

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

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 August 13, 2020

12 **Abstract**

13 **Summary:** The Genome-Wide Association Studies (GWAS) are essential to
14 determine the genetic bases of either ecological or economic phenotypic variation
15 across individuals within populations of wild and domesticated species. For
16 this research question, current practice is the replication of the GWAS testing
17 different parameters and models to validate the reproducibility of results. How-
18 ever, straightforward methodologies that manage both replication and tetra-
19 ploid data are still missing. To solve this problem, we designed the MultiGWAS,
20 a tool that does GWAS for diploid and tetraploid organisms by executing in par-
21 allel four software, two for polyploid data (GWASPoly and SHEsis) and two for
22 diploids data (PLINK and TASSEL). MultiGWAS has several advantages. It runs
23 either in the command line or in an interface. It manages different genotype
24 formats, including VCF. It executes both the full and naïve models using sev-
25 eral quality filters. Besides, it calculates a score to choose the best gene action
26 model across GWASPoly and TASSEL. Finally, it generates several reports that
27 facilitate the identification of false associations from both the significant and the
28 best-ranked association SNP among the four software. We tested MultiGWAS
29 with tetraploid potato data. The execution demonstrated that the Venn diagram
30 and the other companion reports (i.e., Manhattan and QQ plots, heatmaps for
31 associated SNP profiles, and chord diagrams to trace associated SNP by chromo-
32 somes) were useful to identify associated SNP shared among different models
33 and parameters. Therefore, we confirmed that MultiGWAS is a suitable wrap-
34 ping tool that successfully handles GWAS replication in both diploid and tetra-
35 ploid organisms.

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

37 **Keywords:** GWASPoly, PLINK, polyploids, SNP, SHEsis, software, TASSEL

Las palabras clave NO deben ser repetición del título. Por eso borré las dos que ya estaban mencionadas y añadí las de los cuatro software que estamos usando. Por favor revisen si están de acuerdo

38

1 Introduction

39 The Genome-wide association studies (GWAS) are used to identify which variants
40 through the whole genome of a large number of individuals are associated with a
41 specific trait (Begum et al., 2012; Cantor et al., 2010). This methodology started
42 with humans and several model plants, such as rice, maize, and *Arabidopsis* (Cao
43 et al., 2011; Han and Huang, 2013; Korte and Farlow, 2013; Lauc et al., 2010;
44 Tian et al., 2011). Because of the advances in the next-gen sequencing technology
45 and the decline of the sequencing cost in recent years, there is an increase in the
46 availability of genome sequences of different organisms at a faster rate (Ekblom
47 and Galindo, 2011; Ellegren, 2014). Thus, the GWAS is becoming the standard
48 tool to understand the genetic bases of either ecologically or economically relevant
49 phenotypic variation for both model and non-model organisms. This increment
50 includes complex species such as polyploids (Fig. 1) (Ekblom and Galindo, 2011;
51 Santure and Garant, 2018).

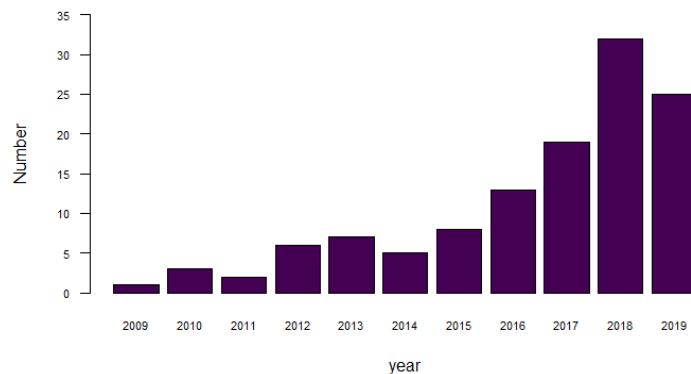


Figure 1: The number of peer-reviewed papers that contains the keywords "GWAS" and "polyploid" in the PubMed database between 2009 and 2019.

52 The GWAS for polyploid species has fourth related challenges. First, replication
53 is key to validate GWAS results and capture real associations. This approach
54 involved the use of different parameters, models, or conditions, to test how consist-
55 ent the results are either in the same software or in different GWAS tools (De et al.,
56 2014; Pearson and Manolio, 2008). However, the performance of different GWAS
57 software could affect the results. For example, the significance threshold for *pvalue*
58 changes through four GWAS software (i.e., PLINK, TASSEL, GAPIT, and FaST-LMM)
59 when the sample size varies (Yan et al., 2019). It means that well-ranked SNPs from
60 one package can be ranked differently in another.

61 Second, there are very few tools focused on the integration of several GWAS
62 software, to make comparisons under different parameters and conditions across

63 them. As far as we are aware, there is only two software with this service in mind,
64 which are iPAT and easyGWAS.

65 The iPAT allows running in a graphic interface three well-known command-line
66 GWAS software such as GAPIT, PLINK, and FarmCPU (Zhang et al., 2018). However,
67 the output from each package is splitted. On the other hand, the easyGWAS allows
68 running a GWAS analysis on the web using different algorithms. This analysis runs
69 independently of both the computer capacity and the operating system. However,
70 it needs either several databases available, or a dataset with a large number of
71 individuals to make replicates, and compare along the algorithms. Moreover, the
72 output from different algorithms is splitted also (Grimm et al., 2017). Thus, for
73 both software iPAT and easyGWAS, the integrative and comparative outputs among
74 software or algorithms are missing.

75 Third, although there are different GWAS software available to repeat the anal-
76 ysis under different conditions (Gumpinger et al., 2018), most of them are designed
77 exclusively for the diploid data matrix (Bourke et al., 2018). Therefore, it is often
78 necessary to "diploidizing" the polyploid genomic data in order to replicate the anal-
79 ysis. The main consequence of this process is missing the complexity of polyploid
80 data(Ferrão et al., 2018).

81 Finally, and related to the previous point, the GWAS on polyploids generates a
82 new level of complexity to understand how allele dosage affect the phenotype ex-
83 pression on quantitative traits. Therefore, any tool that compares among software
84 but also models different gene action will contribute to gain a better understand-
85 ing in how redundancy or complex interaction among alleles affect the phenotype
86 expression and the evolution of new phenotypes among polyploid species (Bourke
87 et al., 2018; Ferrão et al., 2018; Rosyara et al., 2016).

88 In order to overcome these challenges, we developed the MultiGWAS tool that
89 performs GWAS analyses for diploid and tetraploid species using four software in
90 parallel. Our tool includes GWASpoly (Rosyara et al., 2016) and the SHEsis tool
91 (Shen et al., 2016) that accept polyploid genomic data, and PLINK (Purcell et al.,
92 2007) and TASSEL (Bradbury et al., 2007), originally designed for diploids, but
93 that in the case of tetraploid data, they require "diploidizing" genomic matrix. The
94 tool deals with input file formats, data preprocessing, search for associations by
95 running four GWAS tools in parallel, and creation of comparative reports from the
96 output of each software to help the user to distinguish genuine associations from
97 false positives.

98 2 Method

99 The MultiGWAS tool has three main consecutive steps: the adjustment, the multi
100 analysis, and the integration (Fig. 2). In the adjustment step, MultiGWAS processes
101 the configuration file. Then it cleans and filters the genotype and phenotype, and
102 MultiGWAS "diploidize" the genomic data. Next, during the multi analysis, each
103 GWAS tool runs in parallel. Subsequently, in the integration step, the MultiGWAS
104 tool scans the output files from the four packages (i.e., GWASPoly, SHEsis, PLINK,
105 and TASSEL). Finally, it generates a summary of all results that contains score tables,

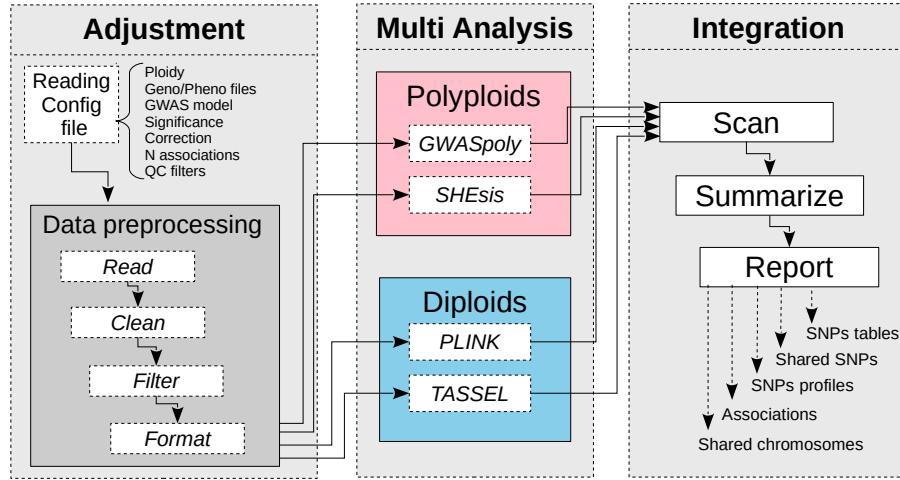


Figure 2: MultiGWAS flowchart has three steps: adjustment, multi analysis, and integration. In the first step, the user uploads the input data management, reading the configuration file, and reading and preprocessing the input data (genotype and phenotype). In the second step, the users perform the GWAS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

106 Venn diagrams, SNP profiles, and Manhattan plots.

107 2.1 Adjustment stage

108 MultiGWAS takes as input a configuration file where the user specifies the genomics
 109 data along with the parameters that will be used by the four tools. Once the config-
 110 uration file is read and processed, the genomic data files (genotype and phenotype)
 111 are preprocessed by cleaning, filtering, and checking data quality. The output of this
 112 stage corresponds to the inputs for the four programs at the Multi Analysis stage.

113 2.1.1 Reading configuration file

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

115 **Ploidy:** Numerical value for the ploidy level of the genotype, currently MultiGWAS
 116 supports diploids and tetraploids genotypes (2: for diploids, 4: for tetraploids).

117 **Genotype and phenotype input files:** MultiGWAS uses two input files, one for
 118 genotype and one for the phenotype. Genotypes files can be either in GWASpoly
 119 format (Rosyara et al., 2016) using SNP markers in rows and samples in columns
 120 (Fig. 3.a) or Variant Call Format (VCF) (Fig.3.b) which is transformed into GWAS-
 121 poly format using NGSEP 4.0.2 (Tello et al., 2019). The phenotype file contains

122 only one trait and uses a matrix format with the first column for the sample names
 123 and the second column for the trait values (Fig. 3.c).

a. <pre>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, ...</pre>	b. <pre>##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</pre>	c. <table border="1" style="border-collapse: collapse; width: 100%;"> <thead> <tr> <th>Individual,Trait</th> </tr> </thead> <tbody> <tr> <td>sample01, 3.59</td> </tr> <tr> <td>sample02, 4.07</td> </tr> <tr> <td>sample03, 1.05</td> </tr> </tbody> </table>	Individual,Trait	sample01, 3.59	sample02, 4.07	sample03, 1.05
Individual,Trait						
sample01, 3.59						
sample02, 4.07						
sample03, 1.05						

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 a 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 (Tello et al., 2019): ACN, ADP or BSDP. c. 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).

124 **GWAS model:** MultiGWAS is designed to work with quantitative phenotypes and
 125 can run GWAS analysis using two types of statistical models that we have called *full*
 126 and *naive* models. The *full model* is known in the literature as the Q+K model (Yu et
 127 al., 2006) and includes control for structure (Q) and relatedness between samples
 128 (K), whereas the *naive model* does not include any type of correction. Both models
 129 are based on linear regression approaches and variations of them are implemented
 130 by the four GWAS packages used by MultiGWAS. The *naive* is modeled with Gener-
 131 alized Linear Models (GLMs, Phenotype + Genotype), and the *full* is modeled with
 132 Mixed Linear Models (MLMs, Phenotype + Genotype + Structure + Kinship). The
 133 default model used by MultiGWAS is the *full model* (Q+K) (Yu et al., 2006), which
 134 is expressed with the following equation:

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

135 where y is the vector of observed phenotypes; β is a vector of fixed effects other
 136 than SNP or population group effects; α is a vector of SNP effects (Quantitative
 137 Trait Nucleotides); ν is a vector of population effects; μ is a vector of polygene
 138 background effects; e is a vector of residual effects; Q , modeled as a fixed effect,
 139 refers to the incidence matrix for subpopulation covariates relating y to ν ; and X ,
 140 S and Z are incidence matrices of 1s and 0s relating y to β , α and μ , respectively.

141 **Genome-wide significance:** GWAS searches SNPs associated with the phenotype
 142 in a statistically significant manner. A threshold or significance level α is specified
 143 and compared with the *p-value* derived for each association score. Standard signif-
 144 icance levels are 0.01 or 0.05 (Gumpinger et al., 2018; Rosyara et al., 2016), and
 145 MultiGWAS uses an α of 0.05 for the four GWAS packages. But the threshold is
 146 adjusted according to each package, as some packages as GWASpoly and TASSEL

147 calculates the SNP effect for each genotypic class using different gene action models
148 (see “Multi analysis stage”). So, the number of tested markers may be different in
149 each model (see below) that results in different *p-value* thresholds.

150 **Multiple testing correction:** Due to the massive number of statistical tests per-
151 formed by GWAS, it is necessary to perform a correction method for multiple hy-
152 pothesis testing and adjusting the *p-value* threshold accordingly. Two common
153 methods for multiple hypothesis testing are the false discovery rate (FDR) and the
154 Bonferroni correction. The latter is the default method used by MultiGWAS, which
155 is one of the most stringent methods. However, instead of adjusting the *p-values*,
156 MultiGWAS adjust the threshold below which a *p-value* is considered significant,
157 that is α/m , where α is the significance level and m is the number of tested markers
158 from the genotype matrix.

159 **Number of reported associations:** Criticism has arisen in considering only sta-
160 tistically significant associations as the only possible correct associations (Kaler and
161 Purcell, 2019; Thompson et al., 2011). Many of low *p-value* associations, closer to
162 being significant, are discarded due to the stringent significance levels, and conse-
163 quently increasing the number of false negatives. To help to analyze both signif-
164 icant and non-significant associations, MultiGWAS provides the option to specify
165 the number of best-ranked associations (lower *p-values*), adding the corresponding
166 *p-value* to each association found. In this way, it is possible to enlarge the number
167 of results, and we can observe replicability in the results for different programs.
168 Nevertheless, we present each association with the corresponding *p-value*.

169 **Quality control filters:** A control step is necessary to check the input data for
170 genotype or phenotype errors or poor quality that can lead to spurious GWAS re-
171 sults. MultiGWAS provides the option to select and define thresholds for the follow-
172 ing filters that control the data quality: Minor Allele Frequency (MAF), individual
173 missing rate (MIND), SNP missing rate (GENO), and Hardy-Weinberg threshold
174 (HWE):

- 175 • **MAF of x:** filters out SNPs with minor allele frequency below x (default 0.01);
- 176 • **MIND of x:** filters out all individuals with missing genotypes exceeding $x*100\%$
177 (default 0.1);
- 178 • **GENO of x:** filters out SNPs with missing values exceeding $x*100\%$ (default
179 0.1);
- 180 • **HWE of x:** filters out SNPs which have Hardy-Weinberg equilibrium exact test
181 *p-value* below the x threshold.

182 MultiGWAS does the MAF filtering, and uses the PLINK package (Gumpinger et al.,
183 2018) for the other three filters: MIND, GENO, and HWE.

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

187 **2.1.2 Data preprocessing**

188 Once the configuration file is processed, the genomic data is read and cleaned by se-
189 lecting individuals present in both genotype and phenotype. Then, individuals and
190 SNPs with poor quality are removed by considering the previous selected quality-
191 control filters and their thresholds,

192 At this point, the format "ACGT" suitable for the polyploid software GWAS-
193 poly and SHEsis, is "diploidized" for PLINK and TASSEL. The homozygous tetra-
194 ploid genotypes are converted to diploid thus: AAAA→AA, CCCC→CC, GGGG→GG,
195 TTTT→TT. Moreover, for tetraploid heterozygous genotypes, the conversion de-
196 pends on the reference and alternate alleles calculated for each position (e.g., AAAT
197 →AT, ... ,CCCG→CG).

198 After this process, the genomic data, genotype and phenotype, are converted to
199 the specific formats required for each of the four GWAS packages.

200 **2.2 Multi analysis stage**

201 MultiGWAS runs in parallel using two types of statistical models specified in the
202 parameters file, the Full model (Q+K) and Naive (i.e., without any control) where
203 Q refers to population structure and K refers to relatedness, calculated by kinship
204 coefficients across individuals (Sharma et al., 2018). The Full model (Q+K) controls
205 for both population structure and individual relatedness. For population structure,
206 MultiGWAS uses the Principal Component Analysis (PCA) and takes the top five PC
207 as covariates. For relatedness, MultiGWAS uses kinship matrices that TASSEL and
208 GWASpoly calculated separately, and for PLINK and SHEsis, relatedness depends on
209 kinship coefficients calculated with the PLINK 2.0 built-in algorithm (Chang et al.,
210 2015).

211 **2.2.1 GWASpoly**

212 GWASpoly (Rosyara et al., 2016) is an R package designed for GWAS in polyploid
213 species used in several studies in plants (Berdugo-Cely et al., 2017; Ferrão et al.,
214 2018; Sharma et al., 2018; Yuan et al., 2019). GWASpoly uses a Q+K linear mixed
215 model with biallelic SNPs that account for population structure and relatedness.
216 Also, to calculate the SNP effect for each genotypic class, GWASpoly provides eight
217 gene action models: general, additive, simplex dominant alternative, simplex dom-
218 inant reference, duplex dominant alternative, duplex dominant, diplo-general, and
219 diplo-additive. As a consequence, the number of statistical test performed can be
220 different in each action model and so thresholds below which the *p*-values are con-
221 sidered significant.

222 MultiGWAS is using GWASpoly version 1.3 with all gene action models available
223 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 (Meng et al., 2019; Qiao et al., 2015).

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 (Power et al., 2016). 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 (Álvarez et al., 2017; Zhang et al., 2018). 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.

261 2.3 Integration stage.

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

265 2.3.1 Calculation of *p-values* and significance thresholds

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

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

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

285 Furthermore, most GWAS packages compute both *p-values* and thresholds differ-
286 ently, with the consequence that significant associations identified by one package
287 do not reach the threshold of significance in the others. This results in the loss of
288 real associations, the so-called false negatives. To overcome these difficulties, Multi-
289 GWAS reports two sets of associations: significant and best-ranked (those closest to
290 being statistically significant), as described below.

291 2.3.2 Selection of significant and best-ranked associations

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

297 In the case of PLINK and SHEsis, which have a unique gene action model, the
298 associations are as described above. But, in the case of GWASpoly and TASSEL,
299 which have eight and three models respectively, MultiGWAS automatically selects
300 the "best gene action model" from each package and takes the associations from it.

301 This selection within GWASPoly and TASSEL has three criteria: the inflation factor
302 (I), the shared SNPs (R) and the significant SNPs (S).

303 Each gene action model is scored using the following equation:

304

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

305 where $score(M_i)$ is the score for the gene action model M_i , with i from 1.. k ,
306 for a GWAS package with k gene action models. I_i is the score for the inflation
307 factor defined as $I_i = 1 - |1 - \lambda(M_i)|$, where $\lambda(M_i)$ is the inflation factor for the
308 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
309 $|M_i \sim M_j|$ is the number of SNPs shared between M_i and M_j models, normalized by
310 the maximum number of SNPs shared between all models. And, S_i is the number of
311 significant SNPs of model M_i normalized by the total number SNPs shared among
312 all models.

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

320 **2.3.3 Integration of results**

321 At this stage, MultiGWAS integrates the results to evaluate reproducible results
322 among tools (Fig 4). However, it still reports a summary of the results of each
323 tool:

- 324
- A Quantile-Quantile (QQ) plots for the resultant *p-values* of each tool and
325 the corresponding inflation factor λ to assess the degree of the test statistic
326 inflation.

327

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

329 To present the replicability, we use two sets: (1) the set of all the significative SNPs
330 provided by each tool and (2) the set of all the best-ranked SNPs. For each set,
331 we present a Venn diagram that shows SNPs predicted exclusively by one tool and
332 intersections that help to identify the SNPs predicted by one, two, three, or all the
333 tools. Also, we provide detailed tables for the two sets.

334 For each SNP identified more than once, we provide what we call the SNP pro-
335 file. That is a heat diagram for a specific SNP, where each column is a genotype
336 state AAAA, AAAB, AABB, ABBC, and BBBB. Moreover, each row corresponds to a
337 sample. Samples with close genotypes form together clusters. Thus to generate
338 the clusters, we do not use the phenotype information. However, we present the
339 phenotype information in the figure as the color. This figure visually provides in-
340 formation regarding genotype and phenotype information simultaneously for the

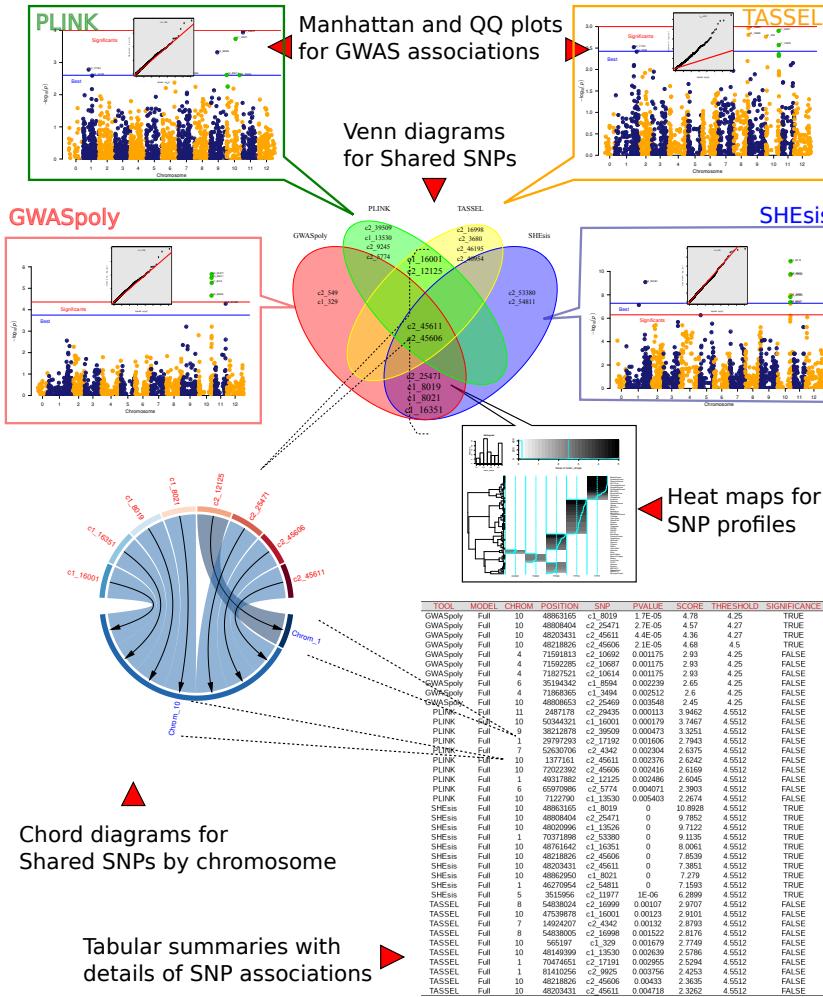


Figure 4: Reports presented by MultiGWAS. For each tool, first a QQ plot that assesses the resultant p -values. Second, a Manhattan plot for each tool with two lines, blue and red, respectively, is the lower limit for the best ranked and significative SNPs. We present two Venn diagrams, one for the significative 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.

341 whole population. We present colors as tones between white and black for color
342 blind people.

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

346 In the following section, we present the results of the functionality of the tool,
347 configured with a Full GWAS model using quality filters, and applied on a open
348 dataset of a diversity panel of a tetraploid potato, genotyped and phenotyped as part
349 of the USDA-NIFA Solanaceae Coordinated Agricultural Project (SolCAP) Hirsch
350 et al., 2013. The complete report of this analysis together with the report of a
351 second analysis using a naive GWAS model without quality filters are presented in
352 the supplementary materials S1 and S2, respectively.

Revisar cómo se pone en la
el formato de la revista el
supplementary material

353 3 Results

354 All four GWAS packages adopted by MultiGWAS use linear regression approaches.
355 However, they often produce different association results for the same input. Com-
356 puted *p-values* for the same set of SNPs are different between packages. Therefore,
357 SNPs with significant *p-values* for one package maybe not significant for the oth-
358 ers. Alternatively, well-ranked SNPs in one package may be ranked differently in
359 another.

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

- 364 • Manhattan and Q-Q plots to show GWAS associations.
- 365 • Venn diagrams to show associations identified by single or several tools.
- 366 • Heat diagrams to show the genotypic structure of shared SNPs.
- 367 • Chord diagrams to show shared SNPs by chromosomes.
- 368 • Score tables to show detailed information of associations for both summary
369 results from MultiGWAS and particular results from each GWAS package

370 The complete reports generated by MultiGWAS for both types of analysis, full
371 and naive, applied to the diversity panel of tetraploid potato are in the supplemen-
372 tary information at <https://github.com/agrosavia-bioinformatics/multiGWAS-Supplementary>.

revisa hipervinculo

374 3.1 Manhattan and QQ plots for GWAS associations

375 MultiGWAS uses classical Manhattan and Quantile–Quantile plots (QQ plots) to
376 visualize the results of each package. In both plots, the points are the SNPs and
377 their *p-values* are transformed into scores like $-\log_{10}(p\text{-values})$ (see Fig. 5). The

378 Manhattan plot shows the strength of association of the SNPs (y-axis) distributed at
 379 their genomic location (x-axis), so the higher the score, the stronger the association.
 380 At the same time, the QQ plot compares the expected distribution of p -values (y-
 381 axis) with the observed distribution (x-axis).

382 MultiGWAS adds distinctive marks to both plots to identify different types of
 383 SNPs: (a) In the Manhattan plots, the significant SNPs are above a red line and the
 384 best-ranked SNPs are above a blue line. Also, SNPs shared between packages are
 385 coloured green (See Fig. 6.b). (b) In the QQ plots, a red diagonal line indicates
 386 the expected distribution under the null hypothesis of no association of SNPs with
 387 the phenotype, both distributions should coincide, and most SNPs should lie on the
 388 diagonal line. Deviations for a large number of SNPs may reflect inflated p -values
 389 due to population structure or cryptic relatedness. But, few SNPs deviate from the
 390 diagonal for a truly polygenic trait (Power et al., 2016).

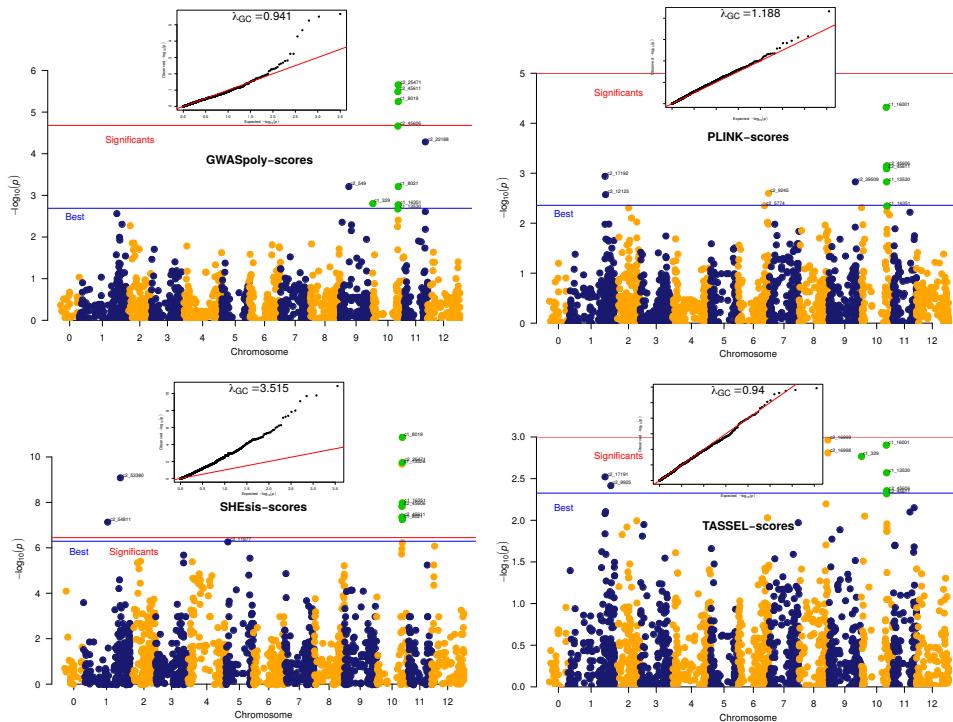


Figure 5: Associations in the tetraploid potato dataset. MultiGWAS shows the associations identified by the four GWAS packages using Manhattan and QQ plots. The tetraploid potato data showed several SNPs shared between the four software (green dots). The best-ranked SNPs are above the blue line, but 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.

391 3.2 Tables and Venn diagrams for single and shared SNPs

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

396 First, best-ranked SNPs correspond to the top-scored N SNPs, whether they were
397 assessed significant or not by its package, and with N defined by the user in the
398 configuration file. These SNPs appears in both a SNPs table (Figure 6.a), and in a
399 Venn diagram (Figure 6.b). The table lists them by package and sorts by decreasing
400 score, whereas the Venn diagram emphasizes if they were best-ranked either in a
401 single package or in several at once (shared).

402 Second, the significant SNPs correspond to the ones valued statistically signifi-
403 cant by each package. They appear in a Venn diagram (Figure 6.c), and in the SNPs
404 table, marked with significance TRUE (T) in the table of the Figure 6.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.

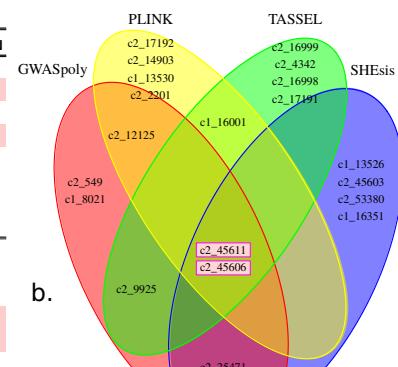
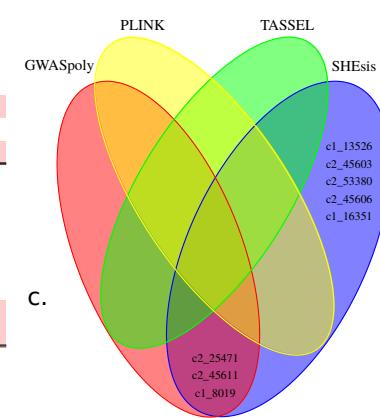
**b.****c.**

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).

405 In this analysis, the polyploid packages GWASpoly and SHEsis co-identified
 406 the SNPs c2_25471, c2_45611, and c1_8019. Of these SNPs, c1_8019, has been
 407 reported in previous studies to be associated with tuber shape and depth of eye
 408 traits (Rosyara et al., 2016; Sharma et al., 2018). Furthermore, in another analy-
 409 sis of MultiGWAS using a naive model without filters (Supplemental Material S2),
 410 the SNP c1_8019 was co-identified by three packages: GWASpoly, SHEsis, and the
 411 diploid PLINK package .

412 3.3 Heat diagrams for the structure of shared SNPs

413 MultiGWAS creates a two-dimensional representation, called the SNP profile, to vi-
 414 sualize each trait by individuals and genotypes as rows and columns, respectively
 415 (Figure 7). At the left, the individuals are grouped in a dendrogram by their geno-
 416 type. At the right, there is the name or ID of each individual. At the bottom, the
 417 genotypes are ordered from left to right, starting from the major to the minor allele
 418 (i.e., AAAA, AAAB, AABB, ABBB, BBBB). At the top, there is a description of the
 419 trait based on a histogram of frequency (top left) and by an assigned color for each
 420 numerical phenotype value using a grayscale (top right). Thus, each individual ap-
 421 pears as a colored line by its phenotype value on its genotype column. For each
 422 column, there is a solid cyan line with the mean of each column and a broken cyan
 423 line that indicates how far the cell deviates from the mean.

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

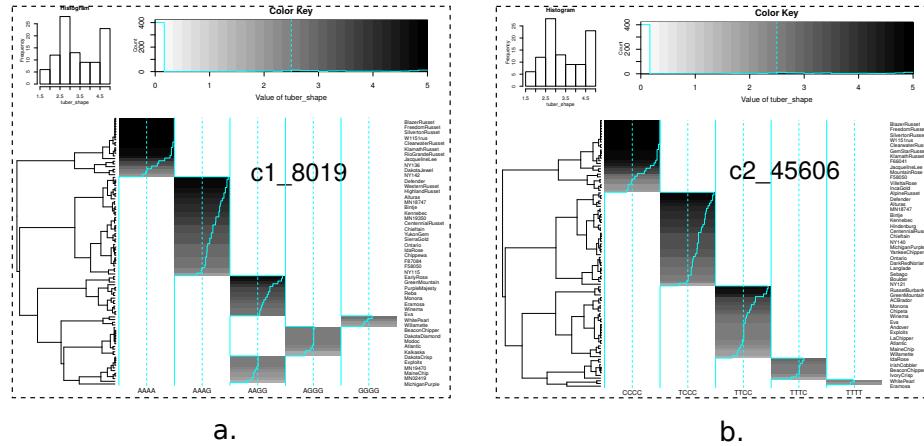


Figure 7: SNP profiles. SNP profiles for two of the best-ranked significant SNPs shown in figure 6.b. (a) SNP c2_45606 best-ranked by the four packages (central intersection of the Venn diagram Figure 6.b) (b) SNP c1_8019 best-ranked by the two tetraploid packages (Figure 6.b), and also identified as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

426 3.4 Chord diagrams for SNPs by chromosome

427 The chord diagrams visualize the location across the genome of the best-ranked
 428 associated SNPs shared among the four packages and described in the table 6.a.
 429 Thus, in the case of the tetraploid potato, we found that they are located mostly in
 430 chromosome 10 (Figure 8.a). This visualization complements the manhattan plots
 431 from each GWAS package (Figure 8.b).

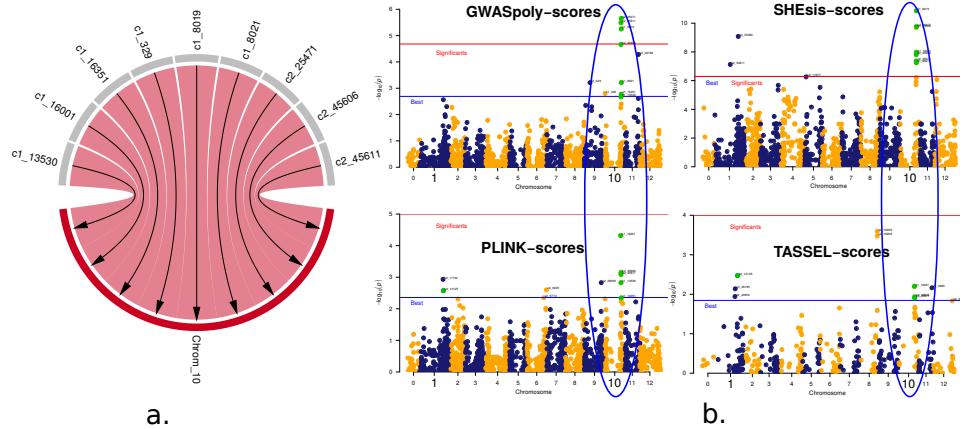


Figure 8: SNPs by chromosome. The position of best-ranked SNPs across chromosomes using two different visualizations. (a) Chord diagram showing that best-ranked SNPs located in chromosome 10. The SNPs are at the top and the chromosomes at the bottom. The arrows connect the best-ranked SNPs with their 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 10). 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 9.

```

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.

447 4.2 Using the command line interface

448 The execution of the CLI tool is simple. It only needs to open a Linux console,
 449 change to the folder where the configuration file was created, and type the name of
 450 the executable tool followed by the the filename of the configuration file, like this:
 451 multiGWAS Test01.config

452 Then, the tool starts the execution, showing information on the process in the
 453 console window. When it finishes, the results are in a new subfolder called “*out-*
 454 *Test01*. The results include a complete HTML report containing the different views
 455 described in the results section, the source graphics and tables supporting the re-
 456 port, and the preprocessed tables from the results generated by the four GWAS
 457 packages used by MultiGWAS.

458 4.3 Using the graphical user interface

459 The interface allows users to save, load or specify the different input parameters
 460 for MultiGWAS in a friendly way (Fig. 10. The input parameters correspond to the
 461 settings included in the configuration file described in subsection 2.1.1. It executes
 462 by calling the following command from a Linux console:

463 jmultiGWAS

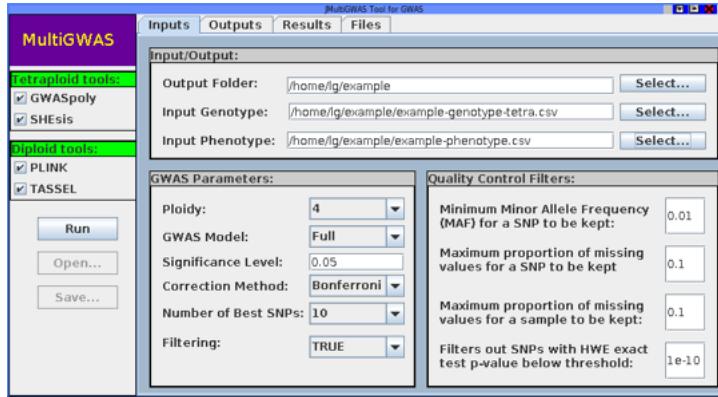


Figure 10: Main view of the MultiGWAS graphical user interface. The interface has a toolbar at the left side and four tabs at the top. In the toolbar, users can select the GWAS packages (Two for tetraploids and two for diploids). The analysis starts with the current parameters or loading a previously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

468

466 5 Discussion

467 The reanalysis of potato data with MultiGWAS showed that this tool is handy to im-
 468 prove the GWAS in tetraploid species. Through MultiGWAS performance, we could
 469 test its effectiveness to answer some of the challenges associating phenotypic in a
 470 polyploid organism. They include the integration and replication among param-
 471 eters and software, the diploidization of polyploid data, and the incorporation of
 472 allele dosage models (Dufresne et al., 2014).

473 The main advantage of MultiGWAS is that replicate the GWAS analysis among
 474 four software and integrate the results obtained across software, models and param-
 475 eters. Depending on the software, users usually have to choose between sensitivity
 476 or specificity. But using MultiGWAS, users do not have to choose between both
 477 approaches because they can observe their effect in the analysis within the same
 478 environment.

479 Another difficulty for replication among software is the variability of structures
 480 for the genomic input data. Currently, the most common format for next-generation
 481 sequencing variant data is the VCF (Variant Call Format) (Danecek et al., 2011;
 482 Ebbert et al., 2014) . One of the advantages of VCF is the versatility to summarize
 483 important genome information for hundreds or thousands of individuals and SNP,
 484 including information about levels of ploidy. MultiGWAS different from most of
 485 GWAS software available allows the VCF as an input (but see VarStats tool in VTC).

486 Moreover, the MultiGWAS is the unique tool as far as we know that allows us to

Corregido, Revisar

Luis, revisa la leyenda de la Fig. 10, si estás conforme con esta descripción de la figura. Traté de hacerla más descriptiva

487 compare the effect of diploidized the tetraploid data in the performance of the anal-
488 ysis directly. The graphic outputs are a handy approach to find similar results. The
489 SNP profile allows identifying what the significant associations detected by more
490 than one software are. Furthermore, although MultiGWAS check for significative
491 SNPs based on the p-value, it is essential to go back to the data and check if the
492 SNP is a real association between the genotype and phenotype. For this purpose,
493 the SNP profile gives visual feedback for the accuracy of the association.

494 Finally, the MultyGWAS allows comparing among the gene action models that
495 offer GWASPoly and TASSEL. GWASPoly (Rosyara et al., 2016) provides models of
496 different types of polyploid gene action including additive, diploidized additive, du-
497 plex dominant, simplex dominant, and general. On the other hand, TASSEL (Brad-
498 bury et al., 2007) also models different types of gene action for diploids general,
499 additive and dominant. To choose among models, We propose an automatic se-
500 lection of the gene action model for both tools based on a balance between three
501 criteria: the inflation factor, the replicability of identified SNPs and the significance
502 of identified SNPs. This inflation index is a new tool for comparison that do not
503 offer either GWASPoly or TASSEL. This automatic strategy will help to understand
504 the gene action model for the trait of interest. Even though the main focus is on
505 the resultant SNPs, the model has assumptions that reflect the gene actions for a
506 specific phenotype.

507 On the other hand, although MultyGWAS does not solve the uncertainty in
508 the allele dosage and null alleles, However the active comparison among models
509 that MultiGWAS addresses the search of the inheritance mechanisms by comparing
510 among two designed for polysomic inheritance software (Rosyara et al., 2016; Shen
511 et al., 2016) with two software for disomic inheritance (Bradbury et al., 2007; Pur-
512 cell et al., 2007). Understanding the inheritance mechanisms for polyploids organ-
513 ism is an open question. For autopolyploids, most loci have a polysomic heritage.
514 However, sections of the genome that did not duplicate lead to disomic inheritance
515 for some loci (Dufresne et al., 2014; Lynch and Conery, 2000; Ohno, 1970). Thus
516 it is a useful tool for researchers because it looks for significative associations that
517 involve both types of inheritance.

518 6 Acknowledgements

519 This research was possible thanks to AGROSAVIA five-years macroproject entitled
520 *Investigación en conservación, caracterización y uso de los recursos genéticos vegetales*.

521 We thanks to the Minister of Science, Technology and Innovation of the re-
522 public of Colombia (previously COLCIENCIAS), for supporting the postdoctoral re-
523 searcher L. Garreta at AGROSAVIA during 2019-2020 under the supervision of ICS
524 and PHRH, (Grant number 811-2019). The editorial of AGROSAVIA gave for fi-
525 nancial supporting for this publication. Finally to Andres J. Cortes for his valuable
526 comments to improve this manuscript.

⁵²⁷ **7 Author Contributions**

⁵²⁸ LG, ICS, and PHRH conceived the idea. LG developed MultiGWAS. MP tested Multi-
⁵²⁹ GWAS. LG, ICS, and PHRH drafted the first version of this manuscript, edited by
⁵³⁰ the other co-authors.

⁵³¹ **References**

- ⁵³² Álvarez, M. F., Angarita, M., Delgado, M. C., García, C., Jiménez-Gomez, J., Geb-
⁵³³ hardt, C., & Mosquera, T. (2017). Identification of Novel Associations of
⁵³⁴ Candidate Genes with Resistance to Late Blight in Solanum tuberosum
⁵³⁵ Group Phureja. *Frontiers in Plant Science*, 8, 1040. <https://doi.org/10.3389/fpls.2017.01040>
- ⁵³⁶ Begum, F., Ghosh, D., Tseng, G. C., & Feingold, E. (2012). Comprehensive literature
⁵³⁷ review and statistical considerations for gwas meta-analysis. *Nucleic acids*
⁵³⁸ *research*, 40(9), 3777–3784.
- ⁵³⁹ Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., & Yock-
⁵⁴⁰ teng, R. (2017). Genetic diversity and association mapping in the colom-
⁵⁴¹ bian central collection of solanum tuberosum L. Andigenum group using
⁵⁴² SNPs markers. *PLoS ONE*, 12(3). <https://doi.org/10.1371/journal.pone.0173039>
- ⁵⁴³ Bourke, P. M., Voorrips, R. E., Visser, R. G. F., & Maliepaard, C. (2018). Tools for Ge-
⁵⁴⁴ netic Studies in Experimental Populations of Polyploids. *Frontiers in Plant*
⁵⁴⁵ *Science*, 9, 513. <https://doi.org/10.3389/fpls.2018.00513>
- ⁵⁴⁶ Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler,
⁵⁴⁷ E. S. (2007). TASSEL: software for association mapping of complex traits
in diverse samples. *Bioinformatics*, 23(19), 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- ⁵⁴⁸ Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing gwas results: A
⁵⁴⁹ review of statistical methods and recommendations for their application.
⁵⁵⁰ *The American Journal of Human Genetics*, 86(1), 6–22.
- ⁵⁵¹ Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D.,
⁵⁵² Lanz, C., Stegle, O., Lippert, C., Et al. (2011). Whole-genome sequencing
⁵⁵³ of multiple arabidopsis thaliana populations. *Nature genetics*, 43(10), 956.
- ⁵⁵⁴ Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J.
⁵⁵⁵ (2015). Second-generation PLINK: Rising to the challenge of larger and
⁵⁵⁶ richer datasets. *GigaScience*, 4(1), arXiv 1410.4803, 1–16. <https://doi.org/10.1186/s13742-015-0047-8>
- ⁵⁵⁷ Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Hand-
⁵⁵⁸ saker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin,
⁵⁵⁹ R., & Group, 1. G. P. A. (2011). The variant call format and VCFtools.
⁵⁶⁰ *Bioinformatics*, 27(15), <https://academic.oup.com/bioinformatics/article-pdf/27/15/2156/1125001/btr330.pdf>, 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>

- 568 De, R., Bush, W. S., & Moore, J. H. (2014). Bioinformatics Challenges in Genome-
569 Wide Association Studies (GWAS). In R. Trent (Ed.), *Clinical bioinformatics*
570 (pp. 63–81). New York, NY, Springer New York. https://doi.org/10.1007/978-1-4939-0847-9_5
- 572 Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and
573 challenges in population genetics of polyploid organisms: An overview of
574 current state-of-the-art molecular and statistical tools. *Molecular Ecology*,
575 23(1), <https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12581>, 40–
576 69. <https://doi.org/10.1111/mec.12581>
- 577 Ebbert, M. T., Wadsworth, M. E., Boehme, K. L., Hoyt, K. L., Sharp, A. R., D O'Fallon,
578 B., Kauwe, J. S., & Ridge, P G. (2014). Variant tool chest: An improved tool
579 to analyze and manipulate variant call format (vcf) files. *BMC bioinformatics*, 15(S7), S12.
- 581 Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in
582 molecular ecology of non-model organisms. *Heredity*, 107(1), 1–15.
- 583 Ellegren, H. (2014). Genome sequencing and population genomics in non-model
584 organisms. *Trends in ecology & evolution*, 29(1), 51–63.
- 585 Ferrão, L. F. V., Benevenuto, J., Oliveira, I. d. B., Cellon, C., Olmstead, J., Kirst, M.,
586 Resende, M. F. R., & Munoz, P. (2018). Insights Into the Genetic Basis
587 of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a
588 GWAS Context. *Frontiers in Ecology and Evolution*, 6, 107. <https://doi.org/10.3389/fevo.2018.00107>
589 - Paper for layout. - Many concepts of GWAS, especially structure popula-
590 tion.
591
- 592 Grimm, D. G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C.,
593 Lippert, C., Stegle, O., Schölkopf, B., Weigel, D., & Borgwardt, K. M. (2017). easyGWAS: A Cloud-Based Platform for Comparing the Results of
594 Genome-Wide Association Studies. *The Plant Cell*, 29(1), 5–19. <https://doi.org/10.1105/tpc.16.00551>
- 597 Gumpinger, A. C., Roqueiro, D., Grimm, D. G., & Borgwardt, K. M. (2018). *Methods and Tools in Genome-wide Association Studies* (Vol. 1819).
- 599 Han, B., & Huang, X. (2013). Sequencing-based genome-wide association study in
600 rice. *Current opinion in plant biology*, 16(2), 133–138.
- 601 Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De
602 Jong, W., Veilleux, R. E., Jansky, S., Bethke, P., Douches, D. S., & Buell, C. R.
603 (2013). Retrospective view of North American potato (*Solanum tuberosum*
604 L.) breeding in the 20th and 21st centuries. *G3: Genes, Genomes, Genetics*,
605 3(6), 1003–1013. <https://doi.org/10.1534/g3.113.005595>
- 606 Kaler, A. S., & Purcell, L. C. (2019). Estimation of a significance threshold for genome-
607 wide association studies. *BMC Genomics*, 20(1), 618. <https://doi.org/10.1186/s12864-019-5992-7>
- 609 Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with
610 gwas: A review. *Plant methods*, 9(1), 29.
- 611 Lauc, G., Essafi, A., Huffman, J. E., Hayward, C., Knežević, A., Kattla, J. J., Polašek,
612 O., Gornik, O., Vitart, V., Abrahams, J. L., Et al. (2010). Genomics meets

- 613 glycomics—the first gwas study of human n-glycome identifies *hnf1α* as a
614 master regulator of plasma protein fucosylation. *PLoS genetics*, 6(12).
- 615 Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of du-
616 plicate genes. *science*, 290(5494), 1151–1155.
- 617 Meng, J., Song, K., Li, C., Liu, S., Shi, R., Li, B., Wang, T., Li, A., Que, H., Li, L., &
618 Zhang, G. (2019). Genome-wide association analysis of nutrient traits in
619 the oyster *Crassostrea gigas*: Genetic effect and interaction network. *BMC*
620 *Genomics*, 20(1), 1–14. <https://doi.org/10.1186/s12864-019-5971-z>
- 621 Ohno, S. (1970). *Evolution by gene duplication*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-86659-3>
- 622 Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association
623 study. *JAMA - Journal of the American Medical Association*, 299(11), 1335–
624 1344. <https://doi.org/10.1001/jama.299.11.1335>
- 625 Power, R. A., Parkhill, J., & De Oliveira, T. (2016). Microbial genome-wide associa-
626 tion studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1),
627 41–50. <https://doi.org/10.1038/nrg.2016.132>
- 628 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller,
629 J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool
630 set for whole-genome association and population-based linkage analyses.
631 *American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- 632 Qiao, H. P., Zhang, C. Y., Yu, Z. L., Li, Q. M., Jiao, Y., & Cao, J. P. (2015). Genetic vari-
633 ants identified by GWAS was associated with colorectal cancer in the Han
634 Chinese population. *Journal of Cancer Research and Therapeutics*, 11(2),
635 468–470. <https://doi.org/10.4103/0973-1482.150346>
- 636 Rosyara, U. R., De Jong, W. S., Douches, D. S., & Endelman, J. B. (2016). Software
637 for Genome-Wide Association Studies in Autopolyploids and Its Applica-
638 tion to Potato. *The Plant Genome*, 9(2), 1–10. <https://doi.org/10.3835/plantgenome2015.08.0073>
- 639 Santure, A. W., & Garant, D. (2018). Wild gwas—association mapping in natural
640 populations. *Molecular ecology resources*, 18(4), 729–738.
- 641 Sharma, S. K., MacKenzie, K., McLean, K., Dale, F., Daniels, S., & Bryan, G. J.
642 (2018). Linkage disequilibrium and evaluation of genome-wide associa-
643 tion mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*,
644 8(10), 3185–3202. <https://doi.org/10.1534/g3.118.200377>
- 645 Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., & Shi, Y. (2016). SHEsisPlus, a toolset
646 for genetic studies on polyploid species. *Scientific Reports*, 6, 1–10. <https://doi.org/10.1038/srep24095>
- 647 Tello, D., Gil, J., Loaiza, C. D., Riascos, J. J., Cardozo, N., & Duitama, J. (2019).
648 NGSEP3: accurate variant calling across species and sequencing protocols.
649 *Bioinformatics*, 35(22), 4716–4723. <https://doi.org/10.1093/bioinformatics/btz275>
- 650 Thompson, J. R., Attia, J., & Minelli, C. (2011). The meta-analysis of genome-wide
651 association studies. *Briefings in Bioinformatics*, 12(3), 259–269. <https://doi.org/10.1093/bib/bbr020>

- 658 Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford,
659 T. R., McMullen, M. D., Holland, J. B., & Buckler, E. S. (2011). Genome-
660 wide association study of leaf architecture in the maize nested association
661 mapping population. *Nature genetics*, 43(2), 159–162.
- 662 Yan, Y. Y., Burbridge, C., Shi, J., Liu, J., & Kusalik, A. (2019). Effects of input data
663 quantity on genome-wide association studies (GWAS). *International Journal*
664 *of Data Mining and Bioinformatics*, 22(1), 19–43. <https://doi.org/10.1504/IJDMB.2019.099286>
- 666 Yu, J., Pressoir, G., Briggs, W. H., Vroh, I. B., Yamasaki, M., Doebley, J. F., McMullen,
667 M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Et al. (2006). A unified
668 mixed-model method for association mapping that accounts for multiple
669 levels of relatedness. *Nature genetics*, 38(2), 203–208.
- 670 Yuan, J., Bizimungu, B., De Koeyer, D., Rosyara, U., Wen, Z., & Lagüe, M. (2019).
671 Genome-Wide Association Study of Resistance to Potato Common Scab.
672 *Potato Research*. <https://doi.org/10.1007/s11540-019-09437-w>
- 673 Zhang, S., Chen, X., Lu, C., Ye, J., Zou, M., Lu, K., Feng, S., Pei, J., Liu, C., Zhou, X.,
674 Ma, P., Li, Z., Liu, C., Liao, Q., Xia, Z., & Wang, W. (2018). Genome-wide
675 association studies of 11 agronomic traits in cassava (*Manihot esculenta*
676 *crantz*). *Frontiers in Plant Science*, 9(April), 1–15. <https://doi.org/10.3389/fpls.2018.00503>