genes

MDPI

Article

Genetic Diversity and Genome-Wide Association Study of Seed Aspect Ratio Using a High-Density SNP Array in Peanut (*Arachis hypogaea* L.)

Kunyan Zou [1,†], Ki-Seung Kim [2,†], Kipoong Kim [3], Dongwoo Kang [1], Yu-Hyeon Park [1], Hokeun Sun [3], Bo-Keun Ha [4], Jungmin Ha [5] and Tae-Hwan Jun [1,6,*]

1 Department of Plant Bioscience, Pusan National University, Miryang 50463, Korea; 601588zky@pusan.ac.kr (K.Z.); kk7ing@pusan.ac.kr (D.K.); eksvnd951@pusan.ac.kr (Y.-H.P.)
2 FarmHannong, Ltd., Nonsan 33010, Korea; leehan26@snu.ac.kr
3 Department of Statistics, Pusan National University, Busan 46241, Korea; kkp7700@gmail.com (K.K.); hsun@pusan.ac.kr (H.S.)
4 Department of Applied Plant Science, Chonnam National University, Gwangju 61186, Korea; bkha@jnu.ac.kr
5 Department of Plant Science, Gangneung-Wonju National University, Gangneung 25457, Korea; j.ha@gwnu.ac.kr
6 Life and Industry Convergence Research Institute, Pusan National University, Miryang 50463, Korea
* Correspondence: thjun76@pusan.ac.kr; Tel.: +82-55-350-5507
† These authors contributed equally to this work.

Abstract: Peanut (*Arachis hypogaea* L.) is one of the important oil crops of the world. In this study, we aimed to evaluate the genetic diversity of 384 peanut germplasms including 100 Korean germplasms and 284 core collections from the United States Department of Agriculture (USDA) using an Axiom_Arachis array with 58K single-nucleotide polymorphisms (SNPs). We evaluated the evolutionary relationships among 384 peanut germplasms using a genome-wide association study (GWAS) of seed aspect ratio data processed by ImageJ software. In total, 14,030 filtered polymorphic SNPs were identified from the peanut 58K SNP array. We identified five SNPs with significant associations to seed aspect ratio on chromosomes Aradu.A09, Aradu.A10, Araip.B08, and Araip.B09. AX-177640219 on chromosome Araip.B08 was the most significantly associated marker in GAPIT and Regularization method. Phosphoenolpyruvate carboxylase (PEPC) was found among the eleven genes within a linkage disequilibrium (LD) of the significant SNPs on Araip.B08 and could have a strong causal effect in determining seed aspect ratio. The results of the present study provide information and methods that are useful for further genetic and genomic studies as well as molecular breeding programs in peanuts.

Keywords: peanut; core collection; genetic diversity; population structure; genome-wide association study; linkage disequilibrium

Citation: Zou, K.; Kim, K.; Kim, K.; Kang, D.; Park, Y.; Sun, H.; Ha, B.; Ha, J.; Jun, T. Genetic Diversity and Genome-Wide Association Study of Seed Aspect Ratio Using a High-Density SNP Array in Peanut (*Arachis hypogaea* L.). *Genes* **2021**, *12*, 2. https://dx.doi.org/10.3390/genes12010002

Received: 22 August 2020
Accepted: 17 December 2020
Published: 22 December 2020

1. Introduction

1.1. Peanut Information

Peanut or groundnut (*Arachis hypogaea* L.) is an important oil and cash crop of the world [1]. Peanut seeds are rich in oil (48–50%) and protein (25–28%) and they contain certain vitamins and minerals which allows them to be used as an energy source for humans [2,3]. In addition, peanuts contain rich functional elements, such as oleic acid, linoleic acid, resveratrol, fiber, and vitamins [4–6].

Since the beginning of agriculture, food grains have been subjected to selection and breeding for size and most of the grains have seeds far larger than their wild relatives [7]. In the United States, peanut seed size is one of the standards used to determine the grade of shelled peanuts and to evaluate the commercial potential of advanced peanut breeding lines prior to the release of varieties [8].

There have been some studies on seed size in peanut. Quantitative trait loci (QTL) study were conducted to identify loci controlling seed size using a 142 backcross population (87 BC3F1 and 55 BC2F2) with two parents under two water regimes in peanut, while several QTLs associated with increased seed width were detected under water-limited treatment [9]. Simple sequence repeat (SSR) marker PM375 associated with seed length was identified in a total of 88 F2:6 recombinant inbred lines (RILs), representing that increase in seed length may influence in an increase in the weight of a hundred seeds, or in the length of the pod [10]. Florida-07 by GP-NC WS 16. A major seed size QTL on chromosome A05 was identified in the US peanut mini core collection using RILs from a cross between Florida-07 and GP-NC WS 16 [8]. However, there are few studies on seed shape in peanuts so far.

1.2. Peanut Germplasms and Core Collection

Various germplasms with large genetic diversity are excellent resources for peanut breeders to broaden the genetic basis of breeding materials and integrate important alleles related to valuable traits [11]. Diverse germplasms in peanuts have been used to enrich genetic resources, introduce resistance to diseases and pests and, finally, to improve the yield potential through continuous breeding programs.

Recently, effective methods to evaluate and introduce a genetic diversity of germplasm resources have been performed in various studies. Core collections were first defined as a limited set of accessions "representing, with a minimum of repetitiveness, the genetic diversity of a crop species and its wild relatives" [12,13]. The use of core collections has many advantages and they also represent a good starting material for association mapping. Recently, core collections have been established in various crops, including rice [14], wheat [15], maize [16], and *Brassica napus* [17]. The peanut core collections were developed from the US germplasm collection [18], and information on the accessions of the core collection are available at the Germplasm Resource Information Network (GRIN) (https://www.ars-grin.gov).

To promote and improve the utilization of germplasm resources in peanut breeding programs, the peanut mini core collection was established by utilizing the stratification strategy of the United States (US) peanut germplasm resource center [19]. The majority of the accessions in the mini core collection were unrelated individuals, which may be a good starting material for initiating the peanut association study. The purpose of establishing a core or mini core collection for any crop is to promote the efficient and economical use of plant materials by end-users and to identify germplasms with desirable characteristics.

The Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences in China established a core collection with 576 *A. hypogaea* genotypes and a mini core collection with 298 accessions representing the majority of the genetic diversity of cultivated peanut in China. They conducted an association study using the mini core collection, and a total of 89 simple sequence repeat (SSR) alleles were identified as associated with 15 agronomic traits. The results showed that there was a great possibility to combine association analysis and marker-assisted breeding using the peanut mini core collection [20,21]. The US mini core collection was evaluated and mapped using quantitative trait loci (QTL) for several traits, such as resistance to Tomato spotted wilt virus (TSWV) [21]. In the ICRISAT mini core collection, several candidate regions associated with non-redundant leaf proteins were identified as being related to tolerance to water deficit stress; however, little has been reported regarding these traits in the US germplasms [22].

1.3. Characteristic of Peanut Genome

Cultivated peanut is allotetraploid (2n = 4× = 40, AABB) with a genome size of 2800 Mb/1C and the genome composition of cultivated peanut was shown to have derived from a recent hybridization of *A. duranensis* (A subgenome) and *A. ipaensis* (B subgenome) [23–26]. As the polyploidization event occurred recently, the genetic diversity of cultivated peanut is extremely low [27]. Peanut subgenomes are very closely related [28,29] and

have an estimated repetition rate of 64% [1], which makes the assembly of peanut genome sequences extremely difficult [1,26,30]. The genome sequences of the diploid ancestors (*A. duranensis* and *A. ipaensis*) of cultivated peanut were reported in 2016, which became the basis for understanding the genome of cultivated peanut [26]. The sequencing results of A. *duranensis* (A genomic progenitor) and A. *ipaensis* (B genome progenitor) provided new insights into the biology, evolution, and genome changes of cultivated peanut and accelerated the molecular breeding of peanut varieties [31].

Recently, the cultivated peanut allotetraploid A. *hypogaea* genome was sequenced in 2019 and compared with the related diploid *A. duranensis* and *A. ipaensis* genomes. A total of 39,888 A subgenome genes and 41,526 B subgenome genes were annotated in the allotetraploid subgenome [32].

1.4. Development of Molecular Markers Using Next Generation Sequencing (NGS) Technology

In 2005, pyrosequencing technology was implemented using large-scale parallel sequencing or deep sequencing, revolutionizing next generation sequencing (NGS) technology and biological genomic research [33]. In the past decade, NGS technology made significant progress, and the cost of sequencing dropped sharply [27]. In addition, there have been innovative improvements in the productivity and accuracy of sequencing data. In particular, genome-wide studies using de novo assembly, resequencing, and a variety of bioinformatic methods have enabled the production of large numbers of single-nucleotide polymorphisms (SNPs) and simple sequence repeats (SSR) in complex genomes [26,34–36]. In recent work, high-throughput genotyping was conducted using NGS technology through double-digest restriction-site-associated DNA sequencing (ddRADseq), a total of 14,663 SNPs were developed, and a genetic linkage map based on SNPs was constructed using 1765 SNP markers in 166 F9 RIL population from a cross between Zhonghua 5 and ICGV86699 [37]. Numerous SNP and cleaved amplified polymorphic sequence (CAPS) markers were developed from the re-sequencing of two Korean peanut germplasms of K-Ol and Pungan, which indicates that the molecular marker information can provide valuable guidance and information for peanut breeding programs [27].

Due to the relatively large genome size and the low genetic diversity in cultivated peanut, developing SNP array chips for high-throughput genotyping is necessary [38]. By DNA resequencing and the RNA sequencing of 41 peanut genetic materials and wild diploid ancestors, a total of 163,782 SNPS were obtained. A total of 58,233 unique SNP sequences with large amounts of information were selected to construct the high-density SNP array Axiom_Arachis with 58K SNPs [39]. The high-density SNP Axiom_Arachis array with 58K SNPs could be used to accelerate the process of high-resolution mapping and molecular breeding in peanuts.

1.5. Applications of High-Density SNP Arrays in Crops

As the most abundant type of DNA sequence variation in the genome, SNPs could be successfully used to associate the genotypic variations with target phenotypes. High-density SNP arrays have been developed for high-resolution mapping of crops and are widely used in many applications that require a large number of molecular markers, such as high-density genetic profiling, genome-wide association study (GWAS), and genomic selection [38,40,41]. One hundred and seven U.S. peanut mini core collections were genotyped using a 58K Affymetrix SNP array and a total of 13,527 highly polymorphic SNP markers were selected for marker-trait associations in arachidic and behenic fatty acid compositions [42]. A total of 2882 polymorphic SNPs retained from the second edition of the Axiom_Arachis array (Axiom_Arachis2) were used to identify loci controlling pod construction trait using 195 F7 recombinant inbred lines (RILs) [43]. The 48K Axiom Arachis2 SNP array was applied to identify single nucleotide polymorphisms (SNP) among the two sets of RILs and the two original Nod+ parental lines to explore the genetic factors and genetic regions controlling nodulation in peanut [44].

Genomic-assisted breeding (GAB) using large amounts of genomic data related to important agronomic traits could be used to develop new varieties faster than when using traditional breeding methods. Detailed genetic maps consisting of thousands of array-based SNPs have been used for the identification of genes controlling target traits [41,45]. GWAS, also known as whole-genome association study, is an observational study of a genome-wide set of genetic variants in different individuals to investigate whether any variant is associated with the target traits [46]. Any phenotypic differences could then be connected back to the underlying causative loci via various mapping approaches, including quantitative trait loci (QTL) mapping. Many research groups have used GWAS to identify associations between genotypes and phenotypes as well as to discover novel biological mechanisms [47]. Currently, most GWAS have been performed using high-throughput SNP data obtained by SNP arrays with a greater density of variants and a wide range of allele frequencies [48–51]. The GWAS format is easy to share and generate, and GWAS can be conducted using various applications and software [46].

1.6. Purpose

In this study, we aimed to (1) evaluate the population structure and genetic diversity of 384 peanut germplasms including 100 Korean germplasms and 284 United States Department of Agriculture (USDA) core collections using Axiom_Arachis array with 58K SNPs, and (2) to conduct GWAS for seed shape and identify candidate genes associated with this trait. Our results could provide useful tools for improving various agronomic traits in molecular breeding programs for peanuts.

2. Materials and Methods

2.1. Plant Materials, DNA Extraction, and Genotyping

A total of 384 peanut accessions were used for the present study (Supplementary Table S1). Among those, 284 peanut accessions were obtained from the core collections of the US Department of Agriculture (USDA) according to the proportion of the number of germplasms, which were widely distributed in East Asia, South Asia, West Asia, East Africa, South Africa, West Africa, North America, South America, Europe, and the Australian continent. In addition, 100 peanut germplasms were obtained from the National Agrobiodiversity Center Korean, RURAL DEVELOPMENT ADMINISTRATION (RDA)-GenBank Information Center, South Korea, including landraces, breeding lines, and cultivars. A young leaf from each individual accession was collected to extract the genomic DNA. A total of 384 peanut genomic DNA were extracted for each accession using the cetyltrimethylammonium bromide (CTAB) protocol with slight modifications [52]. The quality and quantity of the extracted DNA were determined using a NanoDrop ND-1000 (Thermos Fisher Scientific Inc., Wilmington, DE, USA) and 1% agarose gel electrophoresis.

A high-density SNP array Axiom_Arachis with 58K SNPs was used to obtain the genotyping data [39]. Reference genome builds were acquired from arahy.Tifrunner.gnm1.KYV3 (https://www.ncbi.nlm.nih.gov/assembly) to serve as controls in the array design.

2.2. Screening of Seed Aspect Ratio

The seed aspect ratio data (Supplementary Table S2) were obtained by scanning seed images. The scanning images were processed by ImageJ 1.52a software (https://imagej.nih.gov/ij/notes.html) to generate phenotype data for the genome-wide association study. The seed aspect ratio was calculated as the seed major axis divided by the seed minor axis. Ten seeds per accession were scanned at the same time, and the seed aspect ratios of the ten seeds were averaged (Supplementary Figure S1). The phenotype data were analyzed using the R program to conduct a t-test and normal distribution in the accessions.

2.3. Population Structure Analysis

A principle coordinate analysis (PCoA) was conducted using the software GenAlEx V6.503 [53,54]. The population structure of 384 peanut accessions was evaluated by Struc-

ture v2.3.4 software (https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html) under the admixture model. We compared the structures following the same parameters with K-values ranging from 1 to 10, and 20,000 Markov chain Monte Carlo iterations after a burn-in period of 10,000 iterations were carried out for three independent runs per K value. To make a decision for the optimum number for K, the delta K (ΔK) method used the software online "harvester structure".

2.4. Genome-Wide Association Analysis

We analyzed the SNPs in Axiom_Arachis with a 58K array of the cultivated peanut using R software analysis tools. In the present study, the GAPIT package of R software—was used to conduct GWAS, and the enriched compressed mixed linear model (ECMLM) was selected for the analysis of association between SNPs and the phenotype data of interest [55]. The cutoff for significant association was a false discovery rate (FDR) adjusted *p*-value of less than 0.05. Candidate genes covering significantly associated SNPs were selected from the PEANUTBASE website tool (https://www.peanutbase.org) within a 150 kb region upstream or downstream of peak SNPs according to the linkage disequilibrium (LD) decay results.

2.5. Linkage Disequilibrium (LD) Analysis

We performed linkage disequilibrium analysis for all possible pairs of SNPs with a minor allele frequency (MAF) greater than 0.01 in a dataset. To determine the degree of resolution achieved in the association analysis, both the genome and chromosome-wide LDs were estimated [56].

LD blocks were viewed using Haploview4.2, which uses permutation tests to determine the *p*-values for each pairwise correlation. The LD decay was calculated with PopLDdecay [57,58]. The physical distance of the LD decay plot was determined based on the *D'* values and distances between each pair of SNPs on each chromosome using a nonlinear model [59].

The standard descriptive LD parameter D' was estimated as previously described by [60,61]. The average D' value was calculated for each chromosome using Haploview software [60].

2.6. Regularization Method

In human genome-wide association studies, regularization methods based on penalized likelihood are popular regarding their application to identify disease-related genes or genetic regions as they are computationally efficient when used in analysis of high-dimensional genomic data [62–68]. The penalized likelihood function using an elastic-net penalty is defined as

$$Q(\beta) = -l(\beta) + \lambda\alpha \sum_{j=1}^{p} |\beta_j| + \lambda(1-\alpha) \sum_{j=1}^{p} \beta_j^2 \quad (1)$$

where $l(\beta)$ is a log-likelihood function, β is the *p*-dimensional coefficient vector, $\lambda \geq 0$ is a tuning parameter for sparsity, and $\alpha \in [0,1]$ is a tuning parameter for smoothness. When $\alpha = 1$, the coefficient vector β becomes the solution of the least absolute and shrinkage selection operator (LASSO) [69]. The estimated coefficient β consists mostly of zero values and only a few nonzero values. Based on 100 bootstrap samples, the selection probability of individual SNPs was computed where only SNPs with nonzero coefficients were selected for each bootstrap sample. Finally, we were able to identify the top ranked SNPs by their selection probability.

In order to select significant SNPs, we used two types of threshold of selection probability which can control the number of falsely selected SNPs. The first one is the theoretical threshold proposed by [70]. The second one is the empirical threshold [71] which basically computes the quantile value of an empirical distribution of selection probability based on permutation. In their extensive simulation studies, it was demonstrated that the number of falsely selected SNPs can be controlled when the empirical threshold is applied to high-

dimensional genomic data. The theoretical threshold (π_θ) and the empirical threshold (π^*_θ) can be written as:

$$\pi_\theta = \frac{q^2_\Lambda}{2\theta p} + \frac{1}{2} \text{ and } \pi^*_\theta = \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} SP^{[\theta]}_{(b)}(I_b)\,, \tag{2}$$

where θ is the upper bound of the expected number of false discoveries, q_Λ is the average number of selected SNPs, B is the number of permutations and I_b is the b-th random permuted sample. We denote $SP^{[\theta]}_{(b)}$ by the top θ-th ranked selection probability when they were sorted in descending order for the b-th permuted sample such as $SP^{[1]}_{(b)} > \cdots > SP^{[p]}_{(b)}$. We chose the expected number of false discoveries $\theta = 1$, and thereby the number of falsely selected SNPs by each threshold can be guaranteed to be less than $\theta = 1$.

3. Results

3.1. SNP Genotyping

Of the 58K informative SNPs, a total of 47,837 polymorphic SNPs were selected (Supplementary Table S3). Of the 47,837 SNPs, 19,554 and 21,876 SNPs were derived from the subgenomes A and B, respectively, and 6407 SNPs were derived from scaffolds (Supplementary Table S3 and Figure 1a). A total of 14,030 SNPs were selected for association analysis after eliminating SNPs with high levels of missing data (>20%), heterozygosity (>20%), or low a minor allele frequency (MAF) (<0.01). Of the 14030 SNPs, 6623 and 7407 SNPs were derived from A and B subgenomes, respectively. The majority of SNPs were evenly distributed across the chromosomes; however, there were some large gaps between SNPs on the chromosomes Aradu.A09, Aradu.A10, Araip.B05, Araip.B06, Araip.B07, Araip.B09, and Araip.B10 (Figure 1b). The peanut genome had an overall SNP density of 5.91 SNPs/Mb, with the Aradu.A09 (3.45 SNPs/Mb) and Aradu.A08 (9.35 SNPs/Mb) chromosomal densities being the lowest and highest, respectively (Supplementary Table S4).

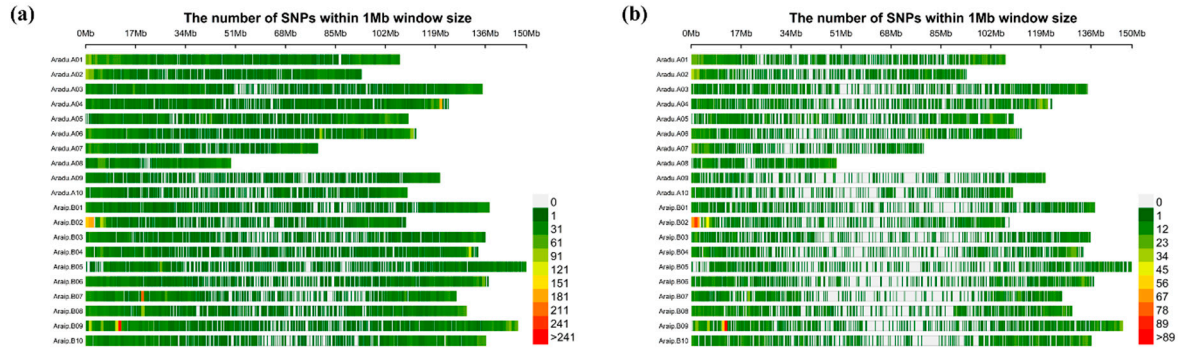

Figure 1. Single nucleotide polymorphisms (SNP) distribution in the 20 chromosomes of the cultivated peanut. The horizontal axis shows chromosome length (Mb), the shades of red represent SNP density. The vertical axis shows the 20 chromosomes. (**a**) Polymorphic SNPs except for scaffold markers; (**b**) Polymorphic SNPs (except for scaffold markers) after filter by GAPIT coding.

3.2. Phenotype Data Analysis

The mean value for the seed aspect ratio was 1.6325 (Figure 2a). The normal distribution test showed that the scatter points of the quantile–quantile (QQ plot) graph (Figure 2b) were clustered around the fixed line; therefore, we assumed that the data were normally distributed (p = 0.05).

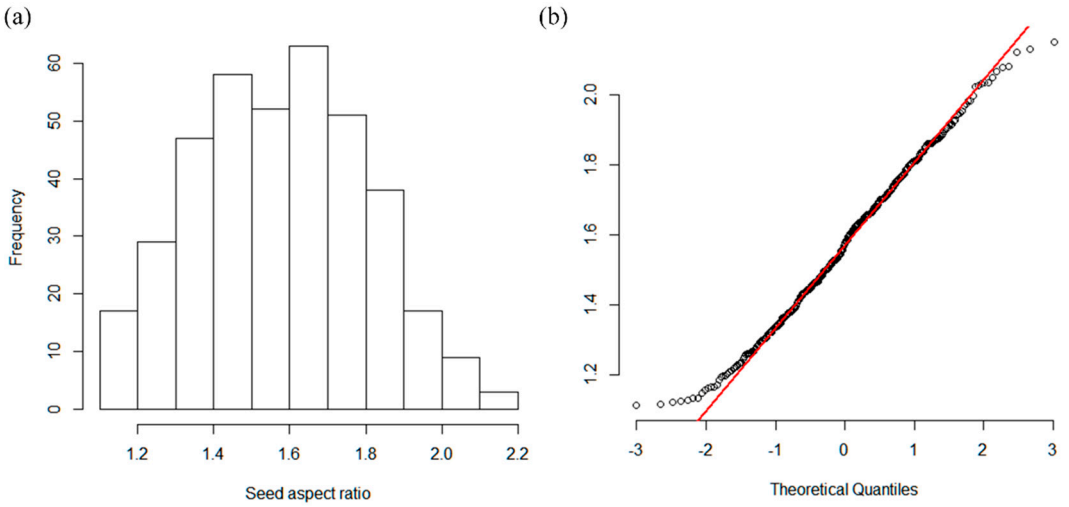

Figure 2. (**a**) Peanut seed aspect ratio data histogram; (**b**) The normal distribution test by the quantile-quantile (QQ plot) graph.

3.3. Genetic Diversity

The pattern of PCoA (Figure 3) showed that the first two axes accounted for 30.19% and 6.91%, respectively, of the total variation and the 384 peanut accessions were divided into three broad groups across the first two axes. The first axes separated the South Korean (clustered filled diamonds) and South American (green filled squares) peanut accessions into two very different parts, and, at the same time, assigned East Asian, South Asian, and West Asian peanut accessions (brown filled triangles, pink filled diamonds, and green filled circles, respectively) to another part. Additionally, the peanut accessions that originated from East Africa, South Africa, West Africa, North America, and Europe formed two concentrated groups by the first and second axes. Interestingly, the accessions from South Korea were genetically very different from those from South America, which is the origin of the cultivated peanut.

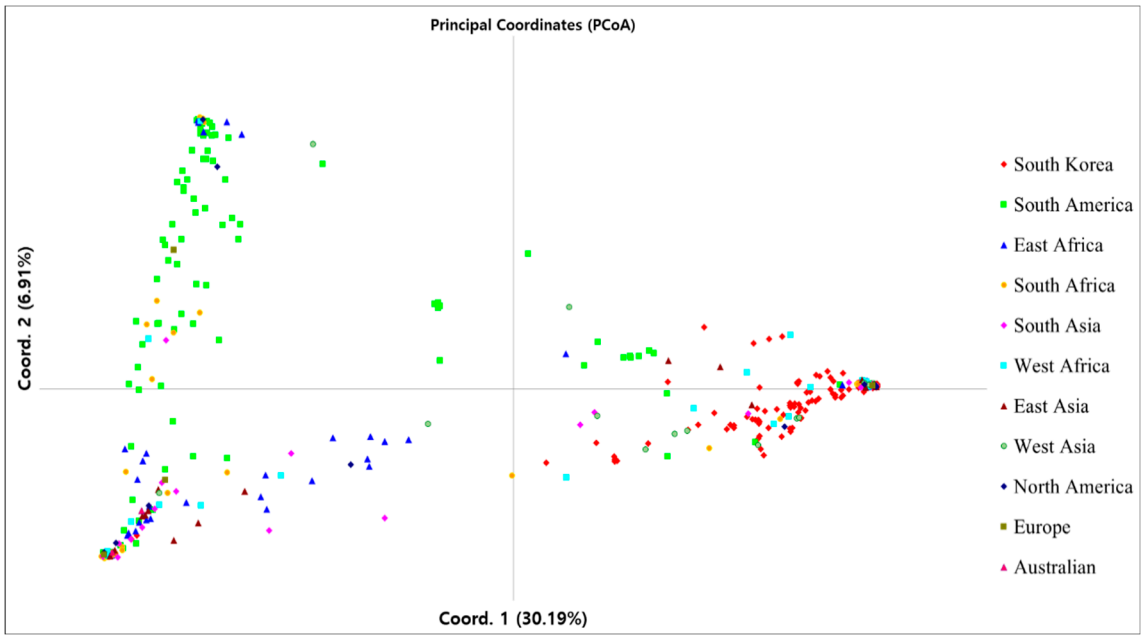

Figure 3. The pattern of Principal Coordinates Analysis (PCoA).

3.4. Genetic Structure

At K = 2, we found maximum Δk values that were plotted against the K to confirm the number of populations, while another lower peak was shown at K = 7 (Supplementary Figure S2). When most individuals were divided into the two subpopulations (K = 2, Figure 4), the peanut accessions, including 64.9% from Asia (of which approximately 74% individuals were from South Korea and 26% from other origins in Asia), 24.4% from Africa, 10.2% from South America, and 0.5% from Europe, belonged to one subgroup (red), while another subgroup (green) revealed features of accessions, including 16.8% from Asia (comprising about 6.7% from South Korea and 93.3% from other Asia origins), 35.2% from Africa, 42.5% from South America, 2.8% from North America, 1.7% from Europe, and 1% from Australia.

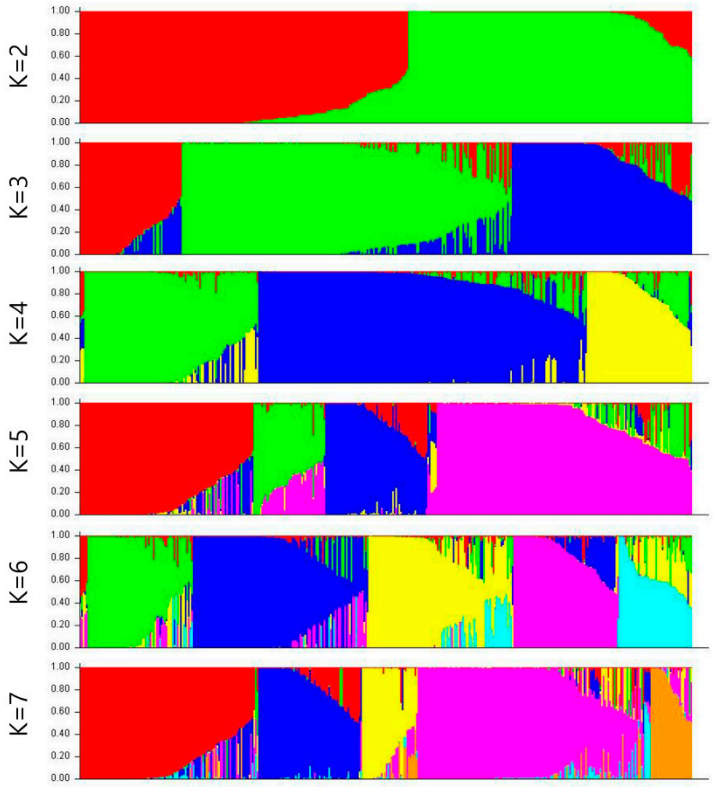

Figure 4. Structure clustering results obtained at K = 2 to K = 7 of the 384 peanut accessions. Each individual is represented by a bar corresponding to the sum of assignment probabilities to the K cluster.

As we continued to divide the subgroups carefully, there were new divisions into the subgroups. The most divergent subgroups were formed at K = 7. Of the peanut accessions, 26.4% originating from Asia (of which approximately 10.3% were from South Korea and 89.7% from other Asia origins), 45.5% from Africa, 20.9% from South America, 3.6% from North America, 1.8% from Europe, and 1.8% from Australia belonged to the red subgroup. The green subgroup revealed features of 50% accessions from South Korea and 50% from Africa. The dark blue subgroup showed features of 1.5% accessions from Asia, 20% from Africa, 75.5% from South America, 1.5% from North America, and 1.5% from Europe. The yellow subgroup showed features of 91.4% accessions from Asia (including about 87.5% from South Korea and 12.5% from other Asia origins), 2.9% from Africa, and 5.7% from South America. The pink subgroup consisted of 56.3% accessions from Asia (of which approximately 63.8% were from South Korea, and 36.2% from other Asia origins), 31% from Africa, 12% from South America, and 0.7% from Europe. The light blue subgroup showed features of accessions from only South America. The orange subgroup showed features of individuals with 76.9% of accessions from Asia (of which approximately 85%

accessions were from South Korea and 15% from other Asia origins), 15.4% from Africa, and 7.7% from South America.

3.5. Genome-Wide Association Study (GWAS)

The genotype data of 14,030 filtered polymorphic SNPs and the phenotypic data of the seed aspect ratios were analyzed for GWAS by GAPIT. A total of five candidate SNPs showing significant associations ($p < 0.0001$) with the seed aspect ratio were identified on chromosomes Aradu.A09, Aradu.A10, Araip.B08, and Araip.B09 (Table 1 and Figure 5a). The distribution of the observed −log10(p) for each SNP was compared with the expected distribution in the QQ plot representing that the population structure and kinship relationship were well controlled in the GWAS (Figure 5b). The significance of the marker–trait associations were determined using the FDR with adjusted p-value ($p = 0.05$). AX-177640219 on chromosome Araip.B08 was significantly associated with the seed trait at the significant threshold (Table 1).

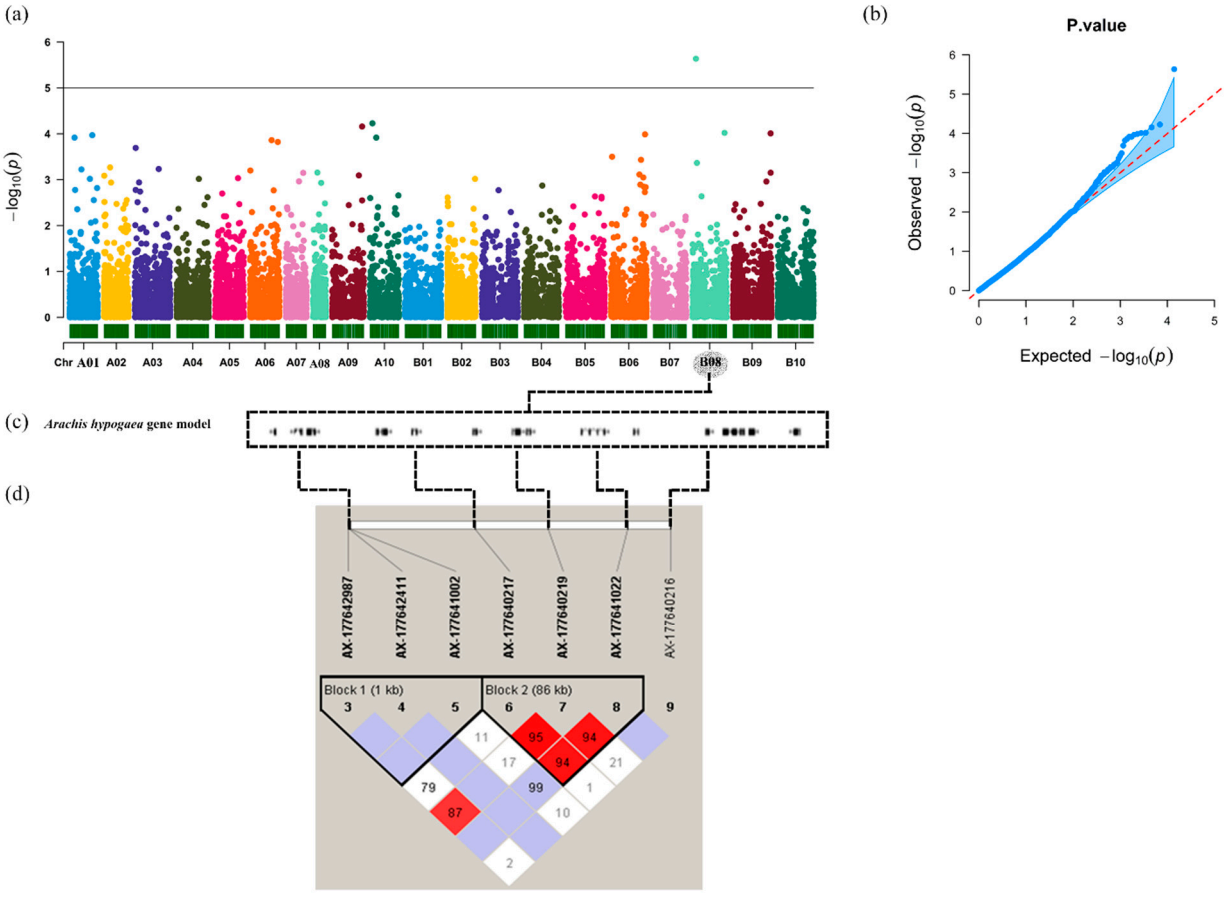

Figure 5. (**a**) Manhattan plot of a genome-wide association analysis by GAPIT; (**b**) Q-Q (quantile-quantile) plot; (**c**) Genes located at the association regions based on the *Arachis hypogaea* Tifrunner 1.0 reference genome; (**d**) Linkage disequilibrium (LD) plot generated using Haploview, D′ values that correspond to SNP pairs are shown within the respective squares. Higher D′ values are indicated with a brighter red color.

Table 1. Significant markers associated with seed aspect ratio of peanut identified using GAPIT analysis.

SNP	Chromosome	Position (bp)	*p*-Value (*p*)	FDR_Adjusted_*p*-Values
AX-177640219	Araip.B08	12829161	2.31×10^{-6}	0.032
AX-147235444	Aradu.A10	8911644	5.91×10^{-5}	NS [a]
AX-176807953	Aradu.A09	113907685	6.95×10^{-5}	NS
AX-176822392	Araip.B08	121783058	9.55×10^{-5}	NS
AX-147262340	Araip.B09	143554366	9.80×10^{-5}	NS

[a] FDR_adjusted_*p*-value is not significant at the level of 0.05.

3.6. LD and Candidate Genes Analysis

Pairwise comparisons were performed between all SNPs for the estimation of LD decay. At a cutoff value of $r^2 = 0.1$, the averaged LD decay distance of the 384 peanuts was approximately 150 kb (Supplementary Figure S3). The pattern of LD across the entire genome presented a number of haplotype blocks containing SNPs that can be used to determine the range of the candidate gene. The genomic locations harboring significant SNPs from the GWAS were investigated to identify putative candidate genes based on the peanut reference genome (*A. hypogaea* Tifrunner 1.0). Strong and extensive pairwise LD was observed among highly significant SNPs around AX-177640219 (*p*-value = 0.000015) on chromosome Araip.B08 from the 12,629,161 to 13,029,161 bp region ($D' > 0.80$) in which D' varied from 0.036 to 1 (Figure 5d and Supplementary Table S5).

Fifteen annotated genes at the association regions flanked by SNP AX-177640219 on chromosome Araip.B08 were identified within the estimated ±150 kb window based on the reference genome (Figure 5c and Supplementary Table S6).

3.7. Regularization Method

Alternatively, we also conducted regularization methods, such as LASSO, to identify candidate regions associated with the seed aspect ratio (Figure 6) [71]. The regularization method was performed using an entire dataset at a time and could select several putative markers most likely related to the trait based on the value of selection probability, whereas the ECMLM analysis only tested one marker at a time. As a result, one SNP locus (AX-177640219 on Araip.B08) was identified as being most likely related to the seed aspect ratio based on the selection probability at the permuted threshold 0.894, and was also found to be highly significantly associated in the GAPIT analysis (Figure 6). When loosening the strict threshold to 0.506, a total of six SNPs were additionally identified, AX-177640938 on chromosome Araip.B08, AX-147218661 on Aradu.A03, AX-147251864 on Araip.B06, AX-176802342 on Araip.B04, AX-176791478 on Aradu.A02, and AX-176800768 on Aradu.A01, which presented significant associations with ECMLM results indicating that the regions flanked with these markers might be candidate regions for possible determination of seed shape in peanuts. Therefore, the use of both methods to conduct association studies is beneficial in (1) boosting confidence in the case where common markers are identified and (2) to maximize the possibility of finding new significant markers associated with a trait of interest.

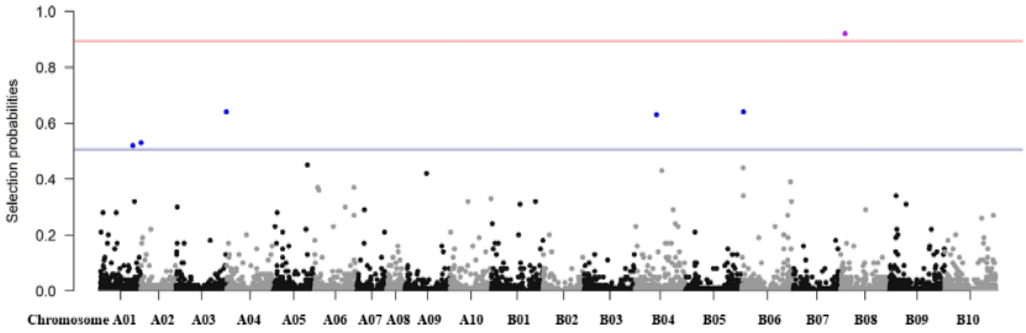

Figure 6. Manhattan plot of a genome-wide association analysis by the least absolute and shrinkage selection operator (LASSO).

3.8. Evaluation of Heterozygous Rate

The same filtering conditions with maximum missing data of 20% and MAF of 0.01%, different heterozygosity rates (starting from 5% and 10% and every 10% until 100% maximum heterozygous SNPs (Figure 7 and Supplementary Table S7) were used to filter the genotype data in our study, and different significance cutoff thresholds were used to assess the effect of the SNPs on the seed aspect ratio.

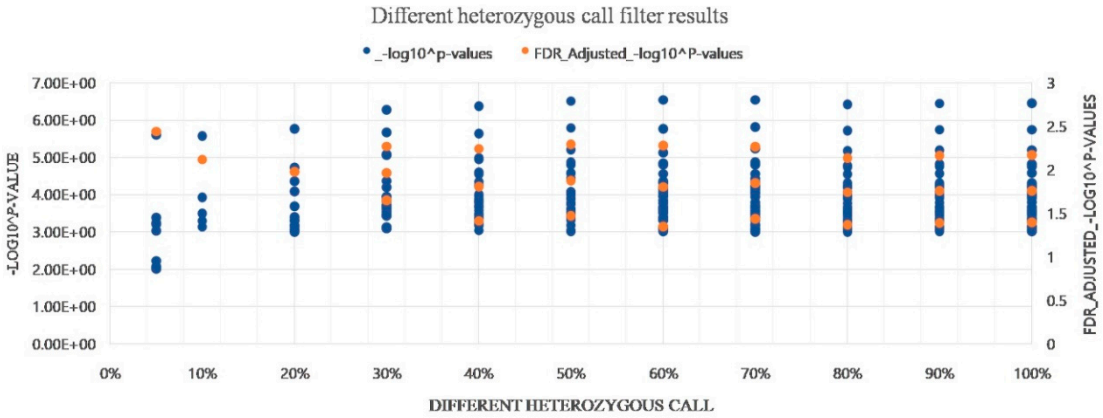

Figure 7. Different heterozygosity rates to filter genotype data and different significance cut-off thresholds used to assess the effect of the SNPs on seed aspect ratio.

When the genotype data filtered by a 5% to 20% maximum heterozygous rate were used for GWAS analysis, a higher specificity of the results was obtained; however, only one significance marker was evaluated at the 0.05 critical threshold for the false discovery rate (FDR) adjusted p-value. On the contrary, when a high heterozygosity rate of 30% to 100% was used for data filtering, additional significant markers were detected; however, those markers require validation.

4. Discussion

The trait of seed ratio (length-width ratio) screened in this study has been reported to have very high broad-sense of heritability in recently published peanut research. Zhang et al. [72] reported that it has a high broad-sense of heritability (0.81) in peanuts. For other legume crops, Hu et al. [73] reported a very high broad-sense of heritability ranged from 92.46 to96.25 in three traits related to seed shape in soybeans. If a phenotypic trait has a high level of heritability, the influence of the environmental factors might be relatively small, and in this case, it could be possible that genes (or QTL) with relatively large effects on the trait could be identified even if the trait were not measured in the same conditions. Of course, even in this case, the influence of the environmental conditions cannot be overlooked.

The genotyping data from the 58K SNP array chip could play an important role in understanding the evolutionary history of peanuts and the domestication of cultivated peanut [74]. The application of the array chip also demonstrated that it is a powerful and reliable tool for peanut germplasm background selection and evolutionary studies [75]. In the present study, it is the first to conduct GWAS analysis using a large number of Korean peanut germplasms as well as the USDA peanut core collection with a high-density SNP chip data that can be used toward increasing the genetic diversity of the US peanut germplasm collection.

The cultivated peanut species (A. *hypogaea*) is known to originate from southern Bolivia to northwestern Argentina based on the occurrence of the two progenitor species, A. *duranensis* and A. *ipaensis*, and archaeological evidence gathered in those regions [76–78]. Researchers also suggested that the eastern slopes of Cordillera may be a possible area for the origin of A. *hypogaea* due to the favorable environment for peanut growth [78,79]. However, the present study showed an interesting result in that South American peanuts, generally regarded as the origin of peanuts, were revealed as having significant genetic differences from peanuts of other regions, including South Korea.

The evaluation results for the evolutionary relationships among the entirety of the 384 peanut germplasms indicated that most of the peanut individuals from South Korea and South America separated into two distinct groups and were also independent from the peanuts from the other origins. This might indicate that there was a great genetic difference between the peanut germplasms from South Korea and South America. Likely, due to the lack of interactions between South Korean peanut germplasms and others, it might be possible that an independent breeding history by human selection and/or environmental influences for a long period have caused these genetic differences.

In human genetic association studies with high-dimensional genomic data, regularization methods, such as LASSO and elastic-net, have been widely applied to identify outcome-related genetic sites and genes as they have certain advantages over univariate analysis. First, regularization methods can easily handle highly correlated genomic measurements and covariate effects as they are based on a regression model. Secondly, the majority of regularization methods have been implemented into very efficient computational algorithms such R package 'glmnet' and 'gglasso'. These packages can detect outcome-related genetic-sites and genes in less than a minute for more than 100K dimensional genomic data. Lastly, there are various types of regularization methods that can be applied to different types of genomic data. For example, we applied LASSO and elastic-net to SNP data in the GWAS or QTL analysis; however, sparse group LASSO [80] and network-based regularization [81] are ideal for group structured genomic data, such as gene expression data and DNA methylation data. Despite these advantages of regularization methods, they have rarely been applied to detect QTLs or genes of interest in crops. In this study, LASSO was able to identify potentially outcome-related SNPs that were not identified in general GWAS methods although further validation studies are required for these SNPs.

Data filtering is the primary process of genome-wide association analysis, which includes huge amounts of data and requires strict quality control standards. Data filtering is divided into two sections, one for marker variables and another for individuals. The former considers the minor allele frequency (MAF) and the degree of missing data and heterozygosity, etc., whereas the latter mostly considers missing levels, population stratification, and independency among individuals [82]. The entire set of heterozygous SNPs are typically used in human GWAS analysis [83]. In peanuts, a high level of heterozygosity may not be expected as peanut is a self-pollinating crop revealed to have a low outcrossing rate ranging from 1.9% to 8% [84]. However, our array chip data showed a large number of heterozygous SNPs, which can affect the GWAS results. According to Figure 7, the significant SNPs identified from using 5% to 20% maximum heterozygous rate showed the same GWAS results, with one significant marker at FDR 0.05, while the results from using 30% to 100% maximum heterozygous rate showed similar GWAS results with three

significant markers. Therefore, we filtered the genotype data with maximum heterozygous SNPs of 20%, and we used less heterozygous SNPs for analysis.

Carbon assimilated by photosynthesis is transported into seeds with multiple purposes, such as the biosynthesis of starch, oil, amino acids, and cellulose. The most important aspect of oil accumulation in developing seeds lies in the activation of metabolic pathways driving incoming carbon into fatty acid biosynthesis at the expense of competitive pathways. Within the genomic region of ~300 kb associated with seed development, phosphoenolpyruvate (PEP) carboxylase (PEPC; Arahy.HT9EWH) was among eleven genes located within the LD of significant SNPs on the chromosome Araip.B08 (Supplementary Table S6). PEP is catalyzed into oxaloacetate (OAA), a protein precursor, by PEPC [85]. OAA can be converted to malate and then to pyruvate (a precursor for oil). PEPC had been reported to regulate the metabolic network of glycolytic carbon into precursors for both oil and protein in soybean seed development [86]. The activation status of PEPC has been reported to play a key role in the partitioning of assimilates into the different storage products in barley (*Hordeum vulgare*), alfalfa (*Medicago sativa*), and fava bean (*Vicia faba*) [87–89]. In peanuts, researchers reported that the expression levels of PEPC genes were significantly associated with lipid accumulation [90]. In the present study, only fifteen annotated genes were identified within the genomic region as being highly associated with seed development through high-throughput GWAS analysis. Among them, the PEPC gene could have a strong causal effect within this region associated with diverse metabolic pathways that includes including protein and oil biosynthesis.

5. Conclusions

Peanut is one of the most important food/oil crops and improving the quality and yield potential of crops is an important challenge in most breeding programs. Our study demonstrated the feasibility of GWAS analysis using the core germplasm from diverse origins and high-density array chips. Five candidate markers with a significant correlation with the aspect ratio of peanut seeds were identified and lay a foundation for further research. The Arahy.HT9EWH, phosphoenolpyruvate carboxylase (PEPC) gene corresponding to the most significantly associated marker was a promising candidate gene that is involved in many metabolic pathways, including those involved in seed development processes. Therefore, the results of the present study provide valuable information and methods for the genetic and genomic study as well as molecular breeding programs in peanuts.

Supplementary Materials: The following are available online at https://www.mdpi.com/2073-4425/12/1/2/s1, Figure S1: ImageJ protocol, Figure S2: Δk values to confirm the number of populations, Figure S3: The estimation of LD decay, Table S1: Information on the 384 peanut accessions, Table S2: Information on the 384 peanut accessions seed aspect ratio, Table S3: A high-density SNP array 'Axiom_Arachis' with 58K SNPs, Table S4: Information of the peanut genome from 58K SNP array chip, Table S5: Pairwise LD on chromosome Araip.B08, Table S6: List of genes in the significant region (chromosome Araip.B08) with annotations identified by *Arachis hypogaea* Tifrunner 1.0 reference, Table S7: Genome-wide association study (GWAS) (p-value < 0.001) of seed aspect ratio were constructed by different heterozygous filter methods.

Author Contributions: Conceptualization, T.-H.J.; methodology, T.-H.J. and H.S.; software, K.Z. and K.K.; formal analysis, K.Z. and K.K.; investigation, K.Z., D.K. and Y.-H.P.; resources, T.-H.J.; data curation, J.H. and K.-S.K.; writing—original draft preparation, K.Z. and K.-S.K.; writing—review and editing, B.-K.H., H.S., K.-S.K. and T.-H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was carried out with the support of the "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ013125022020)" Rural Development Administration, Republic of Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dhillon, S.S.; Rake, A.V.; Miksche, J.P. Reassociation kinetics and cytophotometric characterization of peanut (*Arachis hypogaea* L.) DNA. *Plant Physiol.* **1980**, *65*, 1121–1127. [CrossRef] [PubMed]
2. Win, M.M.; Abdul-Hamid, A.; Baharin, B.S.; Anwar, F.; Sabu, M.C.; Pak-Dek, M.S. Phenolic compounds and antioxidant activity of peanut's skin, hull, raw kernel and roasted kernel flour. *Pak. J. Bot.* **2011**, *43*, 1635–1642.
3. Pasupuleti, J.; Nigam, S.N.; Pandey, M.K.; Nagesh, P.; Varshney, R.K. Groundnut improvement: Use of genetic and genomic tools. *Front. Plant Sci.* **2013**, *4*, 23. [CrossRef]
4. Radhakrishnan, R.; Pae, S.B.; Lee, B.K.; Baek, I.Y. Evaluation of luteolin from shells of Korean peanut cultivars for industrial utilization. *Afr. J. Biotechnol.* **2013**, *12*, 4477–4480. [CrossRef]
5. Musa, Ö.M. Some nutritional characteristics of kernel and oil of peanut (*Arachis hypogaea* L.). *J. Oleo Sci.* **2010**, *59*, 1–5. [CrossRef]
6. Sales, J.M.; Resurreccion, A.V. Resveratrol in peanuts. *Crit. Rev. Food Sci. Nutr.* **2014**, *54*, 734–770. [CrossRef]
7. Sundaresan, V. Control of seed size in plants. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17887–17888. [CrossRef]
8. Chu, Y.; Chee, P.; Isleib, T.G.; Holbrook, C.C.; Ozias-Akins, P. Major seed size QTL on chromosome A05 of peanut (Arachis hypogaea) is conserved in the US mini core germplasm collection. *Mol. Breed.* **2020**, *40*, 6. [CrossRef]
9. Fonceka, D.; Tossim, H.A.; Rivallan, R.; Vignes, H.; Faye, I.; Ndoye, O.; Rami, J.F. Fostered and left behind alleles in peanut: Interspecific QTL mapping reveals footprints of domestication and useful natural variation for breeding. *BMC Plant Biol.* **2012**, *12*, 26. [CrossRef]
10. Gomez Selvaraj, M.; Narayana, M.; Schubert, A.M.; Ayers, J.L.; Baring, M.R.; Burow, M.D. Identification of QTLs for pod and kernel traits in cultivated peanut by bulked segregant analysis. *Electron. J. Biotechnol.* **2009**, *12*, 3–4. [CrossRef]
11. Zhao, Z.; Tseng, Y.C.; Peng, Z.; Lopez, Y.; Chen, C.Y.; Tillman, B.L.; Dang, P.; Wang, J. Refining a major QTL controlling spotted wilt disease resistance in cultivated peanut (*Arachis hypogaea* L.) and evaluating its contribution to the resistance variations in peanut germplasm. *BMC Genet.* **2018**, *19*, 17. [CrossRef] [PubMed]
12. Arber, W.; Illmensee, K.; Peacock, W.J.; Starlinger, P. (Eds.) *Genetic Manipulation: Impact on Man and Society (No. 1)*; Cambridge University Press: Cambridge, UK, 1984.
13. Frankel, O.H.; Brown, A.H.D. *Current Plant Genetic Resources—A Critical Appraisal*; Oxford and IBH Publishing Co.: New Delhi, India, 1984.
14. Zhang, H.; Zhang, D.; Wang, M.; Sun, J.; Qi, Y.; Li, J.; Han, L.; Qiu, Z.; Tang, S.; Li, Z. A core collection and mini core collection of *Oryza sativa* L. in China. *Theor. Appl. Genet.* **2011**, *122*, 49–61. [CrossRef] [PubMed]
15. Hao, C.Y.; Zhang, X.Y.; Wang, L.F.; Dong, Y.S.; Shang, X.W.; Jia, J.Z. Genetic diversity and core collection evaluations in common wheat germplasm from the Northwestern Spring Wheat Region in China. *Mol. Breed.* **2006**, *17*, 69–77. [CrossRef]
16. Yang, X.; Gao, S.; Xu, S.; Zhang, Z.; Prasanna, B.M.; Li, L.; Li, J.; Yan, J. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* **2011**, *28*, 511–526. [CrossRef]
17. Xiao, Y.; Cai, D.; Yang, W.; Ye, W.; Younas, M.; Wu, J.; Liu, K. Genetic structure and linkage disequilibrium pattern of a rapeseed (*Brassica napus* L.) association mapping panel revealed by microsatellites. *Theor. Appl. Genet.* **2012**, *125*, 437–447. [CrossRef]
18. Holbrook, C.C.; Anderson, W.F.; Pittman, R.N. Selection of a core collection from the US germplasm collection of peanut. *Crop Sci.* **1993**, *33*, 859–861. [CrossRef]
19. Jiang, H.F.; Ren, X.P.; Huang, J.Q.; Liao, B.S.; Lei, Y. Establishment of peanut mini core collection in China and exploration of new resource with high oleat. *Chin. J. Oil Crop Sci.* **2008**, *30*, 294–299.
20. Hui-Fang, J.; Xiao-Ping, R.; Bo-Shou, L.; Jia-Quan, H.; Yong, L.; Ben-Yin, C.; Guo, B.Z.; Holbrook, C.C.; Upadhyaya, H.D. Peanut core collection established in China and compared with ICRISAT mini core collection. *Acta Agron. Sin.* **2008**, *34*, 25–30. [CrossRef]
21. Holbrook, C.C.; Dong, W. Development and evaluation of a mini core collection for the US peanut germplasm collection. *Crop Sci.* **2005**, *45*, 1540–1544. [CrossRef]
22. Kottapalli, K.R.; Rakwal, R.; Shibato, J.; Burow, G.; Tissue, D.; Burke, J.; Puppala, N.; Burow, M.; Payton, P. Physiology and proteomics of the water-deficit stress response in three contrasting peanut genotypes. *Plant Cell Environ.* **2009**, *32*, 380–407. [CrossRef]
23. Smartt, J.; Gregory, W.C.; Gregory, M.P. The genomes of *Arachis hypogaea*. 1. Cytogenetic studies of putative genome donors. *Euphytica* **1978**, *27*, 665–675. [CrossRef]
24. Seijo, G.; Lavia, G.I.; Fernández, A.; Krapovickas, A.; Ducasse, D.A.; Bertioli, D.J.; Moscone, E.A. Genomic relationships between the cultivated peanut (*Arachis hypogaea*, L.) and its close relatives revealed by double GISH. *Am. J. Bot.* **2007**, *94*, 1963–1971. [CrossRef] [PubMed]
25. Robledo, G.; Lavia, G.I.; Seijo, G. Species relations among wild *Arachis* species with the A genome as revealed by FISH mapping of rDNA loci and heterochromatin detection. *Theor. Appl. Genet.* **2009**, *118*, 1295–1307. [CrossRef] [PubMed]
26. Bertioli, D.J.; Cannon, S.B.; Froenicke, L.; Huang, G.; Farmer, A.D.; Cannon, E.K.; Liu, X.; Gao, D.; Clevenger, J.; Dash, S.; et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **2016**, *48*, 438–446. [CrossRef]

27. Kim, K.S.; Lee, D.; Bae, S.B.; Kim, Y.C.; Choi, I.S.; Kim, S.T.; Lee, T.H.; Jun, T.H. Development of SNP-Based Molecular Markers by Re-Sequencing Strategy in Peanut. *Plant Breed. Biotechnol.* **2017**, *5*, 325–333. [CrossRef]
28. Moretzsohn, M.C.; Gouvea, E.G.; Inglis, P.W.; Leal-Bertioli, S.C.; Valls, J.F.; Bertioli, D.J. A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Ann. Bot.* **2012**, *111*, 113–126. [CrossRef] [PubMed]
29. Nielen, S.; Vidigal, B.S.; Leal-Bertioli, S.C.; Ratnaparkhe, M.; Paterson, A.H.; Garsmeur, O.; D'Hont, A.; Guimaraes, P.M.; Bertioli, D.J. Matita, a new retroelement from peanut: Characterization and evolutionary context in the light of the *Arachis* A–B genome divergence. *Mol. Genet. Genom.* **2012**, *287*, 21–38. [CrossRef] [PubMed]
30. Temsch, E.M.; Greilhuber, J. Genome size variation in *Arachis hypogaea* and A. *monticola* re-evaluated. *Genome* **2000**, *43*, 449–451. [CrossRef] [PubMed]
31. Chen, X.; Li, H.; Pandey, M.K.; Yang, Q.; Wang, X.; Garg, V.; Li, H.; Chi, X.; Doddamani, D.; Hong, Y.; et al. Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 6785–6790. [CrossRef]
32. Chen, X.; Lu, Q.; Liu, H.; Zhang, J.; Hong, Y.; Lan, H.; Li, H.; Wang, J.; Liu, H.; Li, S.; et al. Sequencing of cultivated peanut, *Arachis hypogaea*, yields insights into genome evolution and oil improvement. *Mol. Plant* **2019**, *12*, 920–934. [CrossRef]
33. Margulies, M.; Egholm, M.; Altman, W.E.; Attiya, S.; Bader, J.S.; Bemben, L.A.; Berka, J.; Braverman, M.S.; Chen, Y.J.; Chen, Z.; et al. Genome sequencing in micro fabricated high-density picolitre reactors. *Nature* **2005**, *437*, 376–380. [CrossRef] [PubMed]
34. Yang, H.; Tao, Y.; Zheng, Z.; Li, C.; Sweetingham, M.W.; Howieson, J.G. Application of next-generation sequencing for rapid marker development in molecular plant breeding: A case study on anthracnose disease resistance in *Lupinus angustifolius* L. *BMC Genom.* **2012**, *13*, 318. [CrossRef] [PubMed]
35. Lee, J.; Izzah, N.K.; Jayakodi, M.; Perumal, S.; Joh, H.J.; Lee, H.J.; Lee, S.C.; Park, J.Y.; Yang, K.W.; Nou, I.S.; et al. Genome-wide SNP identification and QTL mapping for black rot resistance in cabbage. *BMC Plant Biol.* **2015**, *15*, 32. [CrossRef]
36. Kang, Y.J.; Ahn, Y.K.; Kim, K.T.; Jun, T.H. Resequencing of Capsicum *annuum* parental lines (YCM334 and Taean) for the genetic analysis of bacterial wilt resistance. *BMC Plant Biol.* **2016**, *16*, 235. [CrossRef] [PubMed]
37. Zhou, X.; Xia, Y.; Ren, X.; Chen, Y.; Huang, L.; Huang, S.; Liao, B.; Lei, Y.; Yan, L.; Jiang, H. Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genom.* **2014**, *15*, 351. [CrossRef]
38. Pandey, M.K.; Monyo, E.; Ozias-Akins, P.; Liang, X.; Guimarães, P.; Nigam, S.N.; Upadhyaya, H.D.; Janila, P.; Zhang, X.; Guo, B.; et al. Advances in *Arachis* genomics for peanut improvement. *Biotechnol. Adv.* **2012**, *30*, 639–651. [CrossRef]
39. Pandey, M.K.; Agarwal, G.; Kale, S.M.; Clevenger, J.; Nayak, S.N.; Sriswathi, M.; Chitikineni, A.; Chavarro, C.; Chen, X.; Upadhyaya, H.D.; et al. Development and evaluation of a high density genotyping 'Axiom_Arachis' array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* **2017**, *7*, 40577. [CrossRef]
40. Varshney, R.K.; Mohan, S.M.; Gaur, P.M.; Gangarao, N.V.P.R.; Pandey, M.K.; Bohra, A.; Sawargaonkar, S.L.; Chitikineni, A.; Kimurto, P.K.; Janila, P.; et al. Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnol. Adv.* **2013**, *31*, 1120–1134. [CrossRef]
41. Pandey, M.K.; Roorkiwal, M.; Singh, V.K.; Ramalingam, A.; Kudapa, H.; Thudi, M.; Chitikineni, A.; Rathore, A.; Varshney, R.K. Emerging genomic tools for legume breeding: Current status and future prospects. *Front. Plant Sci.* **2016**, *7*, 455. [CrossRef]
42. Otyama, P.I.; Wilkey, A.; Kulkarni, R.; Assefa, T.; Chu, Y.; Clevenger, J.; Anglin, N.L. Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. *BMC Genom.* **2019**, *20*, 481. [CrossRef]
43. Patil, A.S.; Popovsky, S.; Levy, Y.; Chu, Y.; Clevenger, J.; Ozias-Akins, P.; Hovav, R. Genetic insight and mapping of the pod constriction trait in Virginia-type peanut. *BMC Genet.* **2018**, *19*, 93. [CrossRef] [PubMed]
44. Peng, Z.; Zhao, Z.; Clevenger, J.P.; Chu, Y.; Paudel, D.; Ozias-Akins, P.; Wang, J. Comparison of SNP Calling Pipelines and NGS Platforms to Predict the Genomic Regions Harboring Candidate Genes for Nodulation in Cultivated Peanut. *Front. Genet.* **2020**, *11*, 222. [CrossRef] [PubMed]
45. Li, X.; Singh, J.; Qin, M.; Li, S.; Zhang, X.; Zhang, M.; Khan, A.; Zhang, S.; Wu, J. Development of an integrated 200K SNP genotyping array and application for genetic mapping, genome assembly improvement and genome wide association studies in pear (*Pyrus*). *Plant Biotechnol. J.* **2019**, *17*, 1582–1594. [CrossRef] [PubMed]
46. Tam, V.; Patel, N.; Turcotte, M.; Bossé, Y.; Paré, G.; Meyre, D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **2019**, *20*, 467–484. [CrossRef]
47. Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* **2013**, *9*, 29. [CrossRef]
48. Salem, M.; Al-Tobasei, R.; Ali, A.; Lourenco, D.; Gao, G.; Palti, Y.; Leeds, T.D. Genome-wide association analysis with a 50K transcribed gene SNP-chip identifies QTL affecting muscle yield in rainbow trout. *Front. Genet.* **2018**, *9*, 387. [CrossRef]
49. Bayer, M.M.; Rapazote-Flores, P.; Ganal, M.; Hedley, P.E.; Macaulay, M.; Plieske, J.; Waugh, R. Development and evaluation of a barley 50k iSelect SNP array. *Front. Plant Sci.* **2017**, *8*, 1792. [CrossRef]
50. Comadran, J.; Kilian, B.; Russell, J.; Ramsay, L.; Stein, N.; Ganal, M.; Hedley, P. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat. Genet.* **2012**, *44*, 1388–1392. [CrossRef]

51. Allen, A.M.; Winfield, M.O.; Burridge, A.J.; Downie, R.C.; Benbow, H.R.; Barker, G.L.; Scopes, G. Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* **2017**, *15*, 390–401. [CrossRef]
52. Saghai-Maroof, M.A.; Soliman, K.M.; Jorgensen, R.A.; Allard, R.W.L. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 8014–8018. [CrossRef]
53. Ya, N.; Raveendar, S.; Bayarsukh, N.; Ya, M.; Lee, J.R.; Lee, K.J.; Shin, M.J.; Cho, G.T.; Ma, K.H.; Lee, G.A. Genetic diversity and population structure of Mongolian wheat based on SSR markers: Implications for conservation and management. *Plant Breed. Biotechnol.* **2017**, *5*, 213–220. [CrossRef]
54. Singh, N.; Choudhury, D.R.; Singh, A.K.; Kumar, S.; Srinivasan, K.; Tyagi, R.K.; Singh, N.K.; Singh, R. Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS ONE* **2013**, *8*, e84136. [CrossRef] [PubMed]
55. Tang, Y.; Liu, X.; Wang, J.; Li, M.; Wang, Q.; Tian, F.; Buckler, E.S. GAPIT version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome* **2016**, *9*. [CrossRef] [PubMed]
56. Yu, J.; Buckler, E.S. Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* **2006**, *17*, 155–160. [CrossRef] [PubMed]
57. 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]
58. Zhang, C.; Dong, S.S.; Xu, J.Y.; He, W.M.; Yang, T.L. PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **2019**, *35*, 1786–1788. [CrossRef]
59. Chu, C.W.; Zhang, G.P. A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *Int. J. Prod. Econ.* **2003**, *86*, 217–231. [CrossRef]
60. Barrett, J.C.; Fry, B.; Maller, J.; Daly, M.J. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **2005**, *21*, 263–265. [CrossRef]
61. Lewontin, R.C. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **1964**, *49*, 49–67.
62. Wu, T.T.; Chen, Y.F.; Hastie, T.; Sobel, E.; Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **2009**, *25*, 714–721. [CrossRef]
63. Zhou, H.; Sehl, M.E.; Sinsheimer, J.S.; Lange, K. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **2010**, *26*, 2375. [CrossRef] [PubMed]
64. Alexander, D.H.; Lange, K. Stability selection for genome-wide association. *Genet. Epidemiol.* **2011**, *35*, 722–728. [CrossRef] [PubMed]
65. Sun, H.; Wang, S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics* **2012**, *28*, 1368–1375. [CrossRef] [PubMed]
66. Sun, H.; Wang, S. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat. Med.* **2013**, *32*, 2127–2139. [CrossRef]
67. Okser, S.; Pahikkala, T.; Airola, A.; Salakoski, T.; Ripatti, S.; Aittokallio, T. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet.* **2014**, *10*, e1004754. [CrossRef]
68. Sun, H.; Wang, Y.; Chen, Y.; Li, Y.; Wang, S. pETM: A penalized Exponential Tilt Model for analysis of correlated high-dimensional DNA methylation data. *Bioinformatics* **2017**, *33*, 1765–1772. [CrossRef]
69. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
70. Nicolai, M.; Bühlmann, P. Stability Selection: Stability Selection. *J. R. Stat. Soc. Ser. B* **2010**, *72*, 417–473. [CrossRef]
71. Kim, K.; Koo, J.; Sun, H. An empirical threshold of selection probability for analysis of high-dimensional correlated data. *J. Stat. Comput. Simul.* **2020**, *90*, 1606–1617. [CrossRef]
72. Zhang, S.; Hu, X.; Miao, H.; Chu, Y.; Cui, F.; Yang, W.; Wang, C.; Shen, Y.; Xu, T.; Zhao, L.; et al. QTL identification for seed weight and size based on a high-density SLAF-seq genetic map in peanut (*Arachis hypogaea* L.). *BMC Plant Biol.* **2019**, *19*, 537. [CrossRef]
73. Hu, Z.; Zhang, H.; Kan, G.; Ma, D.; Zhang, D.; Shi, G.; Hong, D.; Zhang, G.; Yu, D. Determination of the genetic architecture of seed size and shape via linkage and association analysis in soybean (*Glycine max* L. Merr.). *Genetica* **2013**, *141*, 247–254. [CrossRef] [PubMed]
74. Bertioli, D.J.; Jenkins, J.; Clevenger, J.; Dudchenko, O.; Gao, D.; Samoluk, S.S. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nature genetics* **2019**, *51*, 877–884. [CrossRef] [PubMed]
75. Otyama, P.I.; Kulkarni, R.; Chamberlin, K.; Ozias-Akins, P.K.; Chu, J.; Fernández-Baca, D.F. Genotypic characterization of the US peanut core collection. *BioRxiv* **2020**. [CrossRef]
76. Hammons, R.O.; Herman, D.; Stalker, H.T. Origin and early history of the peanut. In *Peanuts*; AOCS Press: Urbana, IL, USA, 2016; pp. 1–26. [CrossRef]
77. Simpson, C.E.; Krapovickas, A.; Valls, J.F.M. History of Arachis including evidence of A. *hypogaea L. progenitors. Peanut Sci.* **2001**, *28*, 78–80. [CrossRef]
78. Stalker, H.T.; Dhesi, J.S.; Kochert, G. Genetic diversity within the species *Arachis duranensis* Krapov. & W. C. Gregory, a possible progenitor of cultivated peanut. *Genome* **1995**, *38*, 1201–1212. [CrossRef]

79. Stalker, H.T.; Tallury, S.P.; Seijo, G.R.; Leal-Bertioli, S.C. Biology, speciation, and utilization of peanut species. In *Peanuts*; AOCS Press: Urbana, IL, USA, 2016; pp. 27–66. [CrossRef]
80. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [CrossRef]
81. Li, C.; Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **2008**, *24*, 1175–1182. [CrossRef]
82. Anderson, C.A.; Pettersson, F.H.; Clarke, G.M.; Cardon, L.R.; Morris, A.P.; Zondervan, K.T. Data quality control in genetic case-control association studies. *Nat. Protoc.* **2010**, *5*, 1564–1573. [CrossRef]
83. Rao, X.; Thapa, K.S.; Chen, A.B.; Lin, H.; Gao, H.; Reiter, J.L.; Hargreaves, K.A.; Ipe, J.; Lai, D.; Xuei, X.; et al. Allele-specific expression and high-throughput reporter assay reveal functional genetic variants associated with alcohol use disorders. *Mol. Psychiatry* **2019**, 1–10. [CrossRef]
84. Oliveira, J.C.D.; Rufino, P.B.; Azêvedo, H.S.F.D.S.; Sousa, A.C.B.D.; Assis, G.M.L.D.; Silva, L.M.D.; Campos, T.D. Inferring mating system parameters in forage peanut, *Arachis pintoi*, for Brazilian Amazon conditions. *Acta Amazonica* **2019**, *49*, 277–282. [CrossRef]
85. Baud, S. Seeds as oil factories. *Plant Reprod.* **2018**, *31*, 213–235. [CrossRef]
86. Smith, A.J.; Rinne, R.W.; Seif, R.D. Phosphoenolpyruvate carboxylase and pyruvate kinase involvement in protein and oil biosynthesis during soybean seed development. *Crop Sci.* **1989**, *29*, 349–353. [CrossRef]
87. Feria, A.B.; Alvarez, R.; Cochereau, L.; Vidal, J.; García-Mauriño, S.; Echevarría, C. Regulation of phosphoenolpyruvate carboxylase phosphorylation by metabolites and abscisic acid during the development and germination of barley seeds. *Plant Physiol.* **2008**, *148*, 761–774. [CrossRef] [PubMed]
88. Aivalakis, G.; Dimou, M.; Flemetakis, E.; Plati, F.; Katinakis, P.; Drossopoulos, J.B. Immunolocalization of carbonic anhydrase and phosphoenolpyruvate carboxylase in developing seeds of Medicago sativa. *Plant Physiol. Biochem.* **2004**, *42*, 181–186. [CrossRef] [PubMed]
89. Golombek, S.; Rolletschek, H.; Wobus, U.; Weber, H. Control of storage protein accumulation during legume seed development. *J. Plant Physiol.* **2001**, *158*, 457–464. [CrossRef]
90. Pan, L.J.; Yang, Q.L.; Chi, X.Y.; Chen, M.N.; Zhen, Y.A.N.G.; Na, C.H.E.N.; Yu, S.L. Functional analysis of the phosphoenolpyruvate carboxylase on the lipid accumulation of peanut (*Arachis hypogaea* L.) seeds. *J. Integr. Agric.* **2013**, *12*, 36–44. [CrossRef]