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

L. Garreta[1], I. Cerón-Souza[1], M.R. Palacio[2], and P.H. Reyes-Herrera[1]

[1]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI Tibaitatá, Kilómetro 14, Vía a Mosquera, 250047, Colombia

[2]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), CI El Mira, Kilómetro 38, Vía Tumaco Pasto, Colombia

May 4, 2020

Abstract

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

Keywords: GWAS, tetraploids, SNPs,XXX

1 Introduction

The GWAS (Genome-Wide Association Study) is used to identify which variants through the whole genome of a large number of individuals are associated with a specific trait (CITES). This methodology started with humans and several model plants, such as rice, maize, and Arabidopsis (CITES). Because of the advances in the next-gen sequencing technology and the decreasing of the sequencing cost in recent years, there is an increase in genome sequences in non-model organisms at a faster rate (CITES). Therefore, several research projects are intended for the first time a GWAS analysis for non-model wild plants and crops that often are polyploids (CITES).

One of the main challenges in the GWAS analysis is to identify real associations. A reliable method to validate the results is by replicating the study using different software. This replication process is a challenge if our study organism is polyploid. Currently, the GWAS software to analyze polyploids is limited. Therefore, to confirm the GWAS of polyploids species, it is often necessary to "diploidizing" the data in order to use software designed exclusively for a diploid data matrix. Thus, the replication process is time-consuming. Each software has its characteristics, such as different user interfaces (GUI or command-line based), genotype-phenotype formats, models, algorithm assumptions, and outputs.

To solve this problem, we developed the MultiGWAS tool that performs GWAS analyses for tetraploid species using four software in parallel. Our tool include GWASpoly [17] and the SHEsis tool [19] that accept polyploid genomic data, and PLINK [15] and TASSEL [4] with the use of a "diploidized" genomic matrix. The tool deals with preprocessing data, running four GWAS tools in parallel, and create reports to help the user decide more intuitively the possible true or false associations.

2 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, 13]. 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.

3 Material and Methods

3.1 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 [17] and SHEsis [20], and another two designed for diploids species and extensively used in humans and plants: PLINK [15, 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.

3.2 GWASpoly

GWASpoly is a recent R package designed for GWAS in polyploid species that has been used in several studies in plants [2, 7, 18, 21]. 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.

3.3 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 [16, 12].

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).

3.4 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 [14]. 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.

3.5 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, 22], 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 built in the TASSEL built in algorithms.

3.6 Method

The MultiGWAS tool has three main steps, the adjustment state, the multi analysis stage, and the integration step (Fig. 1).

3.6.1 Adjustment stage

MultiGWAS takes as input a configuration file where the user specifies the genomics data along with the parameters that will be used by the four tools. It starts by preprocessing the genomic data by selecting individuals present in both genotype

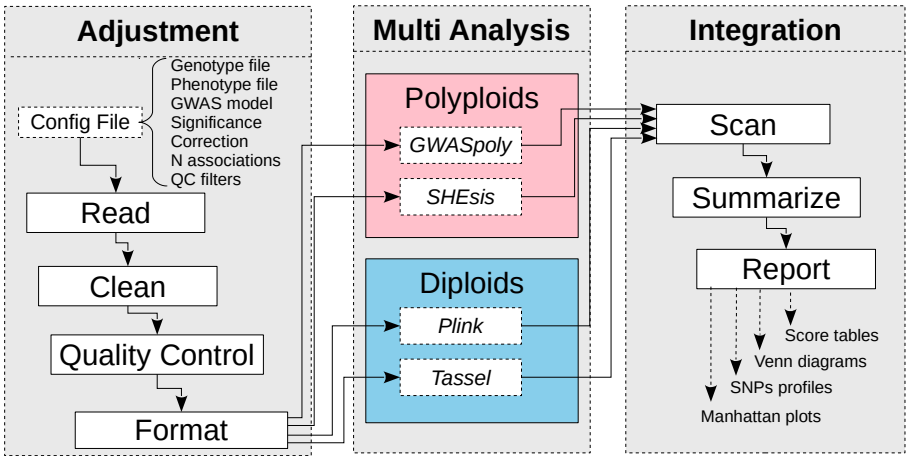


Figure 1: MultiGWAS flowchart has three stages: adjustment, multi analysis, and integration. In the first stage, we process the configuration file. It includes the genotype/phenotype filenames, genome-wide significance threshold, multiple testing correction methods, GWAS model, number of associations to be reported, and TRUE or FALSE whether to use quality control (QC) filters or not. Then the genotype and phenotype are cleaned and filtered using the QC filters. In the second stage, each GWAS tool runs in parallel. In the last stage, after the output files scanning, a summary of results is generated in a report that contains score tables, Venn diagrams, SNP profiles, and Manhattan plots. The QC filters are minor allele frequency, individual missing rate, SNP missing rate, HardyWeinberg threshold.

and phenotype and excluding individuals and SNPs that have poor quality. Moreover, the format "ACGT" suitable for the polyploid software GWASpoly and SHEsis, is "diploidized" for PLINK and TASSEL. The homozygous tetraploid genotypes are converted to diploid thus: (e.g.,AAAA→AA, CCCC→CC, GGGG→GG, TTTT→TT). Moreover, for tetraploid heterozygous genotypes, the conversion depends on the reference and alternate alleles calculated for each position (e.g. AAAT→AT, ... ,CCCG→CG). After this process, MultiGWAS transform the genomic data to the formats required for each software.

3.6.2 Multi analysis stage

MultiGWAS runs in parallel using two types of statistical models specified in the parameters file, the Full model (Q+K) and Naive (i.e., without any control) [18]. The Full model (Q+K) controls for both population structure and individual relatedness. For population structure, MultiGWAS uses the Principal Component Analysis (PCA) and takes the top ten PC as covariates. For relatedness, the tool uses kinship matrices that TASSEL and GWASpoly calculated separately, and for PLINK and SHEsis depends on the King software [11].

3.6.3 Integration stage.

The outputs resulting from the four software are scanned and processed to identify both significant and best-ranked associations. Based on the specification of the configuration file about the correction method and the significant level, MultiGWAS corrects the p-values and calculates the threshold value for each associated marker. The calculation uses the number of valid genotype calls (i.e., non-missing genotypes, phenotypes, and covariates). Then, MultiGWAS summarize the results in tables, Venn diagrams, SNP profiles, and Manhattan plots.

4 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 2), 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 3), 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`)

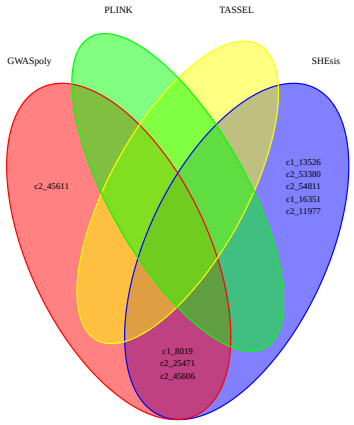
4.1 Visualization of significant SNPs

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 consequently possible true associations, when their $pvalue$ fall below a predefined significance level, usually 0.01 or 0.05.

Here, the MultiGWAS reports the SNPs considered statistically significant by each GWAS package. For that purpose, it provides two views: a tabular and Venn diagram. The table shows detailed information of the SNPs, where both $pvalues$ and significance levels have been scaled as $-log_{10}(pvalue)$, score (SCR) and threshold (THR) in the Figure 2.A. Whereas, Venn diagram visually shows the same SNPs but emphasizing if these were significant either for a single package or for more than one. As an example, Figure 2 shows the significative SNPs resulting from running MultiGWAS on a tetraploid potato dataset.

PKG	MDL	CHR	POS	SNP	SCR	THR	SGN
GWASpoly	Full	10	488631	c1_8019	4.78	4.25	TRUE
GWASpoly	Full	10	488084	c2_25471	4.57	4.27	TRUE
GWASpoly	Full	10	482034	c2_45611	4.36	4.27	TRUE
GWASpoly	Full	10	482188	c2_45606	4.68	4.50	TRUE
SHEsis	Full	2	136974	c1_8019	9.47	3.30	TRUE
SHEsis	Full	1	308379	c1_13526	8.45	3.29	TRUE
SHEsis	Full	5	460460	c2_53380	8.24	3.26	TRUE
SHEsis	Full	3	392552	c2_25471	7.82	3.29	TRUE
SHEsis	Full	5	498044	c2_54811	6.96	3.26	TRUE
SHEsis	Full	1	698098	c1_16351	6.02	3.28	TRUE
SHEsis	Full	4	693115	c2_45606	5.95	3.29	TRUE

(a)



(b)

Figure 2: **MultiGWAS views for significant SNPs.** (a) Table with detailed information of significant SNPs found by package and sorted by decreasing score (computed as $-log_{10}(pvalue)$). The information includes: package reporting the SNP, GWAS model used by the package, chromosome, position in the genome, name or ID, score, threshold to consider the SNP as significant, and TRUE of FALSE wether the SNP is statistically significant or not (score > threshod). (b) Venn diagram with the SNPs found significant either for one package or for more than one. SNPs found by only one package are at the top of its ellipse, while SNPs found by more than one package (shared) are at the intersections of their ellipses. For example, the SNP c2_45611 at the top left was found significant only by one tool: GWASpoly, but the three SNPs c1_8019, c2_25471, and c2_45606, were found significative by both packages GWASpoly and SHEsis. However, the other packages, PLINK and TASSEL, did not report any significant SNP.

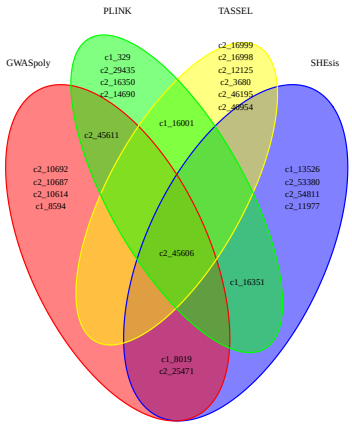
4.2 Visualization of best-ranked SNPs

Most GWAS packages compute differently both *pvalues* and significance levels, and these values may be computed either to high or to low, respectively. This results in SNPs with low *pvalues* but that did not reach the significance levels defined by the packages. Consequently, as it is important to know the significant SNPs, it is equally important to know these SNPs closer to being statistically significant, as they may represent important associations to consider for posterior analysis (e.g. false negatives).

As in the previous section, MultiGWAS tool provides a table and a Venn diagram to report these best-ranked SNPs for each GWAS package, wheter these SNPs were assesed significant or not. The number of SNPs to be reported is defined by the user in the configuration file, and these SNPs are listed in the table by tool in order of decreasing score, whereas the Venn diagram visually shows the same SNPs but emphasizing if these were best-ranked either in a single package or in several at once. Figure 3 shows the MultiGWAS best-ranked SNPs resulting from GWAS analysis on the tetraploid potato dataset.

PKG	MDL	CHR	POS	SNP	SCR	THR	SGN
GWASpoly	Full	10	48863165	c1_8019	4.7800	4.2500	TRUE
GWASpoly	Full	10	48808404	c2_25471	4.5700	4.2700	TRUE
GWASpoly	Full	10	48203431	c2_45611	4.3600	4.2700	TRUE
GWASpoly	Full	10	48218826	c2_45606	4.6800	4.5000	TRUE
....	....	....	....	....	....	....	....
PLINK	Full	10	67293176	c1_16001	1.7693	3.2601	FALSE
PLINK	Full	10	69323144	c2_45611	1.0229	3.2553	FALSE
PLINK	Full	2	41814861	c2_16350	0.9598	3.3010	FALSE
PLINK	Full	10	69311500	c2_45606	0.8489	3.2923	FALSE
....	....	....	....	....	....	....	....
SHEsis	Full	2	13697423	c1_8019	9.4711	3.3010	TRUE
SHEsis	Full	1	30837971	c1_13526	8.4501	3.2923	TRUE
SHEsis	Full	3	39255236	c2_25471	7.8241	3.2923	TRUE
SHEsis	**Full**	**4**	**69311500**	**c2_45606**	**5.9557**	**3.2923**	**TRUE**
....	....	....	....	....	....	....	....
TASSEL	Full	10	47539878	c1_16001	2.2143	3.8943	FALSE
TASSEL	Full	1	64259758	c2_46195	2.1478	3.8943	FALSE
TASSEL	Full	1	63756796	c2_40954	1.9548	3.8943	FALSE
TASSEL	**Full**	**10**	**48218826**	**c2_45606**	**1.9443**	**3.8943**	**FALSE**

(a)

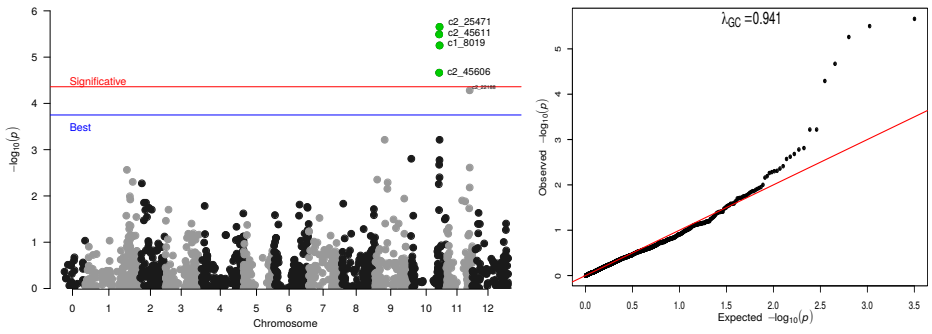


(b)

Figure 3: **MultiGWAS views for best-ranked SNPs.** The views are similar to the ones described in the Figure 2, but now they present a list of high-scored SNPs by package and ordered by their score. (a) Table view listing by package their first N=8 best-ranked SNPs (but only shown here four SNPs by package). (b) Venn diagram showing by package their N=8 best-ranked SNPs. One SNP was best-ranked by the four packages, c2_45606 (central intersection of the diagram and blue highligthed row in the table). Whereas, other SNPs were best-ranked by more than one tool, c1_8019 and 25471 (at the bottom of the diagram), were best-ranked by two packages: GWASpoly and SHEsis..

4.3 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 4). 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.



(a) (b)

Figure 4: **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 3) 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.

5 Discussion

XXXXXXXXXXXXXXXXXXXXXXX

6 Availability and implementation:

The tool is in R. Source code, examples, documentation and installation instructions are available at https://github.com/agrosavia/multiGWAS-min

References

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

[2] Jhon Berdugo-Cely et al. “Genetic diversity and association mapping in the colombian central collection of solanum tuberosum L. Andigenum group using SNPs markers”. In: *PLoS ONE* 12.3 (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0173039.

[3] Peter M. Bourke et al. “Tools for Genetic Studies in Experimental Populations of Polyploids”. In: *Frontiers in Plant Science* 9 (2018), p. 513. ISSN: 1664-462X. DOI: 10.3389/fpls.2018.00513. URL: http://journal.frontiersin.org/article/10.3389/fpls.2018.00513/full.

[4] Peter J Bradbury et al. “TASSEL: software for association mapping of complex traits in diverse samples”. In: *Bioinformatics* 23.19 (2007), pp. 2633–2635. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm308. URL: https://doi.org/10.1093/bioinformatics/btm308.

[5] Christopher C. Chang et al. “Second-generation PLINK: Rising to the challenge of larger and richer datasets”. In: *GigaScience* 4.1 (2015), pp. 1–16. ISSN: 2047217X. DOI: 10.1186/s13742-015-0047-8. arXiv: 1410.4803.

[6] Rishika De, William S Bush, and Jason H Moore. “Bioinformatics Challenges in Genome-Wide Association Studies (GWAS)”. In: *Clinical Bioinformatics*. Ed. by Ronald Trent. New York, NY: Springer New York, 2014, pp. 63–81. ISBN: 978-1-4939-0847-9. DOI: 10.1007/978-1-4939-0847-9_5. URL: https://doi.org/10.1007/978-1-4939-0847-9%7B%5C_%7D5.

[7] Luís Felipe V. Ferrão et al. “Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context”. In: *Frontiers in Ecology and Evolution* 6 (2018), p. 107. ISSN: 2296-701X. DOI: 10.3389/fevo.2018.00107. URL: https://www.frontiersin.org/articles/10.3389/fevo.2018.00107/full.

[8] Dominik G Grimm et al. “easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies”. In: *The Plant Cell* 29.1 (2017), pp. 5–19. ISSN: 1040-4651. DOI: 10.1105/tpc.16.00551. URL: http://www.plantcell.org/content/29/1/5.

[9] Anja C Gumpinger et al. *Methods and Tools in Genome-wide Association Studies*. Vol. 1819. 2018. ISBN: 9781493986187.

[10] Candice N. Hirsch et al. “Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries”. In: *G3: Genes, Genomes, Genetics* 3.6 (2013), pp. 1003–1013. ISSN: 21601836. DOI: 10.1534/g3.113.005595.

[11] Ani Manichaikul et al. “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* 26.22 (2010), pp. 2867–2873. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq559.

[12] Jie Meng et al. “Genome-wide association analysis of nutrient traits in the oyster Crassostrea gigas: Genetic effect and interaction network”. In: *BMC Genomics* 20.1 (2019), pp. 1–14. ISSN: 14712164. DOI: 10.1186/s12864-019-5971-z.

[13] Thomas A. Pearson and Teri A. Manolio. “How to interpret a genome-wide association study”. In: *JAMA - Journal of the American Medical Association* 299.11 (2008), pp. 1335–1344. ISSN: 00987484. DOI: 10.1001/jama.299.11.1335.

[14] Robert A. Power, Julian Parkhill, and Tulio De Oliveira. “Microbial genome-wide association studies: lessons from human GWAS”. In: *Nature Reviews Genetics* 18.1 (2016), pp. 41–50. ISSN: 14710064. DOI: 10.1038/nrg.2016.132.

[15] Shaun Purcell et al. “PLINK: A tool set for whole-genome association and population-based linkage analyses”. In: *American Journal of Human Genetics* 81.3 (2007), pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795.

[16] Hui Ping Qiao et al. “Genetic variants identified by GWAS was associated with colorectal cancer in the Han Chinese population”. In: *Journal of Cancer Research and Therapeutics* 11.2 (2015), pp. 468–470. ISSN: 19984138. DOI: 10.4103/0973-1482.150346.

[17] Umesh R. Rosyara et al. “Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato”. In: *The Plant Genome* 9.2 (2016), pp. 1–10. ISSN: 1940-3372. DOI: 10.3835/plantgenome2015.08.0073. URL: https://dl.sciencesocieties.org/publications/tpg/abstracts/9/2/plantgenome2015.08.0073.

[18] Sanjeev Kumar Sharma et al. “Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato”. In: *G3: Genes, Genomes, Genetics* 8.10 (2018), pp. 3185–3202. ISSN: 21601836. DOI: 10.1534/g3.118.200377.

[19] Jiawei Shen et al. “SHEsisPlus, a toolset for genetic studies on polyploid species”. In: *Scientific Reports* 6 (2016), pp. 1–10. ISSN: 20452322. DOI: 10.1038/srep24095. URL: http://dx.doi.org/10.1038/srep24095.

[20] 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)”. In: *Cell Research* 16.10 (2006), p. 851. ISSN: 10010602. DOI: 10.1038/sj.cr.7310101.

[21] Jiazheng Yuan et al. “Genome-Wide Association Study of Resistance to Potato Common Scab”. In: *Potato Research* (2019). ISSN: 18714528. DOI: 10.1007/s11540-019-09437-w.

[22] Shengkui Zhang et al. “Genome-wide association studies of 11 agronomic traits in cassava (Manihot esculenta crantz)”. In: *Frontiers in Plant Science* 9.April (2018), pp. 1–15. ISSN: 1664462X. DOI: 10.3389/fpls.2018.00503.