1 Abstract-Intro

2 Motivation

3 Background

One of the major steps regarding GWAS is the replication of the results, generally achieved by finding the identified associations in an independent population sample or using an independent technology [6, 12]. One way of replicating a scientific analysis achieved by computational tools, is running a new analysis using different tools with different parameters, models, or conditions, and see to what extent the scientific conclusions are affected. Although, there are many software available that perform GWAS [9], most of them for diploids than polyploid species [3], few had worked in the integration of the tools and their results.

Some works have been oriented to the evaluation or the usage of GWAS packages. For example, Y. Yan et al. (2019) evaluated four GWAS packages for diploid species, PLINK, TASSEL, GAPIT, and FaST-LMM, to examine the effects of amount of input data (sample size) on GWAS results in the context of plants. They conclude that results depend of both the package and measurement, and the threshold $pvalue$ for SNP significance is different for each package. It means that well-ranked SNPs from one package can be ranked differently in another, causing difficulty to select the most plausible associations when results from each package are analyzed separately. Chen and Zhang (2018) developed a graphical user oriented sofware called iPAT to facilitate the use of three popular command-line GWAS packages: GAPIT, PLINK, and FarmCPU. However, results from the execution of each package are shown separately and the problem of interpreting and selecting the best associations persists. A more elaborated GWAS software is the easyGWAS cloud platform to perform, share and compare the results of GWAS [8]. For comparing results, easeGWAS offers two types of analysis: the first is an intersection analysis that searches associations that were found significant in more than one dataset, and the second is a meta-analysis that searches associations mutually soported by several datasets. Both types are based on different datasets with same GWAS parameters to confirm or search for new associations, which is desiderable but sometimes it is difficult to achieve when there is only a unique or small sample of individuals.

4 Methods

4.1 Tools

We have selected four GWAS software tools to be integrated in our multiGWAS tool, two designed specifically for polyploid species as many important crops are polyploids: GWASpoly [16] and SHEsis [18], and another two designed for diploids species and extensively used in humans and plants: PLINK [14, 5] and TASSEL [4], respectively.

As MultiGWAS implements two types of GWAS analysis, naive and full, each tool is called in two different ways. The naive without any additional parameter, but the full with two parameters that take into account for population structure (Q) and relatedness (K) to prevent false associations.

4.1.1 GWASpoly

GWASpoly is a recent R package designed for GWAS in polyploid species that has been used in several studies in plants [2, 7, 17, 19]. It is based on the Q+K linear mixed model with biallelic SNPs that accounts for population structure and relatedness. In addition, to calculate the SNP effect for each genotypic class, GWASpoly provides a general gene action model along with four additional models: additive, simplex dominant, and duplex dominant.

MultiGWAS is using GWASpoly version 1.3. The population structure and relatedness, used in the full model, are estimated using the first five principal components and the kinship matrix, respectively, both calculated with the algorithms built in GWASpoly. For both, naive and full models, all gene action models are tested for detecting associations.

4.1.2 SHEsis

SHEsis is another program designed for polyploid species that includes single locus association analysis, among others. It is based on a linear regresion model, and it has been used in some studies of animals and humans [15, 11].

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

4.1.3 PLINK

PLINK is one of the most extensively used programs for GWAS in diploids species. It was developed for humans but it is applicable to any species [13]. PLINK includes a range of analysis, 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 is used to achieve both types of analysis, naive and full. For the full analysis, population structure is estimated using the first five principal components calculated with the PLINK 1.9 built in algorithm. But relatedness is estimated from the kinship coefficients calculated with the PLINK 2.0 built in algorithm, removing the close relatives or individuals with first-degree relatedness.

4.1.4 TASSEL

TASSEL is another common GWAS program based on the Java software. It was developed for maize and it has been used in several studies in plants [1, 20], but like PLINK, it is applicable to any species. For association analysis, TASSEL includes the general lineal model (GLM) and mixed linear model (MLM) that accounts for population structure and relatedness.

MultiGWAS is using TASSEL 5.0, with naive GWAS achieved by the GLM, and full GWAS achieved by the MLM with two parameters: one for population structure, using the first five principal components, and another for relatedness, using the kinship matrix with centered IBS method, both calculated with TASSEL built in algorithms.

5 Results

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

Here, we show the reports resulting from running MultiGWAS tool in the genomic data from a tetraploid potato diversity panel, genotyped and phenotyped as part of the USDA-NIFA Solanaceae Coordinated Agricultural Project (SolCAP) [10]. The reports include: significant SNPs, best-ranked SNPs, profile SNPs, and visualization of associations.

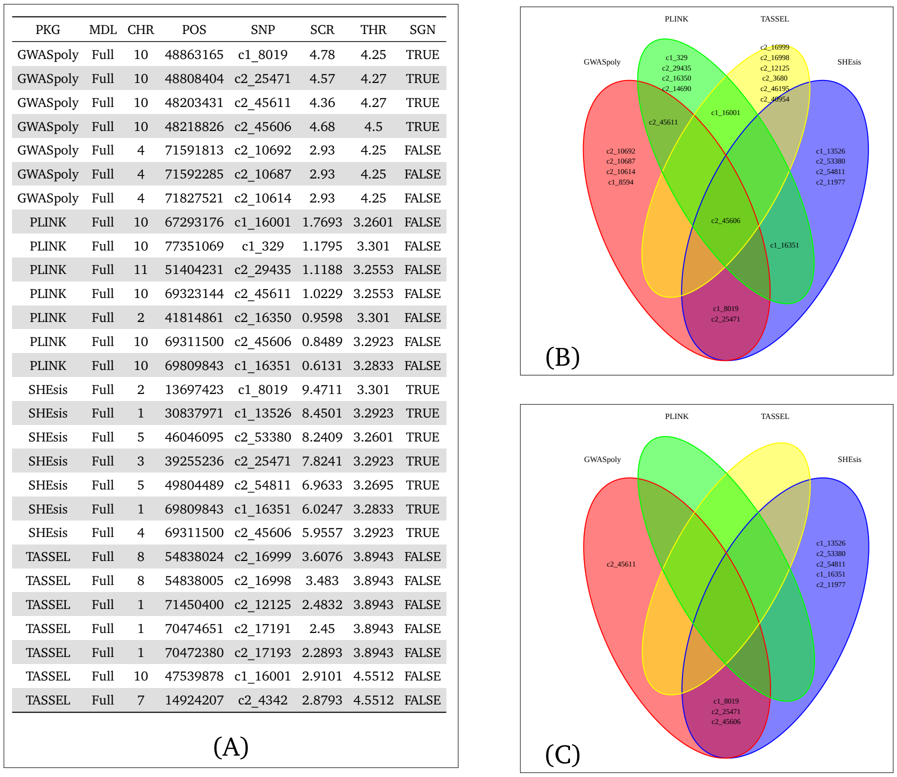
First, the significant SNPs (Figure **??**), where the two polyploid software, GWASpoly and SHEsis, found as significant three SNPs, c1_8019, c2_25471, and c2_45606, of which the c1_8019 was also the most significant association found in the same potato dataset analyzed by Rosyara et al. (2016). Second, the best-ranked SNPs (Figure **??**), where the SNP c2_45606 was evaluated with a high score by the four packages, but other SNPs were also ranked with high scores by almost two packages. Third, the SNP profiles(Figure XX), where for each significant association, a heat map figure is generated to summarize the genotype associated with a trait for each individual. And fourth, the visualization of associations (Figure YY), where for each package, a Manhattan and QQ plots are generated using special marks to help to identify significative, best-ranked, and shared SNPs (found by more than one tool).

The complete report from MultiGWAS for the naive and full model is in the Supplementary information (`https://www.overleaf.com/project/5e8b8de6ae23ed0001a9a14f`)

5.1 Visualization of shared SNPs

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

MultiGWAS provides tabular and graphic views to report in an integrated way both the best-ranked and significant SNPs identified by the four different packages (see Figure 1). Both $pvalues$ and significance levels have been scaled as $-log_{10}(pvalue)$ to give high scores to the best statistically evaluated SNPs. First, the best-ranked SNPs correspond to the top-scored *N* SNPs, wheter these SNPs were assesed significant or not by the package reporting them, and with *N* defined by the user in the configuration file. These SNPs are shown both in a table (Figure 1.A) and in a Venn diagram (Figure 1B). The table lists them by package and sorted by decreasing score, whereas the Venn diagram shows them emphasizing if these were best-ranked either in a single package or in several at once (shared). And second, the significant SNPs correspond to the ones assesed statistically significant by each package (score (SCR) > threshold (THR), table in Figure1.A).



PKG	MDL	CHR	POS	SNP	SCR	THR	SGN
GWASpoly	Full	10	48863165	c1_8019	4.78	4.25	TRUE
GWASpoly	Full	10	48808404	c2_25471	4.57	4.27	TRUE
GWASpoly	Full	10	48203431	c2_45611	4.36	4.27	TRUE
GWASpoly	Full	10	48218826	c2_45606	4.68	4.5	TRUE
GWASpoly	Full	4	71591813	c2_10692	2.93	4.25	FALSE
GWASpoly	Full	4	71592285	c2_10687	2.93	4.25	FALSE
GWASpoly	Full	4	71827521	c2_10614	2.93	4.25	FALSE
PLINK	Full	10	67293176	c1_16001	1.7693	3.2601	FALSE
PLINK	Full	10	77351069	c1_329	1.1795	3.301	FALSE
PLINK	Full	11	51404231	c2_29435	1.1188	3.2553	FALSE
PLINK	Full	10	69323144	c2_45611	1.0229	3.2553	FALSE
PLINK	Full	2	41814861	c2_16350	0.9598	3.301	FALSE
PLINK	Full	10	69311500	c2_45606	0.8489	3.2923	FALSE
PLINK	Full	10	69809843	c1_16351	0.6131	3.2833	FALSE
SHEsis	Full	2	13697423	c1_8019	9.4711	3.301	TRUE
SHEsis	Full	1	30837971	c1_13526	8.4501	3.2923	TRUE
SHEsis	Full	5	46046095	c2_53380	8.2409	3.2601	TRUE
SHEsis	Full	3	39255236	c2_25471	7.8241	3.2923	TRUE
SHEsis	Full	5	49804489	c2_54811	6.9633	3.2695	TRUE
SHEsis	Full	1	69809843	c1_16351	6.0247	3.2833	TRUE
SHEsis	Full	4	69311500	c2_45606	5.9557	3.2923	TRUE
TASSEL	Full	8	54838024	c2_16999	3.6076	3.8943	FALSE
TASSEL	Full	8	54838005	c2_16998	3.483	3.8943	FALSE
TASSEL	Full	1	71450400	c2_12125	2.4832	3.8943	FALSE
TASSEL	Full	1	70474651	c2_17191	2.45	3.8943	FALSE
TASSEL	Full	1	70472380	c2_17193	2.2893	3.8943	FALSE
TASSEL	Full	10	47539878	c1_16001	2.9101	4.5512	FALSE
TASSEL	Full	7	14924207	c2_4342	2.8793	4.5512	FALSE

Figure 1: **Visualization of shared SNPs.** Tabular and graphical views of the best-ranked and significant SNPs identified by the four packages. **(A)** Tabular view with detailed information of each SNPs, including: package name (PKG), GWAS model used (MDL), chromosome (CHR), position in the genome (POS), ID (SNP), score (SCR), threshold (THR), and significance flag (SGN), wether the SNP was evaluated statistically significant or not (score > theshold). **(B)** Venn diagram with the best-ranked SNPs, showing that one SNP was shared by the four packages (c2_45606), other two only by the two polyploid packages GWASpoly and SHEsis (c1_8019 and c2_25471), and other one only by the two diploid packages PLINK and TASSEL (c1_16001). **(C)** Venn diagram with the significant SNPs, showing that only three SNPs (c1_8019, c2_25471, and c2_45606) were evaluated as significant by the two polyploid packages GWASpoly and SHEsis.

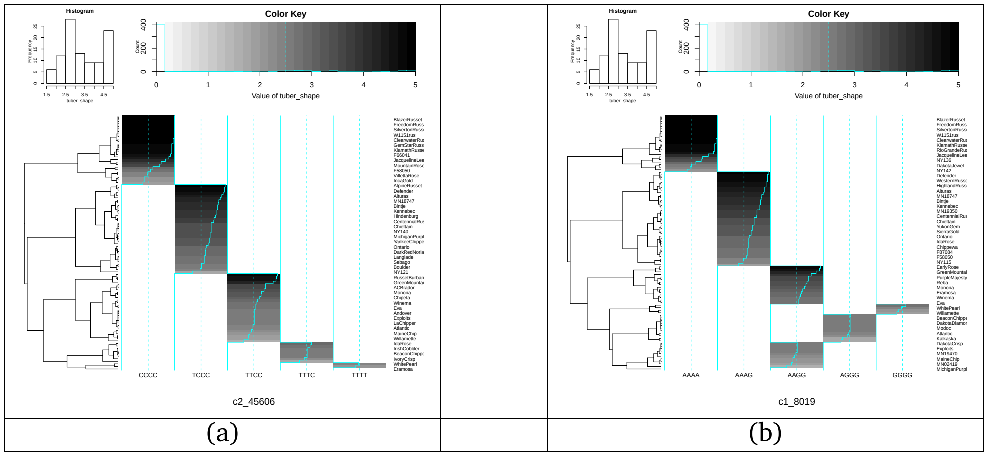
5.2 Visualization of Associations

MultiGWAS uses classical Manhattan and Quantile–Quantile plots (QQ plots) to visualize the results of GWAS analysis from each package. In both plots, SNPs are represented by dots and their $pvalues$ are transfomed to scores as $-log_{10}(pvalue)$ (see Figure 3). The Manhattan plot displays the SNP association strength (y-axis) distributed in their genomic location (x-axis), so the higher the score the stronger the association. Whereas the QQ plot is used to visually compare the expected distribution of $pvalues$ (y-axis) vs. the observed distribution (x-axis), so under the null hypothesis of no association of SNPs with the phenotype, both distributions should coincide, and most SNPs should lie on a diagonal line.

5.3 Visualización of common significative SNPs

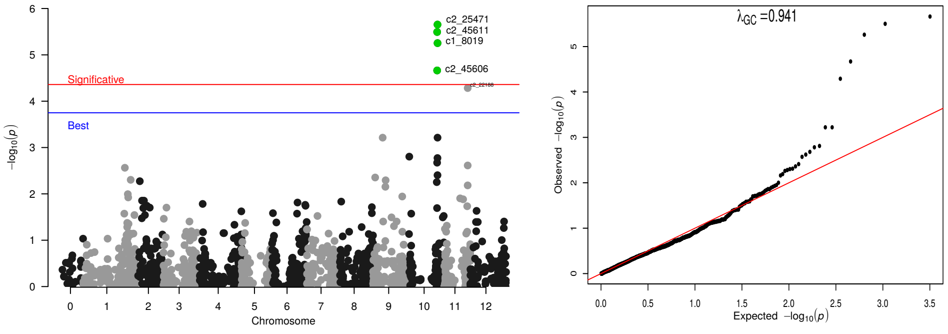
MultiGWAS creates a two-dimensional representation, called SNP profile, to visualize each trait by individuals and genotypes as rows and columns, respectively (Figure 2). At the left, the individuals are grouped in a dendrogram by their genotype. At the right, there is the name or ID of each individual. At the bottom, the genotypes are ordered from left to right, starting from the major to the minor allele (i.e., AAAA, AAAB, AABB, ABBB, BBBB). At the top, there is a description of the trait based on a histogram of frequency (top left) and by an assigned color for each numerical phenotype value using a grayscale (top right). Thus, each individual appears as a colored line by its phenotype value on its genotype column. For each column, there is a solid cyan line with the mean of each column and a broken cyan line that indicates how far the cell deviates from the mean.

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



(a) (b)

Figure 2: **SNP profiles.** SNP profiles for two of the best-ranked significant SNPs shown in the figure **??**.b. (a) SNP c2_45606 best-ranked by the four packages (central intersection of the Venn diagram Figure **??**.b) (b) SNP c1_8019 best-ranked by the two tetraploid packages (Figure **??**.b), and also identified as significant by the same packages (at the bottom of the Figure **??**.a).



(a) (b)

Figure 3: **MultiGWAS visualization of associations.** MultiGWAS adds special marks to the Manhattan and QQ plots to help identify different types of SNPs: (a) In Manhattan plots, significant SNPs are above a red line, best-ranked SNPs are above a blue line, and shared SNPs (See Figure **??**) are colored in green (b) In QQ plots, a red diagonal line indicates the expectation, so potential associations can be observed when the number of SNPs deviating from the diagonal is small, as in the case of monogenic traits, or when this number is somewhat higher, as in the case of truly polygenic traits. However, deviations for a high number of SNPs could reflect inflated $pvalues$ owing to population structure or cryptic relatedness.

6 Availability and Implementation

The core of the MultiGWAS tool was developed in R, it uses as the only input a simple configuration text file where users set the values for the main parameter that drives the GWAS process, 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 4).

6.1 Input parameters

The input parameters include: the output folder where results will be written, input genotype/phenotype filenames, genome-wide significance threshold, method for multiple testing correction, GWAS model, number of associations to be reported, and TRUE or FALSE whether to use quality control filters or not. The filters are: minor allele frequency, individual missing rate, SNP missing rate, and Hardy-Weinberg threshold.

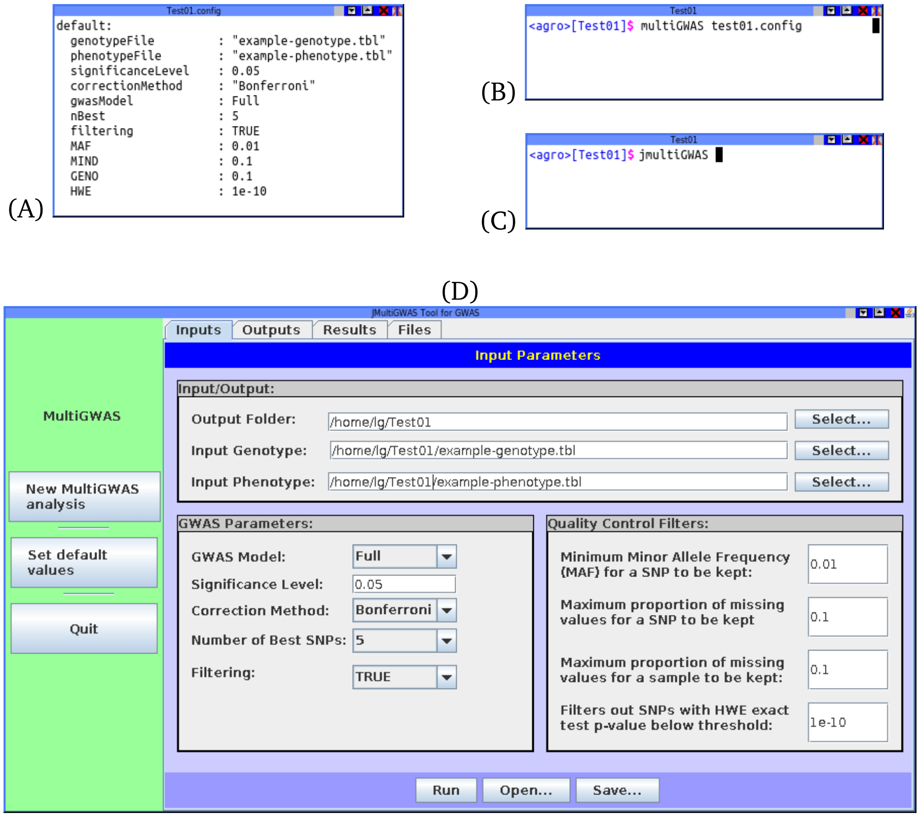
The configuration file can be created either using a general text editor or using the GUI application. In the first case, the file must have the structure shown in the Figure 4.A, where parameter names and values are separated by colon, filenames are enclosed in quotation marks, and TRUE or FALSE indicates wheter filters are applied or not. In the second case, the user creates the config file in a simple and straightforward way using the input parameter view from the GUI application (Figure 4.D) and clicking the "Save" button.

6.2 Using the command line interface

The execution of the CLI tool is very simple, it only needs to open a linux console, change to the folder where the configuration file was created, and type the name of the executable tool **"multiGWAS"** followed by the filename of the configuration file (See Figure 4B). Then, the tool starts the execution, showing information of the process in the console window, and after the execution finishes the results are saved to a new subfolder called *"outgwas/reports"*. Results include a full html report containing the different views described in the results section, the original graphics and summary tables used to create the html report in pdf/png and tab-delimited tsv/txt formats, respectively, and also includes the preprocessed result tables from the four GWAS packages used by MultiGWAS.

6.3 Using the graphical user interface

The MultiGWAS GUI application can be executed in two ways: running from a Linux console the script *"jmultiGWAS"* (see Figure 4.C) or by clicking from a file browser on the Java application file *"JMultiGWAS.jar"* located in the *"multiGWAS/sources"* subfolder. After it opens, it shows a main frame with four tabs at the top: "Inputs", "Outputs", "Results", and "Files". The "Inputs" tab shows the form to create the configuration file and run the application (see Figure 4D). The "Outputs" tab shows the messages from the running process after it starts the execution. The "Results" tab shows the full html report described above. And the "Files" tab show an embbeded file browser pointing to the subfolder that contains the files of the graphics and tables used to create the report and describe above.



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

```
<agro>[Test01]$ multiGWAS test01.config
```

```
<agro>[Test01]$ jmultiGWAS
```

Input Parameters

Input/Output:

Output Folder: /home/lg/Test01
Input Genotype: /home/lg/Test01/example-genotype.tbl
Input Phenotype: /home/lg/Test01/example-phenotype.tbl

GWAS Parameters:

GWAS Model: Full
Significance Level: 0.05
Correction Method: Bonferroni
Number of Best SNPs: 5
Filtering: TRUE

Quality Control Filters:

Minimum Minor Allele Frequency (MAF) for a SNP to be kept: 0.01
Maximum proportion of missing values for a SNP to be kept: 0.1
Maximum proportion of missing values for a sample to be kept: 0.1
Filters out SNPs with HWE exact test p-value below threshold: 1e-10

Figure 4: **MultiGWAS inputs and interaction**. MultiGWAS uses as input a simple configuration text file and can be executed using either a command line interface script in R (CLI) or a graphic user interface application in Java (GUI). **(A)** An example of a configuration text file named *"Test01.config"* including the parameters that drives the GWAS process. It can be created using a general text editor or using the GUI application (see below) **(B)** Call to the CLI program from a Linux console by typing the program name *"multiGWAS"* followed by a configuation file *"test01.config"*. **(C)** Call to the GUI application named *"jmultiGWAS"* from a Linux console **(D)** Input view of the MultiGWAS GUI application where setting input parameters can be done in a straightforward way.

References

[1] María F. Álvarez, Myrian Angarita, María C. Delgado, Celsa García, José Jiménez-Gomez, Christiane Gebhardt, and Teresa Mosquera. Identification of Novel Associations of Candidate Genes with Resistance to Late Blight in Solanum tuberosum Group Phureja. *Frontiers in Plant Science*, 8:1040, 2017.

[2] Jhon Berdugo-Cely, Raúl Iván Valbuena, Erika Sánchez-Betancourt, Luz Stella Barrero, and Roxana Yockteng. Genetic diversity and association mapping in the colombian central collection of solanum tuberosum L. Andigenum group using SNPs markers. *PLoS ONE*, 12(3), 2017.

[3] Peter M. Bourke, Roeland E. Voorrips, Richard G. F. Visser, and Chris Maliepaard. Tools for Genetic Studies in Experimental Populations of Polyploids. *Frontiers in Plant Science*, 9:513, 2018.

[4] Peter J Bradbury, Zhiwu Zhang, Dallas E Kroon, Terry M Casstevens, Yogesh Ramdoss, and Edward S Buckler. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635, 2007.

[5] Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):1–16, 2015.

[6] Rishika De, William S Bush, and Jason H Moore. *Bioinformatics Challenges in Genome-Wide Association Studies (GWAS)*, pages 63–81. Springer New York, New York, NY, 2014.

[7] Luís Felipe V. Ferrão, Juliana Benevenuto, Ivone de Bem Oliveira, Catherine Cellon, James Olmstead, Matias Kirst, Marcio F. R. Resende, and Patricio Munoz. Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context. *Frontiers in Ecology and Evolution*, 6:107, 2018.

[8] Dominik G Grimm, Damian Roqueiro, Patrice A Salomé, Stefan Kleeberger, Bastian Greshake, Wangsheng Zhu, Chang Liu, Christoph Lippert, Oliver Stegle, Bernhard Schölkopf, Detlef Weigel, and Karsten M Borgwardt. easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies. *The Plant Cell*, 29(1):5–19, 2017.

[9] Anja C Gumpinger, Damian Roqueiro, Dominik G Grimm, and Karsten M Borgwardt. *Methods and Tools in Genome-wide Association Studies*, volume 1819. 2018.

[10] Candice N. Hirsch, Cory D. Hirsch, Kimberly Felcher, Joseph Coombs, Dan Zarka, Allen Van Deynze, Walter De Jong, Richard E. Veilleux, Shelley Jansky, Paul Bethke, David S. Douches, and C. Robin Buell. Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries. *G3: Genes, Genomes, Genetics*, 3(6):1003–1013, 2013.

[11] Jie Meng, Kai Song, Chunyan Li, Sheng Liu, Ruihui Shi, Busu Li, Ting Wang, Ao Li, Huayong Que, Li Li, and Guofan Zhang. Genome-wide association analysis of nutrient traits in the oyster Crassostrea gigas: Genetic effect and interaction network. *BMC Genomics*, 20(1):1–14, 2019.

[12] Thomas A. Pearson and Teri A. Manolio. How to interpret a genome-wide association study. *JAMA - Journal of the American Medical Association*, 299(11):1335–1344, 2008.

[13] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1):41–50, 2016.

[14] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. De Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, 2007.

[15] Hui Ping Qiao, Chun Yang Zhang, Zhi Long Yu, Qi Min Li, Yang Jiao, and Jian Ping Cao. Genetic variants identified by GWAS was associated with colorectal cancer in the Han Chinese population. *Journal of Cancer Research and Therapeutics*, 11(2):468–470, 2015.

[16] Umesh R. Rosyara, Walter S. De Jong, David S. Douches, and Jeffrey B. Endelman. Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome*, 9(2):1–10, 2016.

[17] Sanjeev Kumar Sharma, Katrin MacKenzie, Karen McLean, Finlay Dale, Steve Daniels, and Glenn J. Bryan. Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3: Genes, Genomes, Genetics*, 8(10):3185–3202, 2018.

[18] Yong Yong Shi and Lin He. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci (Cell Research (2005) 15, (97-98) DOI: 10.1038/sj.cr.7290272). *Cell Research*, 16(10):851, 2006.

[19] Jiazheng Yuan, Benoît Bizimungu, David De Koeyer, Umesh Rosyara, Zixiang Wen, and Martin Lagüe. Genome-Wide Association Study of Resistance to Potato Common Scab. *Potato Research*, 2019.

[20] Shengkui Zhang, Xin Chen, Cheng Lu, Jianqiu Ye, Meiling Zou, Kundian Lu, Subin Feng, Jinli Pei, Chen Liu, Xincheng Zhou, Ping'an Ma, Zhaogui Li, Cuijuan Liu, Qi Liao, Zhiqiang Xia, and Wenquan Wang. Genome-wide association studies of 11 agronomic traits in cassava (Manihot esculenta crantz). *Frontiers in Plant Science*, 9(April):1–15, 2018.