

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 14, 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: GWAS on polyploids, GWASPoly, PLINK, SNP, SHEsis, software,
38 TASSEL

39

1 Introduction

40 The Genome-Wide Association Studies (GWAS) comprise statistical tests that iden-
41 tify which variants through the whole genome of a large number of individuals are
42 associated with a specific trait ([cantor2010prioritizing](#); [begum2012comprehensive](#)).
43 This methodology started with humans and several model plants, such as rice,
44 maize, and *Arabidopsis* ([lauc2010genomics](#); [tian2011genome](#); [cao2011whole](#);
45 [korte2013advantages](#); [han2013sequencing](#)). Because of the advances in the
46 next-gen sequencing technology and the decline of the sequencing cost in recent
47 years, there is an increase in the availability of genome sequences of different or-
48 ganisms at a faster rate ([ekblom2011applications](#); [ellegren2014genome](#)). Thus,
49 the GWAS is becoming the standard tool to understand the genetic bases of either
50 ecologically or economically relevant phenotypic variation for both model and non-
51 model organisms. This increment includes complex species such as polyploids (Fig.
52 1) ([ekblom2011applications](#); [santure2018wild](#)).

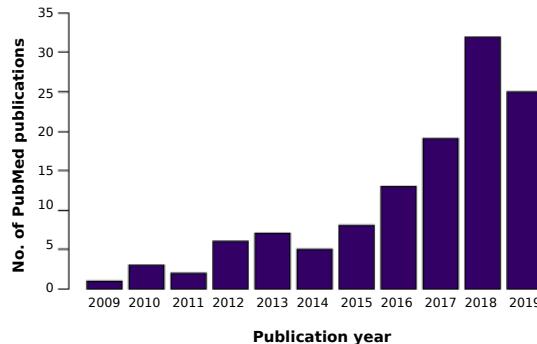


Figure 1: The number of peer-reviewed papers that contains the keywords "GWAS" and "polyploid" in the PubMed database between 2009 and 2019.

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

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

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

66 The iPAT allows running in a graphic interface three well-known command-line
67 GWAS software such as GAPIT, PLINK, and FarmCPU (**Zhang2018**). However, the
68 output from each package is separated. On the other hand, the easyGWAS allows
69 running a GWAS analysis on the web using different algorithms and combining sev-
70 eral GWAS results. This analysis runs independently of both the computer capacity
71 and the operating system. Nevertheless, it needs either several datasets to obtain
72 the different GWAS results to make replicates or GWAS results already computed. In
73 either case, the results from different algorithms are also separated (**Grimm2017**).
74 Thus, although both software iPAT and easyGWAS integrate with different programs
75 or algorithms, an output that allows them to compare similitudes and differences
76 in the association is missing.

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

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

87 To overcome these challenges, we developed the MultiGWAS tool that performs
88 GWAS analyses for both diploid and tetraploid species using four software in par-
89 allel. Our tool includes GWASPoly (**Rosyara2016**) and the SHEsis tool (**Shen2016**)
90 that accept polyploid genomic data, and PLINK (**Purcell2007**) and TASSEL (**Bradbury2007**),
91 designed exclusively for diploids, but that in the case of tetraploid data, their use re-
92 quire "diploidizing" genomic matrix. This wrapping tool deals with different input
93 file formats, including VCF. Besides, manage data preprocessing, search for asso-
94 ciations by running four GWAS software in parallel, and create a score to choose
95 between gene action models in GWASPoly and TASSEL. Moreover, create compar-
96 ative reports from the output of each software to help the user distinguish genuine
97 associations from 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 pro-
101 cesses the configuration file. Then it cleans and filters the genotype and phenotype
102 datasets, and in case of tetraploids, MultiGWAS "diploidize" the genomic data. Next,
103 during the multi analysis, each GWAS tool runs in parallel. Subsequently, in the in-
104 tegration step, the MultiGWAS tool scans the output files from the four packages
105 (i.e., GWASPoly, SHEsis, PLINK, and TASSEL). Finally, it generates a summary of
106 all results that contains score tables, Venn diagrams, SNP profiles, and Manhattan

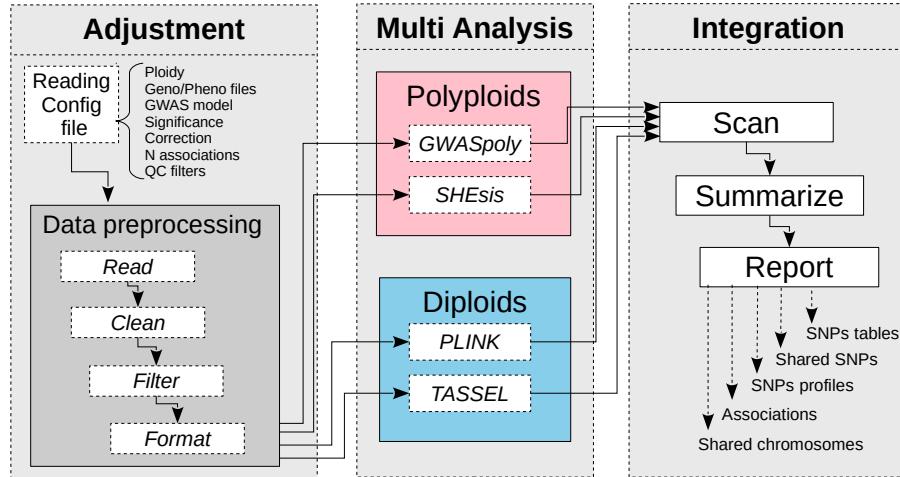


Figure 2: MultiGWAS flowchart has three steps: adjustment, multi analysis, and integration. In the first step, after the input data management upload, MultiGWAS read the configuration file, and preprocess the input data (genotype and phenotype dataset). The second step is the GWAS analysis, where MultiGWAS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

107 plots.

108 2.1 Adjustment stage

109 MultiGWAS takes as input a configuration file where the user specifies the genomics
110 data and the parameters used by the four tools. Once the configuration file is read
111 and processed, the genomic data files (genotype and phenotype) are then cleaned,
112 filtered, and checked for data quality. The output of this stage corresponds to the
113 inputs for the four programs at the Multi Analysis stage.

114 2.1.1 Reading configuration file

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

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

118 **Genotype and phenotype input files:** MultiGWAS uses two input files, one for
119 the genotype and one for the phenotype. Genotype data can be input in three differ-
120 ent formats, including a matrix format (Fig. 3.a), a GWASpoly format (Rosyara2016)
121 (Fig. 3.b), and Variant Call Format (VCF) (Fig.3.c) which is transformed into GWAS-
122 poly format using NGSEP 4.0.2 (Duitama2019). The phenotype file contains only

123 one trait with the first column for the sample names and the second column for the
 124 trait values (Fig. 3.c).

a.	Marker, sample01, sample02, sample03, ... c2_41437, AAAG, AAGG, AAGG, ... c2_24258, AAGG, AGGG, GGGG, ... c2_21332, TTCC, TTCC, TTCC, ...	b.	Marker, Chrom, Pos, sample01, sample02, sample03, ... c2_41437, 0, 805179, AAAG, AAGG, AAGG, ... c2_24258, 0, 1252430, AAGG, AGGG, GGGG, ... c2_21332, 0, 3499519, TTCC, TTCC, TTCC, ...
c.	#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	d.	Individual,Trait sample01, 3.59 sample02, 4.07 sample03, 1.05

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

125 **GWAS model:** MultiGWAS is designed to work with quantitative phenotypes and
 126 can run GWAS analysis using two types of statistical models that we have called
 127 *full* and *naive* models. The *full model* is known in the literature as the Q+K model
 128 (Yu2006) and includes a control for structure (Q) and relatedness between samples
 129 (K). In contrast, the *naive model* does not include any type of correction. Both
 130 models are linear regression approaches, and each one of the four GWAS packages
 131 used in MultiGWAS has some variations of those models. The *naive* is modeled with
 132 Generalized Linear Models (GLMs, Phenotype + Genotype), and the *full* is modeled
 133 with Mixed Linear Models (MLMs, Phenotype + Genotype + Structure + Kinship).
 134 The default model used by MultiGWAS is the *full model* (Q+K) (Yu2006), following
 135 the equation:
 136

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

136 In this equation, the y is the vector of observed phenotypes. Moreover, the β is a
 137 vector of fixed effects other than SNP or population group effects, the α is a vector of
 138 SNP effects (Quantitative Trait Nucleotides), the ν is a vector of population effects,
 139 the μ is a vector of polygene background effects, and the e is a vector of residual
 140 effects. Besides, Q , modeled as a fixed effect, refers to the incidence matrix for
 141 subpopulation covariates relating y to ν , and X , S and Z are incidence matrices of
 142 1s and 0s relating y to β , α and μ , respectively.

143 **Genome-wide significance:** GWAS searches SNPs associated with the phenotype
 144 in a statistically significant manner. A threshold or significance level α is specified

145 and compared with the *p-value* derived for each association score. Standard significance
146 levels are 0.01 or 0.05 (**Gumpinger2018; Rosyara2016**), and MultiGWAS
147 uses an α of 0.05 for the four GWAS packages. However, in GWASpoly and TASSEL,
148 which calculates the SNP effect for each genotypic class using different gene action
149 models (see “Multi analysis stage”), the threshold is adjusted according to each of
150 those two packages. Therefore, the number of tested markers may be different in
151 each model (see below), impacting the *p-value* thresholds.

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

161 **Number of reported associations:** Criticism has arisen, considering only sta-
162 tistically significant associations as possible correct associations (**Thomson2011;**
163 **Kaler2019**). Many low *p-value* associations, closer to being significant, are dis-
164 carded due to the stringent significance levels, which consequently increases the
165 number of false negatives. To avoid this problem, MultiGWAS provides the option
166 to specify the number of best-ranked associations (lower *p-values*), adding the cor-
167 responding *p-value* to each association found. In this way, it is possible to enlarge
168 the number of results, and their replicability across the different programs. Never-
169 theless, the report displays each association with its corresponding *p-value*.

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

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

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

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

188 **2.1.2 Data preprocessing**

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

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

199 After this process, MultiGWAS converts the genomic data, genotype, and pheno-
200 type datasets to the specific formats required for each of the four GWAS packages.

201 **2.2 Multi analysis stage**

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

211 **2.2.1 GWASpoly**

212 GWASpoly (**Rosyara2016**) is an R package designed for GWAS in polyploid species
213 used in several studies in plants (**Berdugo2017; Ferrao2018; Sharma2018; Yuan2019**).
214 GWASpoly uses a Q+K linear mixed model with biallelic SNPs that account for pop-
215 ulation structure and relatedness. Also, to calculate the SNP effect for each geno-
216 typic class, GWASpoly provides eight gene action models: general, additive, simplex
217 dominant alternative, simplex dominant reference, duplex dominant alternative,
218 duplex dominant, diplo-general, and diplo-additive. Consequently, the number of
219 statistical tests performed can be different in each action model and so thresholds
220 below which the *p-values* are considered significant.

221 MultiGWAS is using GWASpoly version 1.3 with all gene action models available
222 to find associations. The MultiGWAS reports the top *N* best-ranked (the SNPs with
223 lowest *p-values*) that the user specified in the *N* input configuration file. The *full*
224 model used by GWASpoly includes the population structure and relatedness, which

225 are estimated using the first five principal components and the kinship matrix, re-
226 spectively, both calculated with the GWASpoly built-in algorithms.

227 2.2.2 SHEsis

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

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

235 2.2.3 PLINK

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

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

245 2.2.4 TASSEL

246 TASSEL is another standard GWAS program based on the Java software developed
247 initially for maize but currently used in several species (**Alvarez2017; Zhang2018**).
248 For the association analysis, TASSEL includes the general linear model (GLM) and
249 mixed linear model (MLM) that accounts for population structure and relatedness.
250 Moreover, as GWASPoly, TASSEL provides three-gene action models to calculate the
251 SNP effect of each genotypic class: general, additive, and dominant. Hence, the
252 significance threshold depends on each action model.

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

259 2.3 Integration stage.

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

263 **2.3.1 Calculation of *p*-values and significance thresholds**

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

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

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

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

289 **2.3.2 Selection of significant and best-ranked associations**

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

295 PLINK and SHEsis have a unique gene action model (see section 2.2.2 and
296 2.2.3). However, in the case of GWASpoly and TASSEL, which have eight and three
297 models respectively, MultiGWAS automatically selects the "best gene action model"
298 from each package and takes the associations from it. This selection within GWAS-
299 Poly and TASSEL has three criteria: the inflation factor (I), the shared SNPs (R),
300 and the significant SNPs (S).

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

302
$$score(M_i) = I_i + R_i + S_i$$

303 where $score(M_i)$ is the score for the gene action model M_i , with i from $1..k$,
304 for a GWAS package with k gene action models. I_i is the score for the inflation

Figure 4: Reports presented by MultiGWAS. For each tool, first, a QQ plot that assesses the resultant *p*-values. Second, a Manhattan plot for each tool with two lines, blue and red, represents the lower limit for the best ranked and significative SNPs, respectively. We present two Venn diagrams, one for the significative SNPs and one for N best-ranked SNPs of each tool. We show the results for GWAspoly, PLINK, TASSEL, and SHEsis in red, green, yellow, and blue. For each SNP that is in the intersection, thus, that is predicted by more than one tool, we provide an SNP profile. SNPs by chromosome chord diagrams show that the strongest associations are limited to few chromosomes. Furthermore, we present tabular summaries with details of significant and best-ranked associations.

305 factor defined as $I_i = 1 - |1 - \lambda(M_i)|$, where $\lambda(M_i)$ is the inflation factor for the
 306 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
 307 $|M_i \sim M_j|$ is the number of SNPs shared between M_i and M_j models, normalized by
 308 the maximum number of SNPs shared between all models. And, S_i is the number of
 309 significant SNPs of model M_i normalized by the total number SNPs shared among
 310 all models.
 311 The score is high when an M_i model has an inflation factor λ close to 1, identifies
 312 a high number of shared SNPs, and contains one or more significant SNPs.
 313 Conversely, the score is low when the M_i model has an inflation factor λ either low
 314 (close to 0) or high ($\lambda > 2$), which identifies a small number of shared SNPs, and
 315 contains 0 or few significant SNPs. In any other case, the score results from the
 316 balance among the inflation factor, the number of shared SNPs, and the number of
 317 significant SNPs.

318 **2.3.3 Integration of results**

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

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

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

332 For each SNP identified more than once, MultiGWAS provides its SNP profile.
 333 That is a heat diagram for a specific SNP where each column is a genotype state

334 AAAA, AAAB, AABB, ABBB, and BBBB. Moreover, each row corresponds to a sample.
335 Samples with close genotypes form together clusters. Thus to generate the clusters, we do not use the phenotype information. However, we present the phenotypic information in the figure as the color. This figure visually provides information regarding genotype and phenotype information simultaneously for the whole population. We present colors as tones between white and black for color blind people.

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

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

351 3 Results

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

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

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

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

372 3.1 Manhattan and QQ plots for GWAS associations

373 MultiGWAS uses classical Manhattan and Quantile–Quantile plots (QQ plots) to
374 visualize each package’s results. In both plots, the points are the SNPs and their
375 p-values are transformed into scores like $-\log_{10}(p\text{-values})$ (see Fig. 5). The Man-
376 hattan plot shows the strength of association of the SNPs (y-axis) distributed at their
377 genomic location (x-axis), so the higher the score, the stronger the association. At
378 the same time, the QQ plot compares the expected distribution of *p*-values (y-axis)
379 with the observed distribution (x-axis).

380 MultiGWAS adds distinctive marks to both plots to identify different types of
381 SNPs: (a) In the Manhattan plots, the significant SNPs are above a red line, and
382 the best-ranked SNPs are above a blue line. Also, SNPs shared between packages
383 are colored green (See Fig. 6.b). (b) In the QQ plots, a red diagonal line indicates
384 the expected distribution under the null hypothesis of no association of SNPs with
385 the phenotype. Both distributions should coincide, and most SNPs should lie on the
386 diagonal line. Deviations for a large number of SNPs may reflect inflated *p*-values
387 due to population structure or cryptic relatedness. Nevertheless, few SNPs deviate
388 from the diagonal for a truly polygenic trait (**Power2016**).

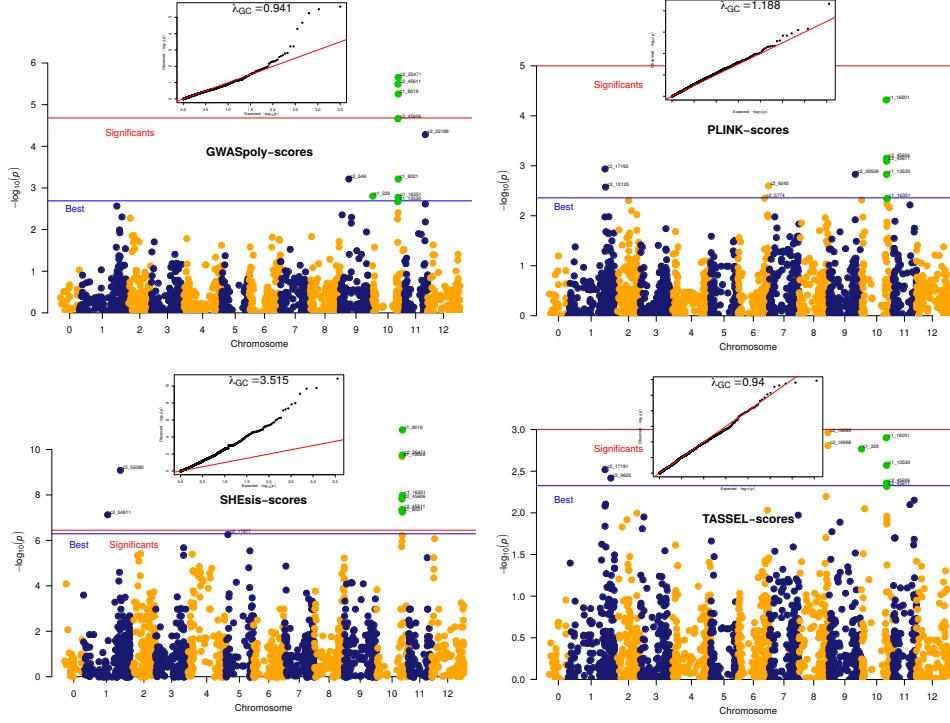


Figure 5: Associations in the tetraploid potato dataset. MultiGWAS shows the associations identified by the four GWAS packages using Manhattan and QQ plots. The tetraploid potato data showed several SNPs shared between the four software (green dots). The best-ranked SNPs are above the blue line, but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

389 3.2 Tables and Venn diagrams for single and shared SNPs

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

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

400 Second, the significant SNPs correspond to the ones valued statistically signifi-
 401 cant by each package. They appear in a Venn diagram (Figure 6.c), and in the SNPs
 402 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.

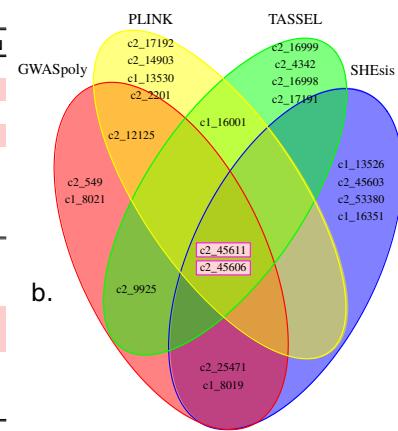
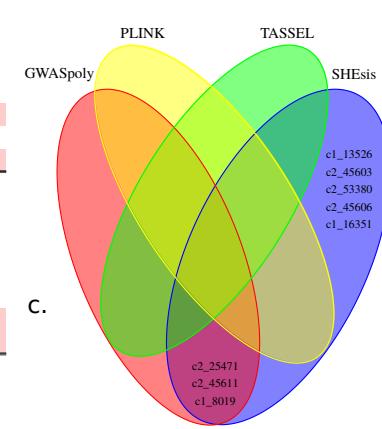
**b.****c.**

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).

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

410 3.3 Heat diagrams for the structure of shared SNPs

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

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

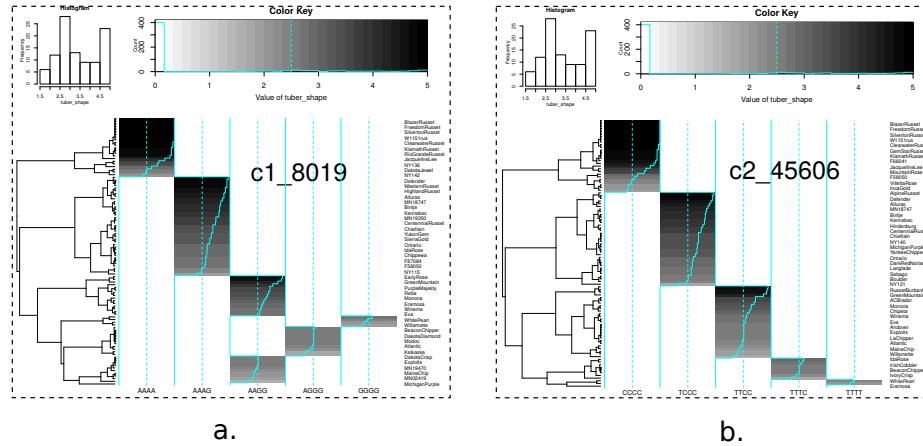


Figure 7: SNP profiles. SNP profiles for two of the best-ranked significant SNPs shown in the figure 6.b. a. SNP c2_45606 best-ranked by the four packages (central intersection of the Venn diagram Figure 6.b) b. SNP c1_8019 best-ranked by the two tetraploid packages (Figure 6.b), and also identified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

424 3.4 Chord diagrams for SNPs by chromosome

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

430

Luis: Xq no se ven los del
cromosoma 1, en el material
suplementario si estan

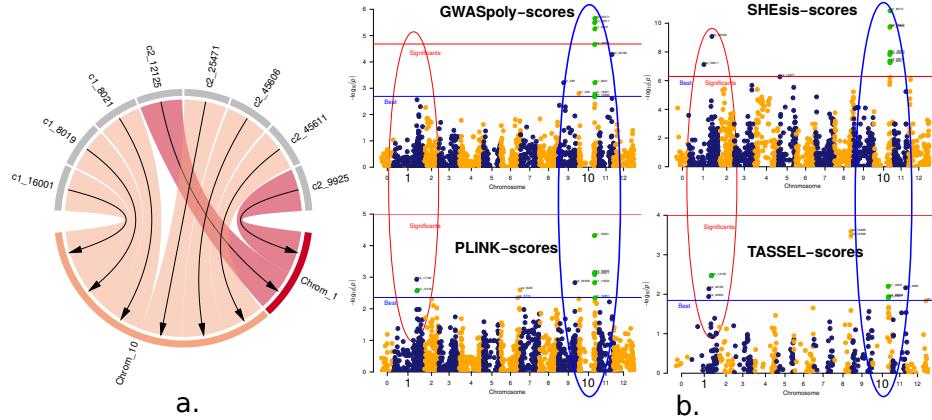


Figure 8: SNPs by chromosome. The position of best-ranked SNPs across chromosomes using two different visualizations. **a.** Chord diagram showing that best-ranked SNPs located in chromosome 10. The SNPs are at the top and the chromosomes at the bottom. The arrows connect the best-ranked SNPs with their 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 showed in 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 organism's ploidy level (2: for diploids, 4: for tetraploids). The input genotype/phenotype filenames. The genome-wide significance threshold. The method for multiple testing correction. The GWAS model. The number of associations to report. The quality control filters choosing TRUE or FALSE. The filters are minor allele frequency, individual missing rate, SNP missing rate, and Hardy-Weinberg threshold. Finally, the GWAS packages selected for the analysis.

446 4.2 Using the command line interface

447 The execution of the CLI tool is easy. It only needs to open a Linux console, change
 448 to the folder where is the configuration file, and type the executable tool's name,
 449 followed by the filename of the configuration file, like this:

450 multiGWAS Test01.config

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

457 4.3 Using the graphical user interface

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

462 jmultiGWAS

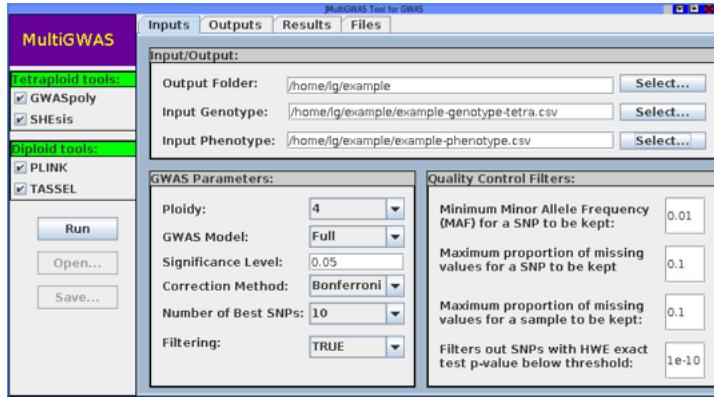


Figure 10: Main view of the MultiGWAS graphical user interface. The interface has a toolbar at the left side and four tabs at the top. In the toolbar, users can select the GWAS packages (Two for tetraploids and two for diploids). The analysis starts with the current parameters or loading a previously saved configuration. In the Input tab, users can set the parameters and quality control filters. The Output tab shows the execution of each process. In the Results tab, users can browse the HTML report of the current analysis generated by the tool. Finally, in the Files tab, users can browse the source files of each software and access the produced data across the analysis.

463 5 Discussion

464 The reanalysis of potato data with MultiGWAS showed that this wrapping tool is
 465 handy to improve the GWAS in tetraploid species. Through MultiGWAS perfor-
 466 mance, we could test its effectiveness to answer some of the challenges analyzing
 467 polyploid organisms. They include the integration and replication among param-
 468 eters and software, the diploidization of polyploid data, and the incorporation of
 469 different inheritance mechanisms (**dufresne2014**).

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

475 Another difficulty for replication among software is the variability of structures
 476 for the genomic input data. Currently, the most common format for next-generation
 477 sequencing variant data is the VCF (Variant Call Format) (**Danecek2011; Ebbert2014**).
 478 One of the advantages of VCF is its versatility in summarizing important genome
 479 information for hundreds or thousands of individuals and SNP, including informa-
 480 tion about levels of ploidy. MultiGWAS simplifies the use of the GWAS software
 481 available because it allows the VCF files as an input (but see VarStats tool in VTC).

482 Moreover, the MultiGWAS is the unique wrapping tool we are aware of that
 483 facilitates to understand the effect of diploidizing the tetraploid data in the per-
 484 formance of the analysis directly. The graphical outputs are a handy approach to find

485 similar results. The SNP profile allows identifying what the significant associations
486 detected by more than one software are. Furthermore, although MultiGWAS checks
487 for significative SNPs based on the *p-value*, it is essential to go back to the data and
488 check if the SNP is a real association between the genotype and phenotype. For this
489 purpose, the SNP profile gives visual feedback for the accuracy of the association.

490 Furthermore, the MultiGWAS allows comparing among the gene action mod-
491 els that offer GWASPoly and TASSEL. GWASpoly (**Rosyara2016**) provides models
492 of different types of polyploid gene action, including additive, diploidized additive,
493 duplex dominant, simplex dominant, and general. On the other hand, TASSEL
494 (**Bradbury2007**) also models different types of gene action for general, additive,
495 and dominant diploids. To choose among models, we propose an automatic se-
496 lection of the gene action model for both tools based on a balance between three
497 criteria: the inflation factor, the replicability of identified SNPs, and the significance
498 of identified SNPs. This inflation index is a new tool for comparison that does not
499 offer either GWASPoly or TASSEL. This automatic strategy will help to understand
500 the gene action model for the trait of interest. Although the main focus is on the re-
501 sultant SNPs, the model has assumptions that reflect the gene actions for a specific
502 phenotype.

503 Finally, MultiGWAS through the active comparison among models addresses
504 the search of the inheritance mechanisms by comparing among two designed for
505 polysomic inheritance software (**Rosyara2016; Shen2016**) with two software for
506 disomic inheritance (**Purcell2007; Bradbury2007**). Understanding the inheritance
507 mechanisms for polyploids organism is an open question. For autopolyploids, most
508 loci have a polysomic heritage. However, sections of the genome that did not dupli-
509 cate lead to disomic inheritance for some loci (**ohno1970; lynch2000; dufresne2014**).
510 Thus it is a useful tool for researchers because it looks for significative associations
511 that involve both types of inheritance.

512 5.1 Future remarks

513 The evolution and population genomics of polyploids is an exciting novel area of
514 research. Because of the advance of next-gen sequencing techniques, it is more
515 common the generation of more empirical polyploid data in several model and non-
516 model organisms (REF).

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

523 For MutiGWAS, we started with the most simple ploidy, such as tetraploids.
524 Moreover, we did not filter data yet for HWE in tetraploids before GWAS as MultiG-
525 WAs do it for diploid data. Nevertheless, future MultiGWAS versions should include
526 more complex ploidies further than tetraploids, as well as the explicit calculation
527 of parameters either for filtering polyploid data before GWAS analysis or comple-
528 menting other population genomics' parameters of the data analyzed in the output

529 reports.

530 POR HACER.. On the other hand, although MultyGWAS does not solve the
531 uncertainty in the allele dosage and null alleles.

532 FUTUROS desarrollos o extensiones y desafios en poliploides. (heterocigosidad,
533 allele dosage, poder de software en poliploides, hitchiking, HW). Aunque shesis y
534 gwaspoly manejan plodias mayores, hemos visto que hacer gwas para un organismo
535 poliploide requiere tener en cuenta otros factores y por eso aun queremos estudiar
536 esto (HW, Inheritance) antes de extender la herramienta a otras ploidias.

537

538

539

540

estaba encadenado antes con un pedacito que mencionaba que estos eran algunos de los problemas de poliploides y se sugería revisar el efecto de estos en el GWAS pero ese pedacito se perdió y creo que queda fuera de contexto a menos que digamos lo anterior

541 6 Acknowledgements

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

544 We thanks to the Minister of Science, Technology and Innovation of the
545 republic of Colombia (previously COLCIENCIAS), for supporting the postdoctoral re-
546 searcher L. Garreta at AGROSAVIA during 2019-2020 under the supervision of ICS
547 and PHRH, (Grant number 811-2019). The editorial of AGROSAVIA gave for
548 financial supporting for this publication. Finally to Andres J. Cortes for his valuable
549 comments to improve this manuscript.

550 7 Author Contributions

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

Andrés sugiere un plot sencillo de heterocigosidad alineado con el de Manhattan para inspeccionar estos casos

Comentario de Andrés: Podría haber un caso valioso que solo se captura una vez? (teniendo en cuenta que no todos los métodos tienen el mismo poder, y por ello no necesariamente tendrían el mismo peso)

si pero se muestran todos los que se capturan una vez y los que se capturan varias veces Discusión

Esta es una sugerencia de Andrés: Tener en cuenta a la hora del calcular el score que los valores de SNPs compartidos y su significancia si bien pueden depender del poder de la asociación, también pueden relacionarse con la arquitectura genética por se, y el efecto de hitchiking. Este efecto es mayor en zonas del genoma de baja recombinación (e.g. cerca a los centrómeros, que es donde menos genes suelen encontrarse). por ello, quizás valdría la pena comentar sobre pesos distintos para la significacia y para los SNPs compartidos? (mayor para el primero). en otras palabras, hay señales "anchas" o "delgadas" pero esto más que darle validez a la asociación, depende de procesos intrínsecos