Post-GWAS Prioritization of Genome-Phenome Associations in Sorghum

1 **Debasmita Pal[1*], Kevin Schaper[2], Addie Thompson[3], Jessica Guo[4], Pankaj Jaiswal[5], Curtis**

2 **Lisle[6], Laurel Cooper[5], David LeBauer[4], Anne Thessen[2†], Arun Ross[1*†]**

3 [1]Department of Computer Science and Engineering, Michigan State University, East Lansing,

4 Michigan, USA

5 [2]Translational & Integrated Sciences Laboratory, University of Colorado Anschutz Medical Campus,

6 Department of Biomedical Informatics, Aurora, Colorado, USA

7 [3]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing,

8 Michigan, USA

9 [4]Arizona Experiment Station, University of Arizona, Tucson, Arizona, USA

10 [5]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

11 [6]KnowledgeVis, LLC., Altamonte Springs, Florida, USA

12 [†]These authors share last authorship

13 *** Correspondence:** paldebas@msu.edu, rossarun@cse.msu.edu

14 **Keywords: plant phenotype, maximum canopy height, maximum growth rate, *Sorghum bicolor*,**

15 **SNP, GWAS, pattern recognition, feature engineering**

16 **Abstract**

This study presents an innovative approach for understanding the genetic underpinnings of two key phenotypes in *Sorghum bicolor*: maximum canopy height and maximum growth rate. Genome-Wide Association Studies (GWAS) are widely used to decipher the genetic basis of traits in organisms, but the challenge lies in selecting an appropriate statistically significant threshold for analysis. Our goal was to employ GWAS to pinpoint the genetic markers associated with the phenotypes of interest using specific permissive-filtered threshold values that allows the inclusion of broader collections of explanatory candidate genes. Then, we utilized a pattern recognition technique to prioritize a set of informative genes, which hold potential for further investigation and could find applications in Artificial Intelligence systems. Utilizing a subset of the Sorghum Bioenergy Association Panel cultivated at the Maricopa Agricultural Center in Arizona, we sought to unveil patterns between phenotypic similarity and genetic proximity among accessions in order to organize Single Nucleotide Polymorphisms (SNPs) which are likely to be associated with the phenotypic trait. Additionally, we explored the impact of this method by considering all SNPs versus focusing on SNPs classified through the GWAS pre-filter. Experimental results indicated that our approach effectively prioritizes SNPs and genes influencing the phenotype of interest. Moreover, this methodology holds promise in the feature selection from genomic data for predicting complex phenotypic traits influenced by numerous genes and environmental conditions and could pave the way for further research in this field.

Author Summary

Analyzing the relationships of an organism's phenotypes with their genotypes and environments stands as a critical aspect in plant biology. Over the past few decades, researchers have focused on identifying impactful genes controlling the plant phenotypes in diverse environmental conditions, a pursuit especially relevant in the face of global climate change. Applied field experiments and

quantitative genetics have identified numerous regions of the genome influencing complex traits. However, the substantial time and cost required for physical experiments have prompted the use of computational methods to facilitate this field of research. Our work aims to unravel the patterns between phenotypic similarity and genetic proximity across a panel of sorghum accessions to prioritize genes influencing the phenotype of interest. Our results suggest that this method is an effective way to prioritize informative genes for further analysis, and for genomic features selection that can be used by machine learning and deep learning algorithms for predicting complex phenotypic traits.

1. Introduction

To mitigate the effects of climate change and population growth, we need to better understand the dynamic interplay between an organism's genome, phenome, and environmental effects. Artificial Intelligence (AI)-based systems using machine learning (ML) and deep learning (DL) algorithms have shown great promise in capturing the complex and nonlinear response of natural systems and genomics (Lin & Ngiam, 2023, Ghannam & Techtmann, 2021; Hsieh, 2022; Thessen, 2016), but several barriers still exist (Danilevicz et al, 2022), including the absence of AI-ready data for most biological applications. This is especially true for genomic data, which contains millions of potential features that can quickly overwhelm an AI model. Therefore, genomic data needs filtering through an automated pattern recognition process to identify the most distinguishing components for use in training the ML/DL algorithms. For model organisms, this can be a simple process of selecting data from genome regions that are known to be causal for the phenotype of interest, but it is challenging for complex phenotypes, such as growth rate and biomass accumulation, which are regulated by molecular interactions and biological pathways involving many genes, and greatly influenced by environmental conditions.

63 Genome-wide association study (GWAS, Visscher et al., 2012) is a widely-used technique to deduce
64 the genetic basis of phenotypic traits in organisms. However, the challenge lies in selecting an
65 appropriate statistically significant threshold (p-value) for the analysis, considering the false
66 discovery rate (Hebbring, 2014). A stringent significance threshold for associations would possibly
67 miss causative variance present at lower frequencies in the population. Other methods to decrease
68 genomic feature set size, for example, correlation with the trait of interest, would be potentially
69 confounded with population structure. Though these features would potentially be highly predictive
70 of the phenotypes, they would likely not lend insight into underlying biology and gene function -
71 only to breeding history or source of accessions in the population. Thus, our approach was to use
72 certain permissive-filtered GWAS thresholds that retain many more features than what the GWAS
73 analysis itself would have identified as significantly associated with the phenotypic trait. This would
74 likely increase false positives as well as true positives and lead to subsets of potentially informative
75 SNPs, whose association with the phenotypes was presumably not confounded with the population
76 structure.

77 On top of the genetic markers identified by permissive-filtered GWAS thresholds (allowing more
78 explanatory candidate genes), we utilized a pattern recognition technique based on statistical
79 correlation coefficient to establish genome-phenome associations in the context of environmental
80 variation. The idea was to compare the resemblance of phenotypic similarity determined by the
81 *normalized* phenotypic trait measurements among accessions with their genetic relatedness. Through
82 this process, our objective was to prioritize Single Nucleotide Polymorphisms (SNPs) and genes,
83 which are likely to be correlated with the phenotype under study. Our algorithm also incorporated a
84 kind of ordering for the prioritized SNPs based on their impact on phenotypic traits, which alleviates
85 the hurdle of discovering SNPs with lower effect estimates through GWAS (Moore et al, 2010). We
86 considered the phenotypes, viz. maximum canopy height and maximum growth rate of *Sorghum*

87 *bicolor,* to examine the usefulness of our method. Empirical evidence in the form of ontology-based
88 gene functions, spatial and temporal expression, and similarity to known homologs were curated,
89 describing that our methodology shows a promising avenue to extract the genes on precedence that
90 may be informative for the studied phenotype. We hypothesize that these SNPs and genes could
91 potentially be considered for new plant breeding and genomics projects to empirically validate their
92 functional annotations.

93 Moreover, the prioritized SNPs and genes extracted by our approach could be utilized to improve the
94 performance of ML/DL algorithms. At the outset of this work, we envisaged to develop ML models
95 to predict the phenotype given information about genomes and environments. Our study focused on a
96 subset of the Sorghum Bioenergy Association Panel, grown at the Maricopa Agricultural Center in
97 Arizona, as part of the TERRA-REF project. Throughout the growing season, various phenotype
98 measurements were collected regularly by an automated field scanner system and field scientists. The
99 difficulty to train ML models based on this dataset was the size of an entire genome compared to the
100 availability of a limited number of phenotypic observations. In general, it is hugely expensive to
101 cultivate at such a large scale to obtain a comparable number of phenotypic measurements with
102 genetic makeup, necessitating some method of genomic feature selection. In GWAS, prefiltering
103 (e.g., on missingness, error probability, linkage disequilibrium, minor allele frequency, or other
104 metrics) could be a favorable way to reduce the dimensionality of the datasets (the number of SNPs),
105 however, they manifested mixed outcomes (Danilevicz et al, 2022, Rice and Lipka, 2019), and the
106 best approach for doing so is still a subject of research.

107 We primarily intended to leverage GWAS in our experiment as it has already shown its capability in
108 finding causal SNPs for complex phenotypic traits and posed our method as post-GWAS analysis
109 upon the selection of lower stringent thresholds. Our approach aims to establish the associations

between genotypes and phenotypes based on the statistical correlation in contrast to regression, which was mostly used by genome-phenome association studies, e.g., GWAS and any BLUP-based methods (Zhang et al., 2022). Regression analysis typically derives a specific equation to express the relationship between independent and dependent variables, whereas correlation helps in determining if at all a relationship exists between two independent distributions, and their outcomes might vary depending on the patterns present in the dataset. Furthermore, we analyzed the relationships at the granularity of every pair of accessions rather than at an accession level. Additionally, the use of *normalized* phenotypic trait measurements introduced more resilience to outliers in our experiments and made the extracted results more comparable across different environmental conditions. We also explored the prospect of our method by investigating the difference in results with applying GWAS pre-filter and without GWAS pre-filter.

Background:

The most intensive and comprehensive studies of genotypes, environments, and phenotypes are being conducted in agricultural research, where plant breeding programs and seed and livestock companies try to find genes or combinations of genes for desired phenotypes (Lobos & Hancock, 2015; Lopes et al., 2015). The most important traits of interest in agricultural plant breeding, such as growth, plant height and yield are generally determined by suites of genes, rather than a single gene, and are influenced by environmental factors. For example, in soybean, Cao et al. (2019) found a number of quantitative trait loci (QTL) implicated in controlling plant height with environmental and epistatic interaction effects. This complicates the process of identifying informative areas of the genome and leads to the development of novel approaches.

In sorghum, QTL analysis was used to identify genomic regions important for plant height (Brown et al., 2008), early-season cold tolerance (Burow et al., 2011), and Ma2, a regulator of flowering time

133 (Casto et al., 2019), among others. Using a biocuration strategy, Mace et al. (2019) compiled a
134 comprehensive QTL Atlas (https://aussorgm.org.au/sorghum-qtl-atlas/) with the projected locations
135 of ~6000 QTL on the sorghum reference genome by data mining approximately 150 peer-reviewed
136 publications in sorghum from 1995 to the present. Another approach involved the development of
137 populations of sorghum plants with induced mutations, for example, the 6400 ethyl methane
138 sulfonate-mutagenized lines developed by Jiao et al. (2016), which was utilized to identify genes
139 involved in drought tolerance (Jiao et al., 2016) and spikelet fertility (Gladman et al., 2019).

140 With the advent of whole genome sequencing in sorghum (Mace et al., 2013; McCormick et al.,
141 2018; Boyles et al., 2019; Cooper et al., 2019), Single Nucleotide Polymorphisms (SNPs) and
142 insertions/deletions were identified and can be associated with variations in traits or phenotypes of
143 interest (Rafalski, 2002). This led to newer approaches such as association mapping (Zhu et al.,
144 2008) and GWAS (Visscher et al., 2012). In sorghum, GWAS has been utilized for determining the
145 genetic basis for such complex phenotypes as root system architecture (Zheng et al., 2020), grain
146 polyphenol content (Rhodes et al., 2014), grain size (Tao et al., 2020), disease responses (Ahn et al.,
147 2019), and grain yield (Boyles et al., 2019).

148 Association studies in specialized panels or populations led to advances in our understanding of plant
149 height (Brown et al., 2008), grain quality (Sukumaran et al., 2012), and bioenergy traits (Brenton et
150 al., 2016). Other methods based on whole genome analysis, including such approaches as genomic
151 selection (Meuwissen et al., 2001; Budhlakoti et al., 2022) and pan-genome analysis (Tao et al.,
152 2021, Naithani et al., 2023), were also utilized.

153 Further, to improve phenotyping for the genomic prediction model, Watanabe et al. (2017) suggested
154 a high-throughput phenotyping system for sorghum plant height, a complex quantitative trait by
155 using unmanned aerial vehicle (UAV) imagery. Researchers also attempted to apply ML approaches

156 for genomic selection that primarily rely on the availability of genomic data to predict phenotypic
157 traits (Tong and Nikoloski, 2021). Costa et al. (2022) proposed genomic prediction for traits
158 modulated by epistatic effect by applying ML and artificial neural networks. Their comparative study
159 indicated better performance with neural networks for complex quantitative traits. Further, Jubair and
160 Domaratzki (2023) presented a comprehensive overview of deep learning approaches in the genomic
161 selection process by incorporating weather data. Similarly, Adak et al. (2021), using their UAV-
162 based weekly observance of plant height phenotype for a growing season of the maize crop, observed
163 that the two SNP markers previously identified by GWAS methods show a greater effect on early-
164 season growth, depicting a more dynamic relationship, that is often not observed in classical
165 experiments. These approaches seek to identify important genes by accounting for the variability in
166 the environment.

167 2. Materials and Methods

168 **Our Approach:**

169 To establish genome-phenome associations, we compared the phenotypic similarity of each pair of
170 accessions with their genetic relatedness in terms of matching alleles at a particular SNP location on
171 the chromosome. We utilized the differences among genetically diverse accessions in the normalized
172 phenotypic trait measurements to build a "target" similarity matrix describing the proximity of each
173 pair of accessions. This is based on the assumption that pairs of accessions showing small differences
174 in phenotype are most likely to be genetically similar. The normalization was obtained by fitting a
175 logistic growth curve model to the measurements. The next step was to construct a similarity matrix
176 corresponding to each SNP identified across the same set of accessions to discover which similarity
177 matrices are analogous to the "target" similarity matrix. This was achieved by feature engineering
178 using Pearson correlation coefficient, which helped in prioritizing a set of SNPs likely to be

179 informative about the phenotype under study. Figure 1 demonstrates the pipeline of our approach.
180 We utilized permissive-filtered GWAS thresholds at various p-value levels resulting in multiple sets
181 of potentially informative SNPs to be included in our experiment. We also compared these results
182 with that obtained considering all the variants across all chromosomes without applying GWAS pre-
183 filter. In the following, we have provided a detailed description of the datasets and the methods used
184 in our experiments. Our code and experimental results are available on GitHub
185 (https://github.com/genophenoenvo/phenotype-to-genotype).

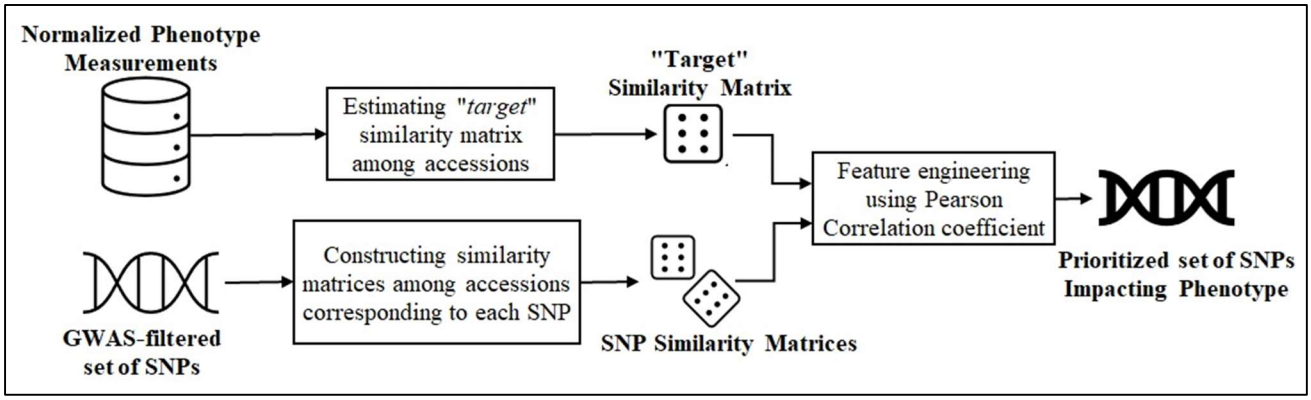

Figure 1. Method Pipeline for Prioritizing SNPs. This pipeline begins with *normalized* phenotypic trait data and a set of GWAS-filtered SNPs in separate files and ends with lists of prioritized SNPs impacting the phenotypic traits.

186 **Plant Material and Phenotype Dataset Description:**

187 The multi-year TERRA Phenotyping Reference Platform (TERRA-REF) program (Burnette et al.,
188 2018; LeBauer et al., 2020) studied how the environment and other factors affected the growth of
189 agricultural products. An underlying goal of the program is to study how to maximize growth and, as
190 a result, maximize agricultural bioenergy output. Season 6 of the program was held in the summer of
191 2017. *Sorghum bicolor* was chosen as the target organism for this study because of its drought
192 tolerance, variety of food, fuel, and fiber applications, and potential improvement through targeted
193 breeding. A subset of 336 sorghum accessions from the Bioenergy Association Panel (BAP) (Brenton
194 et al., 2016) were planted in the study field in Arizona (Maricopa Agricultural Center). An automatic

195 large field scanner system observed these plants for phenotype measurements every day or two
196 throughout the growth season. In addition to comparing automated observations, a small number of
197 common phenotypes were also recorded by field scientists.

198 A suite of measurements, including photographs, canopy height, leaf characteristics, and soil
199 environmental factors, were captured regularly by automated systems to analyze each accession's
200 growth patterns (Burnette et al., 2018). Two replicate plots of each accession were planted within a
201 200m x 20m field. Data collected by TERRA-REF is available in the public domain, along with
202 experimental design, measurement methods, and other study metadata (LeBauer et al., 2020). Our
203 analysis used canopy height phenotype and whole-genome resequencing data. Plot-level canopy
204 height measurements were estimated from laser scans using an algorithm calibrated to hand
205 measurements. Canopy height measurements are available for most days during the growing season.
206 Genomics data are from whole-genome resequencing of the BAP genotypes included in the study.
207 To test the impact of a different environment, we used end-of-season height data from the same
208 accessions grown in Florence County, South Carolina (Brenton et al., 2016).

209 **Phenotype Data Preparation and Normalization:**

210 Automated measurements of canopy height over the growing season were fitted with logistic growth
211 curves to obtain the phenotypic traits of maximum canopy height and maximum growth rate relative
212 to growing degree days (gdd), defined as the accumulation of heat energy above 10°C following the
213 "Method 1" described by McMaster & Wilhelm (1997). This approach can yield phenotypic traits
214 that are less sensitive to outliers, more comparable across environmental conditions, and permit
215 greater biological inference.

216 First, we applied a cleaning algorithm that excluded accessions having a number of automatic
217 measurements per season of less than 35 or having the measurements on fewer than 40% of days,

which resulted in a total of 326 accessions. This preprocessing yielded a time series that could robustly estimate the parameters of maximum canopy height and maximum growth rate.

Time series of canopy height and growing degree days (gdd) for each accession were modeled in a Bayesian framework with the likelihood:

$$height_i \sim Normal(\mu_i, \sigma^2) \qquad (1)$$

where, μ is the expected value of height, σ^2 is the measurement error variance, and i indexes each observation. The logistic model was defined as:

$$\mu_i = \frac{c}{(1+ae^{b*gdd})} \qquad (2)$$

with the logistic parameters a, b, and c and the covariate *gdd* associated with each observation. However, we re-parameterized the model to obtain more biologically meaningful parameters by defining:

$$Y_{max} = c \qquad (3)$$

$$Y_{min} = \frac{c}{(1+a)} \qquad (4)$$

$$R_{half} = -\frac{c*b}{4} \qquad (5)$$

Finally, we transformed the logistic parameters to the whole real line, which allowed for faster mixing and improved model convergence:

$$c = e^{\theta_c} \qquad (6)$$

$$a = e^{\theta_a} - 1 \qquad (7)$$

236 $$b = -e^{\theta_b} \quad (8)$$

237 All root nodes were given wide, relatively non-informative standard priors, including
238 *Normal(0,1000)* for all transformed logistic parameters (θ_a, θ_b, θ_c) and *Gamma(0.1, 0.1)* for the
239 observation precision $\left(\frac{1}{\sigma^2}\right)$, where *Normal*() and *Gamma*() refer to probability distributions.

240 We implemented the above model in JAGS 4.3.0 (Plummer, 2003) via R/rjags (Plummer et al. 2022;
241 R Core Team, 2021). Three parallel Markov Chain Monte Carlo (MCMC) sequences were assigned
242 with dispersed starting values. The models were run until convergence was achieved at a Gelman and
243 Rubin (1992) diagnostic parameter of < 1.2. An additional run of 10,000 iterations with the thinning
244 parameter as 10 yielded a total of 3,000 relatively independent posterior samples for each parameter
245 of interest. The posterior median of Y_{max} (maximum canopy height, cm) and R_{half} (maximum growth
246 rate, cm/*gdd*) of the accessions were considered to estimate the "target" similarity matrix for the
247 corresponding phenotypic traits and in the GWAS analysis (described below). Figure 2 shows the
248 logistic growth curve obtained for two sample accessions. The data and code for the logistic growth
249 curve model are publicly available (Guo & LeBauer, 2022).

250 **Estimation of "Target" Similarity Matrix:**

251 Based on the above-mentioned, *normalized* phenotypic measurements of the *Sorghum* accessions, we
252 computed the "target" similarity matrix for each of the phenotypes (maximum canopy height and
253 maximum growth rate) *separately*. We assumed that the accessions having close genetic proximity
254 with each other will exhibit a smaller difference in their phenotypic observations. Hence, this matrix
255 helped us estimate the potential correlation coefficient values among accessions. The rows and
256 columns of this matrix correspond to the accessions and the values indicate the "target" similarity

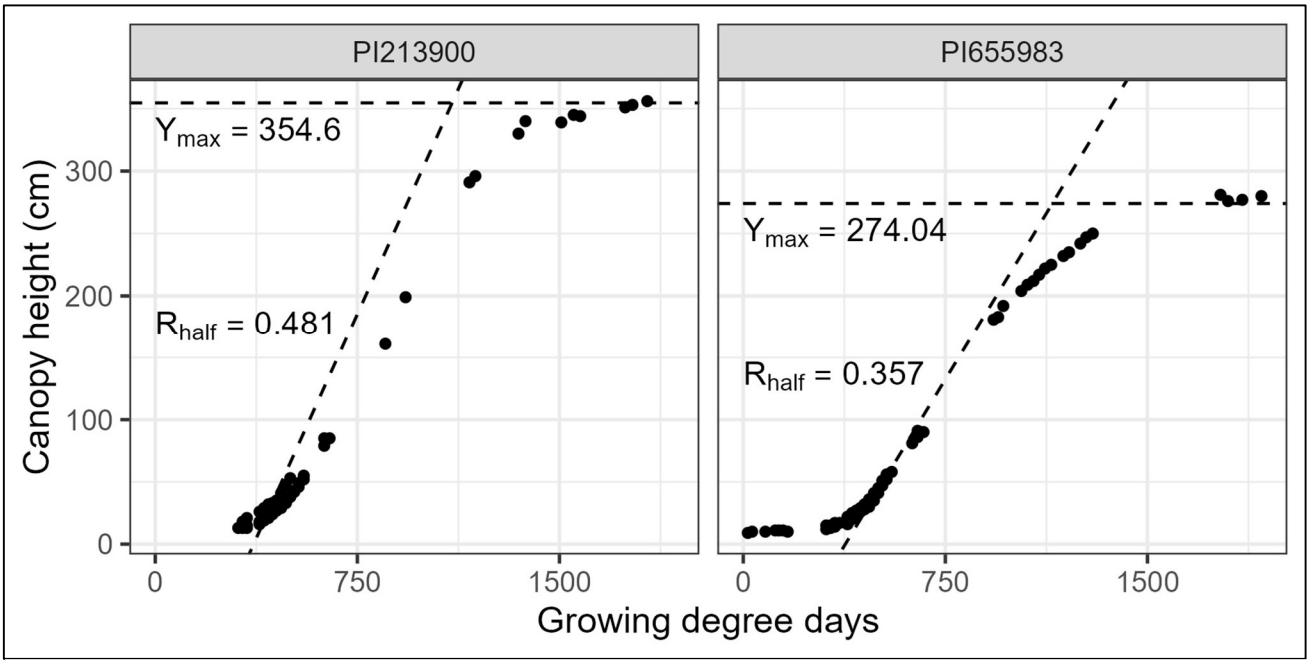

Figure 2. Example of logistic growth curves to illustrate phenotypes of two sorghum accessions. Using automated canopy heights for each accession, logistic growth curves were fitted and produced maximum canopy height (Y_{max}) and maximum growth rate (R_{half}). The cleaning algorithm applied ensured a sufficient number of observations to produce confident estimations of the parameters.

value (potential correlation coefficient), with a range of [0,1]. The higher the similarity value, the greater is the similarity between the accessions, based on the phenotype under study.

The "target" similarity value between the accessions *'i'* and *'j'* was calculated as follows:

1) First, the difference (x_{ij}) in their phenotypic measurement was computed.
2) Min-max scaling was applied to the difference (x_{ij}), using the following equation, in order to scale the difference in the range [0,1]:

$$x_{ij}^{scaled} = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}} \qquad (9)$$

where, x_{min} and x_{max}, respectively, denote the minimum and maximum values obtained over all x_{ij}'s.

3) The scaled difference was subtracted from 1 ($x_{ij}' = 1 - x_{ij}^{scaled}$), so that a higher value indicates a greater similarity between accessions.

Genomic Data Preparation and GWAS:

Throughout the project, we used the Sorghum BTX623 reference genome v3.1, and the analysis of function (described in the "SNP Annotation and Function" subsection) was carried out on either Phytozome (Goodstein et al., 2012), Gramene (Tello-Ruiz et al., 2021), Planteome (Cooper et al., 2018) or the Gene Ontology (The Gene Ontology Consortium, 2021) online databases. Whole genome sequencing data was obtained from the TERRA-REF project (LeBauer et al., 2020, https://docs.terraref.org/protocols/genomic-data), where 384 sorghum accessions from the sorghum Bioenergy Association Panel (BAP) were sequenced with ~25x coverage. Shotgun sequencing (127-bp paired-end) was done using an Illumina X10 instrument at the HudsonAlpha Institute for Biotechnology. After aligning the sequence reads (BTX623 reference genome v3.1 available from Phytozome), GVCF files with variant calls were downloaded from CyVerse, merged, and filtered using protocols described in the TERRA-REF documentation. SNPs were filtered using GATK Variant Filtration, and accessions without phenotypic observations were removed, and only biallelic sites were considered.

GWAS was performed using rMVP (Yin et al., 2021) with the FarmCPU algorithm (Liu et al., 2016) for each of the normalized phenotypes (maximum canopy height and maximum growth rate) using a kinship matrix and 1 principle component to account for relatedness and structure within the population. The results were filtered with *p*-values of .0001, .0005, and .001 to create six different lists of SNPs (Supplementary File). Additionally, these six lists were filtered against QTL regions known to be associated with height and growth obtained from the Sorghum QTL Atlas (Mace et al.,

287 2019). After this step, we had 12 lists of SNPs, which were utilized to filter the GVCF file to obtain

288 12 VCF files, hereafter, referred to as experimental conditions (Table 1 - Total #SNPs).

Table 1. Experimental Summary and Resulting SNP Lists using Permissive-Filtered GWAS Thresholds

p-value	Phenotype	QTL Filtering	Total #SNPs[1]	Prioritized #SNPs[2]	Prioritized #SNPs directly map to known genes	Highest Correlation Coefficient[3]
*p*0001	Maximum Growth	w/ QTL	534	77	18	0.41
		No QTL	1,055	107	28	0.49
	Maximum Height	w/ QTL	4,630	286	69	0.61
		No QTL	6,003	323	84	0.63
*p*0005	Maximum Growth	w/ QTL	2,861	174	30	0.55
		No QTL	5,575	198	48	0.60
	Maximum Height	w/ QTL	15,888	538	119	0.67
		No QTL	20,899	537	133	0.68
*p*001	Maximum Growth	w/ QTL	6,106	230	47	0.59
		No QTL	11,698	326	82	0.64
	Maximum Height	w/ QTL	27,536	634	140	0.68
		No QTL	36,033	755	182	0.69

[1] Total number of SNPs identified through GWAS.
[2] Number of prioritized SNPs extracted.
[3] Pearson correlation coefficient of "target" similarity matrix with the summation of SNP similarity matrices corresponding to the prioritized SNPs.

289 **Creation of SNP Similarity Matrices:**

290 The above VCF files were processed separately to create a similarity matrix for each SNP, comparing

291 the allelic similarity of every pair of accessions at that location on the chromosome. The values of

292 these similarity matrices were computed with an addition of 0.5 for each matching allele between any
293 pair of accessions, along with reversed heterozygous SNPs receiving a value of 1.0. For instance, if a
294 single VCF file consists of 'N' number of SNPs across 'a' number of accessions, there would be a
295 total of 'N' SNP similarity matrices; each of these similarity matrices is a square matrix of
296 dimensions 'a x a', and their cell values encoded among three discrete values (0, 0.5, 1.0). Out of the
297 326 accessions obtained as part of the logistic growth model, we used 274 accessions that were
298 common to both Seasons 4 and 6 of the TERRA-REF project for generating the similarity matrices
299 (common set of accessions were considered for future work). Thus, each SNP similarity matrix has
300 dimensions of 274 x 274 in our experiment.

301 **Prioritization of SNPs:**

302 Next, we compared the corresponding SNP similarity matrices (described above) to the phenotype
303 similarity matrix ("target" similarity matrix) using the algorithm described below. The notion was to
304 find out the SNP similarity matrices wherein the variation between the accessions is similar to that in
305 the phenotypic measurements. Pearson correlation coefficient ($\rho_{(X,Y)}$) was used to measure the
306 relationship between two data matrices. It is defined by the covariance of two data matrices X and Y
307 divided by their standard deviations.

308 $$\rho_{(X,Y)} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (10)$$

309 where, μ_X and μ_Y are, respectively, the mean and, σ_X and σ_Y are, respectively, the standard deviations
310 of the data matrices X and Y. We followed the feature selection technique, known as Sequential
311 Forward Selection (SFS) (Pudil et al, 1994) in which the important features are sequentially selected
312 using an evaluation criterion and appended to the empty candidate list until the addition of further
313 features does not reduce the criterion.

314 We prioritized a certain number of SNPs from a set of SNP similarity matrices, each corresponding
315 to a SNP, by using Pearson correlation coefficient (Table 1 - Prioritized #SNPs). To accomplish this,
316 we initiated the process by defining a "*maximum desired number of prioritized SNPs"* (e.g., 1000),
317 and a *null* "candidate" SNP list. In the first iteration, we selected a SNP similarity matrix from the set
318 of SNP similarity matrices that resulted in a maximum correlation coefficient with the "target"
319 similarity matrix and added the corresponding SNP to the candidate SNP list. Starting from the
320 second iteration, we chose the SNP similarity matrix from the remaining set (i.e., the corresponding
321 SNP not present in the candidate SNP list), which, when summed up with the similarity matrices
322 corresponding to SNPs already added to the candidate SNP list, yielded in the maximum correlation
323 coefficient with "target" similarity matrix. Then, the corresponding selected SNP was appended to
324 the candidate SNP list. We continued this process iteratively, seeking the next *cumulatively impactful*
325 SNP and appended it to the candidate SNP list sequentially until the number of SNPs in the candidate
326 SNP list reached the "*maximum desired number of prioritized SNPs"*. At the end, we derived a
327 "prioritized" SNP list by selecting the first 'K' SNPs from the candidate SNP list, whose
328 corresponding similarity matrices, when summed up, results in the maximum correlation coefficient
329 with the "target" similarity matrix. In other words, our method generated a "final" SNP similarity
330 matrix (sum of similarity matrices corresponding to the prioritized SNPs 'K') based on the allelic
331 proximities among accessions maximizing the correlation coefficient (Table 1 - Highest Correlation
332 Coefficient) with the "target" similarity matrix. Our algorithm is described below:

333 ***Input:***

334 - *Similarity matrices among 'a' unique accessions corresponding to 'N' no. of SNPs (S_j,*
335 *where j =1,2,…,N, and dimension of each S_j is a x a, and the cell values of S_j are*
336 *among three discrete values [0.0,0.5,1.0])*
337 - *"Target" similarity matrix among 'a' accessions ('T' of dimension a x a)*

- *Maximum desired number of prioritized SNPs 'n'*

Output:

- *Prioritized SNP list 'L_K', where 'K' is the no. of prioritized SNPs and K<=n*

Initialize:

- *Candidate SNP list 'L_n' =∅*

Algorithm:

- *Step 1: Compute Pearson correlation coefficient of each S_j with T, for j = 1,2,…,N.*
- *Step 2: Select S_{j^*} with highest correlation coefficient found in Step 1, where $j^* \in \{1,2,\ldots,N\}$ and append the corresponding SNP index j^* to L_n*
- *Step 3: Sum the similarity matrices corresponding to the SNP indices present in L_n and S_j together, and compute Pearson correlation coefficient with T, where $j \in \{1,2,\ldots,N\}$ and $j \notin Ln$*
- *Step 4: Select S_j from step 3 resulting in the highest correlation coefficient value and append the corresponding SNP index 'j' to L_n*
- *Step 5: Repeat steps 3 and 4 until the number of SNPs in L_n reaches the maximum desired number of prioritized SNPs 'n'.*
- *Step 6: Select the first K no. of SNPs from L_n , the summation of whose corresponding similarity matrices resulted in the maximum correlation coefficient value with 'T' and assign them to L_K*

We executed the above algorithm independently for each of the experimental conditions described in Table 1. In cases where the total number of SNPs exceeded 1000, the *maximum desired number of prioritized SNPs (n)* was set to 1000. We observed that this limit did not impact the prioritization of

360 informative SNPs as the highest correlation coefficient was consistently achieved at an index below
361 1000 for all experimental conditions. Figure 3 shows the trend of the correlation coefficient over the
362 iterations while adding SNPs to the candidate SNP list for the experiment based on maximum canopy
363 height with a *p*-value of .001 without QTL filtering. Further, our prioritized set of SNPs provides a
364 sort of ranking based on their influence on the phenotypic trait as they have been extracted through a
365 sequential algorithm. Hence, while doing analysis on these prioritized SNP sets (discussed later in
366 this paper), we used the term "Top k", which refers to the first 'k' SNPs in the prioritized SNP list.

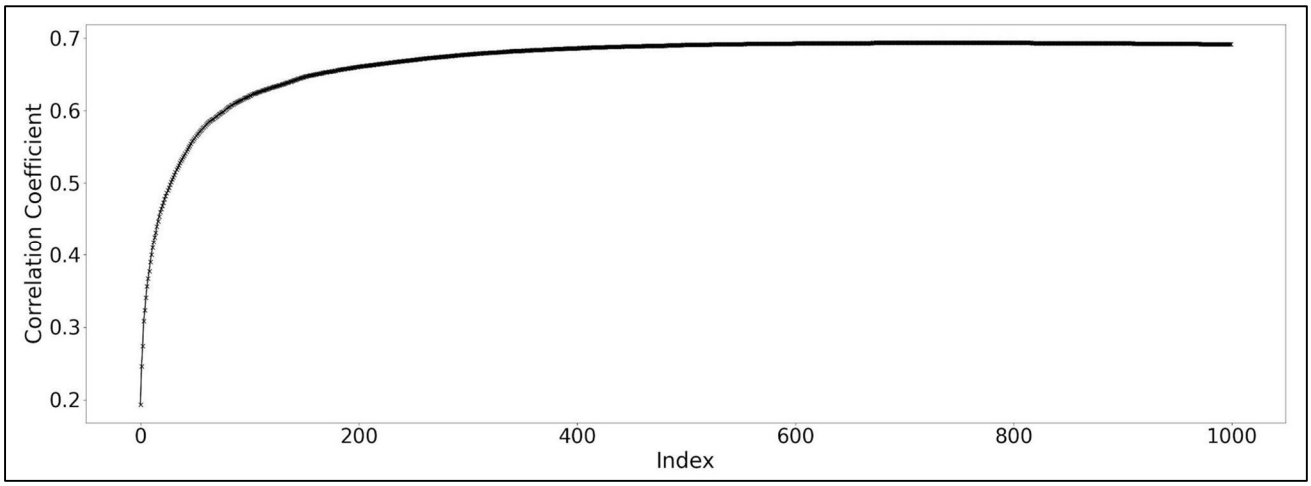

Figure 3. Trend of correlation coefficient while adding SNPs to the candidate SNP list based on their similarity matrices considering the phenotype maximum height with GWAS pre-filtering at p-value of .001.

367 **Prioritizing SNPs without GWAS Pre-filter:**

368 Additionally, we conducted an experiment on the entire set of SNPs without GWAS pre-filter to
369 compare the results with that of the GWAS-filtered sets of SNPs for maximum canopy height. As we
370 were unable to process the whole GVCF file at once due to the larger volume and limited
371 computational resources, 10 VCF files containing SNPs (Table 2 - Total #SNPs) corresponding to
372 each of the chromosomes were created (Supplementary File). The goal was to first process each VCF
373 file individually to prioritize SNPs corresponding to each chromosome with our algorithm (described

374 above). During these executions, we set the *maximum desired number of prioritized SNPs (n)* as
375 1000, similar to the executions with GWAS-filtered SNP set. Next, we merged the prioritized lists of
376 SNPs from each chromosome together and executed the same algorithm again on the combined set of
377 SNPs to obtain a final set of prioritized SNPs for the analysis. While applying the algorithm to the
378 *combined prioritized* set of SNPs from all chromosomes, we iterated through the entire list as the
379 total number of SNPs was already reduced during chromosome-wise execution, i.e., in this case, the
380 *maximum desired number of prioritized SNPs (n)* was set to the total number of SNPs in the set.

Table 2. Experimental Summary and Resulting SNP Lists using all Variants without GWAS pre-filter

Chromosome No.	Phenotype	Total #SNPs[1]	Prioritized #SNPs[2]	Prioritized #SNPs directly map to known genes	Highest Correlation Coefficient[3]
01	Maximum Height	106,824	595	259	0.64
02	Maximum Height	157,148	753	256	0.64
03	Maximum Height	132,661	656	216	0.64
04	Maximum Height	126,273	690	208	0.64
05	Maximum Height	157,930	758	198	0.64
06	Maximum Height	157,808	649	121	0.64
07	Maximum Height	119,573	635	136	0.64
08	Maximum Height	124,899	644	153	0.63
09	Maximum Height	124,843	679	145	0.63
10	Maximum Height	128,622	587	171	0.63
"*Combined prioritized*" SNPs from all chromosomes	Maximum Height	6646	987	247	0.71

[1]Total number of SNPs.
[2] Number of prioritized SNPs extracted.
[3] Pearson correlation coefficient of "target" similarity matrix with the summation of SNP similarity matrices corresponding to the "prioritized" SNPs.

381 **SNP Annotation and Function:**

382 The prioritized SNPs extracted by the above process for each experimental condition were annotated
383 with overlapping or nearest known gene identifiers using the SnpEff package (Cingolani et al., 2012).
384 A list of unique gene identifiers for each phenotype, p-value, GWAS filter, and QTL filter was

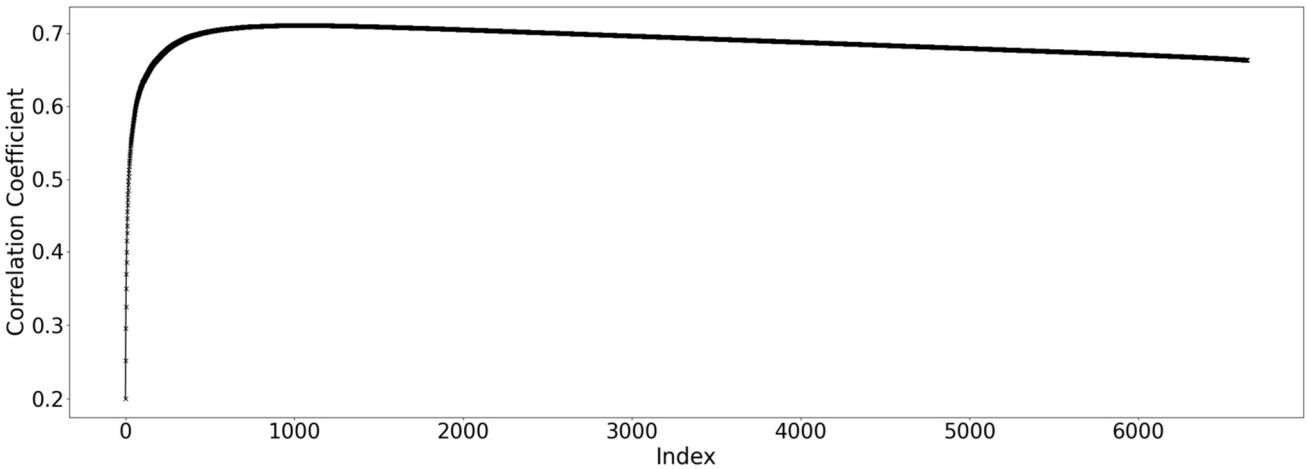

Figure 4. Trend of correlation coefficient while prioritizing SNPs based on their similarity matrices considering the phenotype maximum height without GWAS pre-filter.

385 submitted for Gene Ontology (GO) Enrichment Analysis using PANTHER 17.0 (Thomas et al.,
386 2022) and the Sorghum bicolor reference list (Mi et al., 2019). The lists of genes to be analyzed can
387 be found in GitHub (https://github.com/genophenoenvo/phenotype-to-
388 genotype/tree/main/significant_SNPs_annotated). We did separate enrichment analyses using the
389 complete set of GO Biological Process complete, Molecular Function, and Cellular Component
390 annotation data sets. The specific test performed was Fisher's exact test with FDR correction.

391 The genes identified that neighbor the "Top 10" SNPs (from SnpEff) for both phenotypes with the
392 GWAS filtering (with and without QTL filtering) at p=.0001 were further explored for their
393 functional role by using Gramene and the EMBL-EBI Gene Expression Atlas. This was also done for
394 the "Top 10" SNPs for the height phenotype without the GWAS filtering. InterPro protein domain
395 annotations (Mitchell et al., 2019) were obtained by querying Gramene BioMart (Tello-Ruiz et al.,
396 2018, Tello-Ruiz et al., 2021). The gene IDs were also used to query the EMBL-EBI Gene

397 Expression Atlas (Papatheodorou et al., 2018) to explore their baseline gene expression. The baseline
398 expression data of genes showing a default minimum expression level of 0.1 transcripts per million
399 (TPM) were downloaded for experiments E-MTAB-3839 (Olson et al., 2014), E-MTAB-4400
400 (Davidson et al., 2012), E-GEOD-98817 (Kebrom et al., 2017). Expression heatmaps were generated
401 using ClustVis (Beta) (Metsalu et al., 2015). The Sorghum QTL data was queried from the Sorghum
402 QTL Atlas (Mace et al. 2019) and mapped on Gramene's sorghum genome browser. The SNPs were
403 also queried using Gramene's variant effect prediction (VEP) tool (Naithani et al. 2017; McLaren et
404 al. 2016) to find any overlapping genes or existing SNPs in that database.

405 **Generalizability of Our Method:**

406 The generalizability of this method was tested by examining the predictive value of these prioritized
407 SNPs on the same accessions grown in a different environment. As part of this experiment, we
408 created a "target" similarity matrix based on the phenotypic observations (maximum plant height)
409 collected in Florence County, South Carolina (Clemson University; Brenton et al. 2016) using the
410 process described above. Thereafter, the Pearson correlation coefficient was calculated between this
411 "target" similarity matrix with the "final" SNP similarity matrix (i.e., the summation of similarity
412 matrices corresponding to the prioritized SNPs extracted) obtained for different experimental
413 conditions based on the maximum canopy height of MAC Season 6 dataset.

414 3. Results and Discussion

415 For both phenotypes, a smaller *p*-value resulted in fewer prioritized sets of SNPs, as expected (Table
416 1 - Prioritized #SNPs). Filtering by QTLs reduced the number of SNPs; however, it did not improve
417 the highest correlation coefficient achieved (Table 1 - Highest Correlation Coefficient). We found a
418 little overlap in prioritized SNPs between the two phenotypes and only at the two higher p values

419 (Figure 5). Further, most of the prioritized SNPs extracted without GWAS pre-filter were not

420 identified as informative compared to results obtained from GWAS pre-filter for maximum height

421 (Figure 5-D).

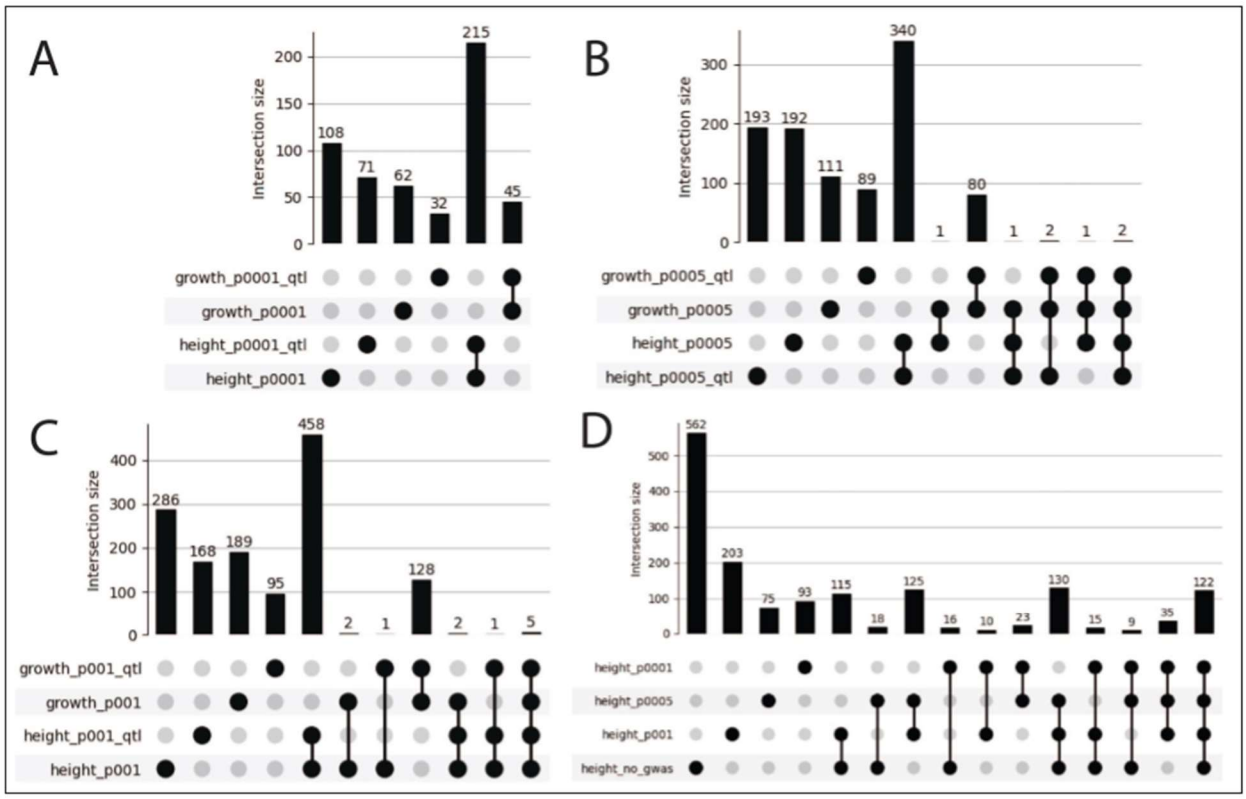

Figure 5. UpSet Plot showing the Overlap of Prioritized SNPs between the growth and height phenotypes. The intersection size is shown by vertical black bars (top). In the matrix, each column corresponds to a set (with the vertical black bar showing the size of the set) and each row (SNP list) corresponds to a possible intersection. The filled cells show which set is part of the intersection. Lists of prioritized SNPs for height with and without QTL filtering have high overlap, whereas the lists of prioritized SNPs for growth with and without QTL filtering have a moderate amount of overlap. At each p-value and for each phenotype, the QTL filtering added between 18% and 31% more unique SNPs. (Lex et al. 2014) Figure made with UpSetPlot v0.8.0 (https://github.com/jnothman/UpSetPlot)

(A). There is very little overlap of prioritized SNPs between the growth and height phenotype. At *p*0001 there is no overlap.

(B). At *p*0005 there are 7 SNPs in common between phenotypes.

(C). At *p*001 there are 11 SNPs in common.

(D). For the height phenotype, the majority of SNPs identified as important without the GWAS pre-filter had not been identified when using the GWAS pre-filter.

422 Few prioritized SNPs were mapped to known genes (Table 1 and Table 2), with the majority

423 mapping to intergenic regions. Enrichment analysis showed that 43 unique GO terms related to

424 biosynthetic processes, metabolic processes, and gene expression were over-represented in the

425 annotations of the genes associated with prioritized SNPs. The triterpenoid biosynthetic process

426 (GO:0016104) and triterpenoid metabolic process (GO:0006722) were the most over-enriched
427 (20.13-fold enrichment) across all the SNP lists and were located in the SNP lists for the growth
428 phenotype with GWAS filtering at p0.001 with and without QTL filtering. Nitrogen compound
429 metabolic process (GO:0006807) was one of the few under-enriched (0.72-fold enrichment) GO
430 annotations and was only present in SNP lists for the height phenotype at p0.0005 with and without
431 QTL filtering. The positive regulation of amide metabolic process (GO:0034250) was the only GO
432 term that was over-enriched without GWAS filtering, but not with GWAS filtering. There were more
433 overrepresented GO annotations for the height phenotype than for the growth phenotype - possibly
434 because we know more about genes for height. We did not see much difference between the
435 annotations enriched in prioritized SNP lists based on QTL filtering and without QTL filtering.

436 Of the "Top 40" prioritized SNPs extracted with GWAS pre-filter (both phenotypes with p0.0001),
437 all except two were in intergenic regions within 700 bp of a known transcript. One SNP,
438 S05_1688138, a deletion event (AATGTG→A) located on Chromosome-5 at 1688138 bp position,
439 was present in the intron of a gene, SORBI_3005G018800, that overlaps with QTLs known to impact
440 dry matter growth rate and plant height (Supplementary Figure 3). Another SNP, S06_45415791, lies
441 between two pre-miRNA (MIR2118) genes, ENSRNA04948199 and ENSRNA049481970, which
442 are conserved in plants and induce the phased small interfering RNAs (phasiRNAs) production
443 known to impact plant development and fertility (Supplementary Figure 4; Araki et al. 2020). A
444 closer look at the 85,709bp region around another intergenic SNP, S01_50208809, shows that it
445 overlaps with QTLs for phenotypes such as plant height, days to flowering, and tiller height. Gene
446 SORBI_3001G265700 flanks this SNP on the right and gene SORBI_3001G265600 on the left.
447 SORBI_3001G265700 encodes a serine-threonine-like kinase, and SORBI_3001G265600 encodes
448 an aldolase-1-epimerase like protein (SbA1ELP). The expression profile of the 51 genes that flank
449 and/or overlap these SNPs (according to SnpEff) showed baseline expression of 44 genes in plant

450 structures important for growth and height, such as the root, vegetative meristem, shoot, and stem
451 internodes (Supplementary Figure 5 (S5A-C); E-MTAB3839, E-MTAB-4400, and E-GEOD-98817).
452 Included in this list of 51 genes are the genes known to impact transcription factors, cytochrome
453 P450, sugar transporters, protein kinases, Leucine-rich Ser/Thr-kinase membrane proteins, and
454 proteins involved in regulating microtubule binding and assembly that are key to regulating mitosis
455 and growth during the cell cycle and thus height and growth. Most of these genes are expressed
456 across experiments (Supplementary Figure 5 (S5D)).

457 Of the "Top 10" prioritized SNPs obtained based on all variants without GWAS pre-filter, all but
458 three were in intergenic regions within 700 bp of a known transcript. The genes most closely
459 mapping to these prioritized SNPs were different from the genes identified with the GWAS pre-filter
460 (SORBI_3001G279900, SORBI_3002G119500, SORBI_3005G116400, SORBI_3001G262900,
461 SORBI_3002G153900, SORBI_3002G212200, SORBI_3006G040401, SORBI_3002G132000,
462 SORBI_3007G086800, SORBI_3002G119600, SORBI_3005G116500, SORBI_3007G115500,
463 SORBI_3001G262950, SORBI_3002G154000, SORBI_3006G040500, SORBI_3007G086900). All
464 these genes overlapped with known QTLs for plant height and most also overlapped with panicle
465 length and panicle emergence QTLs. These panicle traits may also regulate plant height in the
466 reproductive phase. The expression profile of these genes showed baseline expression in plant
467 structures important for growth and height, such as the root, vegetative meristem, shoot, and stem
468 internodes (Supplementary Figure 6 (S6A-C); E-MTAB-4400, E-MTAB-3839.and E-GEOD-98817).
469 Two of these genes (SORBI_3001G262950 and SORBI_3006G040401) did not show expression in
470 any of these studies. One gene (SORBI_3005G116400) showed expression in only one study (E-
471 MTAB-4400) where it was expressed in leaf tissues.

472 Priority SNPs were found near many of the genes known to affect height, but only without the
473 GWAS pre-filter and most of the prioritized SNPs for height did not map closely to any of the known
474 height genes in sorghum (Quinby and Karper, 1953; Figure 6A&B). With the GWAS pre-filter, the
475 only known height gene with a nearby SNP is *Ma1* (SOBIC_006G057866), which has three SNPs
476 nearby and is known to regulate flowering time in *Sorghum* (Quinby and Karper, 1945, Figure 6A). It
477 reflects the influence of flowering time on plant height. Without the GWAS pre-filter, prioritized
478 SNPs are identified near Ma5, Ma2, Ma1, Dw2, and Dw1 (Figure 6B). Recent work on *Sorghum*
479 grown in Ethiopia (Enyew et al. 2022) gives the seven most impactful SNPs on plant height, which
480 fall within known genes (Figure 6A&B). With the GWAS pre-filter, four of these are also near SNPs
481 we identified as significant for plant height (SOBIC.003G119600, SOBIC.006G111800,
482 SOBIC.010G085400, and SOBIC.010G066100), two are near known height genes (*DW1*,
483 SOBIC.009G223500 and *Ma5*, SOBIC.001G017500), and one is not near either
484 (SOBIC.006G235400) (Figure 6A). Without the GWAS pre-filter, an additional SNP identified in the
485 Ethiopian study is also near a SNP identified in this study (Figure 6B).

486 Based on the above data mining and literature review on known genes to support our findings, we
487 observed that our method is capable of prioritizing the informative genes influencing the trait of
488 interest leveraging GWAS, that can be utilized for the purpose of phenotype prediction. Further, the
489 experiment without GWAS pre-filter was able to locate certain SNPs, which are known to be
490 associated with the plant height phenotype. These SNPs were not retrieved after applying GWAS
491 with permissive-filtered thresholds. Though the experiment without GWAS pre-filter does not
492 explicitly account for the population structure, this analysis demonstrated the efficacy of our
493 approach in finding genes that may be predictive for the phenotype. Our method effectively reduced
494 the number of SNPs to be considered as input to ML/DL models, and population structure can be

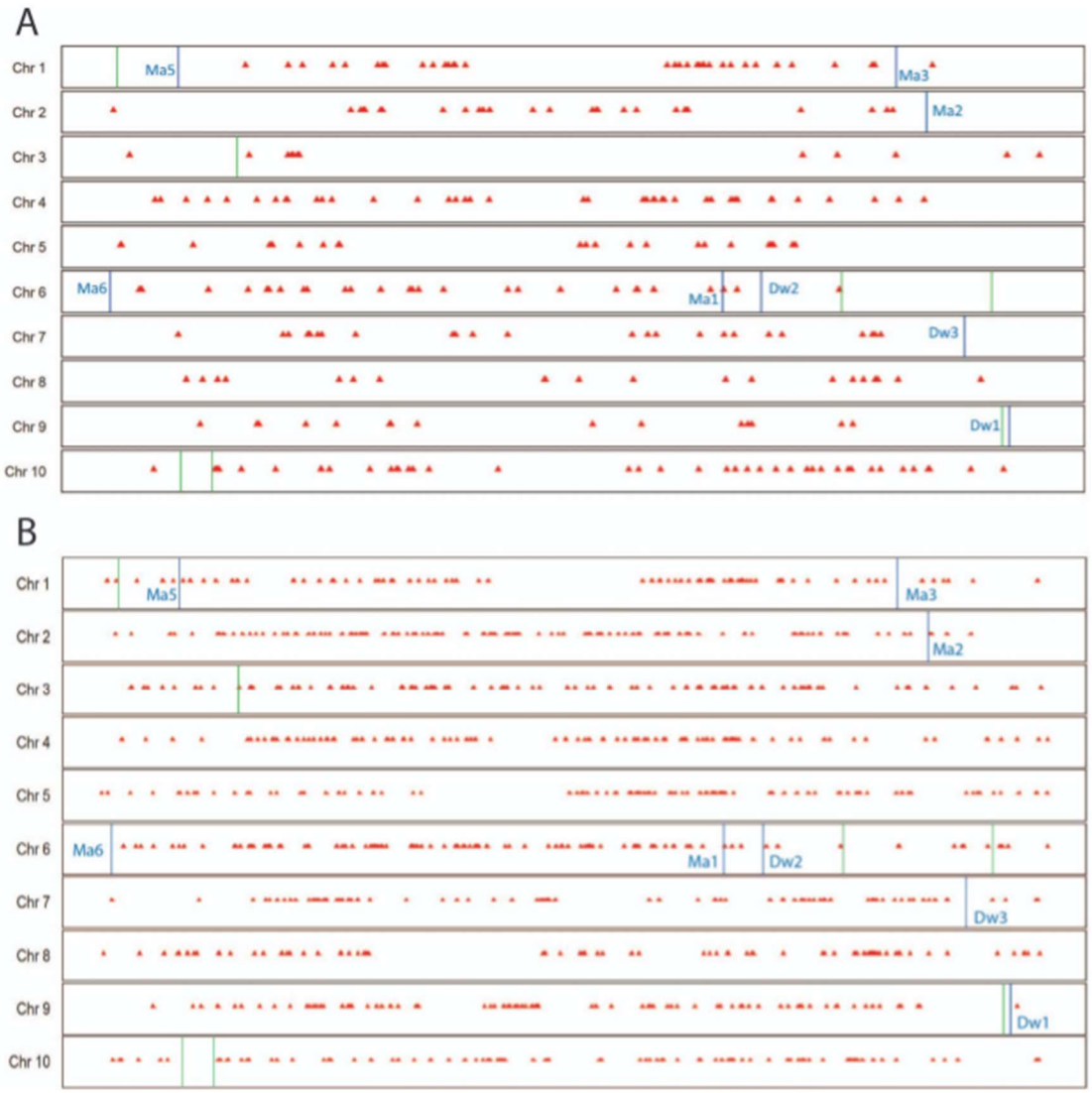

Figure 6. Chromosomal location of prioritized SNPs with known height genes. Prioritized SNPs (red triangles) are scattered throughout the ten chromosomes. The known plant height-associated genes are labeled in blue vertical lines. The green vertical lines show the location of SNPs important for height in sorghum grown in Ethiopia (Enyew et al. 2022)

(A) The 286 SNPs identified for maximum height based on GWAS-filtered set of SNPs with QTL filtering (p-value of .0001)

(B) The 987 SNPs identified for maximum height considering all variants without the GWAS pre-filter.

accounted for later by incorporating certain techniques like clustering and principal component analysis in the models (López-Cortés et al., 2020).

While the lists of prioritized SNPs explained about half of the variation in the phenotype in the TERRA-REF *Sorghum* data set (Table 1 and Table 2), they had very little explanatory power for the

499 same phenotypes in the same accessions grown in a very different environment. There were 271
500 accessions in common between the Clemson and TERRA-REF datasets, and the correlation
501 coefficient is found to be approximately 0.07 and 0.06, respectively, for all the cases in Table 1 and
502 Table 2 with the Clemson "target" similarity matrix while tested with the prioritized SNPs extracted
503 based on MAC Season 6 dataset.

504 4. Conclusions

505 This method is promising for prioritizing the SNPs and genes for further analysis in the domain of
506 functional genetics as well as identifying areas of the genome that are cumulatively informative for
507 predicting complex phenotypes using ML/DL methods. The highest correlation coefficient we found
508 was 0.69 with GWAS pre-filter (Table 1- Highest Correlation Coefficient) and 0.71 without GWAS
509 pre-filter (Table 2- Highest Correlation Coefficient), which are moderately strong and unsurprising
510 because of the influence of the environment on height and growth. We hypothesize that the total
511 number of SNPs that can determine height or growth is much larger than what was found in this
512 study, but depending on what environmental pressures the plant faces, different sub-groups of SNPs
513 will prove to be predictive. Our results from the South Carolina data set (Clemson) support this. In
514 addition, the partial overlap in important SNPs with the Ethiopia study could reflect the hot and arid
515 climate, which is more similar to Arizona than South Carolina despite using different accessions. A
516 logical next step is to find a data set from a similar environment using different accessions to see if
517 the significant SNPs in this study are similarly predictive.

518 In the absence of empirical data on gene function from sorghum, we can only conclude that this
519 method has the potential to identify specific areas of the genome that can help predict height and
520 growth phenotypes. We found gene homology and expression data supporting the importance of
521 some of the identified sorghum SNPs for the height and growth phenotypes. These SNPs are present

522 in the vicinity of gene homologs. For example, tobacco aldolase-1-epimerase (*NbA1ELP*) is known
523 to regulate pectin methylesterase (*NbPME*) antagonistically (Sheshukova et al. 2017). Lower
524 expression levels of tobacco *NbA1ELP* resulted in the dwarf plant phenotype. It gives us the
525 hypothesis that the sorghum ortholog of this tobacco gene (SORBI_3001G265600 *SbAiELP*) may be
526 associated with plant height and biomass yield in sorghum. Existing GO annotations do not include
527 these traits due to a lack of evidence from sorghum. However, this homologous sorghum gene is
528 expressed in embryo, shoot internode, root, emerging inflorescence, and stem internodes
529 (Supplementary Figure 5 (S5A-C)) suggesting that its expression in these tissues are likely
530 contributors for plant height and growth, that needs further validation by functional genomics studies.

531 Our original goal was to use biological knowledge graphs that link genes to phenotypes to help with
532 feature selection in genomic data. We quickly discovered that functional annotation of the *Sorghum*
533 genome is either insufficient or inferred from *Arabidopsis*, which tends to be grown in the laboratory
534 under standardized conditions. While genes are annotated as being useful for drought resistance or
535 salt tolerance, etc., there are no functional annotations of a genome that list a function in a specific
536 environmental context. This method could help add contextualized functional annotations to
537 genomes. These results have implications for using genotype-to-phenotype graphs in knowledge
538 guided machine learning (KGML) for phenotype prediction. Specifically, these graphs will need to
539 include genotype or SNP information (some biological graphs do include limited genotype
540 information) linked to the gene, and the functional annotations should include environmental context.
541 Functional annotations discovered in the laboratory can have limited use in the natural environment.

542 Additionally, we observed that the phenotypic dataset used in our experiment consists of accessions
543 having limited variation in their phenotypic measurements. Hence, our approach resulted in a "target"
544 similarity matrix where most of the accessions are highly correlated with each other considering the

phenotype under study. We hypothesize that the extraction of informative genes from phenotypes by our approach would be more prominent if accessions with larger variations in their phenotypic traits are selected.

Our method is currently limited in capturing linear relationships between genotypes and phenotypes, and the environmental variations were guarded by normalizing the inputs itself rather than explicitly incorporating in the procedure. In future, this algorithm could potentially be extended by using weighted correlation metrics to capture non-linearity as well as incorporating environmental variations explicitly in the procedure. In addition, we utilized a sequential algorithm for selecting the significant SNPs, wherein an SNP once identified as informative, cannot be deleted from the prioritized SNP list. In future, the process of prioritizing SNPs can be improved by employing other feature selection techniques, viz., Sequential Floating Forward Selection (SFFS). We believe that the work reported in this paper provides a novel path leveraging pattern recognition techniques for prioritization of informative SNPs that can be further studied and exploited to predict the plant phenotypes.

Acknowledgements

We thank the TERRA-REF project for sharing their high throughput phenomics and genomics data.

Funding

NSF Awards 1940059, 1940062, 1940330 Harnessing the Data Revolution

Author Contributions

Project PIs AR, AET, PJ, DL and AT conceived the idea. DL and JG collected, analyzed and assembled the TERRA-REF project data. DP, KS, AET, JG, PJ and AT analyzed the data. LC helped

566 in project coordination and CL helped in data visualization. All authors contributed to combined data
567 analysis, writing, and editing the manuscript. All authors approved the manuscript.

568 **Conflicts of Interest**

569 Author Curtis Lisle is employed by KnowledgeVis LLC. The remaining authors declare that the
570 research was conducted in the absence of any commercial or financial relationships that could be
571 construed as a potential conflict of interest.

572 **References**

573 Adak A, Conrad C, Chen Y, Wilde SC, Murray SC, Anderson II SL,Subramanian NK (2021)
574 'Validation of functional polymorphisms affecting maize plant height by unoccupied aerial systems
575 discovers novel temporal phenotypes', G3 Genes|Genomes|Genetics, 11(6), doi:
576 10.1093/g3journal/jkab075

577 Ahn E, Hu Z, Perumal R, Prom LK, Odvody G, Upadhyaya HD, Magill C (2019) 'Genome wide
578 association analysis of sorghum mini core lines regarding anthracnose, downy mildew, and head
579 smut', PLOS ONE, 14: e0216671, doi:10.1371/journal.pone.0216671

580 Araki S, Le NT, Koizumi K et al. (2020) 'miR2118-dependent U-rich phasiRNA production in rice
581 anther wall development', Nat Commun, 11, 3115, doi:10.1038/s41467-020-16637-3

582 Blears MJ, De Grandis SA, Lee H, Trevors JT (1998) 'Amplified fragment length polymorphism
583 (AFLP): a review of the procedure and its applications', J Ind Microbiol Biotech 21:99–114,
584 doi:10.1038/sj.jim.2900537

585 Boyles RE, Brenton ZW, Kresovich S (2019) 'Genetic and genomic resources of sorghum to connect
586 genotype with phenotype in contrasting environments', The Plant Journal 97: 19–39,
587 doi:10.1111/tpj.14113

588 Brenton ZW, Cooper EA, Myers MT, Boyles, RE, Shakoor N, Zielinski KJ, Rauh BL, Bridges WC,
589 Morris GP, Kresovich S (2016) 'A Genomic Resource for the Development, Improvement, and
590 Exploitation of Sorghum for Bioenergy', Genetics 204: 21–33, doi:10.1534/genetics.115.183947

591 Brown PJ, Rooney WL, Franks C, Kresovich S (2008) 'Efficient Mapping of Plant Height
592 Quantitative Trait Loci in a Sorghum Association Population With Introgressed Dwarfing Genes',
593 Genetics, 180: 629–637, doi:10.1534/genetics.108.092239

594 Budhlakoti N, Kushwaha AK, Rai A, Chaturvedi KK, Kumar A, Pradhan AK, Kumar U, et al. (2022)
595 'Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development
596 of Climate-Resilient Crops', Frontiers in Genetics, vol. 13, doi: 10.3389/fgene.2022.832153

597 Burnette M, Kooper R, Maloney JD, Rohde GS, Terstriep JA, Willis C, et al. (2018) 'TERRA-REF
598 Data Processing Infrastructure', Proceedings of the Practice and Experience on Advanced Research
599 Computing, USA, pp 1–7, doi: 10.1145/3219104.3219152

600 Burow G, Burke JJ, Xin Z, Franks CD (2011) 'Genetic dissection of early-season cold tolerance in
601 sorghum (Sorghum bicolor (L.) Moench)', Mol Breeding, 28:391-402, doi: 10.1007/s11032-010-
602 9491-4

603 Cao Y, Li S, Chen G, Wang Y, Bhat JA, Karikari B, Kong J, Gai J, Zhao T (2019) 'Deciphering the
604 Genetic Architecture of Plant Height in Soybean Using Two RIL Populations Sharing a Common
605 M8206 Parent', Plants, 8(10):373, doi: 10.3390/plants8100373

606 Casto AL, Mattison AJ, Olson SN, Thakran M, Rooney WL, Mullet JE (2019) 'Maturity2, a novel
607 regulator of flowering time in Sorghum bicolor, increases expression of SbPRR37 and SbCO in long
608 days delaying flowering', PLOS ONE, 14: e0212154, doi: 10.1371/journal.pone.0212154

609 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) 'A
610 program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs
611 in the genome of Drosophila melanogaster strain w1118', iso-2; iso-3. Fly, 6(2), 80–92, doi:
612 10.4161/fly.19695

613 Cooper EA, Brenton ZW, Flinn BS, Jenkins J, Shu S, Flowers D, Luo F, Wang Y, Xia, P., et al.
614 (2019) 'A new reference genome for Sorghum bicolor reveals high levels of sequence similarity
615 between sweet and grain genotypes: implications for the genetics of sugar metabolism', BMC
616 Genomics, 20: 420, doi: 10.1186/s12864-019-5734-x

617 Cooper L, Meier A, Laporte M-A, Elser JL, Mungall C, Sinn BT, Cavaliere D, et al. (2018) 'The
618 Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics',
619 Nucleic Acids Research, 46: D1168–D1180, doi: 10.1093/nar/gkx1152

620 Costa WG, Celeri MO, Barbosa IP, Silva GN, Azevedo CF, Borem A, Nascimento M, Cruz CD
621 (2022) 'Genomic prediction through machine learning and neural networks for traits with epistasis',
622 Computational and Structural Biotechnology Journal, 20:5490-5499, doi: 10.1016/j.csbj.2022.09.029

623 Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu S-H, Jiang N, Robin Buell C.
624 (2012) 'Comparative transcriptomics of three Poaceae species reveals patterns of gene expression
625 evolution', The Plant Journal, 71: 492-502, doi: 10.1111/j.1365-313X.2012.05005.x

626 Danilevicz MF, Gill M, Anderson R, Batley J, et al. (2022) Plant Genotype to Phenotype Prediction
627 Using Machine Learning. Front. Genet. 13:822173. doi: 10.3389/fgene.2022.822173

Endelman JB (2011). 'Ridge regression and other kernels for genomic selection with R package rrBLUP', Plant Genome, 4, 250-255

Enyew M, Feyissa T, Carlsson AS, Tesfaye K, Hammenhag C, Seyoum A, Geleta M. (2022) 'Genome-wide analyses using multi-locus models revealed marker-trait associations for major agronomic traits in Sorghum bicolor', Front. Plant Sci, p.3837, doi: 10.3389/fpls.2022.999692

Gelman A, Rubin DB (1992) 'Inference from iterative simulation using multiple sequences', Statistical science, 7(4):457-472, doi: 10.1214/ss/1177011136

Ghannam RB, Techtmann S (2021) 'Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring', Computational and Structural Biotechnology Journal, 19: 1092–1107, doi: 10.1016/j.csbj.2021.01.028

Gladman N, Jiao Y, Lee YK, Zhang L, Chopra R, Regulski M, Burow G, Hayes C, et al. (2019) 'Fertility of Pedicellate Spikelets in Sorghum Is Controlled by a Jasmonic Acid Regulatory Module', International Journal of Molecular Sciences, 20: 4951, doi: 10.3390/ijms20194951

Goodstein DM, Shu S, Howson R, Neupane R, et al. (2012) 'Phytozome: a comparative platform for green plant genomics', Nucleic Acids Research, 40: D1178–D1186, doi: 10.1093/nar/gkr944

Guo, J., LeBauer, D. (2022). 'genophenoenvo/JAGS-logistic-growth: v0.1.0', doi: 10.5281/zenodo.7114677

Hebbring SJ (2014) 'The challenges, advantages and future of phenome-wide association studies', Immunology, 141(2):157-65. doi: 10.1111/imm.12195

Hsieh WW (2022) 'Evolution of machine learning in environmental science—A perspective', Environmental Data Science, 1: e3, doi: 10.1017/eds.2022.2

649 Jiao Y, Burke J, Chopra R, Burow G, et al. (2016) 'A Sorghum Mutant Resource as an Efficient
650 Platform for Gene Discovery in Grasses', The Plant Cell, 28: 1551–1562, doi: 10.1105/tpc.16.00373

651 Jubair S, Domaratzki M (2023) 'Crop genomic selection with deep learning and environmental data:
652 A survey', Front. Artif. Intell, 5:1040295, doi: 10.3389/frai.2022.1040295

653 Kearsey MJ, Farquhar AGL (1998) 'QTL analysis in plants; where are we now?' Heredity 80: 137–
654 142, doi: 10.1046/j.1365-2540.1998.00500.x

655 Kebrom TH, McKinley B, Mullet JE (2017) 'Dynamics of gene expression during development and
656 expansion of vegetative stem internodes of bioenergy sorghum', Biotechnol Biofuels 10, 159 (2017).
657 doi: 10.1186/s13068-017-0848-3

658 LeBauer D, Maxwell B, Demieville J, Fahlgren N, French A, Garnett R, Hu Z, Huynh K, Kooper R,
659 Li Z, et al. (2020) 'Data From: TERRA-REF, An open reference data set from high resolution
660 genomics, phenomics, and imaging sensors', Dryad, Dataset, doi: 10.5061/dryad.4b8gtht99

661 Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H (2014) 'UpSet: Visualization of
662 Intersecting Sets', IEEE Transactions on Visualization and Computer Graphics (InfoVis), 20: 1983–
663 1992, doi: 10.1109/TVCG.2014.2346248

664 Lin J, Ngiam KY (2023) 'How data science and AI-based technologies impact genomics', Singapore
665 Med J.,64(1):59-66, doi: 10.4103/singaporemedj.SMJ-2021-438

666 Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) 'Iterative Usage of Fixed and Random Effect
667 Models for Powerful and Efficient Genome-Wide Association Studies', PLoS Genet 12(2):
668 e1005767, doi: 10.1371/journal.pgen.1005767

669 Lobos GA, Hancock JF (2015) 'Breeding blueberries for a changing global environment: a review',
670 Frontiers in Plant Science, doi: 10.3389/fpls.2015.00782

671 Lopes MS, El-Basyoni I, Baenziger PS, Singh S, Royo C, Ozbek K, Aktas H, Ozer E, Ozdemir F, et
672 al. (2015) 'Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate
673 change', Journal of Experimental Botany, 66: 3477–3486, doi: 10.1093/jxb/erv122

674 López-Cortés XA, Matamala F, Maldonado C, Mora-Poblete F, Scapim CA (2020) 'A Deep
675 Learning Approach to Population Structure Inference in Inbred Lines of Maize', Front. Genet., vol.
676 11, doi: 10.3389/fgene.2020.543459

677 Mace ES, Tai S, Gilding EK, Li Y, Prentis PJ, et al. (2013) 'Whole-genome sequencing reveals
678 untapped genetic potential in Africa's indigenous cereal crop sorghum', Nat Commun, 4: 2320, doi:
679 10.1038/ncomms3320

680 Mace E, Innes D, Hunt C, Wang X, Tao Y, Baxter J, et al. (2019) 'The Sorghum QTL Atlas: a
681 powerful tool for trait dissection, comparative genomics and crop improvement', TAG. Theoretical
682 and Applied Genetics, 132(3), 751–766, doi: 10.1007/s00122-018-3212-5

683 McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al (2018) 'The Sorghum
684 bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and
685 signatures of genome organization' The Plant Journal, 93: 338–354, doi: 10.1111/tpj.13781

686 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F (2016)
687 'The Ensembl Variant Effect Predictor', Genome Biology, 17: 122, doi: 10.1186/s13059-016-0974-4

688 McMaster GS, Wilhelm WW (1997) 'Growing degree-days: one equation, two interpretations',
689 Agricultural and Forest Meteorology, 87: 291–300, doi: 10.1016/S0168-1923(97)00027-0

690 Metsalu T, Vilo J (2015) 'ClustVis: a web tool for visualizing clustering of multivariate data using
691 Principal Component Analysis and heatmap', Nucleic Acids Research, Volume 43, Issue W1, W566–
692 W570, doi: 10.1093/nar/gkv468

693 Meuwissen THE, Hayes BJ, Goddard ME (2001) 'Prediction of Total Genetic Value Using Genome-
694 Wide Dense Marker Maps', Genetics, 157: 1819–1829, doi: 10.1093/genetics/157.4.1819

695 Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2019) 'PANTHER version 14: more
696 genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools', Nucleic Acids
697 Research, 47(D1), D419–D426, doi: 10.1093/nar/gky1038

698 Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, et al. (2019)
699 'InterPro in 2019: improving coverage, classification and access to protein sequence annotations',
700 Nucleic Acids Research, 47(D1), D351–D360, doi: 10.1093/nar/gky1100

701 Moore JH, Asselbergs FW, Williams SM (2010), 'Bioinformatics challenges for genome-wide
702 association studies', Bioinformatics, 26(4):445-55. doi: 10.1093/bioinformatics/btp713

703 Naithani S, Cecilia HD, Sahu SK, Jaiswal P (2023) "Exploring Pan-Genomes: An Overview of
704 Resources and Tools for Unraveling Structure, Function, and Evolution of Crop Genes and
705 Genomes", Biomolecules 13(9): 1403, doi: 10.3390/biom13091403

706 Naithani S, Geniza M, Jaiswal P (2017) 'Variant Effect Prediction Analysis Using Resources
707 Available at Gramene Database', Methods Mol Biol.,1533:279–297, doi: 10.1007/978-1-4939-6658-
708 5_17

Olson A, Klein RR, Dugas DV, Lu Z, Regulski M, Klein PE, Ware D (2014) 'Expanding and vetting sorghum bicolor gene annotations through transcriptome and methylome sequencing', The Plant Genome, 7(2), lantgenome2013.08.0025.

Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. (2018) 'Expression Atlas: gene and protein expression across multiple studies and organisms', Nucleic Acids Research, Volume 46, Issue D1, D246–D251, doi: 10.1093/nar/gkx1158

Plummer M (2003) 'JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling', Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, 1-10

Plummer M (2022) 'rjags: Bayesian Graphical Models using MCMC. R package version 4-13', url: https://CRAN.R-project.org/package=rjags

Pudil P, Novovičová J, Kittler J (1994) 'Floating search methods in feature selection', Pattern Recognition Letters, Vol. 15, Issue 11, pp. 1119-1125, doi: 10.1016/0167-8655(94)90127-9

Quinby J, Karper R (1945) 'The inheritance of three genes that influence time of floral initiation and maturity date in milo', J Am Soc Agron, 37:916-936, doi: 10.2134/agronj1945.00021962003700110006x

Quinby J, Karper R (1953) 'Inheritance of height in sorghum. Inherit. height. Sorghum', Abstracts of the Annual Meetings of the American Society of Agronomy, Dallas, Texas, pp. 98–99

R Core Team (2021) 'R: A language and environment for statistical computing. R Foundation for Statistical Computing', Vienna, Austria. url: https://www.R-project.org/

729 Rafalski A (2002) 'Applications of single nucleotide polymorphisms in crop genetics', Current
730 Opinion in Plant Biology, 5:94-100, doi: 10.1016/s1369-5266(02)00240-6

731 Rhodes DH, Hoffmann L, Rooney WL, Ramu P, Morris GP, Kresovich S (2014) 'Genome-Wide
732 Association Study of Grain Polyphenol Concentrations in Global Sorghum [Sorghum bicolor (L.)
733 Moench] Germplasm', J Agric Food Chem, 62(45):10916–10927, doi: 10.1021/jf503651t

734 Rice B, Lipka AE (2019), Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak
735 Genome-Wide Association Study Signals in Maize and Sorghum. The Plant Genome, 12: 180052.
736 doi: 10.3835/plantgenome2018.07.0052

737 Shavrukov YN (2016) 'CAPS markers in plant biology', Russ J Genet Appl Res, 6: 279–287, doi:
738 10.1134/S2079059716030114

739 Sheshukova EV, Komarova TV, Pozdyshev DV, Ershova NM, Shindyapina AV, et al. (2017) 'The
740 Intergenic Interplay between Aldose 1-Epimerase-Like Protein and Pectin Methylesterase in Abiotic
741 and Biotic Stress Control', Front Plant Sci. 8:1646. doi: 10.3389/fpls.2017.01646

742 Sukumaran S, Xiang W, Bean SR, Pedersen JF, Kresovich S, Tuinstra MR, Tesso TT, Hamblin MT,
743 Yu J (2012) 'Association Mapping for Grain Quality in a Diverse Sorghum Collection', The Plant
744 Genome. doi: 10.3835/plantgenome2012.07.0016

745 Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) 'RFLP Mapping in Plant Breeding:
746 New Tools for an Old Science', Nat Biotechnol, 7: 257–264, doi: 10.1038/nbt0389-257

747 Tao Y, Zhao X, Wang X, Hathorn A, Hunt C, Cruickshank AW, Oosterom EJ, Godwin ID, Mace ES,
748 Jordan DR (2020) 'Large-scale GWAS in sorghum reveals common genetic control of grain size
749 among cereals', Plant Biotechnology Journal, 18: 1093–1105, doi: 10.1111/pbi.13284

750 Tao Y, Luo H, Xu J, Cruickshank A, Zhao X, Teng F, Hathorn A, Wu X, Liu Y, Shatte T, et al.
751 (2021) 'Extensive variation within the pan-genome of cultivated and wild sorghum', Nat Plants, 7:
752 766–773, doi: 10.1038/s41477-021-00925-x

753 Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, Wei S, Preece J, et al. (2018),
754 'Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic
755 Acids Research, 46(D1), D1181–D1189, doi: 10.1093/nar/gkx1111

756 Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, Preece J, Jiao Y, Wang B, et al. (2021)
757 'Gramene 2021: harnessing the power of comparative genomics and pathways for plant research',
758 Nucleic Acids Research, Volume 49, Issue D1, D1452–D1463, doi: 10.1093/nar/gkaa979

759 The Gene Ontology Consortium (2021) 'The Gene Ontology resource: enriching a GOld mine',
760 Nucleic Acids Research, 49: D325–D334, doi: 10.1093/nar/gkaa1113

761 Thessen A (2016) 'Adoption of Machine Learning Techniques in Ecology and Earth Science', One
762 Ecosystem, 1: e8621, doi: 10.3897/oneeco.1.e8621

763 Thomas PD, Ebert D, Muruganujan A, et al.(2022) 'PANTHER: Making genome-scale
764 phylogenetics accessible to all', Protein Science, 31: 8–22, doi: 10.1002/pro.4218

765 Tong H, Nikoloski Z (2021) 'Machine learning approaches for crop improvement: Leveraging
766 phenotypic and genotypic big data', Journal of Plant Physiology, vol. 257, doi:
767 10.1016/j.jplph.2020.153354.

768 Visscher PM, Brown MA, McCarthy MI, Yang J (2012) 'Five Years of GWAS Discovery', The
769 American Journal of Human Genetics, 90: 7–24, doi: 10.1016/j.ajhg.2011.11.029

770 Watanabe K, Guo W, Arai K, Takanashi H, Kajiya-Kanegae H, Kobayashi M, et al. (2017) 'High-
771 Throughput Phenotyping of Sorghum Plant Height Using an Unmanned Aerial Vehicle and Its
772 Application to Genomic Prediction Modeling', Front. Plant Sci., 8:421, doi: 10.3389/fpls.2017.00421

773 Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. (2021) 'rMVP: A Memory-efficient,
774 Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study',
775 Genomics, Proteomics & Bioinformatics, 19(4):619-628, doi: 10.1016/j.gpb.2020.10.007.

776 Zhang Q, Zhang Q, Jensen J (2022), Association Studies and Genomic Prediction for Genetic
777 Improvements in Agriculture, Front. Plant Sci.,vol. 13, doi: 10.3389/fpls.2022.904230

778 Zheng Z, Hey S, Jubery T, Liu H, et al. (2020) 'Shared Genetic Control of Root System Architecture
779 between Zea mays and Sorghum bicolor', Plant Physiology, 182: 977–991, doi: 10.1104/pp.19.00752

780 Zhu C, Gore M, Buckler ES, Yu J (2008) 'Status and Prospects of Association Mapping in Plants',
781 The Plant Genome, doi: 10.3835/plantgenome2008.02.0089