Title

Genomic Analysis of Vavilov's Historic Chickpea Landraces Reveals Footprints of Environmental and Human Selection

Journal

International Journal of Molecular Sciences, 21(11)

ISSN

1661-6596

Authors

Sokolkova, Alena
Bulyntsev, Sergey V
Chang, Peter L
et al.

Publication Date

2020

DOI

10.3390/ijms21113952

Peer reviewed

International Journal of
Molecular Sciences

MDPI

Article

Genomic Analysis of Vavilov's Historic Chickpea Landraces Reveals Footprints of Environmental and Human Selection

Alena Sokolkova [1], Sergey V. Bulyntsev [2], Peter L. Chang [3], Noelia Carrasquilla-Garcia [4], Anna A. Igolkina [1], Nina V. Noujdina [1,5], Eric von Wettberg [6], Margarita A. Vishnyakova [2], Douglas R. Cook [4,*], Sergey V. Nuzhdin [1,3,*] and Maria G. Samsonova [1,*]

1 Department of Applied Mathematics, Peter the Great St. Petersburg Polytechnic University, 195251 St. Petersburg, Russia; alyonasok@yandex.ru (A.S.); igolkinaanna11@gmail.com (A.A.I.); nnoujdina@gmail.com (N.V.N.)
2 Federal Research Centre All-Russian N.I. Vavilov Institute of Plant Genetic Resources (VIR), 190000 St. Petersburg, Russia; s_bulyntsev@mail.ru (S.V.B.); m.vishnyakova@vir.nw.ru (M.A.V.)
3 Dornsife College of Letters Arts & Sciences, Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA; peterc@usc.edu
4 Department of Plant Pathology, University of California Davis, Davis, CA 95616, USA; noecarras@ucdavis.edu
5 Department of Geography, University of California Los Angeles, Los Angeles, CA 90095, USA
6 Department of Plant and Soil Science, University of Vermont, Burlington, VT 05405, USA; Eric.Bishop-Von-Wettberg@uvm.edu
* Correspondence: drcook@ucdavis.edu (D.R.C.); snuzhdin@usc.edu (S.V.N.); m.g.samsonova@gmail.com (M.G.S.); Tel.: +1-530-754-6561 (D.R.C.); +7-812-2909645 (M.G.S.)

Received: 19 April 2020; Accepted: 28 May 2020; Published: 31 May 2020

Abstract: A defining challenge of the 21st century is meeting the nutritional demands of the growing human population, under a scenario of limited land and water resources and under the specter of climate change. The Vavilov seed bank contains numerous landraces collected nearly a hundred years ago, and thus may contain 'genetic gems' with the potential to enhance modern breeding efforts. Here, we analyze 407 landraces, sampled from major historic centers of chickpea cultivation and secondary diversification. Genome-Wide Association Studies (GWAS) conducted on both phenotypic traits and bioclimatic variables at landraces sampling sites as extended phenotypes resulted in 84 GWAS hits associated to various regions. The novel haploblock-based test identified haploblocks enriched for single nucleotide polymorphisms (SNPs) associated with phenotypes and bioclimatic variables. Subsequent bi-clustering of traits sharing enriched haploblocks underscored both non-random distribution of SNPs among several haploblocks and their association with multiple traits. We hypothesize that these clusters of pleiotropic SNPs represent co-adapted genetic complexes to a range of environmental conditions that chickpea experienced during domestication and subsequent geographic radiation. Linking genetic variation to phenotypic data and a wealth of historic information preserved in historic seed banks are the keys for genome-based and environment-informed breeding intensification.

Keywords: bioclimatic analysis; chickpea; GBS; GWAS; haploblock; SNP

1. Introduction

Landraces dominated agriculture for millennia, until the advent of intensive modern breeding in the mid 20th century, when reduced sets of elite cultivated varieties largely displaced the wider diversity

of local genotypes [1]. Although the shift away from landraces was neither systematic nor synchronous, it is generally accepted that the subsequent convergence on a limited set of elite germplasm removed considerable useful variation [2]. In the early 20th century (1911–1940), N.I. Vavilov led a systematic effort to collect and preserve crop diversity, now maintained within the Vavilov Institute of Plant Genetic Resources (VIR) collection in St. Petersburg, Russia [3]. The geographic distribution and genetic diversity of most crops collected during this time frame are likely to reflect their historic patterns of cultivation established over the preceding millennia. Exploring these unique genetic resources provides an opportunity to revisit hypotheses about the radiation and secondary diversification of crops, not possible using later collections. Moreover, the expanded diversity of these early collections likely contains 'genetic gems' with the potential to enhance modern breeding efforts [4].

Here, we focus on biodiversity of *Cicer arietinum*, chickpea, which is among the world's most widely grown grain legumes and provides a vital source of dietary protein for ~15% of the world's population. Chickpea was first domesticated ~10 KYA, initially in southeastern Turkey, and then spread regionally throughout the Fertile Crescent. Although exact dates are unknown, archeological evidence suggests chickpea moved to India ~6000 years ago and to Ethiopia and North Africa ~3000 years ago [5]. Millennia of cultivation in these new areas, largely in isolation from each other, led to the establishment of new centers of secondary diversity, with accompanying differentiation of regionally specific landraces. Despite this generally accepted scenario, the relationships among the chickpea crops at these historic centers of cultivation are not fully resolved.

Chickpea domestication and breeding imposed a severe genetic bottleneck on the crop, with an estimated >95% of diversity lost between the crop wild progenitor and modern elite varieties [6]. Landraces represent an intermediate step to modern germplasm. An implicit, yet untested assumption is that chickpea landraces will have increased genetic diversity relative to modern elite germplasm. Moreover, we posit that geographic patterns of landrace diversity were shaped by post-domestication selection to adapt the crop to different agro-ecological environments and cultural preferences. Although Vavilov was unable to quantify the extent of diversity and differentiation, he and his contemporaries recognized the value of landraces as reserves of agriculturally-relevant traits, which motivated these early efforts in collection and conservation. Thus, chickpea landraces are expected to contain beneficial alleles, not segregating among modern elite varieties, which can be accessed and prioritized for crop improvement using genomics, phenotyping, and computational methods.

Here, we combine genomics, phenotyping, and computational biology to understand chickpea's agricultural variation one century ago, and from that analysis to infer the breadth and genetic bases of trait variation in the pre-modern era. Such knowledge can prioritize landrace haplotypes that contributed to diversification of chickpea as a crop, particularly haplotypes missing from modern breeding programs, thereby facilitating their use for crop improvement.

2. Results

2.1. Germplasm Resources and Phenotyping

To fully cover the biogeographic range of historic chickpea cultivation, we assembled 407 accessions collected between 1911 and 1940. Text descriptions of sampling locations, which were often local markets in small towns, were converted to geographic coordinates (Figure 1a). This set of accessions is enriched for genotypes under cultivation a minimum of one century ago in Turkey, India, Ethiopia, Uzbekistan, and Morocco, representing the major centers of post-domestication chickpea diversification and comprising 55% of the 407 analyzed accessions. Beyond the 147 Turkish and Ethiopian genotypes analyzed in an earlier study [4], we genotyped and/or phenotyped an additional 260 accessions spanning a total of 30 countries, with adjacent countries occasionally representing single extended historic agricultural systems (for examples, Ethiopia and Eritrea in eastern Africa, and several countries from the Fertile Crescent) (Table S1). The entire set of accessions was phenotyped under field conditions, genotyped, and used for further analysis.

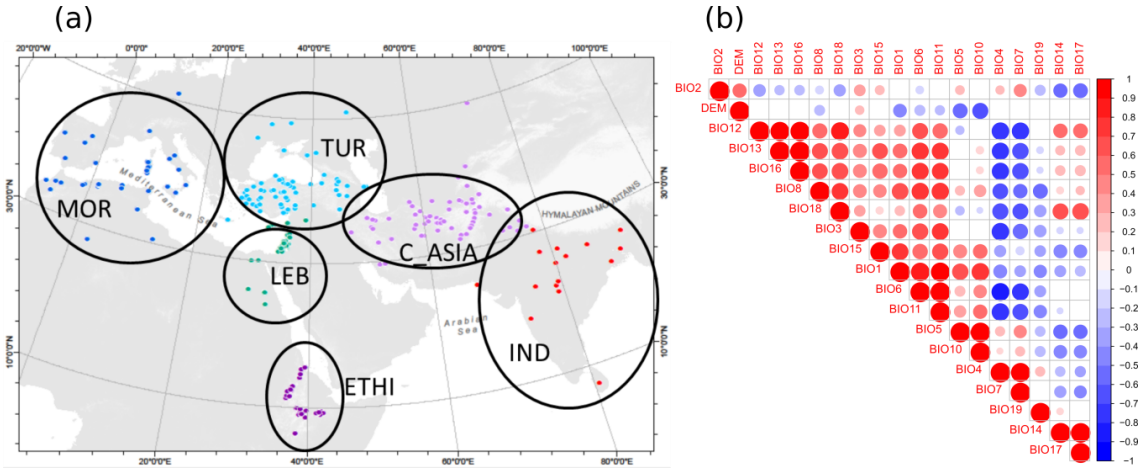

Figure 1. Sample distribution and correlation of bioclimatic variables. (**a**) Location of the chickpea samples around the world that were split into six geographically distinct groups. (**b**) The correlation between nineteen bioclimatic variables (bioclimatic variables and their abbreviations are presented in Table S2). Color intensity and the size of the asterisk are proportional to the correlation coefficients. ETHI, Ethiopia; IND, India; LEB, Lebanon; MOR, Morocco; TUR, Turkey; C_ASIA, Central Asia.

Correlation analyses of nineteen bioclimatic variables (bioclimatic variables and their abbreviations are presented in Table S2) from the range of chickpea collection sites revealed five groups of correlated variables (Figure 1b; Table S3). Three bioclimatic variables (BIO_2, BIO_{19}, DEM) were not strongly correlated to other variables. The first, third, and fifth groups (Table S3) correspond to temperature traits. The second and fourth groups (Table S3) consist of precipitation variables. While the first group (Table S3) consists of traits with moderate positive correlation (pairwise Spearman correlation coefficient, $r > 0.4$, Figure 1b), traits in the second group (Table S3) have stronger positive correlations (pairwise Spearman correlation coefficient, $r > 0.7$, Figure 1b), and traits in the remaining groups (Table S3) have the strongest positive correlations (pairwise Spearman correlation coefficient, $r > 0.9$, Figure 1b).

All 407 landraces accessions were phenotyped for thirty-six traits under field conditions in Kuban, Russia. The scored phenotypes and their abbreviations are presented in Table S4. Correlation analyses identified three groups of correlated traits (Figure 2). Phenotypic traits related to the color of plant organs and tissues were moderately correlated (pairwise Spearman correlation coefficient, $r > 0.5$, Figure 2) and form a single group. Quantitative traits characterizing the weights and sizes of whole plants and pods, as well as leaf size, also had moderate positive correlations (pairwise Spearman correlation coefficient, $r > 0.4$, Figure 2) and form two groups. Two phenological traits describing the duration of flowering and the duration of pod maturation had strong negative correlation (Spearman correlation coefficient, $r = -0.76$, Figure 2). Pod shape (PodSH) had moderate negative correlation with pod length (PDL) (Spearman correlation coefficient, $r = -0.53$, Figure 2) and pod width (PDW) (Spearman correlation coefficient, $r = -0.55$, Figure 2). Pod shape also had moderate negative correlation with thousand seeds weight (TSW) (Spearman correlation coefficient, $r = -0.47$, Figure 2). Phenotypic traits related to organ and tissue coloration had moderate negative correlation with traits describing the weights and sizes of plant and pods (pairwise Spearman correlation coefficient, $r < -0.4$, Figure 2).

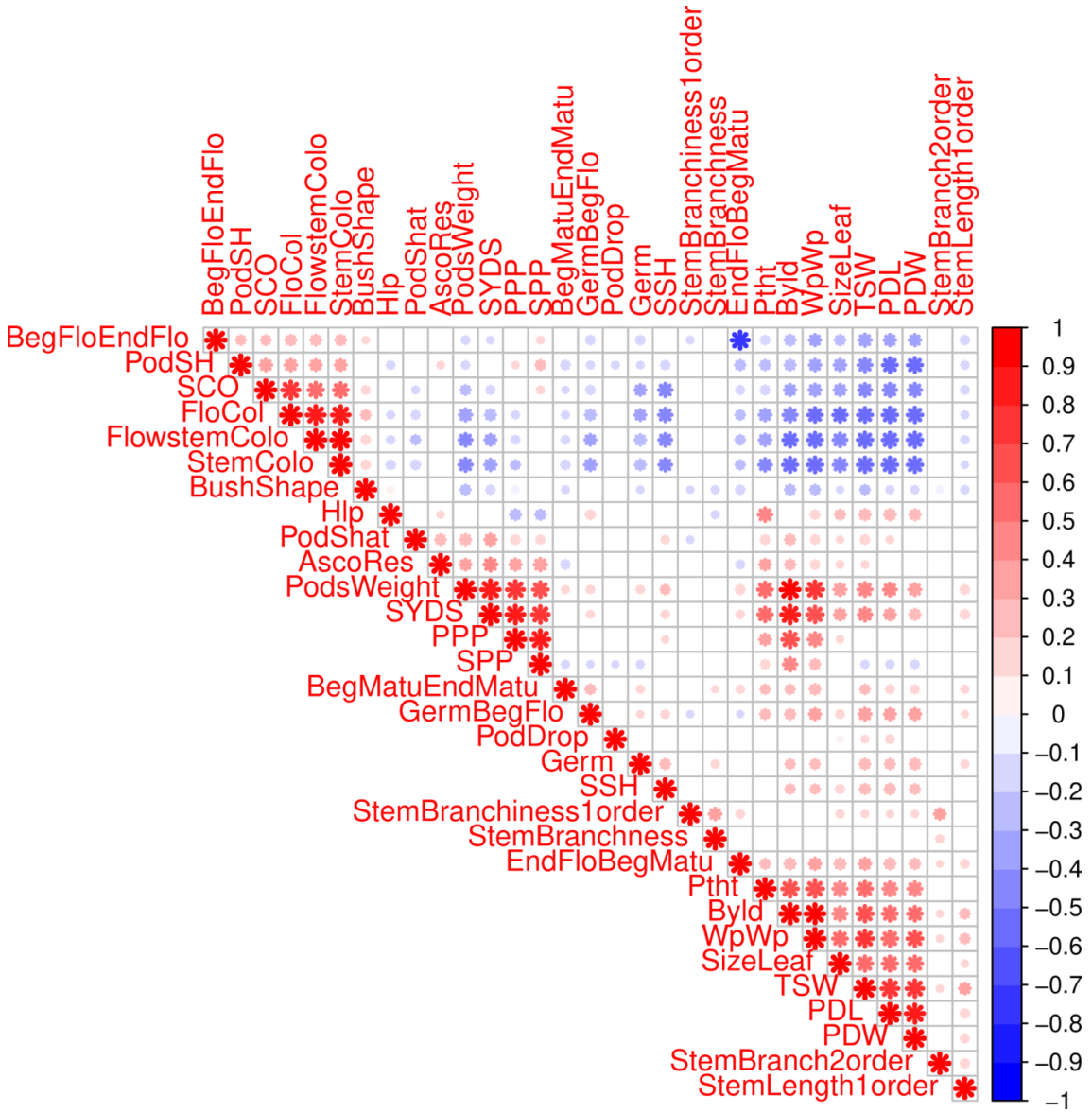

Figure 2. Correlation of thirty-one phenotypic traits. The scored phenotypes and their abbreviations are presented in Table S4. *Ascochyta*, the degree of damage (AsoDes) trait, was excluded from correlation analysis because it is the opposite value of *Ascochyta* resistance (AscoRes) trait. Moreover, we excluded overlapping time periods traits. Color intensity and the size of the asterisk are proportional to the correlation coefficients. PodSH, pod shape; SCO, seed color; SSP, number of seeds per plant; SSH, seed shape; TSW, thousand seeds weight; PDW, pod width; PDL, pod length.

2.2. Marker Polymorphism Analysis

Restriction site associated genotyping by sequencing (RAD-GBS) was used to survey polymorphism within the genomes of 407 accessions. SNPs were filtered to retain polymorphisms present in at least 90% of genotypes with a minor allele frequency of at least 3%. The resulting 2579 polymorphisms are distributed among all chromosomes, but with variable density that is especially elevated on chromosome 4 (Figure 3a). The elevated polymorphism content of chickpea chromosome 4 has been observed in previous studies (e.g., [4]). We hypothesized that selection and introgression via inadvertent hybridization between more and less advanced morphotypes might have resulted in agricultural improvement genes being aggregated to genomic 'agro islands', and in genotype-to-phenotype relationships resembling widespread pleiotropy.

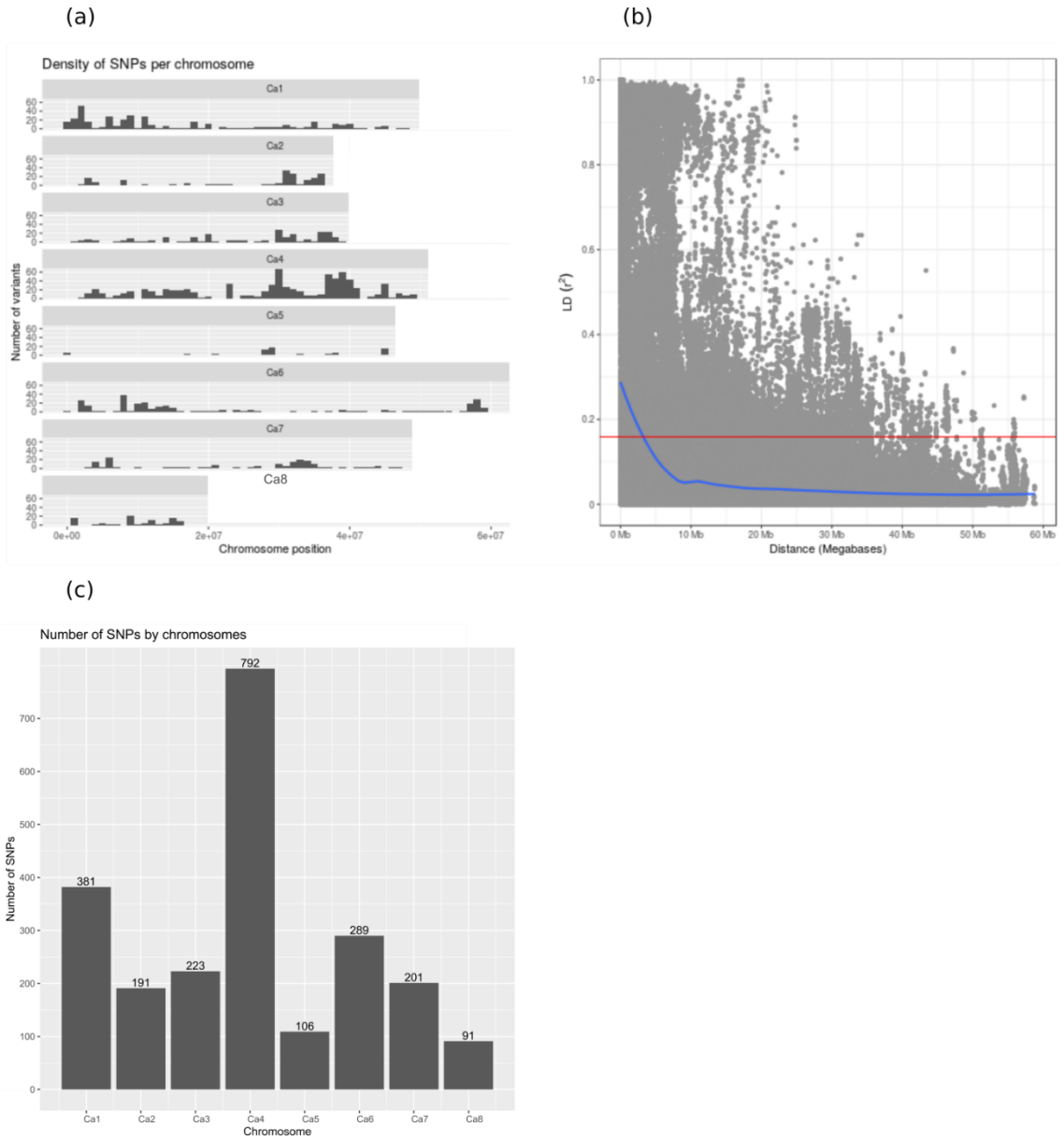

Figure 3. (**a**) Density of SNPs across the chickpea genome. Chromosome Ca6 is the longest chromosome in the chickpea genome (59.46 Mb) and chromosome Ca8 is the shortest (16.48 Mb). (**b**) Linkage disequilibrium (LD) (r^2) plots of the whole chickpea genome. The horizontal red line indicates the 95th percentile of the distribution of the unlinked r^2, which gives the critical value of r^2. (**c**) Distribution of SNPs along the eight chromosomes of the chickpea genome.

The sufficiency of this marker set for genetic tests depends in part on the scale of linkage disequilibrium (LD), because the relationship between physical distance and recombination frequency determines the precision of genetic association tests. LD is the non-random association between polymorphisms and can originate from demographic processes (e.g., shared ancestry and drift) or from selection (i.e., selective sweeps). In smaller populations of predominantly selfing organisms (including those that are the product of breeding), drift and selection typically have stronger effects than recombination, and thus LD extends to large genomic regions. Landraces are expected to exhibit especially extended LD. In line with these expectations, LD in chickpea landraces is very slow to decay (Figure 3b; Figure S1). Moreover, the marker density is uneven between chromosomes: from 91 SNPs

on chromosome Ca8 to 792 SNPs on chromosome Ca4 (Figure 3c). Our sample size is comparable with other recent GWAS crop publications, hopefully resulting in adequate power.

2.3. Geographic Analyses

Patterns of population differentiation were analyzed using principle components (PCA) and visualized with unrooted trees. Figure 4 depicts the PCA plot for genetic data of the first versus second components and Figure S2 depicts a summary of variation and covariation attributed to the first five principle components. Interestingly, the accessions from the center of domestication, Turkey, are mainly divided into two clusters with light seeded Kabuli and Desi, which are smaller with dark seeds and purple flowers market classes intermixed with each cluster (Figure 4). The lack of distinctiveness between Desi and Kabuli adds further support to the same conclusion reached by Penmetsa et al. [7]. All groups containing Turkish accessions also contain minor representation from other regions, with the exception of a preponderance of landraces from North Africa in one of the Turkish groups. Notably, landraces from India and Ethiopia, which represent two of Vavilov's major sites of secondary diversification [8], are well resolved, though not exclusive of one another. Turkish accessions are absent from the group of Ethiopian landraces and constitute only a minor component of the Indian group, which is instead enriched in landraces from Central Asia. A portion of Central Asian accessions also occur in a distinct grouping dominated by the ancestral Desi form (Figure 4).

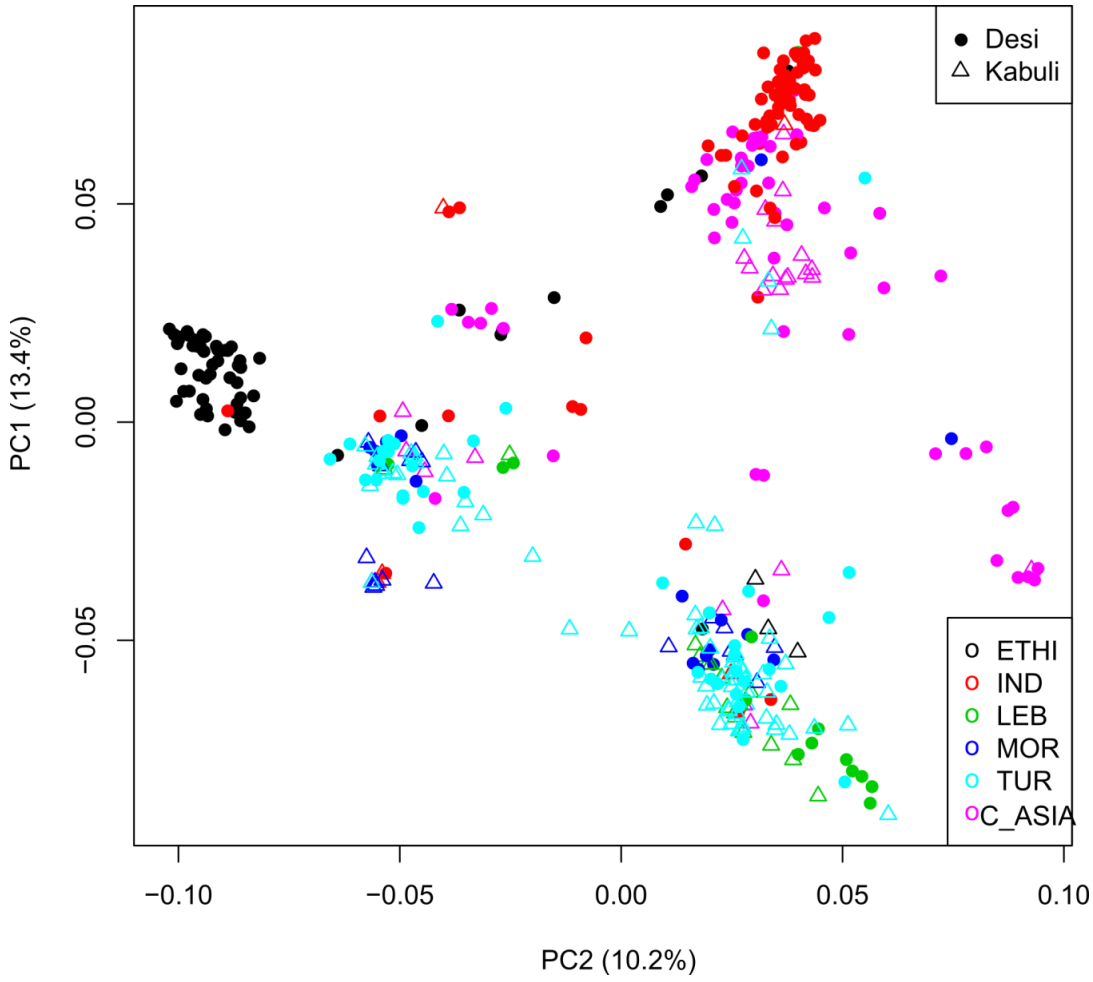

Figure 4. Scatter plots of the first two principal components of the principal component analysis (PCA) based on 2579 SNPs. Each dot represents an accession. Desi varieties are shown as asterisks and Kabuli as triangles.

These observations are consistent with the deduced pattern of molecular evolution. Maximum likelihood phylogenetic trees constructed with genome-wide SNP (Figure 5a) support inferences from the PCA analysis. Central Asian and Turkish accessions are broadly distributed throughout the tree, but notably absent from groups predominated by India and Ethiopia, consistent with more extensive diversity (Table S5) at the Turkish center of origin for the species, and with longstanding, but distinct secondary diversification in India, Central Asia, and Ethiopia. Chromosome 4 is known to have excess diversity relative to the rest of the genome [9,10], as indeed we observe here. Interestingly, certain of the relationships observed using genome-wide SNP are obscured in the tree constructed from chromosome 4 SNPs (Figure 5b). In particular, the previously coherent group of Ethiopian genotypes is divided more broadly within the tree and there is both greater subdivision within the Indian group and less distinction from the Central Asian landraces.

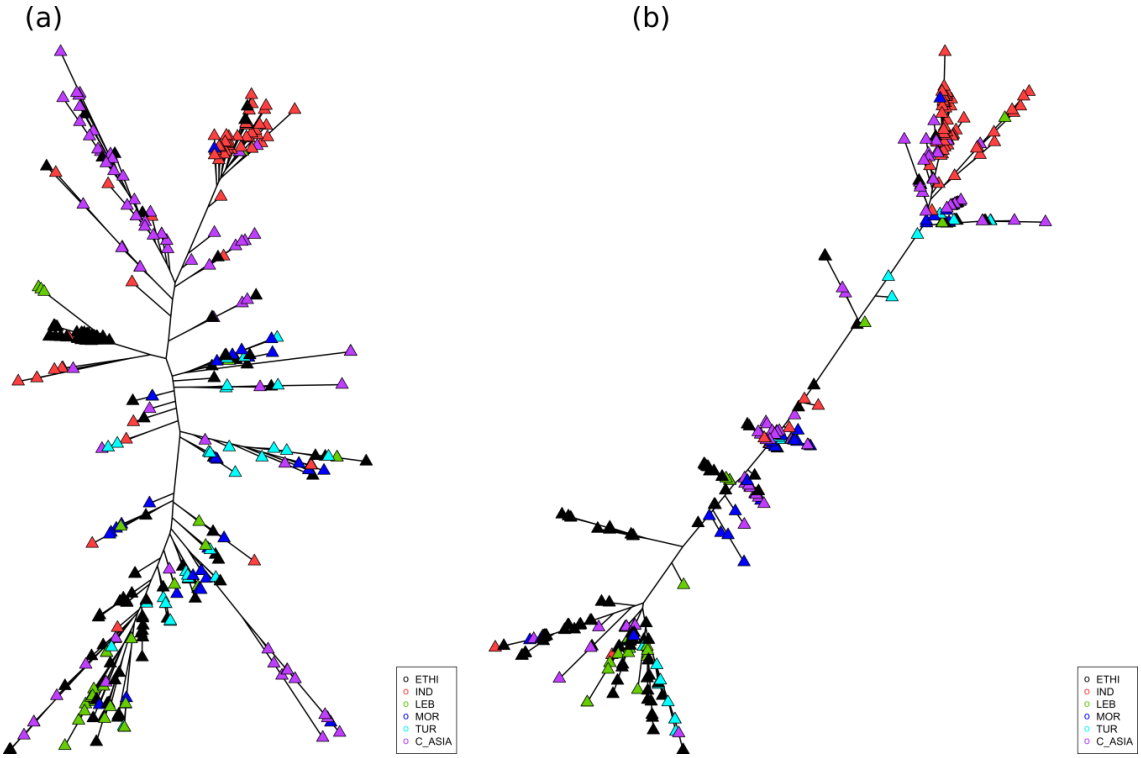

Figure 5. (**a**) Maximum likelihood phylogenetic tree showing relationships among accessions based on the whole genome SNPs and (**b**) on chromosome 4 SNPs.

2.4. Single Trait Associations

Genetic and phenotypic data were strongly concordant, as described in Table S6, which shows co-variances between genetic and phenotypic data.

To account for these effects, GWAS analysis was implemented with the first eight PCA axes scores used as covariates for all phenotypic and bioclimatic data (Figures S3–S18), revealing multiple significant associations among 70 SNPs with bioclimatic and phenotypic traits (Figures 6 and 7; Table S7). Twelve of 70 markers were found to have significant associations with two or more traits. SNP Ca2: 17161867 is associated with plant weight without pods (WpWp) as well as isothermality (BIO_3) and mean temperature of the warmest quarter (BIO_{10}) (see Table S2 and Table S4 for a full list of bioclimatic variables and phenotypes abbreviations). These genetic findings are supported by WpWp weakly negatively correlated with BIO_3 and BIO_{10}. SNP Ca3: 20549509 and SNP Ca6: 2908823 are associated with mean diurnal range (BIO_2) and BIO_3, which are themselves weakly positively correlated (Figure 1b). Three SNPs, two on the 8th chromosome (SNP Ca8: 9098790 and Ca8: 10314452) and one on the 4th chromosome (SNP Ca4: 30948593), are associated with two phenotypic variables:

biological yield (Byld) and plant weight without pods (WpWp), which are very strongly correlated and appear to derive from common genetic capacities (r = 0.92; Figure 2). Also on chromosome 4, Ca4: 33967674 is associated with the correlated group of phenotypes that includes plant weight traits (weight of seeds, pods, and the whole plant). SNP Ca6: 57117312 is associated with flower color (FloCol) and seed shape (SSH), which are themselves moderate negatively correlated (r = −0.45, Figure 2). SNP Ca7: 30930779 is associated with BIO_3, number of seeds per plant (SPP), and the group of phenotypes characterizing plant and organ weights. Three additional SNPs on chromosome 7 (SNP Ca7: 33337524, Ca7: 33340372, Ca7: 33457287) are associated with three bioclimatic variables, BIO_3, BIO_6, and BIO_{11}, which are part of a larger group of correlated variables (Figure 1b).

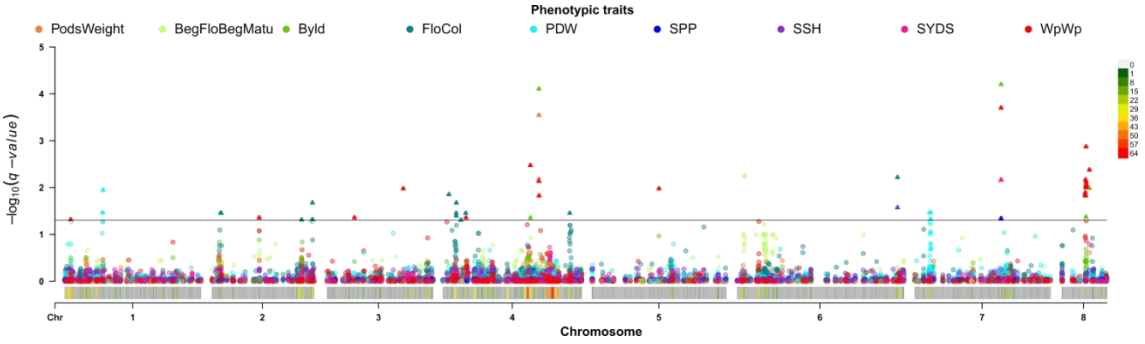

Figure 6. Summary of GWAS analyses with eight PCs as covariates for phenotype data (different colors correspond to different phenotype). SNPs with *q*-value < 0.05 are shown for each chromosome, marked as triangles. Chromosome density is attached on the bottom of the Manhattan plot.

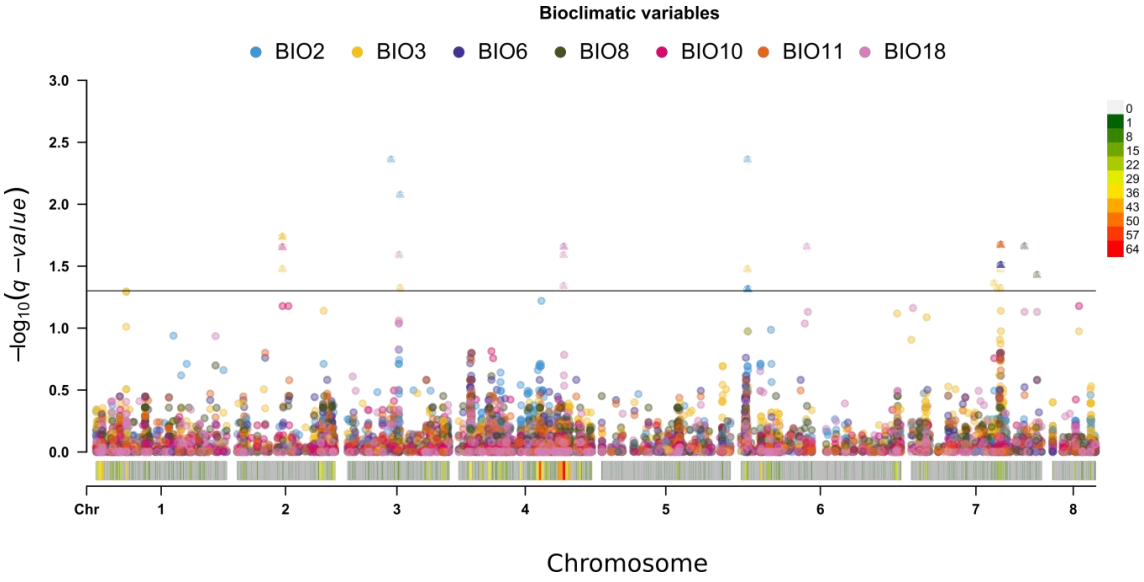

Figure 7. Summary of GWAS analyses with eight PCs as covariates for bioclimatic variables (different colors correspond to different bioclimatic variables). SNPs with *q*-value < 0.05 are shown for each chromosome, marked as triangles.

To incorporate geography explicitly into the analysis, we repeated the above GWAS, but with the addition of the first two axes of PCoA, which derive from the analysis of landrace geographic variation (Figures S19 and S20; Table S7). The results of these analyses were generally consistent with the results described above and are only introduced briefly here. An additional set of significant associations was found. Twelve SNPs are associated with pod length (PDL), nine on chromosome 6 and three on chromosome 7. Ten of these twelve SNPs exhibit significant linkage. Two SNPs on chromosome 7 are associated with secondary branching (StemBranch2order), but without strong linkage.

Because of extended LD, we cannot identify causal relationships between SNPs and phenotypes. Nevertheless, we explored the potential nature of the associated genes and found several important genes that have been reported in previous studies. For example, genes Ca_10410, Ca_10426, and Ca_10428 are present within haploblock Ca6:2541669 … .Ca6:3024335, to which several SNPs associated with the beginning to flowering to the beginning to maturation phenotype and temperature related variables map (see Table S7). Ca_10410 (Ca6:2766285 … .2768999) is involved in floral development and encodes flavin-binding kelch repeat F-box protein with high homology to circadian clock-associated FKF1 gene of soybean. Ca_10426 (Ca6:2881369 … .2884463) encodes a XAP5 protein important for light regulation of the circadian clock that plays a global role in coordinating growth in response to the light environment. SNP Ca2: 17161867 associated with plant weight without pods (WpWp) and temperature related bioclimatic variables BIO_3 and BIO_{10}, as well as Ca2: 17161884 associated with the duration of flowering (BegFloEndFlo) and BIO_3 are all located within intron of gene Ca_16015. This gene encodes phosphoenolpyruvate carboxylase, enzyme involved in carbon fixation, and citric acid cycle biosynthesis flux [11]. The first intron of Ca_11533 gene encoding beta-D-xylosidase contains SNP Ca8: 9098790, which is associated with both WpWp and Byld. beta-D-Xylosidases are involved in the breakdown of xylan, a major component of plant cell-wall hemicelluloses [12]. SNP Ca1: 2218700, which is associated with WpWp, is located in the intergenic region upstream of gene Ca_00278 that encodes protein with polyphenol oxidase activity. In *Clematis terniflora* DC, decreasing activity of this enzyme elevates the plant photosynthesis by activating the glycolysis process, regulating Calvin cycle, and providing adenosine triphosphate (ATP) for energy metabolism. Besides, polyphenol oxidase is involved in the formation of brown melanin pigment in fruits and vegetables, plays a crucial role in the biosynthesis of secondary metabolites, and has a role in plant defense against biotic and abiotic stresses [13]. SNP Ca3: 10855323 associated with WpWp is located upstream of Ca_19358 gene encoding beta-*N*-acetylhexoamidase that catalyzes the hydrolysis of *N*-acetylglucosamine or *N*-acetylgalactosamine from the non-reducing terminal of oligosaccharides, glycoproteins, glycolipids, and other glycoconjugates. b-*N*-acetylhexosaminidase is highly active in dry or germinating seeds, where it participates in the degradation of reserve glycoproteins. Moreover, its activity is induced in the period of ripening in tomato and peaches [14]. The Ca_11539 (Ca8:9151680. … 9159194) intron contains several SNPs associated with WpWp. This gene encodes an oligopeptidase degrading short peptides. SNP Ca4: 2145082 associated with flower color (FloCol) is located upstream of Ca_07836 gene, which is homologue of genes in *Pisum sativum* (protein A) and *Medicago truncatula* (bHLH-A), which are flower color associated genes [15].

2.5. Clustering of Phenotypes and Variables Sharing Enriched Haploblocks

The total number of the Haploview-inferred [16] haploblocks was 224, encompassing 1264 SNPs (mean per haploblock = 5.6) (Table S8). Filtering for more than six SNPs left 74 haploblocks (33% of total) as input to find haploblocks enriched for associated SNPs for each trait and variable using the fast gene set enrichment (FGSEA) method [17] (parameter for permutations = 100,000) (Table S9). Subsequent to bi-clustering of phenotypes and variables sharing enriched haploblocks, we defined several visually distinguished groups (Figure 8, Table S10). The first group contained two consecutive reproductive stages of plant development: the duration of flowering (BegFloEndFlo) and the duration from the end of flowering until the beginning of maturation (EndFloBegMatu). We hypothesize that the same genetic mechanisms influence the duration of both stages. The second group contains pod shattering (PodShat) and pod drop (PodDrop) traits as well as one-third of all bioclimatic factors, related to both temperature and precipitation, exclusive to a well correlated set from Figure 1b ($BIO_{6,8,11,12,13,16}$). Pod-related traits form a subgroup with three temperature-related bioclimatic factors: mean temperature (BIO_1), mean temperature of coldest month (BIO_6), and temperature annual range (BIO_7); this subgroup is similar in a set of enriched haploblocks with the group containing two additional heat-related bioclimatic factors, max temperature of warmest month (BIO_5), and mean temperature of warmest quarter (BIO_{10}). This grouping is consistent with a well-known relationship

between high temperature and pod shattering/retention. A third group includes color-related traits, flower color (FloCol), peduncle color (FlowstemColo), seed color (SCO), and stem color (StemColo), which is expected, because genes in the phenylpropanoid pathway are implicated in the production of pigments in different plant organs. A fourth group aggregates *Ascochyta* blight resistance (AscoRes) and precipitation of the coldest quarter (BIO_{19}), which reflects a well understood relationship between *Ascochyta* incidence and rainfall during periods of reduced temperatures. Also of note is a group containing moisture stress-related covariates ($BIO_{14,17}$, precipitation of the driest month/quarter) and plant height (Ptht), which is expected to depend on moisture availability; interestingly, this group clusters with a group that contains phenotypic traits related to plant size (biological yield and pod size), which are traits related to the duration of vegetative growth and that are limited by moisture availability.

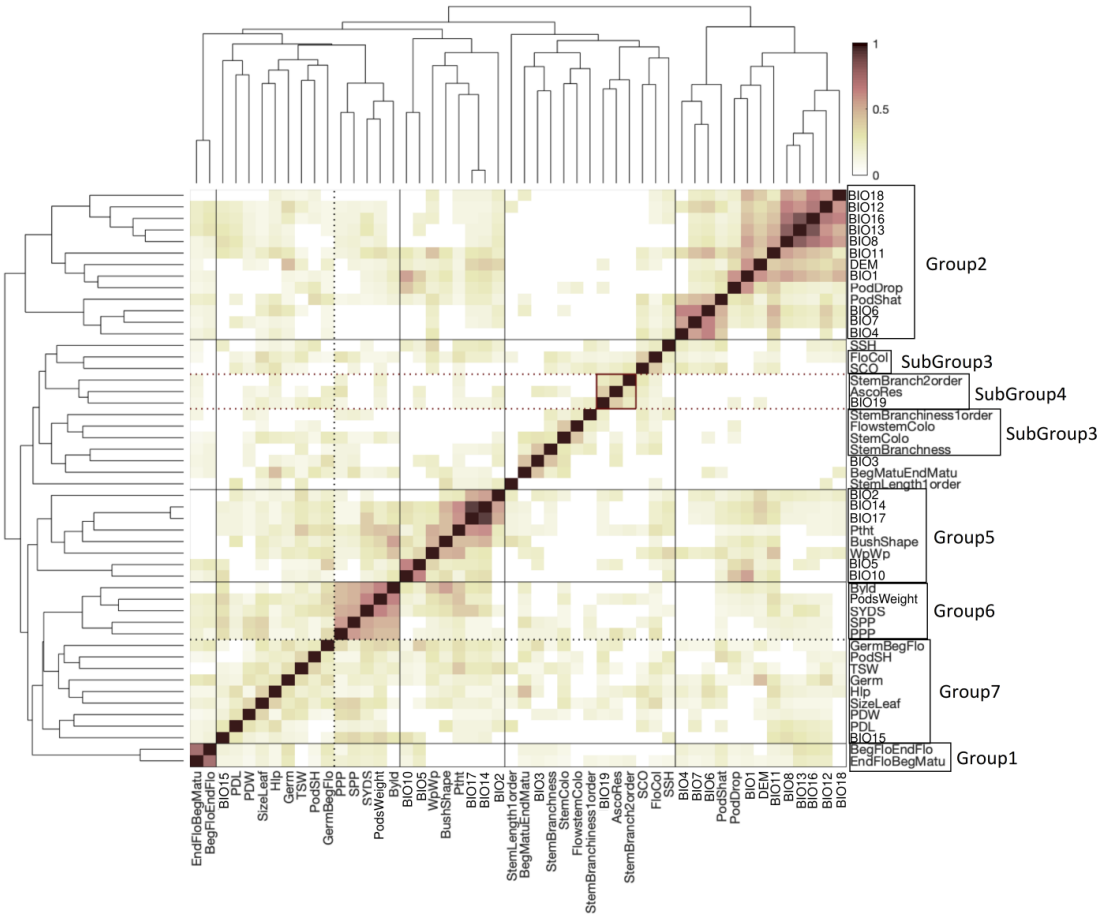

Figure 8. The degree of overlap in haploblocks enriched for SNPs associated with phenotypes and variables. Bi-clustering of similarity scores reveals several visually distinct groups of phenotypes. The haploblock similarity score is defined as a double sum of haploblocks simultaneously enriched for SNPs for both traits normalized to the amount of significantly enriched haploblocks for each trait. The degree of similarity is color coded.

3. Discussion

For many millennia, farmers and breeders have focused on selecting crops with desirable phenotypes [2]. With the successful domestication of numerous crops came the incremental loss of genetic and phenotypic variation. Genetic bottlenecks are especially common in selfing species such as grain legumes (e.g., [18]). Novel sources of variation for biotic and abiotic stress resistance are especially needed in chickpea, because the crop is often grown by resource-poor farmers, on marginal lands, and under low-input conditions. Broadening chickpea's genetic base should facilitate production of

new varieties to address these needs, while also meeting changing consumer demands, new agricultural practices, and anticipated shifts in climatic conditions [6].

Chickpea landraces represent an expanded source of genetic and phenotypic variation that has not been systematically explored and has been used only in an ad hoc manner for modern breeding. The Vavilov Institute of Plant Genetic Resources is one of the world's primary libraries of lost genetic variation in food crops, capturing the genetic and functional diversity of regionally stratified agriculture typical of one century ago. It contains tens of thousands of legume accessions, including approximately one thousand chickpea accessions collected prior to intensive international breeding efforts [3]. The re-introduction of genetic material from the Vavilov Institute's collection into modern elite varieties could be a potent force for future agricultural improvement. To this end, we combine genomics, phenotyping, and computational biology to characterize the chickpea collection of Nikolay Vavilov and his colleagues, linking traits and environments to genes. Our results highlight the collection's currently latent potential of chickpea landraces, and underscore the value of this resource to meet the enormous challenges of 21st century agriculture. However, the identified candidate genes are needed in further validation and functional confirmation owing to such factors as one-year observation of phenotypes and long extend of LD in the germplasm.

Our observations contribute to an increasing understanding of genetic variation of quantitative and categorical traits in chickpea [19–21]. The present work adds a new dimension by incorporating a wider set of historical crop diversity, and by treating bioclimatic data at accession sampling sites as extended crop traits. In doing so, our GWAS hits highlight associations to genomic regions not discovered in prior GWAS and quantitative trait locus (QTL) analyses (Table S7). These hits map in the vicinity of genes involved in floral development, photosynthesis, cell wall or secondary metabolism, and carbohydrate biosynthesis, and some of them are close to already known QTLs. For example, SNP Ca4: 33967674, associated with yield, pod weight, plant weight without pods, and seed weight per plant, is located 752 kb downstream from known QTL (Table S11) governing pod number trait [22] and SNP Ca3: 28094292, associated with plant weight without pods, localizes 96 kb downstream of QTL (Table S11) containing cluster of FLOWERING LOCUS T (FT) genes and controlling phenology and growth habit [23]. SNP Ca4: 30948593 and SNP Ca8: 10314452, associated with yield, are located ~90 kb upstream from previously detected SNP (Table S11) and ~25 kb downstream from previously detected SNP, respectively (Table S11), also associated with yield [24]. SNP Ca6: 3024192, associated with beginning of flowering to the beginning of maturation phenotype, is located in the same haploblock Ca6_Block_3 (~87 kb upstream) as the previously detected SNP (Table S11), associated with days to 50% flowering [24]. Previously, we [25] published a study in which we were looking for associations between SNPs and bioclimatic covariates at collection sites. Two covariates, which include temperature characteristics, were jointly associated with one SNP on chromosome 8 (Ca8: 10314452). This SNP is associated with two phenotypic variables: biological yield (Byld) and plant weight without pods (WpWp) in the current study.

To rigorously test for associations, we implement a novel haploblock-based test that, we believe, will find much use in the crop genomics. The underlying statistics for the test are similar to the gene set enrichment analysis, where each haploblock represents a set of SNPs associated with a trait and all SNPs are ranked according to GWAS *p*-values. This analysis identified eleven haploblocks (Table S12) intersecting with previously reported GWAS hits. Haploblock Ca1_Block_18 and haploblock Ca4_Block_18 are enriched for SNPs associated with several phenotypes and bioclimatic variables, including thousand seeds weight phenotype. These haploblocks covers SNP on chromosome 1 and SNPs on chromosome 4, respectively, reported by Varshney et al. [24], associated with 100 seed weight (Table S12). Haploblock Ca3_Block_4, haploblock Ca4_Block_54 and haploblock Ca5_Block_4 are enriched for SNPs associated with several phenotypes and bioclimatic variables, including seeds weight per plant phenotype. These haploblocks overlay four SNPs on chromosome 3, three SNPs on chromosome 4, and eight SNPs on chromosome 5, respectively, reported by Varshney et al. [24], associated with yield per plant (Table S12). Haploblock Ca3_Block_7 is enriched for SNPs associated

with the duration of vegetative growth, with seeds weight per plant, and with three bioclimatic variables (BIO_5, BIO_{13}, BIO_{16}). This haploblock covers two SNPs on chromosome 3, reported by Varshney et al. [24], associated with days to 50% flowering and with yield per plant, respectively (Table S12). Haploblock Ca3_Block_16 is enriched for SNPs associated with the duration of vegetative growth, as well as with plant height, plant weight without pods, and temperature-related bioclimatic variables BIO_3 and BIO_5. This haploblock intersects with a QTL for days to 50% flowering time (Table S12) reported from the GWAS analysis of Upadhyaya and colleagues [19]; Upadhyaya et al. nominated a particular candidate gene, SBP (SQUAMOSA promoter binding protein), though we advocate a more cautious approach that recognizes limitations of the study design and instead implicates haplotype intervals. Haploblock Ca4_Block_9 is enriched for SNPs associated with the duration of vegetative growth, with pod shattering, and with four bioclimatic variables (BIO_4, BIO_6, BIO_7, BIO_{12}). This haploblock covers SNP on chromosome 4 associated with days to 50% flowering (Table S12), reported by Varshney et al. [24]. Haploblock Ca7_Block_12 is enriched for SNPs associated with the duration of vegetative growth, with number of seeds per plant, with stem branchness, and with temperature-related bioclimatic variable BIO_3. This haploblock covers SNP on chromosome 7 associated with days to maturity (Table S12), reported by Varshney et al. [24]. The last haploblock, Ca8_Block_7, is enriched for traits related for branching and covers SNP on chromosome 8 reported by Bajaj et al. [20], associated with branch number (Table S12).

Previously, we [4] published a pilot study combining historic phenotypic data with reduced representation sequencing to establish a proof-of-principle for the results reported here. We employed a combination of genomics, computational biology, and phenotyping to characterize VIR's 147 chickpea accessions from Turkey and Ethiopia, representing chickpea's center of origin and a major location of secondary diversification, respectively. The majority of SNPs associated with multiple traits localized to a single chromosome 4 region. Here, we observe similar patterns with a larger sample of more diverse landraces and with a more comprehensive phenotypic and environmental dataset. We find multiple SNPs that are non-randomly distributed among several haploblocks, many of which are associated with multiple phenotypes (Table S9). The non-random clustering of phenotypes and variables (Figure 8) exactly arises as a result of such multi-trait associations. Although the grouping of traits and ancestral bioclimatic variables does not necessarily imply co-selection during domestication (e.g., [26]), these clusters may represent genetic complexes co-adapted to a range of environmental conditions that chickpea experienced during domestication and subsequent geographic radiation. Indeed, many of the trait–environment associations reflect well-known interactions between environmental factors and the crop's biology; for example, the relationships between *Ascochyta* blight occurrence and the duration of cool-wet periods, as well as the increased incidence of pod abortion and shattering under conditions of heat stress. Thus, by combining genomics with an explicit biogeographic framework encompassing climatic and phenotype covariates, we are able to suggest concordance between human selection, the crop's known biology, and environmental constraints.

4. Materials and Methods

4.1. Germplasm Resources and Phenotyping

We assembled a collection of VIR's chickpea germplasm originating from a range of countries including Ethiopia, Lebanon, Morocco, Turkey, India, and the broader Central Asia and Mediterranean regions (see Table S1). Phenotyping of the 407 chickpea genotype collection was conducted at the VIR Kuban experimental station with climatic conditions well suited for chickpea cultivation (see Text S1). During the vegetative period, thirty-six phenological, morphological, agronomical, and biological descriptors were measured. The scored phenotypes and their abbreviations are presented in Table S4.

4.2. Genotyping by Sequencing (GBS) and SNP Calling

The restriction site associated (RAD) GBS protocol from von Wettberg et al. [6] was used to generate reduced representation sequence data for 407 accessions (see Text S2). All Illumina data are available from the National Center for Biotechnology database under BioProject PRJNA388691. SNPs were called using the Genome Analysis Tool Kit (GATK) pipeline [27] and further filtered with VCFtools [28]. A total of 2579 SNPs accessions passed all filters, with 407 accessions remaining for further analysis.

4.3. Genetic Data Analyses

Principal component analysis (PCA) was conducted using the "SNPRelate" R library [29]. Custom scripts in Python [30] and R [31] were used to plot depth and distribution of SNPs on chromosomes.

Linkage disequilibrium (LD) was estimated using the squared correlation coefficient (r^2) between genotypes. VCFtools [28] was used to calculate intra-chromosomal and unlinked r^2 values. LD decay was assessed by plotting intra-chromosomal r^2 values against the physical distance (bp) between markers. The parametric 95th percentile of unlinked r^2 values distribution was taken as a critical value. The threshold beyond which the LD was accepted as real physical linkage was estimated to be $r^2 = 0.16$. The intersection of the smothering second degree local regression (LOESS) curve of intra-chromosomal r^2 values with this threshold was considered to be an estimate of the range of LD.

Relationships among genotypes were calculated and maximum likelihood phylogenetic trees were constructed using SNPhylo [32] based on filtered SNPs and drawn using R libraries "phytools" [33] and "ape" [34].

The nucleotide diversity (pi) was estimated from polymorphic sites and separately for each chromosome and geographical group using VCFtools [28]. By considering only polymorphic sites, we overestimate genomic diversity; however, these estimations can be used for between group comparisons. We applied the Mann–Whitney–Wilcoxon test [35] to make between group comparisons.

The Genome-wide complex trait analysis (GCTA) program [36] was used to estimate the proportion of variance in phenotypes explained by all genome-wide SNPs. First, phenotypic data were normalized. Then, the genetic relationships among individuals from genome-wide SNPs were calculated using GCTA-GRM (genetic relationship matrix) analysis. Finally, GCTA-GREML (genome-based restricted maximum likelihood) analysis was performed to estimate the proportion of variance in a phenotype explained by all GWAS SNPs (i.e., the SNP-based heritability).

4.4. Bioclimatic Analysis

Bioclimatic analysis was performed as described in Plekhanova et al. [4]; for details, see Text S3. Nineteen quantitative bioclimatic variables were used in the analysis (Table S2).

Shapiro–Wilk test for normality [37] was implemented to quantitative phenotypic traits and quantitative bioclimatic variables. Spearman correlation coefficients were calculated using the "rcorr" function from the "Hmisc" R library [38].

4.5. Mapping Approaches

GWAS analysis was performed using a single-locus linear mixed model, implemented in FaST-LMM toolset (factored spectrally transformed linear mixed models) [39]. Principal component analysis (PCA) of 2579 SNPs revealed that the first eight significant principal components (PCs) explained 48% of the variance of all markers. The LMM model was implemented with the first eight PCA axes scores used as covariates for all phenotypic and bioclimatic data. Principal coordinate analysis (PCoA), based on geographical distances between the accessions, was performed using the "pco" function from the "labdsv" library [40] in R, and revealed that the first two significant PCs explained 59% of the variance. We repeated the GWAS analysis including the first eight PCA axes scores and the first two PCoA axes scores as covariates for all traits. In both cases, we used genomic

control parameter (λ_{GC}) and a false discovery rate (FDR) [41] of 0.05 to determine significant trait associated loci separately for each trait. Manhattan plots were performed using "CMplot" library [42] in R.

Annotation of significant associated markers was performed using the SNPEff program [43], as well as the legume information system (LIS) [44] and the LegumeIP [45] databases.

4.6. Biogeographic Analyses

In total, 407 accessions were split into six distinct groups reflecting geographic locations (Table S1): Ethiopia ("ETHI"), India ("IND"), Lebanon ("LEB"), Morocco ("MOR"), Turkey ("TUR"), and Central Asia ("C_ASIA"). The Mann–Whitney–Wilcoxon test [35] was used to identify differences among groups for each bioclimatic variable.

4.7. Haploblock Enrichment Analysis and Clustering of Enriched Haploblocks

To divide the genome into haplotype blocks (haploblocks) based on linkage disequilibrium, Haploview tools [16] were applied to the set of 2579 SNPs. Chromosomal regions with strong linkage were identified using default Haploview parameters (confidence interval for LD [0.7, 0.98]). Each haploblock was considered as the set of SNPs located within a given haploblock. We analysed haploblock enrichment for SNPs associated with trait or variable by applying the logic of gene-set enrichment analysis implemented in the FGSEA method [17], which takes as input data the list of all SNPs ranked by increasing GWAS *p*-values and the list of haploblocks. The method returns an enrichment score and FDR corrected *p*-value [41] for each haploblock. We performed FGSEA analysis for each trait (phenotype and bioclimatic variable), and haploblocks significantly enriched for associated SNPs were defined as those having positive enrichment scores and significantly low FDR corrected *p*-values (<0.05). The outcome of this analysis was that each phenotype or bioclimatic variable was characterized by a set of haploblocks significantly enriched with associated SNPs. To obtain groups of phenotypes and variables sharing sets of enriched haploblocks, we applied bi-clustering on the matrix of pairwise similarities between traits. To estimate the degree of overlap between haploblocks enriched for SNPs associated with different traits, we calculated the haploblock simalarity score as a sum of common haploblocks (i.e., haploblocks enriched for SNPs associated with both traits) divided by the sum of all haploblocks significantly enriched for SNPs associated with these two traits.

5. Conclusions

The Vavilov seed bank contains numerous landraces collected nearly one hundred years ago, and thus may contain 'genetic gems' with the potential to enhance modern breeding efforts. Here, we analyze 407 landraces, sampled from major historic centers of chickpea cultivation and secondary diversification. The collection was grown in the southern European part of Russia in 2016 with climatic conditions well suited for chickpea cultivation. GWAS conducted on both phenotypic traits and bioclimatic variables at landraces sampling sites as extended phenotypes resulted in 84 GWAS hits associated to various regions, most of which were not discovered in prior GWAS and QTL analyses. The novel haploblock-based test identified haploblocks enriched for SNPs associated with phenotypes and bioclimatic variables, of which eleven haploblocks intersect with previously reported GWAS hits on chromosomes Ca1, Ca3, Ca4, Ca5, Ca6, Ca7, and Ca8. Subsequent bi-clustering of traits sharing enriched haploblocks underscored both non-random distribution of SNPs among several haploblocks and their association with multiple traits. We suggest that these clusters of pleiotropic SNPs represent co-adapted genetic complexes to a range of environmental conditions that chickpea experienced during domestication and subsequent geographic radiation. We observed significant genomic diversity in Central Asia, which may have been a bridge for subsequent radiation in India and nearby areas. Linking genetic variation to phenotypic data and a wealth of historic information preserved in historic seed banks are the keys for genome-based and environment-informed breeding intensification.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/11/3952/s1.

Author Contributions: Formal analysis, A.S., P.L.C., A.A.I., N.V.N., E.v.W., S.V.N., and M.G.S.; Investigation, S.V.B., N.C.-G., M.A.V., and D.R.C.; Writing—original draft, A.S., E.v.W., D.R.C., S.V.N., and M.G.S. All authors have read and agreed to the published version of the manuscript.

Funding: RSF grant # 16-16-00007 to A.S., S.V.B., A.I., M.A.V., S.V.N. and M.G.S. supports dataset phenotyping, GWAS implementation and analysis, bioclimatic analysis, biogeographic analysis, and enrichment analysis of haploblocks. This work was also supported by a cooperative agreement from the United States Agency for International Development under the Feed the Future Program AID-OAA-A-14-00008 to D.R.C., E.v.W., and S.V.N.; by the Zumberge Foundation to S.V.N.; and by a grant from the U.S. National Science Foundation Plant Genome Program under Award IOS-1339346 to D.R.Cook and E.v.W. We also acknowledge support from the Government of Norway through the Global Crop Diversity Trust CWR14NOR23.3 07 to D.R.C. and E.v.W. E.v.W. is further supported by the USDA Hatch program through the Vermont State Agricultural Experimental Station.

Acknowledgments: We thank Victoria Scobeyeva for helping in DNA extraction.

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

References

1. Fairchild, D. *The World was My Garden: Travels of Plant Explorer*; LWW: New York, NY, USA, 1939; 495p.
2. Maxted, N.; Dulloo, M.E.; Ford-Lloyd, B.V. *Enhancing Crop Genepool Use: Capturing Wild Relative and Landrace Diversity for Crop Improvement*; CABI: Oxfordshire, UK, 2016; 469p.
3. Vishnyakova, M.A.; Burlyaeva, M.O.; Bulyntsev, S.V.; Seferova, I.V.; Plekhanova, E.S.; Nuzhdin, S.V. Chickpea landraces from centers of the crop origin: Diversity and differences. Sel'skokhozyaistvennaya biologiya. *Agric. Biol.* **2017**, *52*, 976–985.
4. Plekhanova, E.; Vishnyakova, M.A.; Bulyntsev, S.; Chang, P.L.; Carrasquilla-Garcia, N.; Negash, K.; Nuzhdin, S.V. Genomic and phenotypic analysis of Vavilov's historic landraces reveals the impact of environment and genomic islands of agronomic traits. *Sci. Rep.* **2017**, *7*, 4816. [CrossRef] [PubMed]
5. Redden, R.J.; Berger, J.D. History and origin of Chickpea. In *Chickpea Breeding & Management*; Yadav, S.S., Redden, R., Chen, W., Sharma, B., Eds.; CABI: Wallingford, UK, 2007; pp. 1–13.
6. Von Wettberg, E.J.; Chang, P.L.; Başdemir, F.; Carrasquila-Garcia, N.; Korbu, L.B.; Moenga, S.M.; Cordeiro, M.A. Ecology and community genomics of an important crop wild relative as a prelude to agricultural innovation. *Nat. Commun.* **2018**, *9*, 1–13. [CrossRef] [PubMed]
7. Varma Penmetsa, R.; Carrasquilla-Garcia, N.; Bergmann, E.M.; Vance, L.; Castro, B.; Kassa, M.T.; Coyne, C.J. Multiple post-domestication origins of kabuli chickpea through allelic variation in a diversification-associated transcription factor. *New Phytol.* **2016**, *211*, 1440–1451. [CrossRef] [PubMed]
8. Vavilov, N.I. The origin, variation, immunity and breeding of cultivated plants (Translated by S.K. Chestitee). *Chron. Botonica* **1951**, *13*, 1–366.
9. Kale, S.M.; Jaganathan, D.; Ruperao, P.; Chen, C.; Punna, R.; Kudapa, H.; Garg, V. Prioritization of candidate genes in 'QTL-hotspot' region for drought tolerance in chickpea (*Cicer arietinum L.*). *Sci. Rep.* **2015**, *5*, 15296. [CrossRef]
10. Thudi, M.; Khan, A.W.; Kumar, V.; Gaur, P.M.; Katta, K.; Garg, V.; Varshney, R.K. Whole genome re-sequencing reveals genome-wide variations among parental lines of 16 mapping populations in chickpea (*Cicer arietinum L.*). *BMC Plant Biol.* **2016**, *16*, 10. [CrossRef]
11. Chollet, R.; Vidal, J.; O'Leary, M.H. PHOSPHOENOLPYRUVATE CARBOXYLASE: A ubiquitous, highly regulated enzyme in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **1996**, *47*, 273–298. [CrossRef]
12. Minic, Z.; Rihouey, C.; Do, C.T.; Lerouge, P.; Jouanin, L. Purification and characterization of enzymes exhibiting beta-D-xylosidase activities in stem tissues of Arabidopsis. *Plant Physiol.* **2004**, *135*, 867–878. [CrossRef]
13. Chen, X.; Yang, B.; Huang, W.; Wang, T.; Li, Y.; Zhong, Z.; Yang, L.; Li, S.; Tian, J. Comparative proteomic analysis reveals elevated capacity for photosynthesis in polyphenol oxidase expression-silenced *Clematis terniflora* DC. Leaves. *Int. J. Mol. Sci.* **2018**, *19*, 3897. [CrossRef]
14. Ryšlavá, H.; Valenta, R.; Hýsková, V.; Křížek, T.; Liberda, J.; Coufal, P. Purification and enzymatic characterization of tobacco leaf β-N-acetylhexosaminidase. *Biochimie* **2014**, *107 Pt B*, 263–269. [CrossRef]

15. Hellens, R.P.; Moreau, C.; Lin-Wang, K.; Schwinn, K.E.; Thomson, S.J.; Fiers, M.W.; Davies, K.M. Identification of mendel's white flower character. *PLoS ONE* **2010**, *5*, e13230. [CrossRef] [PubMed]
16. Barrett, J.C.; Fry, B.; Maller, J.; Daly, M.J. Haploview: Analysis and visualization of, L.D. and haplotype maps. *Bioinformatics* **2005**, *21*, 263–265. [CrossRef] [PubMed]
17. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* **2016**, 060012. [CrossRef]
18. Olsen, K.M.; Wendel, J.F. A bountiful harvest: Genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **2013**, *64*, 47–70. [CrossRef]
19. Upadhyaya, H.D.; Bajaj, D.; Das, S.; Saxena, M.S.; Badoni, S.; Kumar, V.; Parida, S.K. A genome-scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. *Plant Mol. Biol.* **2015**, *89*, 403–420. [CrossRef]
20. Bajaj, D.; Upadhyaya, H.D.; Das, S.; Kumar, V.; Gowda, C.L.L.; Sharma, S.; Parida, S.K. Identification of candidate genes for dissecting complex branchnumber trait in chickpea. *Plant Sci.* **2016**, *245*, 61–70. [CrossRef]
21. Kujur, A.; Upadhyaya, H.D.; Bajaj, D.; Gowda, C.L.L.; Sharma, S.; Tyagi, A.K.; Parida, S.K. Identification of candidate genes and natural allelic variants for QTLs governing plant height in chickpea. *Sci. Rep.* **2016**, *6*, 27968. [CrossRef]
22. Das, S.; Upadhyaya, H.D.; Srivastava, R.; Bajaj, D.; Gowda, C.L.; Sharma, S.; Singh, S.; Tyagi, A.K.; Parida, S.K. Genome-wide insertion-deletion (InDel) marker discovery and genotyping for genomics-assisted breeding applications in chickpea. *DNA Res.* **2015**, *22*, 377–386. [CrossRef]
23. Ortega, R.; Hecht, V.F.G.; Freeman, J.S.; Rubio, J.; Carrasquilla-Garcia, N.; Mir, R.R.; Penmetsa, R.V.; Cook, D.R.; Millan, T.; Weller, J.L. Altered Expression of an, *F.T.* Cluster underlies a major locus controlling domestication-related changes to chickpea phenology and growth habit. *Front. Plant Sci.* **2019**, *10*, 824. [CrossRef]
24. Varshney, R.K.; Thudi, M.; Roorkiwal, M.; He, W.; Upadhyaya, H.D.; Yang, W.; Doddamani, D. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat Genet.* **2019**, *51*, 857–864. [CrossRef] [PubMed]
25. Sokolkova, A.B.; Chang, P.L.; Carrasquila-Garcia, N.; Noujdina, N.V.; Cook, D.R.; Nuzhdin, S.V.; Samsonova, M.G. The signatures of ecological adaptation in the genomes of chickpea landraces. *Biophysics* **2020**, *65*, 237–240. [CrossRef]
26. Van-Oss, R.P.; Gopher, A.; Kerem, Z.; Peleg, Z.; Lev-Yadun, S.; Sherman, A.; Abbo, S. Independent selection for seed free tryptophan content and vernalization response in chickpea domestication. *Plant Breed.* **2018**, *137*, 290–300. [CrossRef]
27. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; DePristo, M.A. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
28. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; McVean, G. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]
29. Zheng, X.; Levine, D.; Shen, J.; Gogarten, S.; Laurie, C.; Weir, B. A High-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **2012**, *28*, 3326–3328. [CrossRef]
30. Python Software Foundation. Python Language Reference, Version 2.7. Available online: http://www.python.org (accessed on 20 June 2018).
31. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. Available online: https://www.R-project.org/ (accessed on 20 June 2018).
32. Lee, T.H.; Guo, H.; Wang, X.; Kim, C.; Paterson, A.H. SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genom.* **2014**, *15*, 162. [CrossRef]
33. Revell, L.J. phytools: An, R. package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **2012**, *3*, 217–223. [CrossRef]
34. Paradis, E.; Schliep, K. Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R. Bioinformatics. 2018. Available online: https://doi.org/10.1093/bioinformatics/bty633 (accessed on 15 June 2018).
35. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]

36. Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **2011**, *88*, 76–82. [CrossRef]
37. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611. [CrossRef]
38. Harrell, F.E., Jr. Hmisc: Harrell Miscellaneous. R Package Version 4.1-1. 2018. Available online: https://CRAN.R-project.org/package=Hmisc (accessed on 15 June 2018).
39. Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C.M.; Davidson, R.I.; Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **2011**, *8*, 833–835. [CrossRef] [PubMed]
40. Roberts, D.W. Labdsv: Ordination and Multivariate Analysis for Ecology. R Package Version 1.8-0. 2016. Available online: http://CRAN.R-project.org/package=labdsv (accessed on 15 June 2018).
41. Storey, J.D. The positive false discovery rate: A Bayesian interpretation and the q-Value. *Source Ann. Stat. Ann. Stat.* **2003**, *31*, 2013–2035. [CrossRef]
42. CMplot: Circle Manhattan Plot. Available online: https://github.com/YinLiLin/R-CMplot (accessed on 20 June 2018).
43. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly Austin* **2012**, *6*, 80–92. [CrossRef]
44. Dash, S.; Campbell, J.D.; Cannon, E.K.; Cleary, A.M.; Huang, W.; Kalberer, S.R.; Weeks, N.T. Legume information system (LegumeInfo. org): A key component of a set of federated data resources for the legume family. *Nucl. Acids Res.* **2016**, *44*, D1181–D1188. [CrossRef]
45. Li, J.; Dai, X.; Liu, T.; Zhao, P.X. LegumeIP: An integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acids Res.* **2012**, *40*, 1221–1229. [CrossRef]