International Journal of
Molecular Sciences

MDPI

Article

Genetic Architecture of Early Vigor Traits in Wild Soybean

Janice Kofsky, Hengyou Zhang † and Bao-Hua Song *

Department of Biological Sciences, University of North Carolina at Charlotte, Charlotte, NC 28223, USA; jkofsky@uncc.edu (J.K.); hengyou.zhang@gmail.com (H.Z.)
* Correspondence: bsong5@uncc.edu; Tel.: +1-704-687-8670
† Current Address: Donald Danforth Plant Science Center, Saint Louis, MO 63132, USA.

Received: 10 April 2020; Accepted: 24 April 2020; Published: 28 April 2020

Abstract: A worldwide food shortage has been projected as a result of the current increase in global population and climate change. In order to provide sufficient food to feed more people, we must develop crops that can produce higher yields. Plant early vigor traits, early growth rate (EGR), early plant height (EPH), inter-node length, and node count are important traits that are related to crop yield. *Glycine soja*, the wild counterpart to cultivated soybean, *Glycine max*, harbors much higher genetic diversity and can grow in diverse environments. It can also cross easily with cultivated soybean. Thus, it holds a great potential in developing soybean cultivars with beneficial agronomic traits. In this study, we used 225 wild soybean accessions originally from diverse environments across its geographic distribution in East Asia. We quantified the natural variation of several early vigor traits, investigated the relationships among them, and dissected the genetic basis of these traits by applying a Genome-Wide Association Study (GWAS) with genome-wide single nucleotide polymorphism (SNP) data. Our results showed positive correlation between all early vigor traits studied. A total of 12 SNPs significantly associated with EPH were identified with 4 shared with EGR. We also identified two candidate genes, *Glyma.07G055800.1* and *Glyma.07G055900.1*, playing important roles in influencing trait variation in both EGR and EPH in *G. soja*.

Keywords: *Glycine soja*; wild soybean; early vigor; early growth; genome-wide association study; GWAS

1. Introduction

The cultivated soybean, *Glycine max*, is an important legume crop that supplies the majority of protein meal and oilseed worldwide [1]. Crop improvement is a continuous necessity as the demands in agriculture increase due to a changing environment and rising population. Modern soybean crops are challenged by abiotic and biotic stress. In order for soybean production to keep up with the growing population, novel modifications must be made to the current crop beyond modern breeding practices [2]. The wild counterpart to the cultivated soybean, *Glycine soja*, is analyzed for genetic associations to early vigor traits that are not present in the cultivated population.

G. max diverged from *G. soja* 0.27 [3] or 0.8 million years ago (mya) [4], leaving much of the wild genetic variability behind [5]. The cultivated soybean further diverged as a result of domestication in China 6000–9000 years ago; however, the wild soybean continues to inhabit a wide range of areas across Eastern Asia [3,6]. The gene pool in wild soybean retains the genes and alleles lost during the process of artificial breeding practices and, more importantly, the initial domestication of the soybean, which contributes to 16.2% and 31% reduction in genetic diversity, respectively [7]. Selection and improvement within *G. max* have been shown to have a significant effect on diversity on

all twenty chromosomes [8,9]. By exploiting the genetic diversity harbored in the wild soybean, we are able to discover novel sources of agronomical superiority.

Early vigor is a combination of traits that impact the eventual success and yield of a plant. The early vigor phenotypic traits studied in *G. soja* are node count, inter-node length, early growth rate, and early plant height. Early growth rate (EGR) and early plant height (EPH) are further analyzed for genotypic dissection using genome-wide association studies (GWAS). These traits have not been studied or used for genetic association in wild soybean thus far. The node count and inter-node length are representative of the density of plant foliage, as each node results in a branching compound leaf. There is a significant correlation between the yield per soybean plant and branch number [10]. Measurements of mature plant height in wild soybeans are relatively challenging due to its fragile stem and vining nature. Therefore, EPH can be used as an analysis of height while the plant is still in early growth stages. EGR is an important complex trait, and often associated with the ultimate success of the plant. Seed vigor is determined partially by an evaluation of seedling growth [11], which was initially established in the 1950s to represent the success and productivity of the resulting plant [12]. All traits analyzed are representative of plant success and a measure of vigor.

A genome-wide association study (GWAS) was used in this study to determine genetic associations with complex agronomically beneficial traits, EPH and EGR. GWAS has significant advantages over the tradition approach of complex candidate loci discovery, quantitative trait loci (QTL) mapping. QTL mapping relies on a crossed mapping population from only two individuals to contain all genetic diversity for the trait being analyzed. In addition, QTL mapping is limited by its resolution in the mapping population due to limited recombination events [13]. Although first demonstrated for use in human disease [14], GWAS has been widely applied to plants for over a decade to identify genes responsible for quantitative traits. While the QTL mapping approach uses just two genotypes for genetic variation discovery, GWAS utilizes an unlimited sample size and takes advantage of the genetic and phenotypic diversity contained in the extensive natural population and thus providing higher mapping resolution than bi-parental populations [15,16]. The wild soybean is an ideal system for GWAS, as it can be maintained by self-fertilization, allowing for repeated screening of genetically similar individuals. GWAS has previously been used to dissect phenotypic traits related to biotic stress resistance [17–19], abiotic stress tolerance [20], and seed composition [21,22] in wild soybean.

The value of crop wild relatives in crop improvement has been progressively recognized in the past decades [23]. The wild soybean, *G. soja*, has been used in the genetic dissection of soybean growth [24,25], yield [26], seed composition [21,27], abiotic stress tolerance [20,28–32], and biotic stress resistance [18,33–42]. However, to our knowledge, using *G. soja* to study early vigor phenotypic traits, such as early growth rate, early plant height, node count, and inter-node length, has never been reported. To explore the genetic diversity harbored in the wild soybean, we used GWAS to effectively characterize the genetic architecture of early vigor traits. This study aims to make use of a cutting-edge quantitative trait loci discovery method to dissect the genetic basis of agronomically beneficial traits, EGR and EPH, and their relationships to inter-node length and node count to further crop-improvement techniques.

2. Results

2.1. Correlation in Phenotypic Traits

All trait relationships studied on 225 ecotypes of *G. soja*, inter-node length to node count, inter-node length to EGR, inter-node length to EPH, node count to EPH, node count to EGR, and EGR to EPH, expressed significant positive correlations, $p < 0.001$, by Spearman's correlation test (Table 1). As a trend, plants with higher growth rate and height had more nodes and the distance between each of the nodes was greater.

Table 1. Phenotypic variation and correlations of studied early vigor traits in wild soybean: Mean, standard deviation (StDev), and coefficient of variance (CV). The units of early plant height (EPH) and inter-node length are mm. Positive significant correlation between traits: early growth rate (EGR), EPH, node count, and inter-node length by Spearman's ρ correlation coefficient.

Trait	Phenotypic Variation			Positive Significant Correlation between Traits		
	Mean	StDev	CV	EPH	Node Count	Inter-Node Length
EGR	16.02	8.34	52.05	0.9861	0.7023	0.8674
EPH	244.68	127.2	51.99	-	0.6911	0.8838
Node Count	4.34	0.93	21.33	-	-	0.3449
Inter-node Length	54.55	24.97	45.76	-	-	-

2.2. Genome-Wide Association Analysis

A total of 31,726 filtered single nucleotide polymorphisms (SNPs) was used for the principle component analysis (PCA) and GWAS. The PCA result was shown in Figure 1C. The Mixed Linear Model (MLM) combined with kinship and structure control resulted in four significant SNPs associated with EGR and 12 significant SNPs associated with EPH, with chromosome-wide false discovery rate (FDR)-adjusted $p < 0.05$ (Table 2). All significant SNPs associated with EGR are also significant for EPH. The most highly associated SNP to EGR and EPH, ss715598271, is significant ($p < 0.05$) by all p-value adjustment and threshold tests (chromosome-wide Bonferroni, genome-wide Bonferroni, and genome-wide FDR). SNP marker ss715598271 explains 10.85% of EGR phenotype variation and 12.42% of EPH phenotype variation. Bracketing markers to ss715598271, ss715598270, and ss715598272 are also shown to be significant by chromosome-wide Bonferroni threshold adjustment in EPH, explaining 8.34% and 8.17% of the phenotypic variation, respectively. Manhattan plots of EGR and EPH (Figure 1A,B) visualize all significant SNPs by chromosome, with corresponding Q-Q plots (Figure 1D,E). The heritability was estimated to be 0.64 for EPH and 0.71 for EGR.

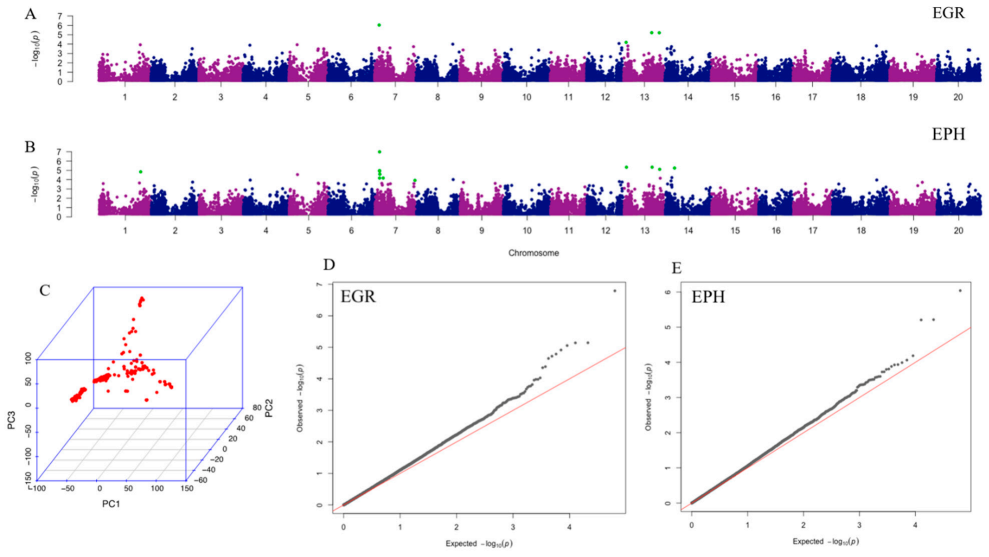

Figure 1. Results of principle component analysis (PCA) and genome-wide association analysis (GWAS). Manhattan plots illustrate the GWAS results of EGR (**A**) and EPH (**B**). X-axis represents single nucleotide polymorphism (SNP) positions across the entire genome by chromosome and the y-axis is the negative logarithm p-value: -log10 (p) of each SNP. Significant SNPs with chromosome-wide false discovery rate (FDR) adjustment are highlighted in green. The corresponding QQ plots for EGR and EPH are shown in (**D**) and (**E**), respectively. For Q-Q plots, x-axis represents expected −log10 (p) and y-axis is observed −log10 (p) of each SNPs. (**C**) Plot of PCA result with all 225 ecotypes. EGR: early growth rate; EPH: early plant height.

Table 2. Significant SNP markers associated with early growth rate (EGR) and early plant height (EPH): The q-value given is the chromosome-wide FDR-adjusted *p*-value. Significance in CB (Chromosome-wide Bonferroni threshold), GB (Genome-wide Bonferroni threshold), and GFDR (Genome-wide FDR adjustment) with $p < 0.05$ are indicated with (*). The position is physical position on the chromosome (Chr.) in base pair.

	SNP	Chr	Position	r^2	Location	Associated Gene	Associated QTL	*p*-value	*q*-value	CB	GB	GFDR
EGR	ss715598271	7	4924020	0.1085	Intron	*Glyma.07G055900.1*	*Plant Height 19-5* [43]	9.14E-07	0.001431	*	*	*
	ss715614175	13	19487316	0.071	Intergenic	-	Plant Height 26-11 [44]	6.50E-05	0.042055	-	-	-
	ss715615103	13	31173270	0.0911	Intergenic	-		6.12E-06	0.006046	*	-	-
	ss715616082	13	39280839	0.0944	5UTR	*Glyma.13G292800.1*		6.23E-06	0.006046	*	-	-
EPH	ss715579500	1	45269059	0.0792	Intergenic	-		2.26E-05	0.028024	*	-	-
	ss715598269	7	4915929	0.0682	Intron	*Glyma.07G055800.1*	Plant Height 19-5 [43]	1.04E-04	0.027144	-	-	-
	ss715598270	7	4918294	0.0834	3UTR	*Glyma.07G055800.1*	Plant Height 19-5 [43]	1.64E-05	0.010022	*	-	-
	ss715598271	7	4924020	0.1242	Intron	*Glyma.07G055900.1*	Plant Height 19-5 [43]	1.62E-07	0.000254	*	*	*
	ss715598272	7	4928272	0.0817	Intron	*Glyma.07G055900.1*	Plant Height 19-5 [43]	1.92E-05	0.010022	*	-	-
	ss715598304	7	5214440	0.0816	Intergenic	-	Plant Height 19-5 [43] Plant Height 3-3 [45] Plant Height 25-6 [46]	4.11E-05	0.016091	-	-	-
	ss715598895	7	8788505	0.0668	Intron	*Glyma.07G094100.1*		1.04E-04	0.027144	-	-	-
	ss715598145	7	42926704	0.0628	CDS	*Glyma.07G251700.1*		1.86E-04	0.041611	-	-	-
	ss715614175	13	19487316	0.09	Intergenic	-	Plant Height 26-11 [44]	7.24E-06	0.007026	*	-	-
	ss715615103	13	31173270	0.0893	Intergenic	-		7.18E-06	0.007026	*	-	-
	ss715616082	13	39280839	0.0874	5UTR	*Glyma.13G292800.1*		1.22E-05	0.007893	*	-	-
	ss715620138	14	9595999	0.0873	Intergenic	-		8.84E-06	0.013826	*	-	-

2.3. In-Depth Candidate Loci Investigation

The SNP ss715598271, which is significantly associated with both EGR and EPH, and its bracketing markers (ss715598269, ss715598270, and ss715598272) associated with EPH, co-localize with a previously identified large QTL, plant height 19-5 [43]. In addition, these four SNPs are within 50 kb adjacent to SNP ss715598304, an intergenic marker significantly associated with EPH, which is located under the known QTLs: plant height 3-3 [45], plant height 25-6 [46], and plant height 19-5 [43] (Table 2). The SNP marker ss715614175 on chromosome 13, significantly associated with both EGR and EPH, located within the previously-described QTL related to plant height: Plant Height 26-11 [44] (Table 2).

The pairwise linkage disequilibrium (LD) analysis of a 65.3-kb window containing the four significant SNPs (ss715598269, ss715598270, ss715598271, and ss715598272) from 4,915,929 bp to 4,928,272 bp on chromosome 7 shows high linkage within this region (Figure 2). Two markers, ss715598269 and ss715598270, are within the intron and 3′ untranslated region of gene *Glyma.07G055800.1*, respectively. Two markers, ss715598271 and ss715598272, are within the intron of gene *Glyma.07G055900.1*. No significant LD is found near the other significant SNPs. These, therefore, were not investigated further for candidate genes.

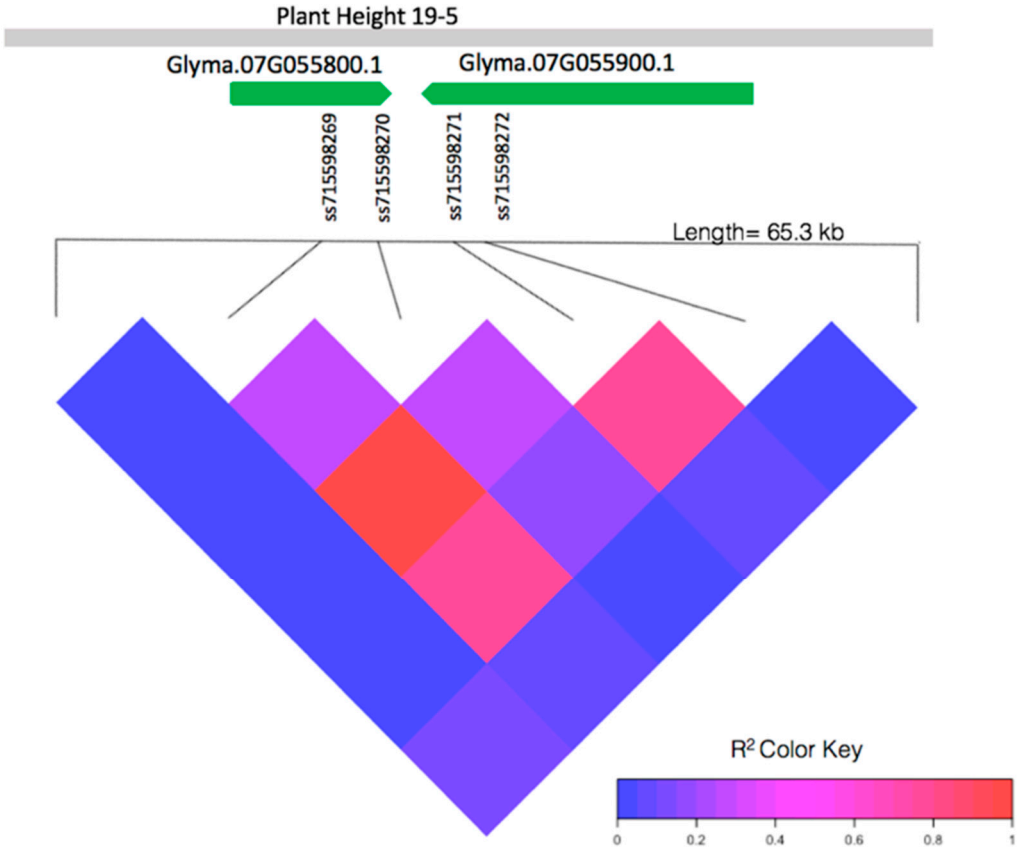

Figure 2. Pairwise linkage disequilibrium (LD) between SNPs in the local genes of interest: *Glyma.07G055800.1*, *Glyma.07G055900.1*, and the known QTL, plant height 19-5 [43], in relation to SNPs ss715598269, ss715598270, ss715598272 (significantly associated with EPH), and ss715598271 (significantly associated with both EPH and EGR).

There is a significant association between ss715598271 allele polymorphism (A/C) and studied traits EPH and EGR. The A allele morph is associated with higher EGR and higher EPH than the C allele morph at this marker ($p < 0.02$) (Figure S1A,B). The significant association between allele and trait suggests that ss715598271 might be located in a region crucial for gene functioning.

Glyma.07G055800.1 and *Glyma.07G055900.1* are promising candidate genes associated with both EPH and EGR. *Glyma.07G055800.1* is predicted to encode a transmembrane protein containing DoH and Cytochrome b-561/ferric reductase transmembrane domains [47]. Its closest homolog in *Arabidopsis* is Cytochrome b561/ferric reductase transmembrane with DOMON-related domain protein (CYB561). This protein is responsible for catalyzing transmembrane electron transfer by ascorbate, which is a key metabolite in growth and development of plants [48]. *Glyma.07G055900.1* is predicted to encode a Tetratricopeptide repeat (TPR)-like super family protein, playing a role in protein binding and translation initiation [47].

3. Discussion

Significant correlations were found between all traits studied, suggesting that EPH, EGR, inter-node length, and node count are all related phenotypic traits (Table 1). The positive relationship between inter-node length and node count, inter-node length and EPH, and internode-length and EGR demonstrates the combined effect of elongation of the inter-nodes along with the addition of new nodes in early growth of the wild soybean. Each node of the wild soybean is accompanied by one branching compound leaf. Therefore, the observed increase in node count with EGR and EPH relates directly to foliage increase and acquisition of available light, having a cumulative effect on growth. The total yield has previously been shown to be positively correlated with the branching of a soybean [10]. A positive correlation between all early vigor traits tested demonstrates that the traits chosen are reliable indicators of early vigor and eventual success of a plant.

We identified a total of 12 significant SNPs for EPH with 4 shared significant SNPs for EPH and EGR. The most compelling results were found on chromosome 7, in which a contiguous set of markers, ss715598269, ss715598270, ss715598271, and ss715598272, are highly significant, revealing the importance of this genomic location to EGR and EPH. This locus also colocalize with a known QTL, plant height 19-5 [43], and is adjacent to two QTLs, plant height 3-3 [45] and plant height 25-6 [46]. Although these consistently mapping QTLs were previously identified, variation in this region might be unique to *G. soja* as previously described in the finding of a *G. soja*-unique salt-tolerant gene *GmCHX1* [49]. Further analysis showed that significant LD is evident in this region. High LD can suggest that this genomic region might have been under positive selection in its natural environment, which is expected as plant height is an ecologically important trait. We were able to determine an allelic correlation with EPH and EGR at marker ss715598271. The A allele morph is correlated with higher EGR and higher EPH than the C allele morph at this marker. This finding suggests that this marker might be linked to genes playing important roles in influencing these trait variations in *G. soja*. This is supported by significant surrounding markers and high LD in this region.

Two candidate genes, *Glyma.07G055800.1* and *Glyma.07G055900.1* on chromosome 7, were identified as significantly related to EPH and EGR. Significant markers are located within the introns of these genes and the untranslated 3' region of *Glyma.07G055800.1*. *Glyma.07G055800.1* encodes a Cytochrome *b561* transmembrane protein, or CYB561 in Arabidopsis, a transmembrane protein involved in electron transport [50]. *Glyma.07G055900.1*, encodes a Tetratricopeptide repeat-like super family protein, homologous to *Reduced Chloroplast Coverage 1* (*REC1*) in Arabidopsis. The Tetratrichopeptide (TPR) domain protein has been shown to be involved in plant height-related phenotypes in Maize and Arabidopsis [51,52] and is a known motif in plant hormone signaling responses, such as auxin, gibberellin, and cytokinin [53,54]. It is not unlikely that multiple individuals of this protein family share a similar influence on traits in various locations in the genome. It has been suggested that *REC1* is involved in establishing and maintaining chloroplast coverage in Arabidopsis and could be manipulated in order to influence energy intake and yield [55]. *G. soja*-type *REC1*, *Glyma.07G055900.1*, is likely involved in similar chloroplast coverage and plant hormone signaling pathways to those found in Maize and Arabidopsis, with a direct relationship to photosynthesis efficiency and growth rate, and therefore is a promising candidate gene to further investigate for crop improvement.

This study identified two candidate genes, *Glyma.07G055800.1* and *Glyma.07G055900.1*, related to early vigor traits, EGR and EPH. These traits are highly beneficial to agriculture and might have been artificially selected for in breeding practices within cultivated soybean *G. max*. However, the extent of adaptation by selection in the cultivated soybean population is limited by the small gene pool. We explored the genetic architecture of these traits in the more diverse wild soybean *G. soja* to further improve our cultivation practices beyond the limited genotype of *G. max*. Further investigation, such as function validation of these candidate genes, will shed light on understanding the molecular mechanisms of these agriculturally and ecologically important traits. Meanwhile, it will provide a foundation for soybean trait improvement by traditional breeding, biotechnology, or genome editing. The long-term goal is to make full use of the wild genetic resources to meet the global challenges of agriculture sustainability and food security.

4. Materials and Methods

4.1. Plant Materials and Phenotyping

In total, 225 *G. soja* accessions obtained from the USDA Soybean Germplasm Collection were used for measurements and analysis. The original geographic distribution of these accessions includes areas in China, Japan, Russia, and South Korea (Figure 3). All seeds were manually scarified and germinated on filter paper for three days, after which 4–5 seedlings per genotype were transplanted into Miracle-Gro soil in separate cells of a 3 × 5 growing tray. Optimal growing conditions were kept constant at 27 °C and 12 h light/day in the greenhouse at University of North Carolina at Charlotte (NC, USA). The plants were watered regularly to keep soil moist.

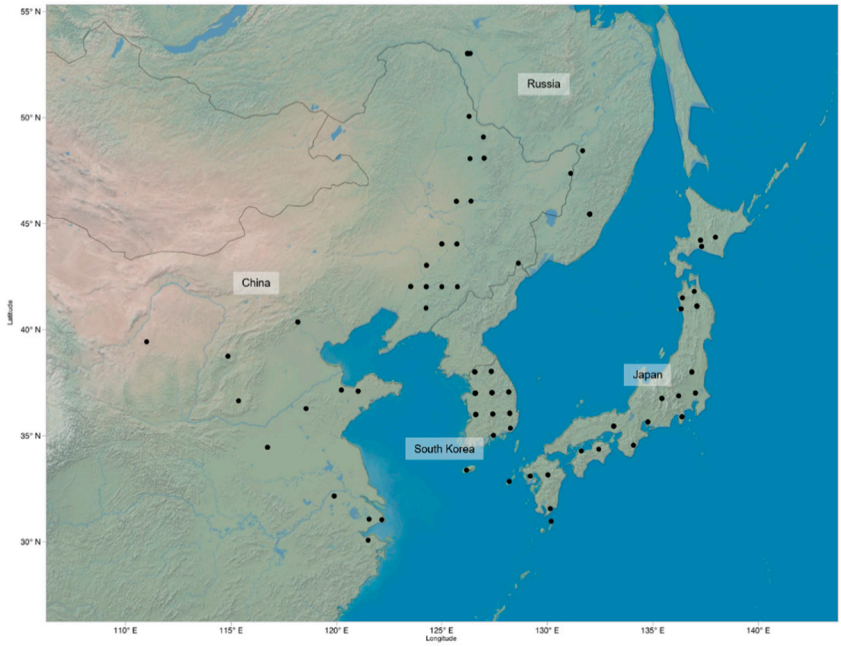

Figure 3. Geographic distribution of known locations of 225 *G. soja* accession: Each point marks a geographic location where the samples were originally collected.

Four phenotypic traits were recorded or measured with a caliper to the nearest 0.1 mm on each accession: node count, internode length, EGR, and EPH. EPH was measured at 20 days after germination, and node count, EGR, and internode length were obtained as the average of three recordings taken at days 7, 14, and 20 after germination. For each accession, measurements from two or three seedlings, quality filtered by coefficient of variance and noted for damage during growth, were averaged for each trait used in the analysis (Table S1).

4.2. Genotypic Data

Previously identified markers, single nucleotide polymorphisms (SNPs), for all 225 *G. soja* accessions were obtained from SoyBase (http://soybase.org) [47]. All markers were originally determined by the use of the Illumina Infinium SoySNP50K iSelect BeadChip, with 52,041 total verified SNP markers [56]. All markers with a minor allele frequency (MAF) <0.05 or missing rate of >10% were filtered out of the analyzed data, leaving 31,726 SNPs. All cleaned data were imputed using BEAGLE (v 3.3.1) [57–59].

4.3. Phenotype Analysis

Pair-wise correlations among the four traits were evaluated with Spearman's analysis due to non-normal distributions to determine the relationship between traits (specifically, the relationship between height and growth traits with the number and frequency of nodes and, therefore, relative foliage quantity). It has been suggested that normalization is not necessary for genome-wide association analysis when the data size is relatively large to avoid error, such as false positives [60]. Thus, the phenotype data was not normalized in this study. The heritability of the quantitative traits was estimated using the following equation: (additive genetic variance)/(additive genetic variance + error variance) [61].

4.4. Genome-Wide Association Analysis

A PCA was conducted using the GAPIT package [62]. A Mixed Linear Model (MLM) in TASSEL [63] was performed to analyze the associations of EGR and EPH with SNPs for all 225 accessions. A principle component (PC) value of three, selected by analysis of quantile-quantile plots at various PC values, and a scaled IBS Kinship matrix was used to control for the population structure. Multiple significance threshold tests were used in order to substantiate the determination of significant SNPs. The significance threshold was determined by chromosome-wide false discovery rate (FDR) resulting in an adjusted p-value (q-value) for each marker [64]. Genome-wide FDR, along with genome-wide and chromosome-wide Bonferroni adjustments [65,66], were also used to validate the significant SNPs at $p < 0.05$ or $q < 0.05$ cutoff.

4.5. Investigation of Candidate Loci

All genes within 50 kb of significant SNPs were evaluated for potential association with each phenotype. The interval was used based on the average linkage disequilibrium reported in *G. soja*. The annotated soybean reference genome, Wm82.a2.v1 (SoyBase, http://soybase.org), was used to determine these genes along with further investigation into the function using Phytozome [67], TAIR [68], and BLAST2GO [69]. Known QTLs for each phenotype from SoyBase [47] in each candidate loci were considered for validation of candidate genes. SNP allele to phenotype comparison was done by nonparametric median tests. Pairwise linkage disequilibrium (LD) was calculated with TASSEL and visualized with the LDheatmap R package [70].

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/21/9/3105/s1.

Author Contributions: B.-H.S. conceived the idea and designed the experiment. J.K. collected data. J.K. and H.Z. analyzed data. J.K., H.Z., and B.-H.S. wrote the manuscript. All authors reviewed the manuscript before submission. All authors have read and agreed to the published version of the manuscript.

Funding: B.-H.S. was funded by the National Institute of General Medical Sciences of the National Institutes of Health, award number R15GM122029; by North Carolina Biotechnology Center, award numbers 2014-CFG-8005 and 2019-BIG-6507; by the North Carolina Soybean Producers Association, award number 18-0252; and by University of North Carolina at Charlotte.

Acknowledgments: We thank the editors and two anonymous reviewers for their insightful and constructive comments. We also thank Leamy L. and Reitzel A. for their invaluable suggestions. We thank Grant D. for his great help with SoyBase.

Conflicts of Interest: The authors declare no competing interests. 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. OECD/FAO. *OECD-FAO Agricultural Outlook 2016–2025*; OECD Publishing: Paris, France, 2016. [CrossRef]
2. Tester, M.; Langridge, P. Breeding technologies to increase crop Production in a changing world. *Science* **2010**, *327*, 818–822. [CrossRef] [PubMed]
3. Kim, M.Y.; Lee, S.; Van, K.; Kim, T.H.; Jeong, S.C.; Choi, I.Y.; Kim, D.S.; Lee, Y.S.; Park, D.; Ma, J.; et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 22032–22037. [CrossRef] [PubMed]
4. Li, Y.H.; Zhou, G.; Ma, J.; Jiang, W.; Jin, L.G.; Zhang, Z.; Guo, Y.; Zhang, J.; Sui, Y.; Zheng, L.; et al. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **2014**, *32*, 1045–1052. [CrossRef] [PubMed]
5. Sedivy, E.J.; Wu, F.Q.; Hanzawa, Y. Soybean domestication: The origin, genetic architecture and molecular bases. *New Phytol.* **2017**, *214*, 539–553. [CrossRef]
6. Carter, T.E.; Hymowitz, T.; Nelson, R.L. Biogeography, local adaptation, vavilov, and genetic diversity in soybean. *Biol. Resour. Migr.* **2004**, 47–59.
7. Li, Y.-H.; Zhao, S.-C.; Ma, J.-X.; Li, D.; Yan, L.; Li, J.; Qi, X.-T.; Guo, X.-S.; Zhang, L.; He, W.-M.; et al. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics.* **2013**, *14*, 579. [CrossRef]
8. Wen, Z.X.; Boyse, J.F.; Song, Q.J.; Cregan, P.B.; Wang, D.C. Genomic consequences of selection and genome-wide association mapping in soybean. *BMC Genomics.* **2015**, *16*. [CrossRef]
9. Kim, M.Y.; Van, K.; Kang, Y.J.; Kim, K.H.; Lee, S.H. Tracing soybean domestication history: From nucleotide to genome. *Breed. Sci.* **2012**, *61*, 445–452. [CrossRef]
10. Akram, S.; Hussain, B.N.; Al Bari, M.A.; Burritt, D.J.; Hossain, M.A. Genetic Variability and Association Analysis of Soybean (*Glycine max* (L.) Merrill) for Yield and Yield Attributing Traits. *Plant. Gene Trait.* **2016**, *7*. [CrossRef]
11. Filho, J.M. Seed vigor testing: An overview of the past, present and future perspective. *Sci. Agr.* **2015**, *72*, 363–374. [CrossRef]
12. Isely, D. Vigor tests. In *Proceedings of the Association of official Seed Analysts*; Society of Commercial Seed Technologists (SCST): Moline, IL, USA; Association of Official Seed Analysts: Ithaca, NY, USA, 1957; pp. 176–182. Available online: https://www.jstor.org/stable/45136792 (accessed on 26 April 2020).
13. Borevitz, J.O.; Nordborg, M. The impact of genomics on the study of natural variation in *Arabidopsis*. *Plant. Physiol.* **2003**, *132*, 718–725. [CrossRef] [PubMed]
14. Klein, R.J.; Zeiss, C.; Chew, E.Y.; Tsai, J.-Y.; Sackler, R.S.; Haynes, C.; Henning, A.K.; SanGiovanni, J.P.; Mane, S.M.; Mayne, S.T.; et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **2005**, *308*, 385–389. [CrossRef] [PubMed]
15. Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: A review. *Plant. Methods* **2013**, *9*, 29. [CrossRef] [PubMed]
16. Brachi, B.; Morris, G.P.; Borevitz, J.O. Genome-wide association studies in plants: The missing heritability is in the field. *Genome Biol.* **2011**, *12*, 232. [CrossRef]
17. Zhang, J.P.; Singh, A.; Mueller, D.S.; Singh, A.K. Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. *Plant. J.* **2015**, *84*, 1124–1136. [CrossRef]
18. Zhang, H.; Li, C.; Davis, E.L.; Wang, J.; Griffin, J.D.; Kofsky, J.; Song, B.H. Genome-wide association study of resistance to soybean cyst nematode (*Heterodera glycines*) HG Type 2.5.7 in wild soybean (*Glycine soja*). *Front. Plant. Sci.* **2016**, *7*, 1214. [CrossRef]
19. Kim, M.; Diers, B. Fine mapping of the SCN resistance QTL *cqSCN-006* and *cqSCN-007* from *Glycine soja* PI 468916. *Crop. Sci.* **2013**, *53*, 775–785. [CrossRef]

20. Anderson, J.E.; Kono, T.J.; Stupar, R.M.; Kantar, M.B.; Morrell, P.L. Environmental association analyses identify candidates for abiotic stress tolerance in *Glycine soja*, the wild progenitor of cultivated soybeans. *G3 (Bethesda)* **2016**, *6*, 835–843. [CrossRef]
21. 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 Genomics.* **2017**, *18*, 18. [CrossRef]
22. Kofsky, J.; Zhang, H.Y.; Song, B.H. The untapped genetic reservoir: The past, current, and future applications of the wild soybean (*Glycine soja*). *Front. Plant. Sci.* **2018**, *9*. [CrossRef]
23. Zhang, H.; Mittal, N.; Leamy, L.J.; Barazani, O.; Song, B.H. Back into the wild—Apply untapped genetic diversity of wild relatives for crop improvement. *Evol. Appl.* **2017**, *10*, 5–24. [CrossRef] [PubMed]
24. Prince, S.J.; Song, L.; Qiu, D.; Maldonado Dos Santos, J.V.; Chai, C.; Joshi, T.; Patil, G.; Valliyodan, B.; Vuong, T.D.; Murphy, M.; et al. Genetic variants in root architecture-related genes in a *Glycine soja* accession, a potential resource to improve cultivated soybean. *BMC Genomics.* **2015**, *16*, 132. [CrossRef] [PubMed]
25. Asekova, S.; Kulkarni, K.P.; Patil, G.; Kim, M.; Song, J.T.; Nguyen, H.T.; Grover Shannon, J.; Lee, J.-D. Genetic analysis of shoot fresh weight in a cross of wild (*G. soja*) and cultivated (*G. max*) soybean. *Mol. Breed.* **2016**, *36*, 103. [CrossRef]
26. Concibido, V.C.; La Vallee, B.; McLaird, P.; Pineda, N.; Meyer, J.; Hummel, L.; Yang, J.; Wu, K.; Delannay, X. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. *Appl. Genet.* **2003**, *106*, 575–582. [CrossRef] [PubMed]
27. Wee, C.-D.; Hashiguchi, M.; Ishigaki, G.; Muguerza, M.; Oba, C.; Abe, J.; Harada, K.; Akashi, R. Evaluation of seed components of wild soybean (*Glycine soja*) collected in Japan using near-infrared reflectance spectroscopy. *Plant. Genet. Resour. Charact. Util.* **2018**, *16*, 94–102. [CrossRef]
28. Zhang, J.; Wang, J.; Jiang, W.; Liu, J.; Yang, S.; Gai, J.; Li, Y. Identification and analysis of NaHCO3 stress responsive genes in wild soybean (*Glycine soja*) roots by RNA-seq. *Front. Plant. Sci* **2016**, *7*, 1842. [CrossRef]
29. Ning, W.; Zhai, H.; Yu, J.; Liang, S.; Yang, X.; Xing, X.; Huo, J.; Pang, T.; Yang, Y.; Bai, X. Overexpression of *Glycine soja WRKY20* enhances drought tolerance and improves plant yields under drought stress in transgenic soybean. *Mol. Breed.* **2017**, *37*, 19. [CrossRef]
30. Lee, J.-D.; Shannon, J.G.; Vuong, T.D.; Nguyen, H.T. Inheritance of salt tolerance in wild soybean (*Glycine soja* Sieb. and Zucc.) accession PI483463. *J. Hered.* **2009**, *100*, 798–801. [CrossRef]
31. Hu, Z.A.; Wang, H.X. Salt tolerance of wild soybean (*Glycine soja*) in populations evaluated by a new method. *Soybean Genet. Newsl.* **1997**, *24*, 79–80.
32. Yang, D.-S.; Zhang, J.; Li, M.-X.; Shi, L.-X. Metabolomics analysis reveals the salt-tolerant mechanism in *Glycine soja*. *J. Plant. Growth Regul.* **2017**, *36*, 460–471. [CrossRef]
33. Zhang, S.; Zhang, Z.; Wen, Z.; Gu, C.; An, Y.C.; Bales, C.; DiFonzo, C.; Song, Q.; Wang, D. Fine mapping of the soybean aphid-resistance genes *Rag6* and *Rag3c* from *Glycine soja* 85-32. *Appl. Genet.* **2017**, *130*, 2601–2615. [CrossRef] [PubMed]
34. Zhang, S.; Zhang, Z.; Bales, C.; Gu, C.; DiFonzo, C.; Li, M.; Song, Q.; Cregan, P.; Yang, Z.; Wang, D. Mapping novel aphid resistance QTL from wild soybean, *Glycine soja* 85-32. *Appl. Genet.* **2017**, *130*, 1941–1952. [CrossRef] [PubMed]
35. Zhang, H.Y.; Song, B.H. RNA-seq data comparisons of wild soybean genotypes in response to soybean cyst nematode (*Heterodera glycines*). *Genom Data* **2017**, *14*, 36–39. [CrossRef] [PubMed]
36. Zhang, H.; Kjemtrup-Lovelace, S.; Li, C.; Luo, Y.; Chen, L.P.; Song, B.H. Comparative RNA-seq analysis uncovers a complex regulatory network for soybean cyst nematode resistance in wild soybean (*Glycine soja*). *Sci. Rep.* **2017**, *7*, 9699. [CrossRef]
37. Yuan, C.; Zhang, L.; Zhao, H.; Wang, Y.; Liu, X.; Dong, Y.; Hartman, G.L. RNA-seq analysis for soybean cyst nematode resistance of *Glycine soja* (wild soybean). *Oil Crop. Sci.* **2019**, *4*, 33–46.
38. Yu, N.; Diers, B.W. Fine mapping of the SCN resistance QTL cqSCN-006 and cqSCN-007 from *Glycine soja* PI 468916. *Euphytica* **2017**, *213*, 54. [CrossRef]
39. Winter, S.M.J.; Shelp, B.J.; Anderson, T.R.; Welacky, T.W.; Rajcan, I. QTL associated with horizontal resistance to soybean cyst nematode in *Glycine soja* PI464925B. *Theor. Appl. Genet.* **2007**, *114*, 461–472. [CrossRef]
40. Wang, D.; Diers, B.W.; Arelli, P.R.; Shoemaker, R.C. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine soja* plant introduction 468916. *Theor. Appl. Genet.* **2001**, *103*, 561–566. [CrossRef]
41. Kabelka, E.A.; Carlson, S.R.; Diers, B.W. Localization of two loci that confer resistance to soybean cyst nematode from *Glycine soja* PI 468916. *Crop. Sci.* **2005**, *45*, 2473–2481. [CrossRef]

42. Hesler, L.S. Resistance to soybean aphid among wild soybean lines under controlled conditions. *Crop. Prot.* **2013**, *53*, 139–146. [CrossRef]
43. Wang, D.; Graef, G.; Procopiuk, A.; Diers, B. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *Theor. Appl. Genet.* **2004**, *108*, 458–467. [CrossRef] [PubMed]
44. Sun, D.; Li, W.; Zhang, Z.; Chen, Q.; Ning, H.; Qiu, L.; Sun, G. Quantitative trait loci analysis for the developmental behavior of soybean (*Glycine max* L. Merr.). *Theor. Appl. Genet.* **2006**, *112*, 665–673. [CrossRef] [PubMed]
45. Mansur, L.; Orf, J.; Chase, K.; Jarvik, T.; Cregan, P.; Lark, K. Genetic mapping of agronomic trait using recombinant inbred lines of soybean. *Crop. Sci.* **1996**, *36*, 1327–1336. [CrossRef]
46. Guzman, P.; Diers, B.W.; Neece, D.; St Martin, S.; LeRoy, A.; Grau, C.; Hughes, T.; Nelson, R.L. QTL associated with yield in three backcross-derived populations of soybean. *Crop. Sci.* **2007**, *47*, 111–122. [CrossRef]
47. Grant, D.; Nelson, R.T.; Cannon, S.B.; Shoemaker, R.C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* **2009**, *38*, D843–D846. [CrossRef]
48. Griesen, D.; Su, D.; Bérczi, A.; Asard, H. Localization of an ascorbate-reducible cytochrome b561 in the plant tonoplast. *Plant. Physiol.* **2004**, *134*, 726–734. [CrossRef]
49. Qi, X.; Li, M.W.; Xie, M.; Liu, X.; Ni, M.; Shao, G.; Song, C.; Kay-Yuen Yim, A.; Tao, Y.; Wong, F.L.; et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nat. Commun* **2014**, *5*, 4340. [CrossRef]
50. Verelst, W.; Asard, H. Analysis of an *Arabidopsis thaliana* protein family, structurally related to cytochromes b561 and potentially involved in catecholamine biochemistry in plants. *J. Plant Physiol.* **2004**, *161*, 175–181. [CrossRef]
51. Weng, J.; Xie, C.; Hao, Z.; Wang, J.; Liu, C.; Li, M.; Zhang, D.; Bai, L.; Zhang, S.; Li, X. Genome-wide association study identifies candidate genes that affect plant height in Chinese elite maize (*Zea mays* L.) inbred lines. *PLoS ONE* **2011**, *6*, e29229. [CrossRef]
52. Lin, Z.; Ho, C.-W.; Grierson, D. AtTRP1 encodes a novel TPR protein that interacts with the ethylene receptor ERS1 and modulates development in *Arabidopsis*. *J. Exp. Bot.* **2009**, *60*, 3697–3714. [CrossRef]
53. Schapire, A.L.; Valpuesta, V.; Botella, M.A. TPR proteins in plant hormone signaling. *Plant. Signal. Behav.* **2006**, *1*, 229–230. [CrossRef] [PubMed]
54. Greenboim-Wainberg, Y.; Maymon, I.; Borochov, R.; Alvarez, J.; Olszewski, N.; Ori, N.; Eshed, Y.; Weiss, D. Cross talk between gibberellin and cytokinin: The Arabidopsis GA response inhibitor SPINDLY plays a positive role in cytokinin signaling. *Plant. Cell* **2005**, *17*, 92–102. [CrossRef] [PubMed]
55. Larkin, R.M.; Stefano, G.; Ruckle, M.E.; Stavoe, A.K.; Sinkler, C.A.; Brandizzi, F.; Malmstrom, C.M.; Osteryoung, K.W. REDUCED CHLOROPLAST COVERAGE genes from *Arabidopsis thaliana* help to establish the size of the chloroplast compartment. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E1116–E1125. [CrossRef] [PubMed]
56. Song, Q.; Hyten, D.L.; Jia, G.; Quigley, C.V.; Fickus, E.W.; Nelson, R.L.; Cregan, P.B. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* **2013**, *8*, e54985. [CrossRef]
57. 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 Genomics.* **2015**, *16*, 217. [CrossRef]
58. Browning, B.L.; Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **2009**, *84*, 210–223. [CrossRef]
59. Browning, B.L.; Browning, S.R. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol. Off. Publ. Int. Genet. Epidemiol. Soc.* **2007**, *31*, 365–375. [CrossRef]
60. Goh, L.; Yap, V.B. Effects of normalization on quantitative traits in association test. *BMC Bioinform.* **2009**, *10*, 415. [CrossRef]
61. Endelman, J.B.; Jannink, J.L. Shrinkage estimation of the realized relationship matrix. *G3 (Bethesda)* **2012**, *2*, 1405–1413. [CrossRef]
62. 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]

63. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef] [PubMed]
64. Leamy, L.J.; Pomp, D.; Eisen, E.; Cheverud, J.M. Pleiotropy of quantitative trait loci for organ weights and limb bone lengths in mice. *Physiol. Genom.* **2002**, *10*, 21–29. [CrossRef] [PubMed]
65. Johnson, R.C.; Nelson, G.W.; Troyer, J.L.; Lautenberger, J.A.; Kessing, B.D.; Winkler, C.A.; O'Brien, S.J. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics.* **2010**, *11*, 724. [CrossRef] [PubMed]
66. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate—A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
67. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **2011**, *40*, D1178–D1186. [CrossRef]
68. Berardini, T.Z.; Reiser, L.; Li, D.; Mezheritsky, Y.; Muller, R.; Strait, E.; Huala, E. The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **2015**, *53*, 474–485. [CrossRef]
69. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef]
70. Shin, J.-H.; Blay, S.; McNeney, B.; Graham, J. LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* **2006**, *16*, 1–10. [CrossRef]