 38.2 (2006),
475 pp. 203–208.
- 476 [34] Jiazheng Yuan et al. "Genome-Wide Association Study of Resistance to Potato
477 Common Scab". In: *Potato Research* (2019). ISSN: 18714528. DOI: 10.1007/
478 s11540-019-09437-w.
- 479 [35] Shengkui Zhang et al. "Genome-wide association studies of 11 agronomic
480 traits in cassava (*Manihot esculenta* crantz)". In: *Frontiers in Plant Science*
481 9.April (2018), pp. 1–15. ISSN: 1664462X. DOI: 10.3389/fpls.2018.
482 00503.