International Journal of
Molecular Sciences

MDPI

Article

Machine-Learning-Based Genome-Wide Association Studies for Uncovering QTL Underlying Soybean Yield and Its Components

Mohsen Yoosefzadeh-Najafabadi [1,†], Milad Eskandari [1,*,†], Sepideh Torabi [1], Davoud Torkamaneh [2], Dan Tulpan [3] and Istvan Rajcan [1]

[1] Department of Plant Agriculture, University of Guelph, Guelph, ON N1G 2W1, Canada; myoosefz@uoguelph.ca (M.Y.-N.); storabi@uoguelph.ca (S.T.); irajcan@uoguelph.ca (I.R.)
[2] Département de Phytologie, Université Laval, Québec City, QC G1V 0A6, Canada; davoud.torkamaneh.1@ulaval.ca
[3] Department of Animal Biosciences, University of Guelph, Guelph, ON N1G 2W1, Canada; dtulpan@uoguelph.ca
* Correspondence: meskanda@uoguelph.ca
† These authors contributed equally to this work.

Abstract: A genome-wide association study (GWAS) is currently one of the most recommended approaches for discovering marker-trait associations (MTAs) for complex traits in plant species. Insufficient statistical power is a limiting factor, especially in narrow genetic basis species, that conventional GWAS methods are suffering from. Using sophisticated mathematical methods such as machine learning (ML) algorithms may address this issue and advance the implication of this valuable genetic method in applied plant-breeding programs. In this study, we evaluated the potential use of two ML algorithms, support-vector machine (SVR) and random forest (RF), in a GWAS and compared them with two conventional methods of mixed linear models (MLM) and fixed and random model circulating probability unification (FarmCPU), for identifying MTAs for soybean-yield components. In this study, important soybean-yield component traits, including the number of reproductive nodes (RNP), non-reproductive nodes (NRNP), total nodes (NP), and total pods (PP) per plant along with yield and maturity, were assessed using a panel of 227 soybean genotypes evaluated at two locations over two years (four environments). Using the SVR-mediated GWAS method, we were able to discover MTAs colocalized with previously reported quantitative trait loci (QTL) with potential causal effects on the target traits, supported by the functional annotation of candidate gene analyses. This study demonstrated the potential benefit of using sophisticated mathematical approaches, such as SVR, in a GWAS to complement conventional GWAS methods for identifying MTAs that can improve the efficiency of genomic-based soybean-breeding programs.

Keywords: data-driven models; FarmCPU; genome-wide association study; MLM; QTL; soybean breeding; support-vector machine

Citation: Yoosefzadeh-Najafabadi, M.; Eskandari, M.; Torabi, S.; Torkamaneh, D.; Tulpan, D.; Rajcan, I. Machine-Learning-Based Genome-Wide Association Studies for Uncovering QTL Underlying Soybean Yield and Its Components. *Int. J. Mol. Sci.* **2022**, *23*, 5538. https://doi.org/10.3390/ijms23105538

Academic Editor: Frank M. You

Received: 5 March 2022
Accepted: 13 May 2022
Published: 16 May 2022

1. Introduction

Soybean (*Glycine max* [L.] Merr.) is known as one of the most important legume crops worldwide with substantial economic value [1]. Despite the importance of genetic improvements in soybean yield, the germplasm has, in general, a narrow genetic basis, especially within North America, which has resulted in the limited progress of genetic gains for seed yield [2]. Therefore, there is a pronounced need for analytical breeding to explore the optimum genetic potential for enhancing yield in soybean [3,4].

An analytical breeding strategy, as an alternate breeding approach, requires a better understanding of the factors, or individual traits, responsible for more complex characteristics such as plant growth, development, and yield [5]. This strategy considers secondary

traits, which are highly correlated with the trait of interest, as the selection criteria to make empirical selections more efficient toward improving the genetic gain [2,5,6]. The yield potential in soybean is mainly determined by its components, such as the total number of pods, seeds, and nodes per plant, as well as seed size [6–8]. Of these traits, the total number of nodes and pods plays a more important role in the final seed yield production [8,9]. Several studies reported a steady increase in the total number of nodes and the total number of pods in soybean cultivars from 1920 to 2010 [2,3,10]. These findings highlight the importance and potential use of the phenotypic and genotypic information in these traits, along with yield per se, as selection criteria in cultivar development programs [10]. The application of analytical approaches to plant-breeding programs has been limited, mainly due to the limited resources available for evaluating several secondary traits that are mostly time- and labor-consuming [5,11]. Most of the analytical breeding studies were conducted on small populations with a limited number of genotypes, and, therefore, the results have limited generalization and limitations in terms of the knowledge of the genome-to-phenome analysis process [9,10,12].

The genetic information of soybean-yield component traits can accelerate the efficiency of cultivar development programs through selecting genotypes with improved genetic gains [13]. Genome-wide association studies (GWASs), as one of the most common genetic approaches, can be implemented on genetically diverse populations to detect the marker-trait associations (MTAs) for soybean-yield components [12]. Up to date, several GWAS approaches, such as mixed linear models (MLM), the multiple loci linear mixed model (MLMM), and fixed and random model circulating probability unification (FarmCPU), have been developed for genetic studies of complex traits [12]. However, due to the narrow genetic base of some plant species, including soybean, these conventional approaches may not have sufficient statistical power to detect reliable MTAs [2,14,15]. Therefore, the development of more sophisticated statistical methods can help to establish effective GWAS methods for plant species with narrow genetic bases.

Machine learning (ML) algorithms, as powerful and reliable mathematical methods, have been considered as an alternative to conventional statistical methods in GWAS analyses [2,16]. Recently, the use of ML algorithms has been reported in different areas such as plant science [14,15,17,18], animal science [19], human science [20], engineering [21], and computer science [22]. The application of ML algorithms in a GWAS was previously investigated in a human-science study by Szymczak, et al. [23], in which different ML algorithms such as artificial neural networks (ANN), Bayesian network analysis (BNA), and random forests (RF) were elucidated for use in GWAS studies focused on human disease studies. One of the most commonly used ML algorithms is RF, developed by Breiman [24], which generates a series of trees from the independent samples and selects the best trees for increasing the prediction performance [25]. The latter algorithm has been widely used in plant genomics [26], phenomics [14], proteomics [27], and metabolomics [28]. The support-vector machine (SVM) is another common algorithm that can detect the behavior and patterns of nonlinear relationships [29–31]. Theoretically, SVM should have high performance due to the use of structural risk-minimization, instead of empirical risk-minimization, inductive principles [32]. There are a significant number of reports on the successful use of SVM in prediction problems [19,33–36]. Support-vector regression (SVR) is known as the regression version of SVM, which is commonly used for continuous variables. There are also reports on the successful use of SVR for addressing plant-prediction problems [37].

In this study, we aimed to (1) gain a better understanding of the genetic relationships between soybean yield and its component traits, and (2) investigate the potential use of RF and SVM algorithms in a GWAS for discovering MTAs for soybean-yield components in comparison with the most commonly used conventional GWAS methods. The results of this study may shed light on the potential use of ML algorithms in soybean GWAS studies and may offer new genomic tools for screening high-yielding genotypes with improved genetic gain in large breeding populations.

2. Results

2.1. Phenotyping Evaluations

The panel consisted of 227 soybean genotypes showing different levels of variations among the genotypes for seed yield, maturity, and yield component traits. The distribution of the phenotypic measures for the target traits across the four environments is presented in Figure 1. The highest heritability was observed for maturity (0.78), followed by NP, RNP, NRNP, and PP, with estimated values of 0.34, 0.33, 0.31, and 0.30, respectively (Figure 1). The lowest heritability value of 0.24 was estimated for yield (Figure 1). The average ± standard deviation values for maturity, yield, NP, NRNP, RNP, and PP in the tested GWAS panel were 106 ± 5 days, 3.5 ± 0.45 t ha^{-1}, 15.21 ± 0.77 nodes, 3.33 ± 0.28 nodes, 11.89 ± 0.98 nodes, and 45.02 ± 8.54 pods, respectively (Figure S1, see Supplementary Materials). The linear correlations (*r*) among all the measured traits were estimated using the Pearson coefficients of correlation (Figure 2). All the traits were found to be positively correlated with each other, except NRNP, which was negatively associated with yield, maturity, RNP, NP, and PP. NP showed the highest correlation with the RNP ($r = 0.97$) and the NRNP ($r = -0.63$). RNP had the highest correlation with yield ($r = 0.86$) among all the tested yield components (Figure 2).

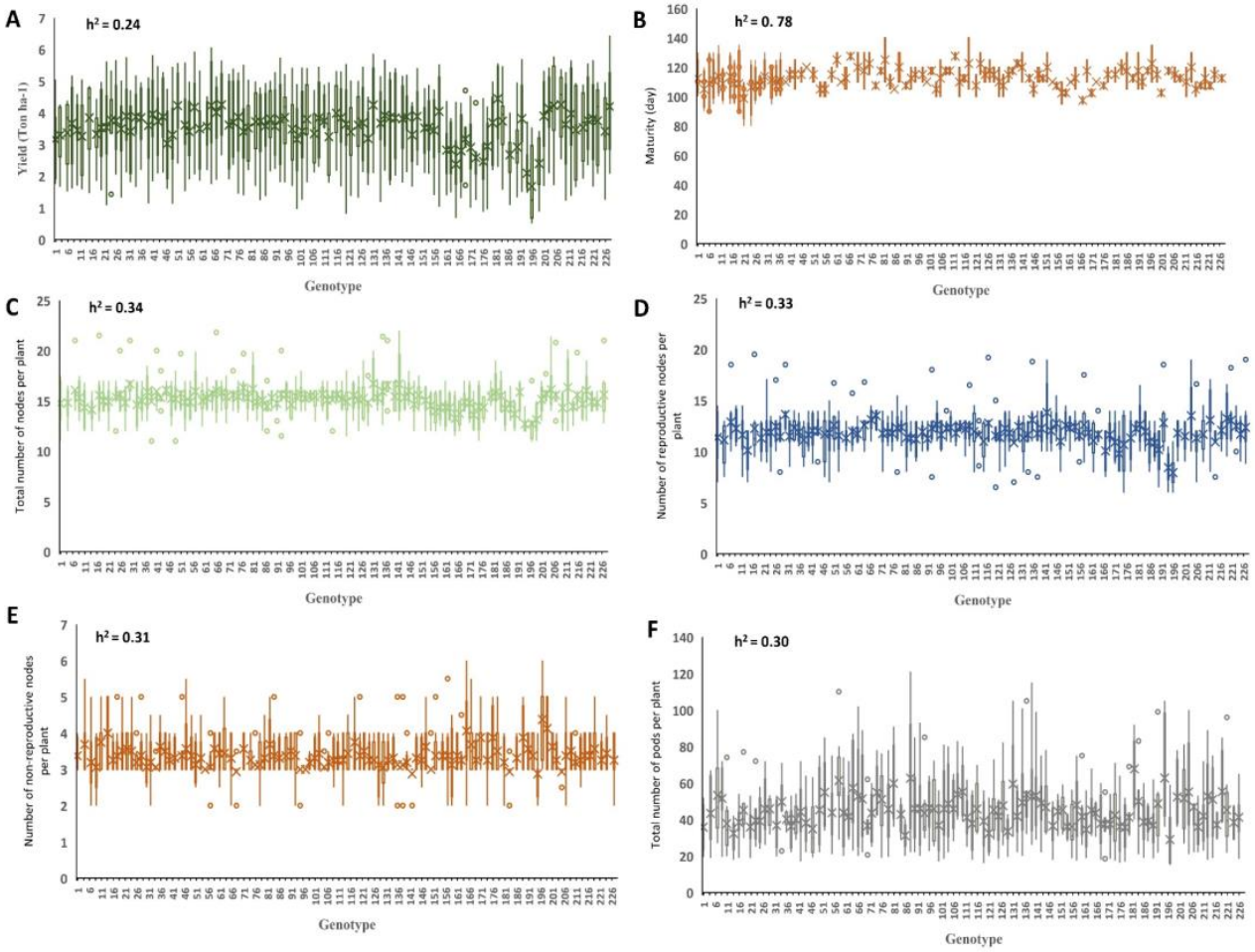

Figure 1. The distribution of seed yield (**A**), maturity (**B**), NP (**C**), NRNP (**D**), RNP (**E**), and PP (**F**) in 227 soybean genotypes across four environments. The estimated heritability is provided for each of the six traits. RNP: the total number of reproductive nodes per plant, NRNP: the total number of non-reproductive nodes per plant, NP: the total nodes per plant, and PP: the total number of pods per plant.

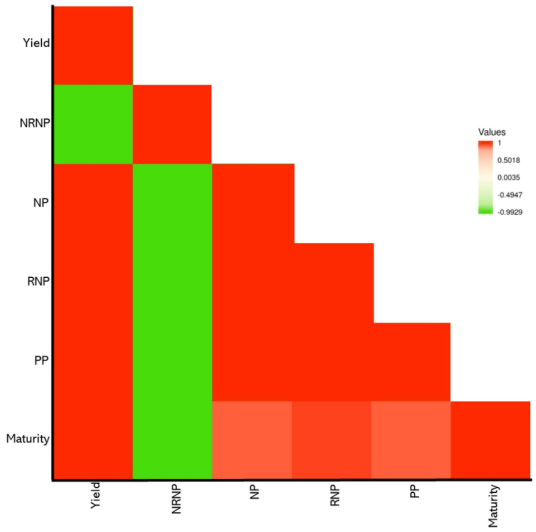

Figure 2. The Pearson correlations among the soybean seed yield, maturity, and yield component traits. RNP: the total number of reproductive nodes per plant, NRNP: the total number of non-reproductive nodes per plant, NP: the total nodes per plant, and PP: the total number of pods per plant. The heat map scale for values is provided by color for the panel.

2.2. Population Structure and Kinship

The structure and kinship profile for the tested population is presented in Figure 3. The result of genotypic evaluations suggested that the tested GWAS panel was composed of four to seven subpopulations. Therefore, we chose to conduct the structure analysis using K = 7 as the appropriate K for the structure profile of the tested GWAS panel (Figure 3).

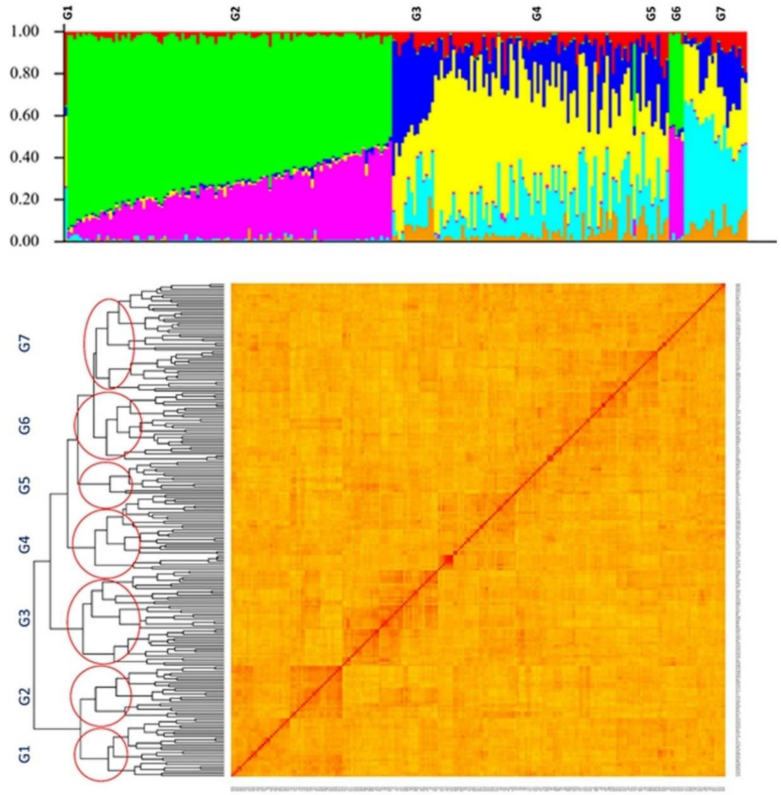

Figure 3. The structure and kinship plots for the 227 soybean genotypes. The x-axis is the number of genotypes used in this GWAS panel, and the y-axis is the membership of each subgroup. G1–G7 stands for the subpopulation.

2.3. GWAS Analysis

The average performance of the tested GWAS methods was compared in Figure S2 (see Supplementary Materials). The association analysis using the MLM method resulted in the identification of nine SNP markers, located on chromosomes 2 and 19, associated with maturity (Table S1). Using FarmCPU resulted in a total of nine maturity-associated SNP markers located on chromosomes 2, 19, and 20 (Figure 4A), of which eight SNPs were also detected by MLM. By using the RF method, a total of three SNP markers on chromosomes 3, 16, and 17 were identified to be associated with this trait, whereas SVR-mediated GWAS detected 12 SNP markers located on chromosomes 2, 6, 10, 16, 19, and 20 (Table S2). For soybean maturity, 3 out of 12 detected MTAs by SVR-mediate GWAS were colocalized with previously reported QTL related to the reproductive period and R8 full maturity (Table 1 and Figure S3). Most of the detected MTAs using MLM and FarmCPU methods were colocalized with previously reported QTL associated with soybean seed weight and Sclerotinia stem rot (Table 1 and Figure S3).

Table 1. The list of MTAs associated with the maturity date, identified using different GWAS methods in this study, which are colocalized with previously reported QTL.

GWAS Method	Chromosome	Peak SNP Position	Co-Located QTL	Reference
MLM	2	2212910	Sclero 3-g31	[38]
		8233782	Seed Weight 6-g1	[39]
FarmCPU	2	2212910	Sclero 3-g31	[38]
		8233766	Seed Weight 6-g1	[39]
	20	37765851	WUE 2-g53	[40]
RF	3	2978272	Leaflet area 1-g2.1	[41]
			Leaflet width 1-g4.1	[41]
			Leaflet area 1-g2.2	[41]
			Leaflet width 1-g4.2	[41]
			Salt tolerance 1-g12	[42]
	16	5730281	Plant height 6-g17	[43]
			Plant height 1-g17	[43]
			First flower 4-g63	[44]
	17	34757372	SDS root retention 1-g6	[45]
SVR	2	695362	Seed linolenic 2-g1	[46]
			Seed linolenic 2-g2	[46]
		720134	SDS 1-g12.1	[47]
			SDS 1-g12.2	[47]
			Ureide content 1-g2	[48]
		827374	SDS 1-g12.3	[47]
	10	1595239	Shoot Cu 1-g8	[49]
		1689395	Seed oil 5-g3	[39]
	16	2438652	Reproductive period 4-g16	[43]
			R8 full maturity 9-g2	[43]
		2460921	Reproductive period 2-g16	[43]
			R8 full maturity 2-g2	[43]
	19	47513536	R8 full maturity 4-g1	[39]
		47513572	First flower 4-g81	[44]

MLM: mixed linear model; FarmCPU: fixed and random model circulating probability unification; RF: random forest; and SVR: support-vector regression.

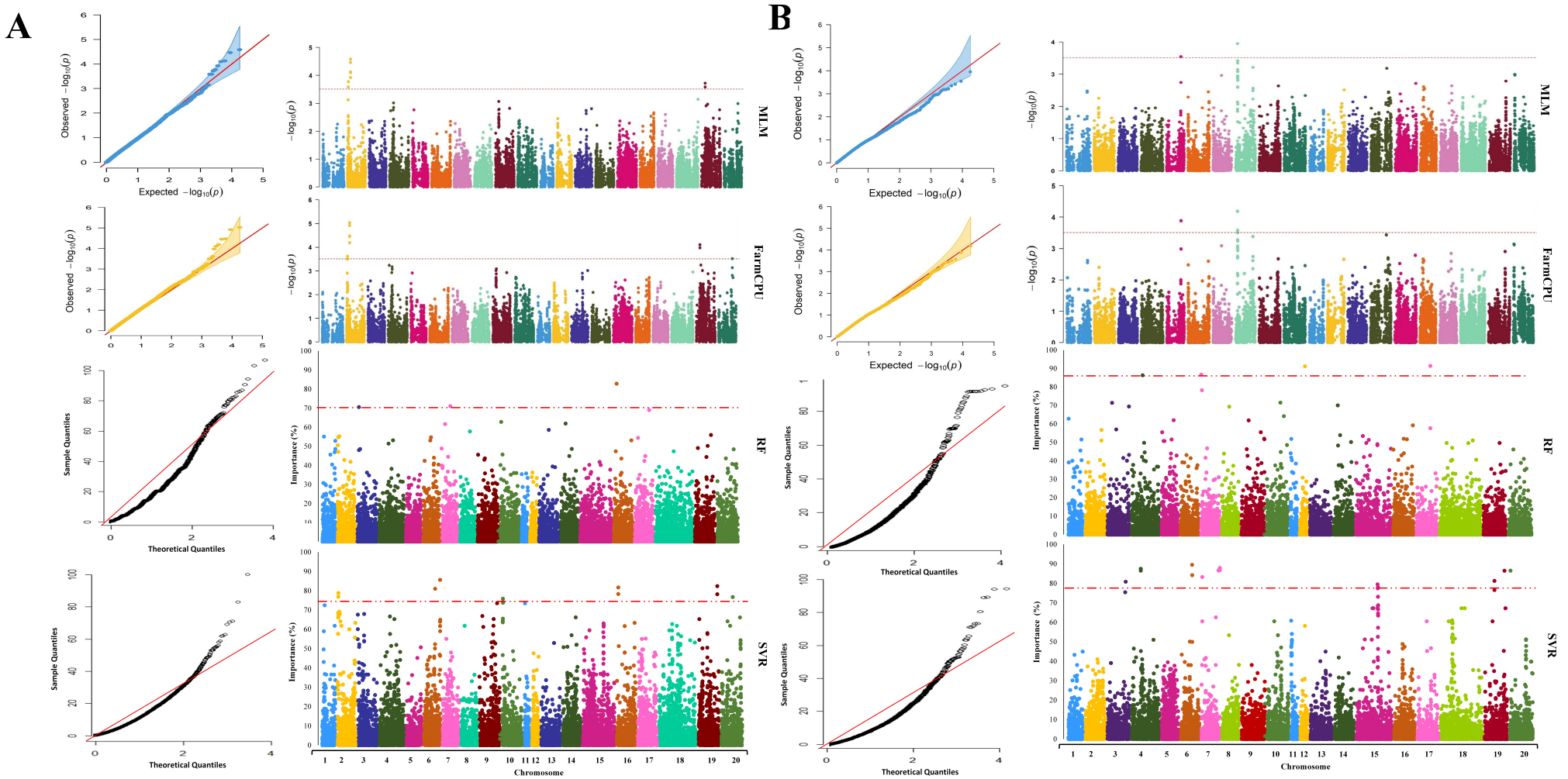

Figure 4. The genome-wide Manhattan and quantile–quantile plots for GWASs of (**A**) maturity and (**B**) seed yield in soybean using MLM, FarmCPU, RF, and SVR methods, from top to bottom, respectively.

Using the MLM, FarmCPU, RF, and SVR approaches, we identified 2, 3, 5, and 18 SNP markers associated with yield, respectively (Tables S3 and S4). The SNP markers identified by MLM and FarmCPU were located on chromosomes 5 and 8. The markers identified through RF were located on chromosomes 4, 7, 12, and 17. The identified markers using the SVR-mediated GWAS method were located on chromosomes 3, 4, 6, 7, 15, 19, and 20 (Figure 4B). In SVR-mediated GWASs, MTAs were colocalized with eight previously reported yield-related QTL such as seed yield, seed weight, and seed set (Table 2 and Figure S3). However, other tested GWAS methods could not find MTAs colocalized with any previously reported QTL associated with seed yield except for ureide content and water-use efficiency (Table 2 and Figure S3).

Table 2. The list of MTAs associated with seed yield, identified using different GWAS methods in this study that are colocalized with previously reported QTL.

GWAS Method	Chromosome	Peak SNP Position	Co-Located QTL	Reference
MLM	5	34391386	Ureide content 1-g16.1	[48]
			Ureide content 1-g16.2	[48]
FarmCPU	5	34391386	Ureide content 1-g16.1	[48]
			Ureide content 1-g16.2	[48]
RF	7	1032587	WUE 2-g18	[40]
SVR	3	36309302	First flower 4-g10	[44]
			First flower 3-g2	[50]
			Seed weight 4-g3	[50]
			Seed yield 4-g2	[50]
			R8 full maturity 3-g3	[50]
		37617293	Plant height 3-g17	[51]
			Leaflet shape 1-g1.1	[41]
			Leaflet shape 1-g1.2	[41]
			Leaflet shape 1-g1.3	[41]
			Seed set 1-g32.1	[41]
			Seed set 1-g32.2	[41]
	7	44488152	Seed yield 4-g4	[50]
		1032587	WUE 2-g18	[40]
	15	34958361	SCN 5-g35	[52]
	19	41385139	Seed weight 5-g20	[53]
			Seed weight 4-g18	[50]
			Seed yield 4-g5	[50]
			Shoot Zn 1-g28.1	[49]
			Shoot Zn 1-g28.2	[49]
			Shoot Zn 1-g29.1	[49]
			Shoot Zn 1-g29.2	[49]
			Shoot Zn 1-g29.3	[49]

MLM: mixed linear model; FarmCPU: fixed and random model circulating probability unification; RF: random forest; and SVR: support-vector regression.

Using the MLM and FarmCPU methods, we respectively detected one and two SNP markers associated with NP (Table S5). Five and ten SNP markers were associated with NP when RF and SVR methods were used, respectively (Table S6). Most of the MTAs detected by MLM and FarmCPU were colocalized with previously reported QTL related to seed set, seed weight, seed long-chain fatty acid, and pubescence density (Table 3). SVR-mediated GWASs identified MTAs colocalized with three previously reported NP-related QTL (Table 3 and Figure S3). A total of 2, 3, 5, and 10 SNP markers were determined to be associated with NRNP using the MLM, FarmCPU, RF, and SVR methods, respectively (Tables S7 and S8). Chromosome numbers 4, 8, and 15 were identified as carrying SNP

markers associated with NRNP using FarmCPU, and the MLM method identified SNP markers located on chromosomes 8 and 15. The detected SNP markers using the SVR method were located on chromosomes 4, 7, 18, 19, and 20, whereas SNP markers identified through RF were located on chromosomes 1, 4, 7, 18, and 19 (Figure 5B). Most of the identified MTAs for NRNP using all GWAS methods were colocalized with previously reported QTL related to seed weight, seed protein, water-use efficiency, first flower, and soybean cyst nematode (Table 4 and Figure S3).

Table 3. The list of MTAs associated with the total number of nodes per plant (NP), identified using different GWAS methods in this study that are colocalized with previously reported QTL.

GWAS Method	Chromosome	Peak SNP Position	Co-Located QTL	Reference
FarmCPU	19	40131952	Pubescence density 1-g17	[54]
			Seed weight 9-g5.1	[55]
RF	4	1205787	Shoot Ca 1-g10	[49]
	6	50570624	Seed set 1-g51.1	[41]
			Seed set 1-g43.1	[41]
			Seed set 1-g25.1	[41]
			Seed set 1-g43.2	[41]
			Seed set 1-g25.2	[41]
			Seed set 1-g51.2	[41]
		50570473	Seed set 1-g43.3	[41]
			Seed set 1-g51.3	[41]
			Seed set 1-g25.3	[41]
			Pod number 1-g3	[41]
			Seed palmitic 2-g2	[41]
			Seed long-chain faty acid 1-g22	[41]
SVR	6	50570624	Seed set 1-g51.1	[41]
			Seed set 1-g43.1	[41]
			Seed set 1-g25.1	[41]
			Seed set 1-g43.2	[41]
			Seed set 1-g25.2	[41]
			Seed set 1-g51.2	[41]
		50570473	Seed set 1-g43.3	[41]
			Seed set 1-g51.3	[41]
			Seed set 1-g25.3	[41]
			Pod number 1-g3	[41]
			Seed palmitic 2-g2	[41]
			Seed long-chain faty acid 1-g22	[41]
	7	1032587	WUE 2-g18	[40]
		1092403	WUE 2-g18	[40]
			First flower 3-g4	[41]
	18	55645699	Leaflet shape 1-g4.1	[41]
			Leaflet shape 1-g4.2	[41]
			Leaflet shape 1-g4.3	[41]
			Seed stearic 4-g5	[56]
			Node number 1-g6.1	[41]
			Node number 1-g6.2	[41]
			Pod number 1-g1.1	[41]
			Pod number 1-g1.2	[41]
			Pode number 1-g1.3	[41]
			WUE 3-g31	[40]
			Seed weight, SoyNAM 14-g28	[57]
			Lodging, SoyNAM 4-g15	[58]
			Branching 1-g1.1	[41]
			Plant height 5-g4.2	[41]
			Plant height 5-g4.3	[41]
			Shoot p 1-g30	[49]
	19	47350110	Node number 1-g2.3	[41]

MLM: mixed linear model; FarmCPU: fixed and random model circulating probability unification; RF: random forest; and SVR: support-vector regression.

Table 4. The list of MTAs associated with the total number of non-reproductive nodes per plant (NRNP), identified using different GWAS methods in this study that are colocalized with previously reported QTL.

GWAS Method	Chromosome	Peak SNP Position	Co-Located QTL	Reference
MLM	15	10193796	Seed protein 6-g2 Seed Arg 1-g4 Seed coat luster 1-g1.3	[59] [59] [41]
FarmCPU	15	10193796	Seed protein 6-g2 Seed Arg 1-g4 Seed coat luster 1-g1.3	[59] [59] [41]
RF	1	54647498	First flower 4-g2	[44]
	7	329800	Phytoph 2-g32 Phytoph 2-g7	[60] [60]
	18	12945778	SCN 4-g14	[61]
	19	40218800	Seed weight 9-g5.1	[55]
SVR	7	1032587	WUE 2-g18	[40]
	19	40218800	Seed weight 9-g5.1	[55]

MLM: mixed linear model; FarmCPU: fixed and random model circulating probability unification; RF: random forest; and SVR: support-vector regression.

Using the MLM and FarmCPU methods, four SNP markers located on chromosomes 8 and 19 were associated with RNP (Table S9). Using the RF method, four associated SNP markers were identified on chromosomes 8, 9, 15, and 20 (Table S10). Using the SVR method, 11 SNP markers were detected associated with RNP, located on chromosomes 4, 7, 8, 15, 18, 19, and 20 (Figure 6A). Regardless of the type of GWAS methods used in this study, we found SNP markers associated with the trait on chromosome 8. The position of the associated SNP marker on chromosome 8 was identical using all GWAS methods (~450 Kbp). The list of detected QTL for RNP is presented in Table 5 and Figure S3.

Table 5. The list of MTAs associated with the total number of reproductive nodes per plant (RNP), identified using different GWAS methods in this study that are colocalized with previously reported QTL.

GWAS Method	Chromosome	Peak SNP Position	Co-Located QTL	Reference
RF	9	40285014	Shoot Fe 1-g8.1 Shoot Fe 1-g8.2 Shoot Fe 1-g8.3 Shoot Fe 1-g9 Shoot Fe 1-g10 Shoot Fe 1-g11 Soybean mosaic virus 2-g5	[49] [49] [49] [49] [49] [49] [62]
	15	34958361	SCN 5-g35	[52]
SVR	7	1032587	WUE 2-g18	[40]
	15	34958361	SCN 5-g35	[52]

MLM: mixed linear model; FarmCPU: fixed and random model circulating probability unification; RF: random forest; and SVR: support-vector regression.

We did not detect any SNP marker associated with PP using the MLM or FarmCPU methods. However, by using the RF method, four SNP markers located on chromosomes 7, 10, 18, and 20 were found to be associated with PP (Table S11). Twelve SNP markers were detected to be associated with PP using SVR. The markers were located on chromosomes 6, 9, 10, 11, 15, 18, and 19 (Figure 6B). The associated SNP markers in chromosome 10 were identified in both RF and SVR with a 4.6 cM distance from each other. Most of the MTAs detected by SVR-mediated GWASs were colocalized with seven previously reported QTL directly related to the pod number (Table 6 and Figure S3).

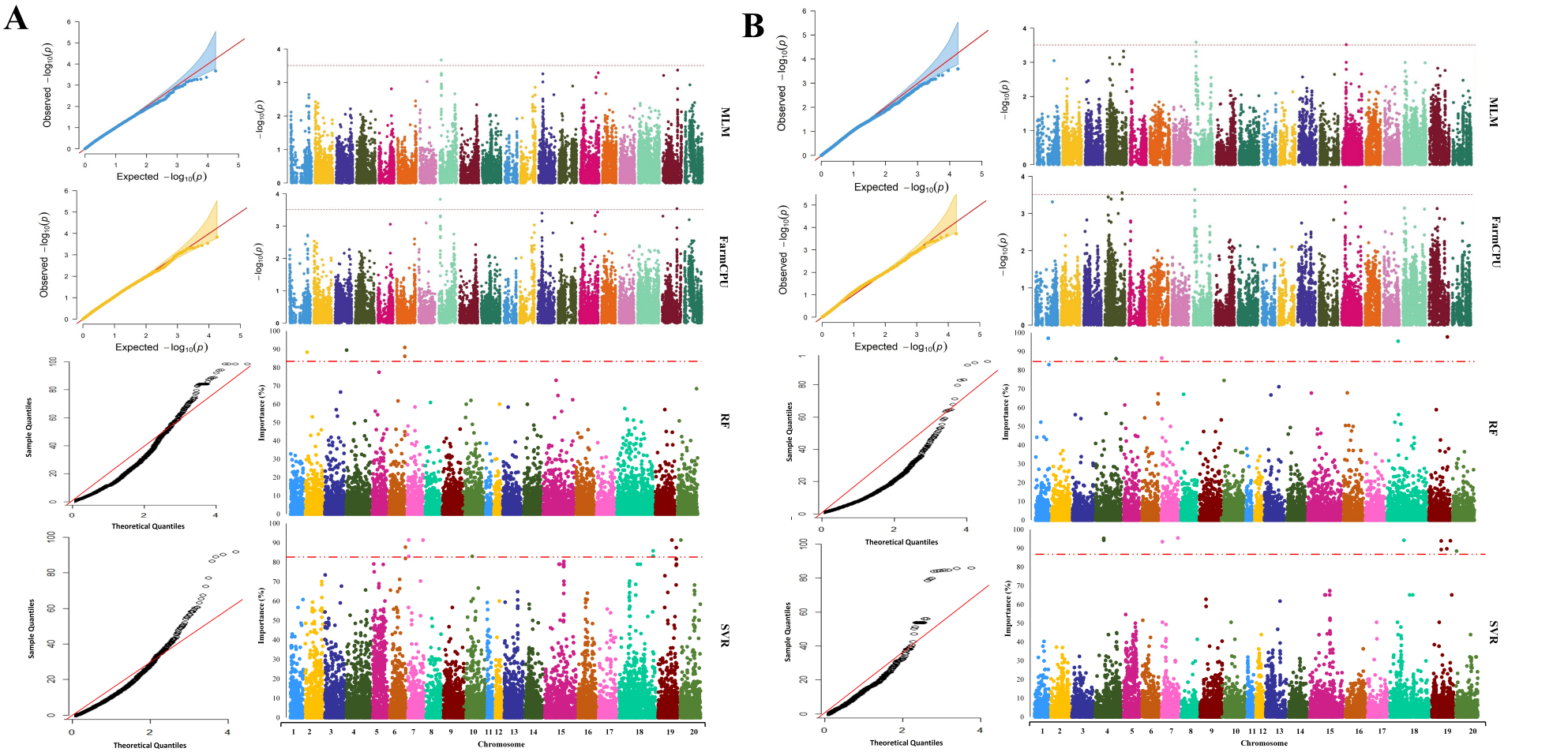

Figure 5. Genome-wide Manhattan and quantile–quantile plots for GWASs of (**A**) the total number of nodes (NP) and (**B**) the total number of non-reproductive nodes (NRNP) in soybean using MLM, FarmCPU, RF, and SVR methods, from top to bottom, respectively.

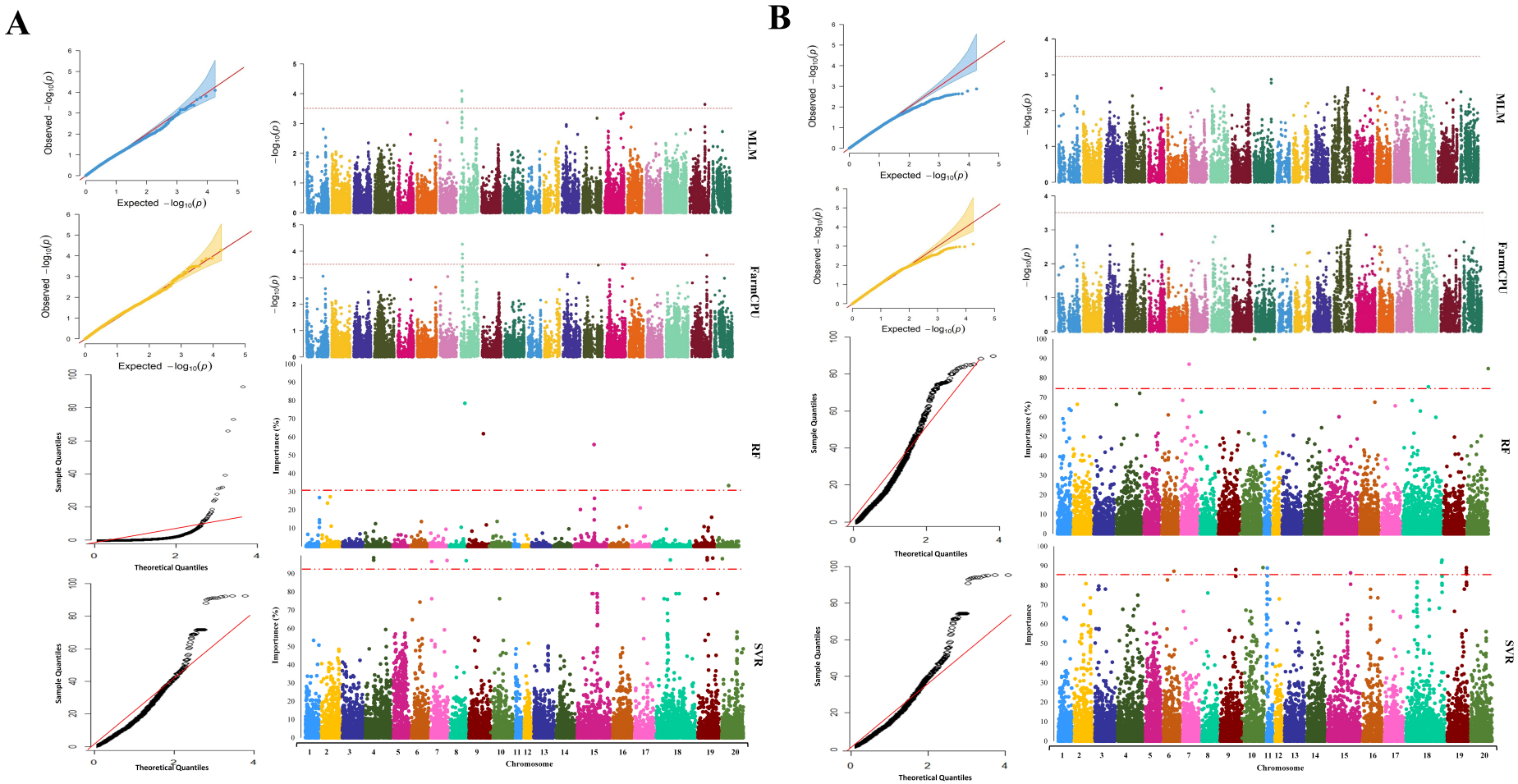

Figure 6. Genome-wide Manhattan and quantile–quantile plots for GWASs of (**A**) the total number of reproductive nodes (RNP) and (**B**) the total number of pods (PP) in soybean using MLM, FarmCPU, RF, and SVR methods, from top to bottom, respectively.

Table 6. The list of MTAs associated with the total number of pods per plant (PP), identified using different GWAS methods in this study that are colocalized with previously reported QTL.

GWAS Method	Chromosome	Peak SNP Position	Co-Located QTL	Reference
RF	7	15331676	Seed weight, SoyNAM 14-g11	[57]
SVR	9	39366957	Pod number 1-g4.1	[41]
			Pod number 1-g4.2	[41]
			Pod number 1-g4.3	[41]
			Seed thickness 2-g4	[41]
	9	39372117	Seed Thr 2-g1	[63]
			Seed Ser 2-g1	[63]
			Seed Tyr 2-g2	[63]
			Seed Lys 2-g2	[63]
			Seed leu 2-g2	[63]
			Seed ile 2-g2	[63]
			Seed Ala 2-g2	[63]
			Seed Gly 2-g2	[63]
	11	5245870	Ureide content 1-g29	[48]
			Pod number 1-g6	[41]
	18	55645699	Leaflet shape 1-g4.1	[41]
		55469601	Leaflet shape 1-g4.2	[41]
			Leaflet shape 1-g4.3	[41]
			Seed stearic 4-g5	[56]
			Node number 1-g6.1	[41]
			Node number 1-g6.2	[41]
			Pode number 1-g1.1	[41]
			Pode number 1-g1.2	[41]
			Pode number 1-g1.3	[41]
			WUE 3-g31	[64]
			Seed weight, SoyNAM 14-g28	[57]
			Lodging, SoyNAM 4-g15	[58]
			Branching 1-g1.1	[41]
			Plant height 5-g4.2	[41]
			Plant height 5-g4.3	[41]
			Shoot p 1-g30	[49]
			Seed yield, SoyNAM 7-g19	[58]
			R8 full maturity, SoyNAM 13-g19	[58]
			Plant height 5-g4.3	[41]
	19	43077182	Seed weight 9-g5.2	[55]
			Seed weight 5-g21	[55]
			First flower 5-g3	[41]
			First flower 5-g17	[41]
		47235604	First flower 4-g77	[44]
			Seed palmitic 1-g19	[65]
		47350110	Leaf carotenoid content 1-g14	[66]
			Ureide content 1-g50.3	[48]
			Ureide content 1-g50.4	[48]
		47224293	Node number 1-g2.3	[41]

MLM: mixed linear model; FarmCPU: fixed and random model circulating probability unification; RF: random forest; and SVR: support-vector regression.

2.4. Extracting Candidate Genes Undelaying Detected QTL

To identify the potential candidate genes of each of the detected MTAs, we used the LD decay distance of the panel and selected 150-kbp upstream and downstream of each SNP's peak as the target regions (Figure 7). The full description of identified candidate

genes is presented in Table S12. The effect of each of the identified peak SNPs in explaining the variance of the tested traits is provided in Figure 8. For soybean maturity, three peak SNPs (Chr2_695362, Chr2_720134, and Chr19_47513536) had the highest allelic effects than other detected peak SNPs (Figure 8A). On the basis of the gene annotation and expression within the QTL, *Glyma.02g006500* (GO:0015996) and *Glyma.19g224200* (GO:0010201), which, respectively, encode the chlorophyll catabolic process and phytochrome A (PHYA)-related genes, were identified as the strong candidate genes for maturity. *Glyma.02g006500* (GO:0015996) was exactly detected in the peak SNP position of Chr2_695362, whereas *Glyma.19g224200* (GO:0010201) was 119 Kbp from the detected peak SNP at Chr19_47513536. The yield-related QTL with the peak SNP positioned on Chr7_1032587 had the highest allelic effect compared to other detected peak SNPs (Figure 8B). Within 77 Kbp away from the detected peak SNP (Chr7_1032587), *Glyma.07G014100* (GO:0010817), which encodes the regulation of hormone levels, was identified as the strongest candidate gene in yield. Two peak SNPs, Chr7_1032587 and Chr7_1092403, had the highest allelic effects for the NP trait among all the detected peak SNPs (Figure 8C). In this study, the Chr7_1032587 SNP was associated with yield, NP, and NRNP. The *Glyma.07G205500* (GO:0009693) and *Glyma.08G065300* (GO:0042546) genes, which encode UBP1-associated protein 2C and cell-wall biogenesis, respectively, were detected as plausible genes influencing both NP and NRNP. Both detected candidate genes were collocated at the corresponding peak SNPs at Chr7_1032587 and Chr8_5005929 (Figure 8D). Regarding peak SNPs associated with RNP, the highest allelic effects were found in the peak SNPs of Chr9_40285014 and Chr15_34958361 (Figure 8E). The *Glyma.15G214600* (GO:0009920) and *Glyma.15G214700* (GO:0009910) genes, which encode cell plate formation involved in plant-type cell-wall biogenesis and acetyl-CoA biosynthetic process, respectively, were nominated as strong candidate genes governing NRNP. *Glyma.15G214600* (GO:0009920) and *Glyma.15G214700* (GO:0009910) were 127 and 90 Kbp far from the peak SNP at Chr15_3495836, respectively. For the PP trait, the highest allelic effects were found in peak SNPs at Chr7_15331676, Chr11_5245870, and Chr18_55469601 (Figure 8F). The *Glyma.07G128100* (GO:0009909) gene, which encodes the regulation of flower development, was the strongest candidate gene that can potentially affect PP. *Glyma.07G128100* (GO:0009909) is located in the peak SNP position, Chr7_15331676.

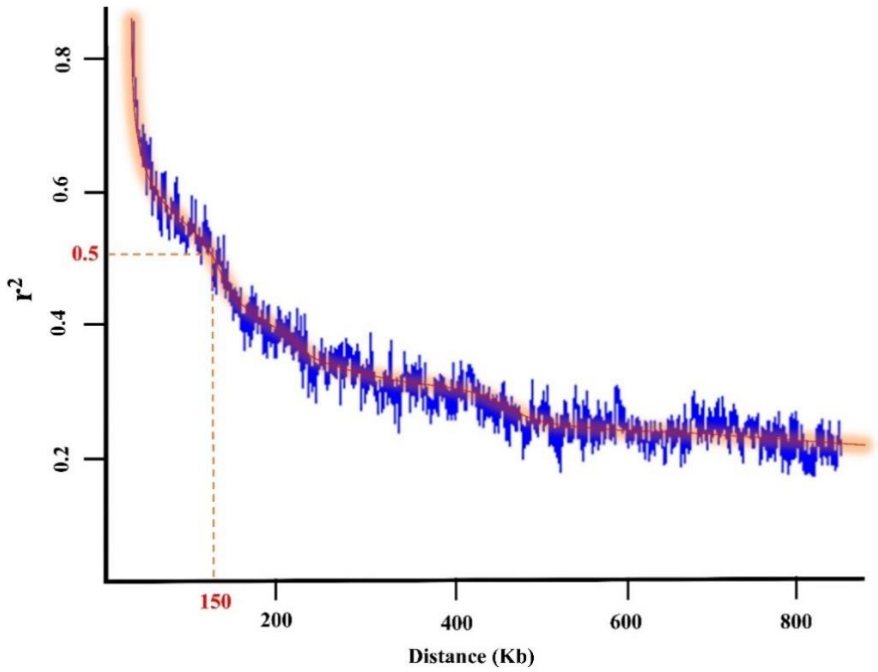

Figure 7. The LD decay distance in the tested 227 soybean genotypes.

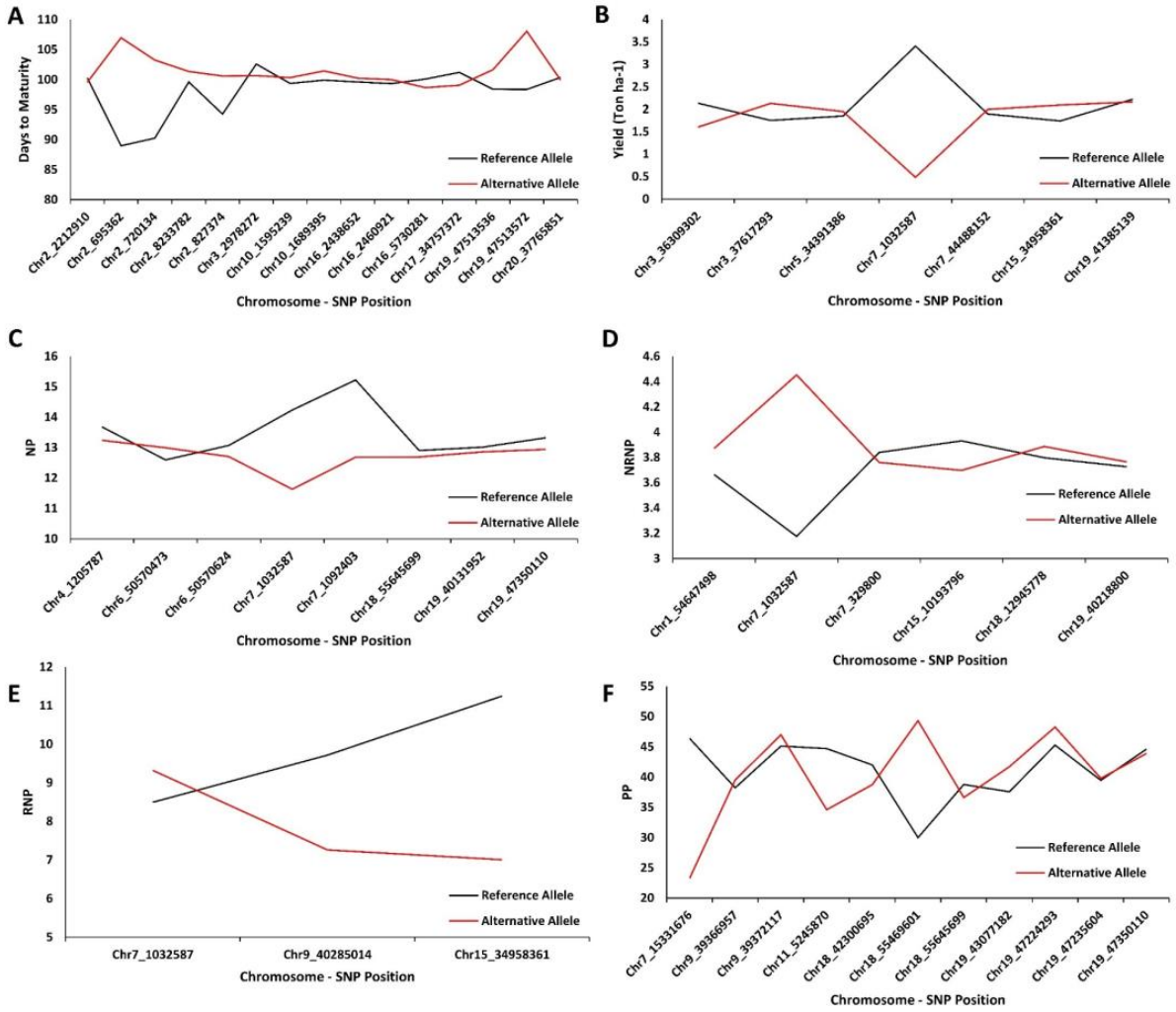

Figure 8. The average effects of the reference allele and the alternative allele from the detected SNP's peak for maturity (**A**), seed yield (**B**), NP (**C**), NRNP (**D**), RNP (**E**), and PP (**F**) in 227 soybean genotypes across 4 environments. RNP: the total number of reproductive nodes per plant; NRNP: the total number of non-reproductive nodes per plant; NP: the total nodes per plant; and PP: the total number of pods per plant.

3. Discussion

One of the objectives of this study was to attain a better understanding of the roles of soybean-yield component traits in the production of total seed yield and how these traits can be used to facilitate the development of high-yielding soybeans with improved genetic gains. The genetic dissection of soybean-yield components and establishing genetic and genomics toolkits can be used for designing crosses and screening large breeding populations for selecting genotypes with improved yield components [67,68]. The results of this study showed high phenotypic variations for yield and PP across the tested environments, whereas maturity and NP had the lowest phenotypic variations. These findings are in line with the results of previous research studies on yield component traits [2,69], in which high variation for total seed yield and total pods per plant were observed. The heritability and correlation analyses showed that NP had the highest heritability and significant linear correlations with RNP and PP. In addition, PP had the highest correlation with yield among all the tested soybean-yield components. The number of nodes and pods in soybean are known as two of the key soybean-yield components that play important roles in determining the final soybean seed yield [69,70]. Previous studies reported low heritability rates for soybean-yield components, especially NP and PP [2,71], as they are significantly affected by

environmental factors [72]. Although, for a given trait, heritability indicates the strength of the relationship between phenotype and genetic variability, it does not necessarily indicate the value of the trait for genetic studies [73]. Different low heritable traits are reported to be highly correlated with significant economic traits [73]. In soybean, for example, yield can be considered as the most important economic trait that is highly determined by its component traits.

The performance of four GWAS methods was compared in this study, and the results showed that all the methods had acceptable performance in detecting MTAs for the tested traits in this particular population. Among all the tested GWAS methods, SVR-mediated GWASs had a higher aptitude to detect SNP markers with high allelic effects associated with the tested traits in this study. The SVR-mediated GWAS method considers the presence of a nonlinear relationship between input and output variables. This ability is used to build an algorithm with greater prediction accuracies [74]. While conventional GWASs are appropriate approaches for detecting SNP markers with large effects on complex traits, they may not consider a wide range of interconnected biological processes and mechanisms that shape the phenotype of complex traits simultaneously [75]. To discover high-resolution variant-trait associations in ML-mediated GWASs, variable importance values can be used [23]. The variable importance methods based on linear and logistic regressions, support-vector machines, and random forests are well established in the literature [14,76–78]. Therefore, MTAs can be discovered by SVR-mediated GWASs as a result of its ability to consider the interaction effects between SNPs rather than the *p*-values for individual SNP-trait GWAS tests.

In this study, several previously reported QTL were colocalized with identified MTAs using all tested GWAS methods. For maturity, for example, five soybean maturity QTL detected by SVR-mediated GWASs were colocalized with previously reported QTL associated with maturity [39,43]. At the same time, none of the MTAs identified using MLM, FarmCPU, or RF were previously reported to be associated with soybean maturity. Additionally, the peak SNP position of Chr19_47513536 detected by SVR-mediated GWASs had the highest allelic effect among all the detected SNPs for soybean maturity, which is consistent with the findings in Sonah, et al. [39]. For soybean seed yield, SVR-mediated GWAS detected MTAs colocalized with five yield-related QTL [50,55], while none of the detected MTAs using other GWAS methods was previously reported for this trait. We did not find any previous study on the genetic structure of NRNP and RNP, and, therefore, all the identified MTAs in this study are considered as novel genomic regions. For PP, conventional GWAS methods were not able to detect any MTAs. However, SVR-mediated GWASs detected MTAs colocalized with seven QTL related to pod numbers [79]. The average allelic effects of the QTL presented in this study (Figure 8) were estimated using the equation developed by Pimentel, et al. [80]. The RF and SVR-mediated GWAS methods do not specifically measure allele effects, and, therefore, the aim of this study was mostly focused on detecting the MTAs, candidate genes, and QTL underlying the soybean yield, maturity, and yield components.

Regarding the results of candidate gene identification within identified QTL, several candidate genes were detected using different GWAS methods. For example, among all the detected candidate genes associated with maturity, gene *Glyma.02g006500* (GO:0015996) is a protein ABC transporter 1 that is annotated as a chlorophyll catabolic process and located exactly in the peak SNP position at Chr02_695362. ATP-binding cassette (ABC) transporter genes play conspicuous roles in different plant-growth and developmental stages by transporting different phytochemicals across endoplasmic reticulum (ER) membranes [81]. Because of the central roles played by ABC transporters in transporting biomolecules such as phytohormones, metabolites, and lipids, they play important roles in plant growth, development, and maturity [81,82]. Moreover, recent studies revealed that ER uses fatty acid building blocks made in the chloroplast to synthesize triacylglycerol (TAG). Therefore, ABC transporter genes are important for the normal accumulation of TAG during the seed-filling stage and during maturity [82,83]. Additionally, *Glyma.19g224200* (GO:0010201)

in E3 locus, which was previously discovered by Buzzell [84] and molecularly characterized as a phytochrome A (PHYA) gene [85], was detected through the SVR-mediated GWAS. Phytochromes, through PHYTOCHROME INTERACTING FACTOR (PIF), regulate the expression of specific genes encoding rate-limiting catalytic enzymes of different plant growth regulators (e.g., abscisic acid, gibberellins, and auxin) and, therefore, play crucial roles in plant maturity [86]. In addition, PHYB is inactivated after imbibition shade signals, which repress PHYA-dependent signaling in the embryo, which results in the maturing of seeds by preventing germination [87,88]. This is obtained by regulating the balance between abscisic acid and gibberellin. Subsequently, abscisic acid is transported from the endosperm to the embryo by the ABC transporter [88].

Among the candidate genes related to NRNP, gene *Glyma.07G205500* (GO:0009693-UBP1-associated protein 2C) that annotated as the ethylene biosynthetic process was located exactly at the peak SNP position at Chr7_37469678. An interaction screen with the heterogeneous nuclear ribonucleoprotein (hnRNP) results in the production of oligouridylatebinding protein 1 (UBP1)-associated protein [89]. It has been well documented that this protein plays an important role in several physiological processes such as responses to abiotic stresses [90], leaf senescence [91], floral development [92], and chromatin modification [93]. In addition, previous studies showed that the production of productive or non-reproductive nodes is completely accompanied by the upregulation or downregulation of this protein [94,95]. In addition, *Glyma.08G065300* (GO:0042546- MADS-box transcription factor), which is associated with cell-wall biogenesis, was located in the SNP position of Chr8_5005929. The genes of the MADS-box family can be considered as the main regulators for cell differentiation and organ determination [96]. The floral organ recognition MADS-box family has been categorized into A, B, C, D, and E classes. Among these classes, class E was shown to be associated with reproductive organ development [97]. Indeed, the activation or repression of this transcription factor leads to the development of nodes to productive or non-productive nodes [98–100].

Gene expression dataset developed by Severin, et al. [101] showed that the detected 20 candidate genes for PP using an SVR-mediated GWAS were expressed in flowers, 1 cm pod (7 DAF), pod shell (10–13 DAF), pod shell (14–17 DAF), and seeds. In PP, most of the genes detected by SVR-mediated GWASs are associated with either the auxin influx carrier or auxin response factors (ARFs), gibberellin synthesis, or the response to brassinosteroid [102,103]. Song, et al. [104] and Li, et al. [105] also reported some genes related to PP that were associated with embryo development, stamen development, ovule development, cytokinin biosynthesis, and response gibberellin that we also identified in this study. Soybean seed yield significantly depends on the number of seeds per plant and the seed size [106,107]. These two factors are determined by different factors, from fertilization to seed maturity. Therefore, soybean seed development can be divided into three stages or phases: pre-embryo or seed set, embryo growth or seed growth, and desiccation stages or seed maturation phases [108,109]. In Arabidopsis, a complex signaling pathway and regulatory networks, including sugar and hormonal signaling, transcription factors, and metabolic pathways, have been reported to be involved in seed development [110,111]. Several key genes and transcription factors (e.g., LEAFY COTYLEDON 1 (LEC1), LEC2, FUSCA3 (FUS3), AGAMOUS-LIKE15 (AGL15), ABSCISIC ACID INSENSITIVE 3 (ABI3), YUCCA10 (YUC10), and ARFs) have been determined to control several downstream plant growth regulators pathways to seed development [112–114]. Indeed, a high ratio of abscisic acid to gibberellic acid can regulate seed development [115,116]. In soybean, RNA seq analyses for seed set, embryo growth, and early maturation stages of developing seeds in two soybeans with contrasting seed size showed that cell division and growth genes, hormone regulation, transcription factors, and metabolic pathways are involved in seed size and numbers [117].

In general, our results showed that ML-mediated GWAS methods are able to complement the conventional GWAS methods for better identification of the MTAs for traits of interest in soybean. However, the effectiveness of using ML methods in a GWAS should be

tested in different soybean populations grown across different environments. In this study, a limited soybean population, which partially covers all the potential genetic variations in the soybean germplasm, was used. Therefore, for further evaluation of the effectiveness of an ML-mediated GWAS, it would be valuable to test the same approaches in a wide range of soybean genotypes using whole-genome sequencing data. In addition, although we used the cross-validation technique and considered several cofactors in our analyses to eliminate the potential false-positive errors, the optimal ML calibrations would be highly recommended to improve capturing the true signals and minimizing the level of errors in ML-based analyses.

4. Materials and Methods

4.1. Population and Experimental Design

A panel of 250 soybean genotypes was grown at the University of Guelph, Ridgetown Campus, in two locations, Palmyra (42°25′50.1″ N 81°45′06.9″ W, 195 m above sea level) and Ridgetown (42°27′14.8″ N 81°52′48.0″ W, 200 m above sea level), in ON, Canada, over the course of two years, 2018 and 2019. The randomized complete block designs (RCBD) with two replications were used for all four environments (two locations × two years). In general, there were 500 and 1000 research plots per environment and year, respectively. Each plot consisted of five 4.2 m long rows with 57 seeds per m^2 seeding rate. The soil type and trials were maintained using standard tillage and cultural practices in both tested locations. No fertilizers were added during the soybean growth and development stages. The herbicides were applied twice before planting and in the middle of the growth period.

4.2. Phenotyping

The soybean seed yield (t ha^{-1} at 13% moisture) for each plot was estimated by harvesting three middle rows and adjusted based on the maturity date. Soybean seed yield components, including the total number of reproductive nodes per plant (RNP), the total number of non-reproductive nodes per plant (NRNP), the total nodes per plant (NP), and the total number of pods per plant (PP), were measured using 10 randomly selected plants from each plot. The maturity was recorded as the number of days from planting to physiological maturity (R7) [118] for each genotype.

4.3. Genotyping

Young trifoliate leaf tissue for each soybean genotype from the first replication of the trial at Ridgetown in 2018 was collected in a 2 mL screw-cap tube. The leaf samples were freeze-dried for 72 h, using the Savant ModulyoD Thermoquest (Savant Instruments, Holbrook, NY, USA). By using the DNA Extraction Kit (SIGMA®, Saint Louis, MO, USA), DNA was extracted for soybean genotypes, and the quantity of DNAs was checked via Qubit® 2.0 fluorometer (Invitrogen, Carlsbad, CA, USA). For genotyping-by-sequencing (GBS), DNA samples were sent to Genomic Analysis Platform at Université Laval (Laval, QC, Canada). The GWAS panel was genotyped via a GBS protocol based on the enzymatic digestion with *ApeKI* [119]. High-quality single-nucleotide polymorphisms (SNPs) were obtained from 210 M single-end Ion Torrent reads that were proceeded with the Fast-GBS.v2 pipeline [120], using the Gmax_275_v2 reference genome. The Markov model was used to impute the missing loci, and SNPs with a minor allele frequency (MAF) less than 0.05 were removed below the threshold. As 23 genotypes did not have sufficient high-quality SNPs, they were eliminated from the experiment. In total, after checking the quality of the reading sequence and removing SNPs with more than 50% heterozygosity, 17,958 SNPs out of 40,712 SNPs were mapped to 20 soybean chromosomes. The minimum number of 403 SNPs was mapped on chromosome 11, and the maximum number of 1780 SNPs was mapped on chromosome 18 (Figure S4). Overall, the average number of SNPs across all the 20 chromosomes was 898, with the mean density of one SNP for every 0.12 cM across the whole genome.

4.4. Statistical Analyses

The best linear unbiased prediction (BLUP) as one of the common linear mixed models [121] was used to estimate the genetic values of each soybean genotype. Additionally, the R package *sommer* was used to analyze yield and yield components with 'environment' as a fixed effect and 'genotype' as a random effect. To control for the possible soil heterogeneity among the plots within a given block and reduce the associated experimental errors, nearest-neighbor analysis (NNA) was used as one of the common error control methods [122–124]. Outliers were determined in the raw dataset based on the protocols proposed by Bowley [124] and treated the same as missing data points in the analysis. Overall, the following statistical model (Equation (1)) was used in this study:

$$Y = A_b + B_g + C_i + \varepsilon \quad (1)$$

where Y stands for the trait of interest (soybean seed yield and yield component traits); b is the vector of block effects that incorporates all the locations and replications, which are added to the overall mean (fixed); g is the vector of random genotype effect, in which $g \sim N(0, \sigma^2_g)$; i is the vector of GxE interaction effects (random), in which $i \sim N(0, \sigma^2_{int})$; and ε_{ij} stands for the residual effect. A, B, and C stand for the incidence matrices of b, g, and i effects, respectively.

The heritability (Equation (2)) was calculated for soybean seed yield and yield components using the *H2cal* function in the *inti* open-source R package (https://inkaverse.com accessed on 1 May 2022) using the following equation:

$$H^2 = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_E} \quad (2)$$

where σ^2_G stands for the genotypic variance and σ^2_E is the environmental variance.

4.5. Analysis of Population Structure

A total of 17,958 high-quality SNPs from 227 soybean genotypes were used to conduct the population structure analysis using fastSTRUCTURE [125]. Five runs were conducted for K set from 1 and 15 to estimate the most appropriate number of subpopulations by using the K tool from the fastSTRUCTURE software. In order to reduce the confounding, the kinship was also estimated between genotypes of the GWAS panel.

4.6. Association Studies

Since different GWAS methods may capture different genomic regions [126], MLM and FarmCPU (the two most common GWAS methods) and RF and SVM (the two most common machine learning algorithms) were used in this study. MLM and FarmCPU were implemented by using the *GAPIT* and *rmvp* packages [127,128], and RF, as well as SVM, were conducted through the *Caret* package [129] in R software version 3.6.1. A brief description of each of the GWAS methods is provided below.

4.7. Mixed Linear Model (MLM)

This GWAS method is based on the likelihood ratio between the full model, consisting of the marker of interest, and the reduced model, which is known as the model without the marker of interest [130]. MLM is broadly used in GWASs as it effectively corrects inflation from small genetic effects caused by polygenic background and controls the possible bias in the population [130–132]. Overall, the equation of MLM would be as follows (Equation (3)):

$$Y = Xa + Z_{MyM} + e_i \quad (3)$$

where Y is the phenotypic value, X is the incident matrix effect, a stands for the vector for the incident matrix, Z_M represents the genotype indicator for the Mth SNPs, yM is equal to

the effect of the SNP_M with an assumed normal distribution and mean zero of variance, and e_i represents the residual.

4.8. Fixed and Random Model Circulating Probability Unification (FarmCPU)

This GWAS takes the advantages of using MLM as the random model and stepwise regression as the fixed model iteratively [133]. FarmCPU takes benefits from the random-effect model (REM) for optimizing the SNPs selection based on the p-values (Equation (4)):

$$Y_i = U_i + e_i \tag{4}$$

where Y_i is the observation on the ith sample, e_i stands for the residual, and U_i represents the total genetic effect of the ith sample.

Additionally, the fixed-effect model (FEM) is used in FarmCPU to test the N number of SNPs simultaneously (Equation (5)):

$$Y_i = N_{i1}F_1 + N_{i2}F_2 + N_{i3}F_3 + \cdots + N_{it}F_t + M_{ij}K_j + e_i \tag{5}$$

where Y_i is the observation on the ith sample; $N_{i1}, N_{i2}, \ldots, N_{it}$ represents the genotypes of the t pseudo-QTNs; $F_1, F_2, F_3, \ldots, F_t$ is equal to the corresponding effect for the pseudo-QTNs; M_{ij} represents the genotype of the jth SNPs and ith sample; K_j stands for the corresponding effect of the jth SNPs; and e_i represents the residual.

4.9. Random Forest (RF)

Random forest (RF) is known as one of the powerful non-parametric regression approaches that is derived from aggregating the bootstrapping in various decision trees [24]. Several decision trees are made based on the training dataset, where the output is the mean of all prediction results from the decision trees (Equation (6)):

$$Y_i = \frac{1}{B}\sum_{b=1}^{B} T_b\left(X_i\right) \tag{6}$$

where Y_i stands for the predicted value of the genotype X_i, T is the total number of constructed trees, and B is the total number of samples. In this experiment, a 1000-set of decision trees was constructed in the forest, and the GWAS analysis was conducted by measuring the importance of each feature [134], which was an SNP in this study.

4.10. Support-Vector Regression (SVR)

Support-vector regression (SVR) is known as one of the common supervised learning methods in prediction problems [135]. This algorithm is based on constructing a set of hyperplanes that can be useful in regression problems [136]. SVR determines the hyperplane by minimizing the difference of squared distances between each datum in the set and its maximum likelihood estimate [137]. In this study, the polynomial kernel was considered in SVR based on the following equation (Equation (7)):

$$L(C_a, C_b) = \left(a + C_1^T + C_2\right)^b \tag{7}$$

where $L(XC_a, C_b)$ represents the polynomial kernel between two data points, b is equal to the degree of the kernel, a is equal to the constant number, and T stands for transpose element.

The association statistics in this algorithm can be achieved by estimating the feature importance that was previously proposed by Weston, et al. [138]. In this experiment, SNP markers were selected as inputs, and the traits were selected as target variables for estimating the feature importance.

4.11. Implementation of ML Algorithms in GWAS

The implementation of ML algorithms in GWASs was reviewed well by Enoma, et al. [139]. In brief, for considering ML algorithms in GWASs, the concept of a GWAS must be seen as a machine learning counterpart. A variable in ML algorithms can be described as genetic information, each possible GWAS covariate as a feature, and phenotypic information as the output or classification, and an individual in the GWAS population can be represented by a single instance of the ML dataset. Additionally, the training, testing, and validation dataset can be considered as the population sample in GWASs.

4.12. Variable Importance Measurement

As one of the common indices for tree-based algorithms, the impurity index was chosen as the metric of the feature importance for the RF algorithm. Regarding the SVR algorithm, the variable importance method for SVR [138] was implemented in this dataset. For both algorithms, the importance of each SNP was scaled based on 0 to 100 percent scale. Since there is no confirmed way of defining the significant threshold in the tested algorithms, the global empirical threshold that provides the empirical distribution of the null hypothesis [140,141] was used for establishing the threshold in this study. The global empirical threshold was estimated based on fitting the ML algorithm, storing the highest variable importance, repeating 1000 times, and selecting the SNPs based on $\alpha = 0.05$. Additionally, the false discovery rate (FDR) is used for setting the threshold both in the FarmCPU and MLM models [142]. To estimate the feature importance in RF and SVR algorithms, a five-fold cross-validation strategy [143] with ten repetitions was applied on the dataset. All of the tested machine learning algorithms were optimized for their parameters for this dataset accordingly.

4.13. Extracting Candidate Genes Undelaying Detected QTL

For each tested GWAS model, the flanking regions of each MTA were determined using LD decay distance (Figure 7), and then potential QTL and candidate genes were retrieved using the *G. max* cv. William 82 reference-genome gene models 2.0 in SoyBase (https://www.soybase.org accessed on 1 May 2022). After listing potential candidate genes in defined windows around each significant SNP, at the peak of each QTL, the gene ontology annotation, the GO term enrichment (https://www.soybase.org accessed on 1 May 2022), and the report from previous studies were used as the criteria to select and report the most relevant candidate genes associated with the identified QTL. The Electronic Fluorescent Pictograph (eFP) browser for soybean (www.bar.utoronto.ca accessed on 1 May 2022) was also used to generate additional information such as tissue- and developmental-stage-dependent expression (based on transcriptomic data from Severin, et al. [101]) for the identified candidate genes. A Venn diagram of the MTAs colocalized with previously reported QTL for the tested traits was created using VennPainter software version 1.2.0 [144].

5. Conclusions

A better understanding of the genetic architecture of the yield component traits in soybean may enable breeders to establish more efficient selection strategies for developing high-yielding cultivars with improved genetic gains through marker-assisted selections within large breeding populations. Major yield component traits such as maturity, NP, NRNP, RNP, and PP play important roles in determining the overall yield production in soybean. Using correlation and distribution analyses, this study showed the importance of those traits in determining the total soybean seed yield. Furthermore, this study demonstrated the potential benefit of exploiting SVR-mediated GWASs for discovering MTAs associated with yield component traits in soybean. SVR-mediated can be recommended to complement conventional GWAS methods with greater power for detecting MTAs for complex traits such as yield and its components in soybean and possibly other crop species. In order to verify the causal relationship between identified MTAs and the target phenotypic traits, we identified candidate genes within each QTL using gene annotation procedures

and information, and the results were promising. Nevertheless, further studies are required to characterize the identified candidate genes in this study and confirm the efficiency of SVR-mediated GWASs for discovering genomic regions with causal relationships with complex traits in plant species.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms23105538/s1.

Author Contributions: Conceptualization, M.E.; methodology, M.E.; software, M.Y.-N.; validation, M.Y.-N., D.T. (Dan Tulpan), D.T. (Davoud Torkamaneh), I.R. and M.E.; formal analysis, M.Y.-N.; investigation, M.Y.-N.; resources, M.E.; data curation, M.Y.-N.; writing—original draft preparation, M.Y.-N. and M.E.; writing—review and editing, M.Y.-N., S.T., D.T. (Dan Tulpan), D.T. (Davoud Torkamaneh), I.R. and M.E.; visualization, M.Y.-N.; gene extraction analysis, S.T.; supervision, M.E.; project administration, M.E.; and funding acquisition, M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded in part by Grain Farmers of Ontario (GFO) and SeCan. The funding bodies did not play any role in the design of the study; in the collection, analysis, and interpretation of data; or in the writing of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets will be freely available upon request.

Acknowledgments: The authors are grateful to the past and current members of the Eskandari laboratory at the University of Guelph, Ridgetown (Bryan Stirling, John Kobler, and Robert Brandt) for their technical support. We would like to thank Maryam Vazin and Mohsen Hesami for their assistance with the field data collection and reviewing the manuscript, respectively. The preprint of this manuscript was previously deposited on bioRxiv as a non-commercial pre-print server with the doi:10.1101/2021.06.24.449776.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Rębilas, K.; Klimek-Kopyra, A.; Bacior, M.; Zając, T. A model for the yield losses estimation in an early soybean (*Glycine max* (L.) Merr.) cultivar depending on the cutting height at harvest. *Field Crop. Res.* **2020**, *254*, 107846. [CrossRef]
2. Xavier, A.; Rainey, K.M. Quantitative genomic dissection of soybean yield components. *G3 Genes Genomes Genet.* **2020**, *10*, 665–675. [CrossRef] [PubMed]
3. Suhre, J.J.; Weidenbenner, N.H.; Rowntree, S.C.; Wilson, E.W.; Naeve, S.L.; Conley, S.P.; Casteel, S.N.; Diers, B.W.; Esker, P.D.; Specht, J.E. Soybean yield partitioning changes revealed by genetic gain and seeding rate interactions. *Agron. J.* **2014**, *106*, 1631–1642. [CrossRef]
4. Mangena, P. Phytocystatins and their Potential Application in the Development of Drought Tolerance Plants in Soybeans (*Glycine max* L.). *Protein Pept. Lett.* **2020**, *27*, 135–144. [CrossRef] [PubMed]
5. Richards, R. Breeding and Selecting for Drought Resistant Wheat. Drought Resistance in Crops with Emphasis on Rice. 1982. Available online: https://agris.fao.org/agris-search/search.do?recordID=XB8110524 (accessed on 1 March 2022).
6. Reynolds, M. *Application of Physiology in Wheat Breeding*; Cimmyt: State of Mexico, Mexico, 2001.
7. Pedersen, P.; Lauer, J.G. Response of soybean yield components to management system and planting date. *Agron. J.* **2004**, *96*, 1372–1381. [CrossRef]
8. Yoosefzadeh-Najafabadi, M.; Tulpan, D.; Eskandari, M. Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. *PLoS ONE* **2021**, *16*, e0250665.
9. Robinson, A.P.; Conley, S.P.; Volenec, J.J.; Santini, J.B. Analysis of high yielding, early-planted soybean in Indiana. *Agron. J.* **2009**, *101*, 131–139. [CrossRef]
10. Ma, B.; Dwyer, L.M.; Costa, C.; Cober, E.R.; Morrison, M.J. Early prediction of soybean yield from canopy reflectance measurements. *Agron. J.* **2001**, *93*, 1227–1234. [CrossRef]
11. Xavier, A.; Jarquin, D.; Howard, R.; Ramasubramanian, V.; Specht, J.E.; Graef, G.L.; Beavis, W.D.; Diers, B.W.; Song, Q.; Cregan, P.B. Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3 Genes Genomes Genet.* **2018**, *8*, 519–529. [CrossRef]

12. Kaler, A.S.; Gillman, J.D.; Beissinger, T.; Purcell, L.C. Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Front. Plant Sci.* **2020**, *10*, 1794. [CrossRef]
13. Yoosefzadeh Najafabadi, M. Using Advanced Proximal Sensing and Genotyping Tools Combined with Bigdata Analysis Methods to Improve Soybean Yield. Ph.D. Thesis, University of Guelph, Guelph, ON, Canada, 2021.
14. Yoosefzadeh-Najafabadi, M.; Earl, H.J.; Tulpan, D.; Sulik, J.; Eskandari, M. Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. *Front. Plant Sci.* **2021**, *11*, 2555. [CrossRef] [PubMed]
15. Hesami, M.; Naderi, R.; Tohidfar, M.; Yoosefzadeh-Najafabadi, M. Development of support vector machine-based model and comparative analysis with artificial neural network for modeling the plant tissue culture procedures: Effect of plant growth regulators on somatic embryogenesis of chrysanthemum, as a case study. *Plant Methods* **2020**, *16*, 112. [CrossRef] [PubMed]
16. Yoosefzadeh-Najafabadi, M.; Torabi, S.; Tulpan, D.; Rajcan, I.; Eskandari, M. Genome-Wide Association Studies of Soybean Yield-Related Hyperspectral Reflectance Bands Using Machine Learning-Mediated Data Integration Methods. *Front. Plant Sci.* **2021**, *12*, 777028. [CrossRef] [PubMed]
17. Hesami, M.; Yoosefzadeh Najafabadi, M.; Adamek, K.; Torkamaneh, D.; Jones, A.M.P. Synergizing off-target predictions for in silico insights of CENH3 knockout in cannabis through CRISPR/CAS. *Molecules* **2021**, *26*, 2053. [CrossRef] [PubMed]
18. Jafari, M.; Shahsavar, A. The application of artificial neural networks in modeling and predicting the effects of melatonin on morphological responses of citrus to drought stress. *PLoS ONE* **2020**, *15*, e0240427. [CrossRef] [PubMed]
19. Tulpan, D. 311 A brief overview, comparison and practical applications of machine learning models. *J. Anim. Sci.* **2020**, *98*, 44–45. [CrossRef]
20. Chen, J.H.; Verghese, A. Planning for the Known Unknown: Machine Learning for Human Healthcare Systems. *Am. J. Bioeth.* **2020**, *20*, 1–3. [CrossRef]
21. Kim, G.B.; Kim, W.J.; Kim, H.U.; Lee, S.Y. Machine learning applications in systems metabolic engineering. *Curr. Opin. Biotechnol.* **2020**, *64*, 1–9. [CrossRef]
22. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
23. Szymczak, S.; Biernacka, J.M.; Cordell, H.J.; González-Recio, O.; König, I.R.; Zhang, H.; Sun, Y.V. Machine learning in genome-wide association studies. *Genet. Epidemiol.* **2009**, *33*, S51–S57. [CrossRef]
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999.
26. Ogutu, J.O.; Piepho, H.-P.; Schulz-Streeck, T. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* **2011**, *5*, S11. [CrossRef]
27. Jamil, I.N.; Remali, J.; Azizan, K.A.; Muhammad, N.A.N.; Arita, M.; Goh, H.-H.; Aizat, W.M. Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology. *Front. Plant Sci.* **2020**, *11*, 944. [CrossRef] [PubMed]
28. Sun, S.; Wang, C.; Ding, H.; Zou, Q. Machine learning and its applications in plant molecular studies. *Brief. Funct. Genom.* **2020**, *19*, 40–48. [CrossRef]
29. Su, Q.; Lu, W.; Du, D.; Chen, F.; Niu, B.; Chou, K.-C. Prediction of the aquatic toxicity of aromatic compounds to tetrahymena pyriformis through support vector regression. *Oncotarget* **2017**, *8*, 49359. [CrossRef]
30. Auria, L.; Moro, R.A. Support Vector Machines (SVM) as a Technique for Solvency Analysis. *SSRN Electron. J.* **2008**, 811.
31. Hesami, M.; Jones, A.M.P. Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. *Appl. Microbiol. Biotechnol.* **2020**, *104*, 9449–9485. [CrossRef]
32. Belayneh, A.; Adamowski, J.; Khalil, B.; Ozga-Zielinski, B. Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. *J. Hydrol.* **2014**, *508*, 418–429. [CrossRef]
33. Duan, K.-B.; Rajapakse, J.C.; Wang, H.; Azuaje, F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans. Nanobiosci.* **2005**, *4*, 228–234. [CrossRef]
34. Denton, S.M.; Salleb-Aouissi, A. A Weighted Solution to SVM Actionability and Interpretability. *arXiv* **2020**, arXiv:2012.03372.
35. Pepe, M.; Hesami, M.; Jones, A.M.P. Machine Learning-Mediated Development and Optimization of Disinfection Protocol and Scarification Method for Improved In Vitro Germination of Cannabis Seeds. *Plants* **2021**, *10*, 2397. [CrossRef] [PubMed]
36. Yoosefzadeh-Najafabadi, M.; Tulpan, D.; Eskandari, M. Using Hybrid Artificial Intelligence and Evolutionary Optimization Algorithms for Estimating Soybean Yield and Fresh Biomass Using Hyperspectral Vegetation Indices. *Remote Sens.* **2021**, *13*, 2555. [CrossRef]
37. Awad, M.; Khanna, R. Support vector regression. In *Efficient Learning Machines*; Springe: New York, NY, USA, 2015; pp. 67–80.
38. Moellers, T.C.; Singh, A.; Zhang, J.; Brungardt, J.; Kabbage, M.; Mueller, D.S.; Grau, C.R.; Ranjan, A.; Smith, D.L.; Chowda-Reddy, R. Main and epistatic loci studies in soybean for Sclerotinia sclerotiorum resistance reveal multiple modes of resistance in multi-environments. *Sci. Rep.* **2017**, *7*, 3554. [CrossRef]
39. Sonah, H.; O'Donoughue, L.; Cober, E.; Rajcan, I.; Belzile, F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* **2015**, *13*, 211–221. [CrossRef] [PubMed]
40. Kaler, A.S.; Dhanapal, A.P.; Ray, J.D.; King, C.A.; Fritschi, F.B.; Purcell, L.C. Genome-wide association mapping of carbon isotope and oxygen isotope ratios in diverse soybean genotypes. *Crop Sci.* **2017**, *57*, 3085–3100. [CrossRef]
41. Fang, C.; Ma, Y.; Wu, S.; Liu, Z.; Wang, Z.; Yang, R.; Hu, G.; Zhou, Z.; Yu, H.; Zhang, M. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* **2017**, *18*, 1–14. [CrossRef]

42. Kan, G.; Zhang, W.; Yang, W.; Ma, D.; Zhang, D.; Hao, D.; Hu, Z.; Yu, D. Association mapping of soybean seed germination under salt stress. *Mol. Genet. Genom.* **2015**, *290*, 2147–2162. [CrossRef]
43. Zhang, J.; Song, Q.; Cregan, P.B.; Nelson, R.L.; Wang, X.; Wu, J.; Jiang, G.-L. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genom.* **2015**, *16*, 1–11. [CrossRef]
44. Mao, T.; Li, J.; Wen, Z.; Wu, T.; Wu, C.; Sun, S.; Jiang, B.; Hou, W.; Li, W.; Song, Q. Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. *BMC Genom.* **2017**, *18*, 415. [CrossRef]
45. Bao, Y.; Kurle, J.E.; Anderson, G.; Young, N.D. Association mapping and genomic prediction for resistance to sudden death syndrome in early maturing soybean germplasm. *Mol. Breed.* **2015**, *35*, 18. [CrossRef]
46. Leamy, L.J.; Zhang, H.; Li, C.; Chen, C.Y.; Song, B.-H. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genom.* **2017**, *18*, 1–15. [CrossRef] [PubMed]
47. Wen, Z.; Tan, R.; Yuan, J.; Bales, C.; Du, W.; Zhang, S.; Chilvers, M.I.; Schmidt, C.; Song, Q.; Cregan, P.B. Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. *BMC Genom.* **2014**, *15*, 1–11. [CrossRef] [PubMed]
48. Ray, J.D.; Dhanapal, A.P.; Singh, S.K.; Hoyos-Villegas, V.; Smith, J.R.; Purcell, L.C.; King, C.A.; Boykin, D.; Cregan, P.B.; Song, Q. Genome-wide association study of ureide concentration in diverse maturity group IV soybean [*Glycine max* (L.) Merr.] accessions. *G3 Genes Genomes Genet.* **2015**, *5*, 2391–2403. [CrossRef] [PubMed]
49. Dhanapal, A.P.; Ray, J.D.; Smith, J.R.; Purcell, L.C.; Fritschi, F.B. Identification of Novel Genomic Loci Associated with Soybean Shoot Tissue Macro and Micronutrient Concentrations. *Plant Genome* **2018**, *11*, 170066. [CrossRef] [PubMed]
50. Hu, Z.; Zhang, D.; Zhang, G.; Kan, G.; Hong, D.; Yu, D. Association mapping of yield-related traits and SSR markers in wild soybean (*Glycine soja* Sieb. and Zucc.). *Breed. Sci.* **2014**, *63*, 441–449. [CrossRef]
51. Contreras-Soto, R.I.; Mora, F.; de Oliveira, M.A.R.; Higashi, W.; Scapim, C.A.; Schuster, I. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. *PLoS ONE* **2017**, *12*, e0171105.
52. Li, Y.h.; Shi, X.h.; Li, H.h.; Reif, J.C.; Wang, J.j.; Liu, Z.x.; He, S.; Yu, B.s.; Qiu, L.j. Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. *Plant Genome* **2016**, 9. [CrossRef]
53. Zhang, J.; Song, Q.; Cregan, P.B.; Jiang, G.-L. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* **2016**, *129*, 117–130. [CrossRef]
54. Chang, H.-X.; Hartman, G.L. Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Front. Plant Sci.* **2017**, *8*, 670. [CrossRef]
55. Copley, T.R.; Duceppe, M.-O.; O'Donoughue, L.S. Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. *BMC Genom.* **2018**, *19*, 167. [CrossRef]
56. Li, Y.-h.; Reif, J.C.; Ma, Y.-s.; Hong, H.-l.; Liu, Z.-x.; Chang, R.-z.; Qiu, L.-j. Targeted association mapping demonstrating the complex molecular genetics of fatty acid formation in soybean. *BMC Genom.* **2015**, *16*, 841. [CrossRef] [PubMed]
57. Xavier, A.; Muir, W.M.; Rainey, K.M. Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. *BMC Bioinform.* **2016**, *17*, 55. [CrossRef] [PubMed]
58. Cook, D.E.; Bayless, A.M.; Wang, K.; Guo, X.; Song, Q.; Jiang, J.; Bent, A.F. Distinct copy number, coding sequence, and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiol.* **2014**, *165*, 630–647. [CrossRef] [PubMed]
59. Zhang, J.; Wang, X.; Lu, Y.; Bhusal, S.J.; Song, Q.; Cregan, P.B.; Yen, Y.; Brown, M.; Jiang, G.-L. Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. *Mol. Plant* **2018**, *11*, 460–472. [CrossRef]
60. Qin, J.; Song, Q.; Shi, A.; Li, S.; Zhang, M.; Zhang, B. Genome-wide association mapping of resistance to Phytophthora sojae in a soybean [*Glycine max* (L.) Merr.] germplasm panel from maturity groups IV and V. *PLoS ONE* **2017**, *12*, e0184613. [CrossRef]
61. Vuong, T.; Sonah, H.; Meinhardt, C.; Deshmukh, R.; Kadam, S.; Nelson, R.; Shannon, J.; Nguyen, H. Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. *BMC Genom.* **2015**, *16*, 593. [CrossRef]
62. Che, Z.; Liu, H.; Yi, F.; Cheng, H.; Yang, Y.; Wang, L.; Du, J.; Zhang, P.; Wang, J.; Yu, D. Genome-Wide Association Study Reveals Novel Loci for SC7 Resistance in a Soybean Mutant Panel. *Front. Plant Sci.* **2017**, *8*, 1771. [CrossRef]
63. Li, X.; Tian, R.; Kamala, S.; Du, H.; Li, W.; Kong, Y.; Zhang, C. Identification and verification of pleiotropic QTL controlling multiple amino acid contents in soybean seed. *Euphytica* **2018**, *214*, 1–14. [CrossRef]
64. Ray, J.D.; Dhanapal, A.P.; Singh, S.K.; Hoyos-Villegas, V.; Smith, J.R.; Purcell, L.C.; King, C.A.; Boykin, D.; Cregan, P.B.; Song, Q. Genome-wide association study (GWAS) of carbon isotope ratio (δ 13 C) in diverse soybean [*Glycine max* (L.) Merr.] genotypes. *Theor. Appl. Genet.* **2015**, *128*, 73–91.
65. Priolli, R.H.G.; Campos, J.; Stabellini, N.; Pinheiro, J.; Vello, N. Association mapping of oil content and fatty acid components in soybean. *Euphytica* **2015**, *203*, 83–96. [CrossRef]
66. Dhanapal, A.P.; Ray, J.D.; Singh, S.K.; Hoyos-Villegas, V.; Smith, J.R.; Purcell, L.C.; King, C.A.; Fritschi, F.B. Association mapping of total carotenoids in diverse soybean genotypes based on leaf extracts and high-throughput canopy spectral reflectance measurements. *PLoS ONE* **2015**, *10*, e0137213. [CrossRef] [PubMed]
67. Cooper, M.; van Eeuwijk, F.A.; Hammer, G.L.; Podlich, D.W.; Messina, C. Modeling QTL for complex traits: Detection and context for plant breeding. *Curr. Opin. Plant Biol.* **2009**, *12*, 231–240. [CrossRef] [PubMed]

68. Hu, D.; Zhang, H.; Du, Q.; Hu, Z.; Yang, Z.; Li, X.; Wang, J.; Huang, F.; Yu, D.; Wang, H. Genetic dissection of yield-related traits via genome-wide association analysis across multiple environments in wild soybean (*Glycine soja* Sieb. and Zucc.). *Planta* **2020**, *251*, 39. [CrossRef] [PubMed]
69. Kahlon, C.S.; Board, J.E. Growth dynamic factors explaining yield improvement in new versus old soybean cultivars. *J. Crop Improv.* **2012**, *26*, 282–299. [CrossRef]
70. Herbert, S.; Litchfield, G. Partitioning Soybean Seed Yield Components 1. *Crop Sci.* **1982**, *22*, 1074–1079. [CrossRef]
71. Sulistyo, A.; Sari, K. Correlation, path analysis and heritability estimation for agronomic traits contribute to yield on soybean. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Banda Aceh, Indonesia, 26–27 September 2018; p. 012034.
72. Price, T.; Schluter, D. On the low heritability of life-history traits. *Evolution* **1991**, *45*, 853–861. [CrossRef]
73. Cassell, B.G. *Using Heritability for Genetic Improvement*; Virginia Cooperative Extension: Blacksburg, VA, USA, 2009.
74. Kaneko, H. Support vector regression that takes into consideration the importance of explanatory variables. *J. Chemom.* **2020**, *35*, e3327. [CrossRef]
75. Lee, S.; Liang, X.; Woods, M.; Reiner, A.S.; Concannon, P.; Bernstein, L.; Lynch, C.F.; Boice, J.D.; Deasy, J.O.; Bernstein, J.L. Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study. *PLoS ONE* **2020**, *15*, e0226157. [CrossRef]
76. Williamson, B.D.; Gilbert, P.B.; Simon, N.R.; Carone, M. A unified approach for inference on algorithm-agnostic variable importance. *arXiv* **2020**, arXiv:2004.03683 2020.
77. Wu, Y.; Liu, Y. Variable selection in quantile regression. *Stat. Sin.* **2009**, 801–817.
78. Grömping, U. Variable importance assessment in regression: Linear regression versus random forest. *Am. Stat.* **2009**, *63*, 308–319. [CrossRef]
79. Zhang, H.; Hao, D.; Sitoe, H.M.; Yin, Z.; Hu, Z.; Zhang, G.; Yu, D. Genetic dissection of the relationship between plant architecture and yield component traits in soybean (*Glycine max*) by association analysis across multiple environments. *Plant Breed.* **2015**, *134*, 564–572. [CrossRef]
80. Pimentel, E.; Edel, C.; Emmerling, R.; Götz, K.-U. How imputation errors bias genomic predictions. *J. Dairy Sci.* **2015**, *98*, 4131–4138. [CrossRef] [PubMed]
81. Hwang, J.-U.; Song, W.-Y.; Hong, D.; Ko, D.; Yamaoka, Y.; Jang, S.; Yim, S.; Lee, E.; Khare, D.; Kim, K. Plant ABC transporters enable many unique aspects of a terrestrial plant's lifestyle. *Mol. Plant* **2016**, *9*, 338–355. [CrossRef]
82. Block, M.A.; Jouhet, J. Lipid trafficking at endoplasmic reticulum–chloroplast membrane contact sites. *Curr. Opin. Cell Biol.* **2015**, *35*, 21–29. [CrossRef]
83. Kim, S.; Yamaoka, Y.; Ono, H.; Kim, H.; Shim, D.; Maeshima, M.; Martinoia, E.; Cahoon, E.B.; Nishida, I.; Lee, Y. AtABCA9 transporter supplies fatty acids for lipid synthesis to the endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 773–778. [CrossRef]
84. Buzzell, R. Inheritance of a soybean flowering response to fluorescent-daylength conditions. *Can. J. Genet. Cytol.* **1971**, *13*, 703–707. [CrossRef]
85. Watanabe, S.; Hideshima, R.; Xia, Z.; Tsubokura, Y.; Sato, S.; Nakamoto, Y.; Yamanaka, N.; Takahashi, R.; Ishimoto, M.; Anai, T. Map-based cloning of the gene associated with the soybean maturity locus E3. *Genetics* **2009**, *182*, 1251–1262. [CrossRef]
86. Legris, M.; Ince, Y.Ç.; Fankhauser, C. Molecular mechanisms underlying phytochrome-controlled morphogenesis in plants. *Nat. Commun.* **2019**, *10*, 5219. [CrossRef]
87. Casal, J.J. Photoreceptor signaling networks in plant responses to shade. *Ann. Rev. Plant Biol.* **2013**, *64*, 403–427. [CrossRef]
88. De Wit, M.; Galvão, V.C.; Fankhauser, C. Light-mediated hormonal regulation of plant growth and development. *Annu. Rev. Plant Biol.* **2016**, *67*, 513–537. [CrossRef]
89. Lambermon, M.H.; Fu, Y.; Kirk, D.A.W.; Dupasquier, M.; Filipowicz, W.; Lorković, Z.J. UBA1 and UBA2, two proteins that interact with UBP1, a multifunctional effector of pre-mRNA maturation in plants. *Mol. Cell. Biol.* **2002**, *22*, 4346–4357. [CrossRef]
90. Li, J.; Kinoshita, T.; Pandey, S.; Ng, C.K.-Y.; Gygi, S.P.; Shimazaki, K.-i.; Assmann, S.M. Modulation of an RNA-binding protein by abscisic-acid-activated protein kinase. *Nature* **2002**, *418*, 793–797. [CrossRef] [PubMed]
91. Kim, C.Y.; Bove, J.; Assmann, S.M. Overexpression of wound-responsive RNA-binding proteins induces leaf senescence and hypersensitive-like cell death. *New Phytol.* **2008**, *180*, 57–70. [CrossRef] [PubMed]
92. Streitner, C.; Danisman, S.; Wehrle, F.; Schöning, J.C.; Alfano, J.R.; Staiger, D. The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in Arabidopsis thaliana. *Plant J.* **2008**, *56*, 239–250. [CrossRef] [PubMed]
93. Liu, F.; Quesada, V.; Crevillén, P.; Bäurle, I.; Swiezewski, S.; Dean, C. The Arabidopsis RNA-binding protein FCA requires a lysine-specific demethylase 1 homolog to downregulate FLC. *Mol. Cell* **2007**, *28*, 398–407. [CrossRef] [PubMed]
94. Bäurle, I.; Dean, C. Differential interactions of the autonomous pathway RRM proteins and chromatin regulators in the silencing of Arabidopsis targets. *PLoS ONE* **2008**, *3*, e2733. [CrossRef]
95. Na, J.-K.; Kim, J.-K.; Kim, D.-Y.; Assmann, S.M. Expression of potato RNA-binding proteins StUBA2a/b and StUBA2c induces hypersensitive-like cell death and early leaf senescence in Arabidopsis. *J. Exp. Bot.* **2015**, *66*, 4023–4033. [CrossRef]
96. Lee, J.H.; Ryu, H.-S.; Chung, K.S.; Posé, D.; Kim, S.; Schmid, M.; Ahn, J.H. Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. *Science* **2013**, *342*, 628–632. [CrossRef]

97. Hussin, S.H.; Wang, H.; Tang, S.; Zhi, H.; Tang, C.; Zhang, W.; Jia, G.; Diao, X. SiMADS34, an E-class MADS-box transcription factor, regulates inflorescence architecture and grain yield in Setaria italica. *Plant Mol. Biol.* **2021**, *105*, 419–434. [CrossRef]
98. Gao, X.; Liang, W.; Yin, C.; Ji, S.; Wang, H.; Su, X.; Guo, C.; Kong, H.; Xue, H.; Zhang, D. The SEPALLATA-like gene OsMADS34 is required for rice inflorescence and spikelet development. *Plant Physiol.* **2010**, *153*, 728–740. [CrossRef]
99. Ditta, G.; Pinyopich, A.; Robles, P.; Pelaz, S.; Yanofsky, M.F. The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity. *Curr. Biol.* **2004**, *14*, 1935–1940. [CrossRef]
100. Liu, C.; Teo, Z.W.N.; Bi, Y.; Song, S.; Xi, W.; Yang, X.; Yin, Z.; Yu, H. A conserved genetic pathway determines inflorescence architecture in Arabidopsis and rice. *Dev. Cell* **2013**, *24*, 612–622. [CrossRef] [PubMed]
101. Severin, A.J.; Woody, J.L.; Bolon, Y.-T.; Joseph, B.; Diers, B.W.; Farmer, A.D.; Muehlbauer, G.J.; Nelson, R.T.; Grant, D.; Specht, J.E. RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biol.* **2010**, *10*, 160. [CrossRef] [PubMed]
102. Yin, Z.; Qi, H.; Mao, X.; Wang, J.; Hu, Z.; Wu, X.; Liu, C.; Xin, D.; Zuo, X.; Chen, Q. QTL mapping of soybean node numbers on the main stem and meta-analysis for mining candidate genes. *Biotechnol. Biotechnol. Equip.* **2018**, *32*, 915–922. [CrossRef]
103. Lin, F.; Wani, S.H.; Collins, P.J.; Wen, Z.; Li, W.; Zhang, N.; McCoy, A.G.; Bi, Y.; Tan, R.; Zhang, S. QTL mapping and GWAS for identification of loci conferring partial resistance to Pythium sylvaticum in soybean (*Glycine max* (L.) Merr). *Mol. Breed.* **2020**, *40*, 1–11. [CrossRef]
104. Song, J.; Sun, X.; Zhang, K.; Liu, S.; Wang, J.; Yang, C.; Jiang, S.; Siyal, M.; Li, X.; Qi, Z. Identification of QTL and genes for pod number in soybean by linkage analysis and genome-wide association studies. *Mol. Breed.* **2020**, *40*, 1–14. [CrossRef]
105. Li, C.; Zou, J.; Jiang, H.; Yu, J.; Huang, S.; Wang, X.; Liu, C.; Guo, T.; Zhu, R.; Wu, X. Identification and validation of number of pod-and seed-related traits QTL s in soybean. *Plant Breed.* **2018**, *137*, 730–745. [CrossRef]
106. Liu, B.; Liu, X.; Wang, C.; Li, Y.; Jin, J.; Herbert, S. Soybean yield and yield component distribution across the main axis in response to light enrichment and shading under different densities. *Plant Soil Environ.* **2010**, *56*, 384–392. [CrossRef]
107. Rotundo, J.L.; Borrás, L.; Westgate, M.E.; Orf, J.H. Relationship between assimilate supply per seed during seed filling and soybean seed composition. *Field Crop. Res.* **2009**, *112*, 90–96. [CrossRef]
108. Weber, H.; Borisjuk, L.; Wobus, U. Molecular physiology of legume seed development. *Annu. Rev. Plant Biol.* **2005**, *56*, 253–279. [CrossRef]
109. Ruan, Y.-L.; Patrick, J.W.; Bouzayen, M.; Osorio, S.; Fernie, A.R. Molecular regulation of seed and fruit set. *Trends Plant Sci.* **2012**, *17*, 656–665. [CrossRef]
110. Orozco-Arroyo, G.; Paolo, D.; Ezquer, I.; Colombo, L. Networks controlling seed size in Arabidopsis. *Plant Reprod.* **2015**, *28*, 17–32. [CrossRef] [PubMed]
111. Le, B.H.; Cheng, C.; Bui, A.Q.; Wagmaister, J.A.; Henry, K.F.; Pelletier, J.; Kwong, L.; Belmonte, M.; Kirkbride, R.; Horvath, S. Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 8063–8070. [CrossRef] [PubMed]
112. Sun, X.; Shantharaj, D.; Kang, X.; Ni, M. Transcriptional and hormonal signaling control of Arabidopsis seed development. *Curr. Opin. Plant Biol.* **2010**, *13*, 611–620. [CrossRef] [PubMed]
113. Lepiniec, L.; Devic, M.; Roscoe, T.; Bouyer, D.; Zhou, D.-X.; Boulard, C.; Baud, S.; Dubreucq, B. Molecular and epigenetic regulations and functions of the LAFL transcriptional regulators that control seed development. *Plant Reprod.* **2018**, *31*, 291–307. [CrossRef]
114. Pelletier, J.M.; Kwong, R.W.; Park, S.; Le, B.H.; Baden, R.; Cagliari, A.; Hashimoto, M.; Munoz, M.D.; Fischer, R.L.; Goldberg, R.B. LEC1 sequentially regulates the transcription of genes involved in diverse developmental processes during seed development. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E6710–E6719. [CrossRef]
115. Figueiredo, D.D.; Köhler, C. Auxin: A molecular trigger of seed development. *Genes Dev.* **2018**, *32*, 479–490. [CrossRef]
116. Wang, L.; Hu, X.; Jiao, C.; Li, Z.; Fei, Z.; Yan, X.; Liu, C.; Wang, Y.; Wang, X. Transcriptome analyses of seed development in grape hybrids reveals a possible mechanism influencing seed size. *BMC Genom.* **2016**, *17*, 898. [CrossRef] [PubMed]
117. Du, J.; Wang, S.; He, C.; Zhou, B.; Ruan, Y.-L.; Shou, H. Identification of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. *J. Exp. Bot.* **2017**, *68*, 1955–1972. [CrossRef] [PubMed]
118. Fehr, W.; Caviness, C.; Burmood, D.T.; Penington, J.S. Development description of soybean, *Glycine max* (L.) Mer. *Crop Sci.* **1971**, *11*, 929–931. [CrossRef]
119. Sonah, H.; Bastien, M.; Iquira, E.; Tardivel, A.; Légaré, G.; Boyle, B.; Normandeau, É.; Laroche, J.; Larose, S.; Jean, M.; et al. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLoS ONE* **2013**, *8*, e54603. [CrossRef]
120. Torkamaneh, D.; Laroche, J.; Belzile, F. Fast-GBS v2.0: An analysis toolkit for genotyping-by-sequencing data. *Genome* **2020**, *63*, 577–581. [CrossRef] [PubMed]
121. Goldberger, A.S. Best linear unbiased prediction in the generalized linear regression model. *J. Am. Stat. Assoc.* **1962**, *57*, 369–375. [CrossRef]
122. Stroup, W.; Mulitze, D. Nearest neighbor adjusted best linear unbiased prediction. *Am. Stat.* **1991**, *45*, 194–200.
123. Katsileros, A.; Drosou, K.; Koukouvinos, C. Evaluation of nearest neighbor methods in wheat genotype experiments. *Commun. Biometry Crop Sci.* **2015**, *10*, 115–123.
124. Bowley, S. *A Hitchhiker's Guide to Statistics in Plant Biology*; Any Old Subject Books: Guelph, ON, Canada, 1999.

125. Raj, A.; Stephens, M.; Pritchard, J.K. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **2014**, *197*, 573–589. [CrossRef]
126. Yang, J.; Yeh, C.-T.E.; Ramamurthy, R.K.; Qi, X.; Fernando, R.L.; Dekkers, J.C.; Garrick, D.J.; Nettleton, D.; Schnable, P.S. Empirical comparisons of different statistical models to identify and validate kernel row number-associated variants from structured multi-parent mapping populations of maize. *G3 Genes Genomes Genet.* **2018**, *8*, 3567–3575. [CrossRef]
127. Lipka, A.E.; Tian, F.; Wang, Q.; Peiffer, J.; Li, M.; Bradbury, P.J.; Gore, M.A.; Buckler, E.S.; Zhang, Z. GAPIT: Genome association and prediction integrated tool. *Bioinformatics* **2012**, *28*, 2397–2399. [CrossRef] [PubMed]
128. Yin, L.; Zhang, H.; Tang, Z.; Xu, J.; Yin, D.; Zhang, Z.; Yuan, X.; Zhu, M.; Zhao, S.; Li, X. rmvp: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genom. Proteom. Bioinform.* **2021**, *19*, 619–628. [CrossRef]
129. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Team, R.C. Package 'caret'. *R J.* **2020**, *223*, 7.
130. Wen, Y.-J.; Zhang, H.; Ni, Y.-L.; Huang, B.; Zhang, J.; Feng, J.-Y.; Wang, S.-B.; Dunwell, J.M.; Zhang, Y.-M.; Wu, R. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* **2018**, *19*, 809. [CrossRef]
131. Wang, S.-B.; Feng, J.-Y.; Ren, W.-L.; Huang, B.; Zhou, L.; Wen, Y.-J.; Zhang, J.; Dunwell, J.M.; Xu, S.; Zhang, Y.-M. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **2016**, *6*, 19444. [CrossRef] [PubMed]
132. Bulik-Sullivan, B.K.; Loh, P.-R.; Finucane, H.K.; Ripke, S.; Yang, J.; Patterson, N.; Daly, M.J.; Price, A.L.; Neale, B.M. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **2015**, *47*, 291–295. [CrossRef] [PubMed]
133. Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **2016**, *12*, e1005767. [CrossRef] [PubMed]
134. Botta, V.; Louppe, G.; Geurts, P.; Wehenkel, L. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS ONE* **2014**, *9*, e93379. [CrossRef]
135. Cortes, C.; Vapnik, V. Support vector machine. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
136. Fletcher, T. *Support Vector Machines Explained*; UCL: London, UK, 2008.
137. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
138. Weston, J.; Mukherjee, S.; Chapelle, O.; Pontil, M.; Poggio, T.; Vapnik, V. Feature selection for SVMs. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 668–674.
139. Enoma, D.O.; Bishung, J.; Abiodun, T.; Ogunlana, O.; Osamor, V.C. Machine learning approaches to genome-wide association studies. *J. King Saud Univ. Sci.* **2022**, *34*, 101847. [CrossRef]
140. Doerge, R.W.; Churchill, G.A. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **1996**, *142*, 285–294. [CrossRef]
141. Churchill, G.A.; Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **1994**, *138*, 963–971. [CrossRef]
142. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
143. Siegmann, B.; Jarmer, T. Comparison of different regression models and validation techniques for the assessment of wheat leaf area index from hyperspectral data. *Int. J. Remote Sens.* **2015**, *36*, 4519–4534. [CrossRef]
144. Lin, G.; Chai, J.; Yuan, S.; Mai, C.; Cai, L.; Murphy, R.W.; Zhou, W.; Luo, J. VennPainter: A tool for the comparison and identification of candidate genes based on Venn diagrams. *PLoS ONE* **2016**, *11*, e0154315. [CrossRef] [PubMed]