Plant Communications
Research Article

CellPress
Partner Journal

High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement

Changyou Liu[1,6], Yan Wang[1,6], Jianxiang Peng[2,6], Baojie Fan[1], Dongxu Xu[3], Jing Wu[4], Zhimin Cao[1], Yunqing Gao[3], Xueqing Wang[1], Shutong Li[3], Qiuzhu Su[1], Zhixiao Zhang[1], Shen Wang[1], Xingbo Wu[5], Qibing Shang[3], Huiying Shi[1], Yingchao Shen[1], Bingbing Wang[2,*] and Jing Tian[1,*]

[1]Institute of Cereal and Oil Crops, Hebei Academy of Agricultural and Forestry Sciences/Hebei Laboratory of Crop Genetics and Breeding, Shijiazhuang 050035, China

[2]Biobin Data Sciences, Changsha 410221, China

[3]Zhangjiakou Academy of Agricultural Sciences, Zhangjiakou 075300, China

[4]Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[5]Tropical Research and Education Center, Department of Environmental Horticulture, University of Florida, 18905 SW 280th St, Homestead, FL 33031, USA

[6]These authors contributed equally to this article.

***Correspondence: Bingbing Wang (bwang@biobin.com.cn), Jing Tian (nkytianjing@163.com)**

https://doi.org/10.1016/j.xplc.2022.100352

ABSTRACT

Mung bean is an economically important legume crop species that is used as a food, consumed as a vegetable, and used as an ingredient and even as a medicine. To explore the genomic diversity of mung bean, we assembled a high-quality reference genome (Vrad_JL7) that was ~479.35 Mb in size, with a contig N50 length of 10.34 Mb. A total of 40,125 protein-coding genes were annotated, representing ~96.9% of the genetic region. We also sequenced 217 accessions, mainly landraces and cultivars from China, and identified 2,229,343 high-quality single-nucleotide polymorphisms (SNPs). Population structure revealed that the Chinese accessions diverged into two groups and were distinct from non-Chinese lines. Genetic diversity analysis based on genomic data from 750 accessions in 23 countries supported the hypothesis that mung bean was first domesticated in south Asia and introduced to east Asia probably through the Silk Road. We constructed the first pan-genome of mung bean germplasm and assembled 287.73 Mb of non-reference sequences. Among the genes, 83.1% were core genes and 16.9% were variable. Presence/absence variation (PAV) events of nine genes involved in the regulation of the photoperiodic flowering pathway were identified as being under selection during the adaptation process to promote early flowering in the spring. Genome-wide association studies (GWASs) revealed 2,912 SNPs and 259 gene PAV events associated with 33 agronomic traits, including a SNP in the coding region of the SWEET10 homolog (jg24043) involved in crude starch content and a PAV event in a large fragment containing 11 genes for color-related traits. This high-quality reference genome and pan-genome will provide insights into mung bean breeding.

Key words: mung bean, long-read sequencing, *de novo* assembly, pan-genome, gene PAV, GWAS

Liu C., Wang Y., Peng J., Fan B., Xu D., Wu J., Cao Z., Gao Y., Wang X., Li S., Su Q., Zhang Z., Wang S., Wu X., Shang Q., Shi H., Shen Y., Wang B., and Tian J. (2022). High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. Plant Comm. **3**, 100352.

INTRODUCTION

Mung bean (*Vigna radiata* L.) is a self-pollinated, fast-growing diploid legume crop species (2n = 2x = 22). As an inexpensive source of carbohydrates, protein (~27% of the dry seed content), folic acid, and iron, mung bean is economically important and widely cultivated, mainly in Asia (Kang et al., 2014). Mostly used as a food (grains), a vegetable (sprouts), and an ingredient

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

(paste), mung bean is also a well-known medicine in China for reducing body heat because of its rich content of flavonoids, vitexin (VITE), and isovitexin (ISOVITE) (Cao et al., 2011). Although high-quality genomic resources are available for many legume crop species, such as soybean (Shen et al., 2019), cowpea (Lonardi et al., 2019), and lima bean (Garcia et al., 2021), and long-read sequencing has enabled high-quality, chromosome-scale assemblies for many plant species (Michael and VanBuren, 2020), there is only a draft genome sequence available for mung bean. This sequence was assembled from Illumina short-read sequences, covering 80% of the estimated genome size, with 239 of 2748 scaffolds organized into 11 pseudochromosomes (Kang et al., 2014).

Recently, several studies on the genetic diversity of mung bean and genome-wide association studies (GWASs) of mung bean germplasm have been performed (Schafleitner et al., 2015; Noble et al., 2017; Breria et al., 2020; Reddy et al., 2020; Sokolkova et al., 2020). A common limitation of these studies was that they were either based on genotyping by sequencing (GBS) or diversity array technology, which revealed only a few thousand markers and was thus insufficient for large-scale gene mining. Whole-genome resequencing of mung bean, as has been applied to soybean (Zhou et al., 2015; Fang et al., 2017), pigeon pea (Varshney et al., 2017), chickpea (Varshney et al., 2019b), and common bean (Wu et al., 2020a), is greatly needed to better understand the genetic variation, nucleotide diversity, population structure, and key genes that govern important agronomic traits.

Owing to variation among different individuals, a single reference genome cannot contain all possible genetic information (Golicz et al., 2016; Hurgobin et al., 2018). The concept of the pan-genome, usually constructed by sequencing dozens to hundreds of individuals, has been proposed to represent the complete genome information of a species (Tettelin, 2005; Golicz et al., 2016). Pan-genomes have been assembled for many plant species, including rice (Qin et al., 2021), rapeseed (Song et al., 2020), soybean (Liu et al., 2020; Torkamaneh et al., 2021), chickpea (Varshney et al., 2021), and pepper (Ou et al., 2018a). Two pan-genome studies have been published for soybean: a graphic pan-genome constructed by *de novo* assembly of 26 representative wild and cultivated accessions using long-read sequencing (Liu et al., 2020) and a pan-genome comprising 204 cultivated soybeans (PanSoy) constructed using next-generation sequencing short reads (Torkamaneh et al., 2021). PanSoy explores the extent of genetic variation in cultivated soybean. The pigeon pea pan-genome was assembled as the first pan-genome of orphan legumes (Zhao et al., 2020). No pan-genome for mung bean has been constructed to date, hindering the progress of genetic discovery.

In this study, we assembled a high-quality reference genome and pan-genome of Chinese mung bean germplasm via deep sequencing of a high-yield variety and 217 accessions. Important agronomic traits, such as yield components, grain composition, morphology, and insect resistance, were measured for these accessions and associated with their genome sequences. Significant single-nucleotide polymorphisms (SNPs) and candidate genes were identified for almost all of the studied traits. These results lay a solid foundation for genomic breeding of mung bean.

RESULTS

Genome assembly and annotation

A high-yielding and early-maturing mung bean variety widely cultivated in China (Jilv 7 [JL7]) was sequenced using the PacBio Sequel II platform (long reads, ~52.83 Gb, 110.21× genome coverage) and Illumina paired-end (PE) read technology (short reads, ~61.28 Gb, 127.85× genome coverage). The estimated genome size was ~479.35 Mb, and the heterozygosity rate was ~0.056%, as calculated by the K-mer (k = 31) method (Supplemental Figure 1); it was ~11.7% smaller than the estimated size of the VC1973A genome (~543 Mb) (Kang et al., 2014). The *de novo*-assembled genome (Vrad_JL7) had a size of 475.19 Mb, which was ~99.13% of the estimated size, with an N50 of 10.34 Mb, and the largest contig was 30.20 Mb (Table 1). Approximately 98.72% (~469.11 Mb) of the genome was anchored onto 11 pseudomolecules according to Hi-C sequencing data (~33.5 Gb) (Figure 1A and Supplemental Figure 2). There were 259 gaps (~0.11 Mb) and 518 contigs (~6.08 Mb) that remained unanchored in the genome (Table 1). Pseudomolecules were named according to those of the VC1973A genome (Kang et al., 2014). Approximately 53.45% of the mung bean genome was composed of repetitive elements (Table 1). Long terminal repeat (LTR) retrotransposons accounted for 33.05% of the genome, and DNA transposons accounted for 4.25%. The LTR/Gypsy and LTR/Copia elements constituted 16.65% and 13.77%, respectively (Supplemental Table 1), which differed from the 25.2% and 11.3% in the VC1973A genome (Kang et al., 2014).

After masking the repetitive regions of the genome, a total of 40 125 protein-coding genes (17 pseudogenes were excluded) and 42 986 transcripts were annotated by combining *ab initio* gene prediction, RNA sequencing (RNA-seq), and protein homology evidence using the BRAKER2 pipeline (Bruna et al., 2021). Among them, 29 114 (72.56%) genes had either RNA-seq or homology evidence (Supplemental Data 1). Possible functions for protein-coding genes were annotated, and 81.60% (32 754) could be assigned functions via the Kyoto Encyclopedia of Genes and Genomes (KEGG) (37.45%), Gene Ontology (GO) (39.83%), Pfam (58.41%), SwissProt (60.26%), or NCBI nonredundant (NR) protein (81.38%) databases. For RNA-encoding genes, 5830 noncoding RNA genes of various types were also predicted using Barrnap and Infernal (version 1.1.4) (Nawrocki and Eddy, 2013) and by searching the Rfam database (Supplemental Table 2).

The overall quality of the Vrad_JL7 assembly was very high. The completeness was >98%, as revealed by Illumina PE read mapping and Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment using eudicots_odb10 (Table 1; Supplemental text; Supplemental Figure 4). Many evaluation scores were better for the Vrad_JL7 assembly than for the VC1973A assembly, including a nearly four-fold greater contig N50 length, fewer gaps, and an LTR assembly index (LAI) of 15.67, demonstrating the reference quality of the assembly (Table 1). The BUSCO completeness for the predicted protein-coding genes was 96.9%, a 20% increase compared with that of the VC1973A assembly (Table 1; Supplemental Figure 4). These results demonstrated that the Vrad_JL7 genome assembly

Genomic feature	Vrad_JL7	VC1973A v1	VC1973A v2
Total assembly size, Mb	475.35	430.88	475.7
No. contigs	632	25 922	1511
Largest contig	30.20 Mb	734.56 kb	12.73 Mb
Contig N50	10.34 Mb	48.83 kb	2.8 Mb
Scaffold N50, Mb	43.79	1.52	47.1
Percentage anchored to chromosomes	98.72	73.09	89.92
No. gaps	259	96 874	1047
Length of gaps, Mb	0.11	33.56	1.81
GC content, %	33.45	33.16	33.27
Complete BUSCOs (genome), %	98.02	96.82	91.36
Complete BUSCOs (protein), %	96.90	81.60	80.01
LAI	15.67	7.86	14.65
Intact LTR-RTs	2725	734	2458
Repetitive sequences, %	53.45	50.10	52.79
Protein-coding genes	40 125	22 427	30 958

Table 1. Summary of assembly and annotations of the Vrad_JL7 and VC1973A genomes

and its annotations were more complete and of higher quality than those of the two versions of VC1973A (Kang et al., 2014; Ha et al., 2021), and it was by far the best quality mung bean genome.

Comparative genomic and evolutionary analysis

To identify evolutionary features of the mung bean genome, the sequences of annotated genes from 12 eudicot plant species were compared with those of the Vrad_JL7 gene set. A total of 32,253 orthogroups/gene families were formed, including 432 single-copy groups. A total of 35 059 (87.4%) mung bean coding genes clustered into gene families, 1532 of which (including 5482 genes) were specific to mung bean (Figure 1B; Supplemental Table 3). The functions of these specific genes were enriched in starch and sucrose metabolism; biosynthesis of amino acids; ribosome biogenesis in eukaryotes; glycine, serine, and threonine metabolism; and more (Figure 1C). These results are consistent with the rich starch and protein content of mung bean grains. Further analysis revealed that 2218 gene families had undergone expansion events and 1093 had undergone contraction events in the mung bean genome (Supplemental Figure 5). The functions of these expanded families were significantly enriched (adjusted $P < 0.05$) in various biological processes and pathways related to the characteristic features of mung bean or enabling adaptation to the environment, including plant–pathogen interactions, biosynthesis of isoflavonoids and terpenoids, and metabolism of unsaturated fatty acids (Supplemental Figure 6; Supplemental Data 2 and 3). Tandem duplication contributed significantly to gene family expansion (Supplemental text). Functions of the contracted gene families were mainly enriched in amino acid and starch metabolism. Interestingly, some contracted family members also had functions in the metabolism of flavonoids or fatty acid pathways (Supplemental Figure 8). Given that there are many different steps in biological processes, the unique, expanded, and contracted gene families likely fine-tuned the path selection in different steps, leading to specific features such as the rich content of certain types of flavonoids (including VITE and ISO-VITE) in mung bean.

Large-scale genome comparison revealed a perfect one-to-one collinear relationship between the chromosomes of mung bean (Vrad_JL7) and adzuki bean (*Vang*) (Figure 1D), including Vrad_JL7 chromosome 3 corresponding to *Vang* chromosome 5. Chromosome 3 of VC1973A v1, however, was probably incorrectly assembled and should be part of chromosome 4 instead (Supplemental Figure 9A). Most chromosomes of cowpea (*Vung*) had good collinearity with mung bean and adzuki bean, with the exception of chromosome 5 of cowpea, which was split into two chromosomes in mung bean and adzuki bean (Figure 1D), suggesting that the event occurred after the divergence of the ancestor of cowpea and the common ancestor of mung bean and adzuki bean.

Population genomic analysis

Genetic diversity in mung bean was evaluated in depth by resequencing 217 accessions: 24 Chinese breeding lines (CBLs), 165 Chinese landraces (CLRs), and 28 non-Chinese lines (NCLs). The average sequencing depth was 12.28× and ranged from 7.06× to 15.31× (Supplemental Table 4). A total of 2 229 343 SNPs, 230 025 short insertions and deletions (indels) (<15 bp), and 56 545 structural variations (SVs) (39 228 large indels [>15 bp], 1991 copy-number variations, 13 937 translocations, and 1389 inversion events) were identified. Their distributions in the genome are summarized in Supplemental Table 5 and Supplemental Table 6 and the Supplemental text. Overall, the different types of variants had similar density distribution patterns in the genome (Figure 2A).

An unrooted tree comprising 207 mung bean accessions (after removing 10 accessions with ambiguous sources) was constructed based on the core SNPs, dividing these accessions into

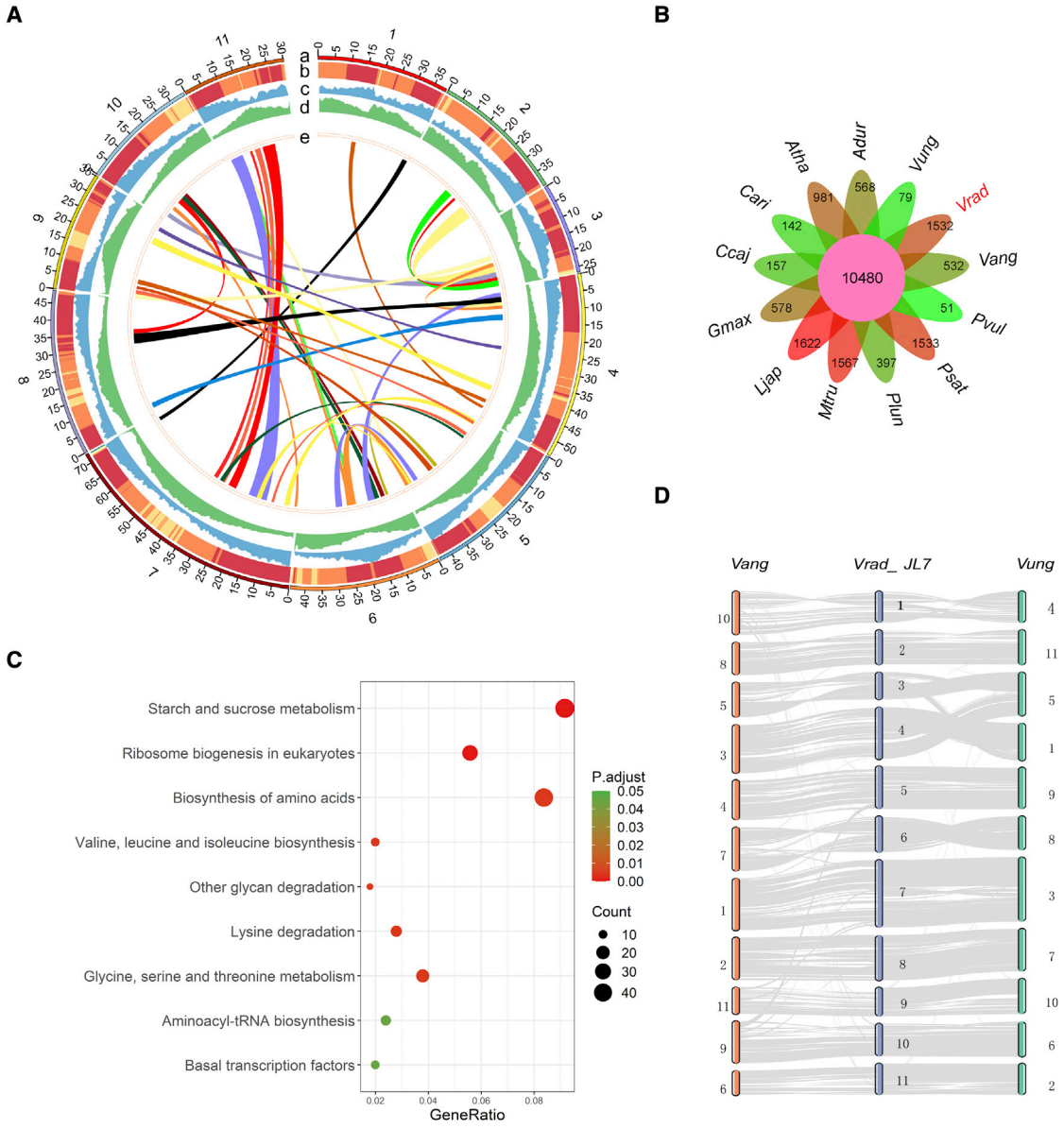

Figure 1. Genome assembly of mung bean and comparative genomic analysis of members of the Leguminosae.

(A) Landscape of genomic features of mung bean. The circles represent, from outermost to innermost, (a) pseudochromosomes, (b) GC content, (c) distribution of genes, (d) DNA transposon and retrotransposon density, and (e) intragenome collinear blocks.

(B) Orthologous gene families among 12 species of Leguminosae and *Arabidopsis thaliana* identified by OrthoFinder (Emms and Kelly, 2019). The numbers represent the gene families identified for each species.

(C) KEGG pathway enrichment of specific gene families in mung bean.

(D) Genomic collinearity between *Vigna radiata* (*Vrad_JL7*), *Vigna angularis* (*Vang*), and *Vigna unguiculata* (*Vung*).

3 groups. The NCL group contained mainly NCL lines plus 3 CBL/CLR lines with a close relationship to NCL lines. The vast majority of CBL and CLR lines clustered within either group 1 or group 2, where group 1 members were genetically closer to those of the NCL Group (Figure 2B). Population structure (K = 3) and principal component analysis (PCA) supported the same grouping structure (Figure 2C). The tree topology based on SNPs was very similar to the results of hierarchical clustering analysis based on gene presence/absence variation (PAV) results (Mantel statistic r: 0.9612; significance: 1e−4) (Supplemental Figure 10). All of the accessions from south China clustered in group 1, and most of the accessions from north China were in group 2 (Figure 2D). The NCL group had the highest nucleotide diversity ($\theta\pi$ = 1.63×10^{-3}) compared with those of group 1 (1.34×10^{-3}) and group 2 (1.02×10^{-3}) or those of south China (1.40×10^{-3}) and north China (1.37×10^{-3}). Pairwise F_{ST} between the NCL group and group 2 was the highest (0.283) compared with those of the NCL group versus group 1 (0.114) and group 1 versus group 2 (0.143), and the F_{ST} of the NCL group versus north China (0.151) was the highest compared with those of NCL group versus south China (0.088) and north China versus south China (0.069). These results indicate that the genetic basis narrowed during the

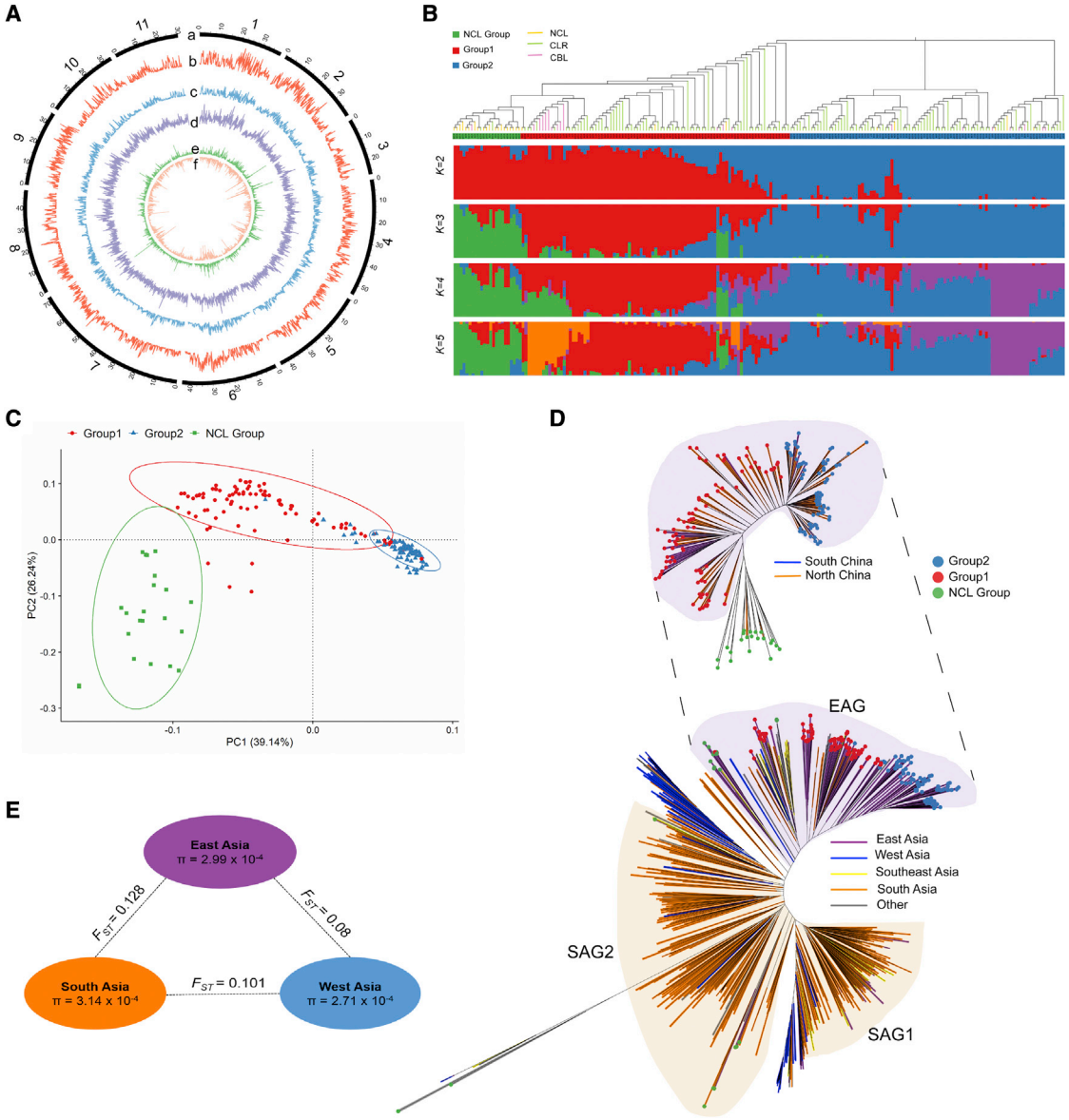

Figure 2. Population genomic analysis of mung bean.

(A) Atlas of variation of 217 accessions. The circles represent, from outermost to innermost, (a) pseudochromosomes; (b–d) SNP, indel, and SV density; and (e and f) nonsynonymous and synonymous SNPs.

(B) Unrooted tree and ADMIXTURE (K = 2–5) plot of 207 accessions inferred from core SNPs. The green, red, and blue strips of the tree represent the NCL group, group 1, and group 2, respectively, and the orange, green, and purple branches represent NCLs, CLRs, and CBLs, respectively.

(C) PCA plot of the first 2 eigenvectors of 207 accessions.

(D) Unrooted tree of 207 and 750 accessions. The red, green, and blue at the tips of the tree represent group 1, group 2, and the NCL group, respectively. The branch colors are shown in the legend.

(E) F-statistics (F_{ST}) and nucleotide diversity (π) of different subgroups in Asia.

process of mung bean adaptation, probably owing to migration from outside China to south China and then to north China.

To view the worldwide diversity landscape of mung bean, public GBS data of 533 accessions from 22 countries were collected and compared with the 217 accessions (Supplemental Figure 11A; Supplemental Data 4). A total of 5671 SNPs common to all 750 accessions from 23 countries were identified and used to construct an unrooted tree. As shown in Figure 2D, accessions from east Asia (~83.82% from China and ~2.90% from South Korea) were clustered into one group (east Asia group [EAG]), and accessions from south Asia (~92.46% from India) were clustered into two groups (south Asia group [SAG]1 and SAG2). Other accessions, including 49 from Iran (west Asia), were closely clustered with either SAG1 or SAG2. Wild mung bean accessions formed an outer group relatively closely related to the members of SAG2. There were two subgroups in the east Asia cluster, corresponding to group 1 and group 2 of the Chinese accessions. The $\theta\pi$ value for south Asia accessions was 3.14×10^{-4}, 5% higher than 2.99×10^{-4} for east Asia. The F_{ST} value for east Asia

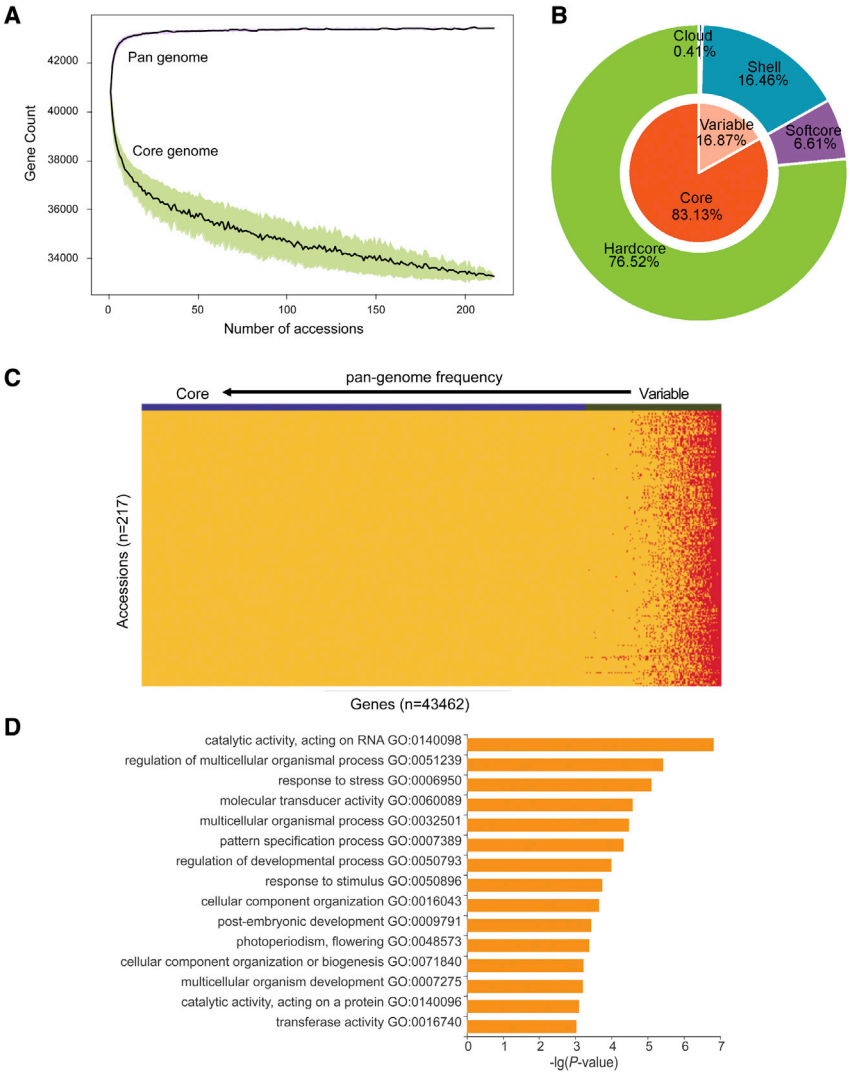

Figure 3. Pan-genome of Chinese mung bean germplasm.

(A) Simulations of the increase in pan-genome size and the decrease in core genome size based on 100 random combinations of each given number of accessions. The pan- and core-genome curves were fitted using data points from all random sub-samples and are indicated by solid black lines. The upper and lower edges of the purple and green areas correspond to the maximum and minimum numbers of genes, respectively.

(B) Composition of the mung bean pan-genome.

(C) Landscape of gene PAVs. The genes are sorted by their occurrence, with the highest frequency of occurrence on the left and the lowest on the right.

(D) GO term enrichment of variable genes.

versus south Asia was 0.128, much higher (60%) than that for east Asia versus west Asia (0.08) and higher (27%) than that for west Asia versus south Asia (0.101) (Figure 2E). These results seemed to support the hypothesis that mung bean originated and was domesticated in south Asia (India) (Fuller, 2007) and spread worldwide from there, including to west Asia and then to east Asia, probably through the Silk Road. After many years of adaptation and selection, mung bean varieties in China have formed two distinct groups.

Pan-genome and PAV analysis

De novo assemblies of all 217 mung bean accessions revealed a total of ~86 Gb contigs, with an average assembly size of ~397 Mb and an average contig N50 of ~3380 bp (Supplemental Table 7). After removing contamination and redundant sequences, a total of ~287.73 Mb of non-reference sequences, consisting of 288 128 contigs, were considered to be additional genome space of mung bean. A total of 3337 additional evidence-supported protein-coding genes were annotated, with an average gene length of 545 bp, much shorter than that of genes in the reference genome (4351 bp), indicating that most of these genes were partial or fragmented. Altogether, the assembled size of the mung bean pan-genome constructed in this study was ~762.92 Mb (including Vrad_JL7), and the total number of annotated genes was 43 462.

A "map-to-pan" strategy (Hu et al., 2017) was used to identify the PAV of genes. Genes with more than 20% of the coding DNA sequence (CDS) region covered by more than two-fold read depth were considered to have "presence" in the accession; the others, "absence." Similar to that of PanSoy (Torkamaneh et al., 2021), the mung bean pan-genome constructed mainly of Chinese accessions seemed closed, as most genes were included when randomly sampling 100 accessions from the collection (Figure 3A). Based on gene frequency, 33 258 (76.5%) genes were defined as hardcore genes, 2872 (6.6%) as softcore genes, 7154 (16.5%) as shell genes, and 178 (0.4%) as cloud genes (Figure 3B; Supplemental Data 5). The high proportion of core gene (83%) content was similar to proportions in the PanSoy (90.6%) (Torkamaneh et al., 2021), tomato (74.2%) (Gao et al., 2019), and pigeon pea (86.6%) pan-genomes (Zhao et al., 2020). Each accession had 15%–20% variable genes (including shell and cloud) (Figure 3C). The average length of variable genes on the reference genome was 2438 bp, with an average of 3 exons, which was significantly shorter than that of core genes (4562 bp), probably owing to fewer exons (3.01 exons per gene) in the soft genes than in the core genes (5.54 exons per gene).

The functions of core genes involved basic biological processes, such as cellular processes, metabolic processes, catalytic activity, and binding (Supplemental Figure 12). The functions of variable genes were enriched in many regulatory and environmental response processes, including catalytic activity, responses to stress and stimuli, regulation of multicellular organismal and developmental processes, and photoperiodic flowering (Figure 3D). The variable genes are likely to have played important roles in enabling mung bean to adapt to

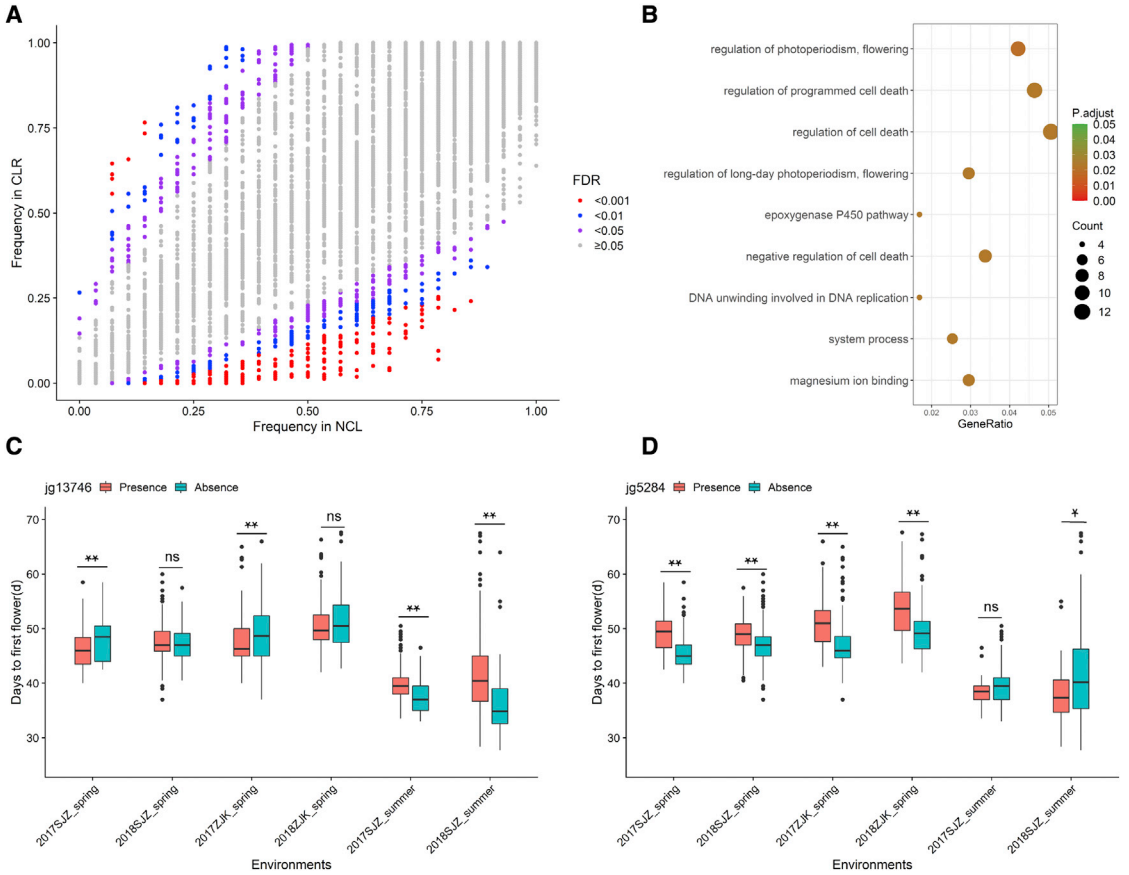

Figure 4. Gene PAV under selection during the adaptation process of mung bean from NCLs to CLRs.

(A) Scatterplots showing gene occurrence frequencies in NCLs and CLRs.

(B) KEGG pathway enrichment of genes that have undergone significant expression changes during adaptation.

(C and **D)** Examples of presence/absence variants of 2 genes (jg13746 and jg5284) that regulate flowering period in different environments. ns, *, and ** indicate statistical significance levels of $P \geq 0.05$, <0.05, and <0.01, respectively.

different environments from the tropics to the temperate zones. Interestingly, the distribution of variable genes was not even, and certain regions, such as 3.0–4.5 Mb on chromosome 2 and 23.0–29.5 Mb on chromosome 1, were hot spots for variable gene clustering (Supplemental Figure 13).

Gene PAV under selection during adaptation

Gene counts and frequencies in the NCLs, CLRs, and CBLs were compared in detail (see Supplemental text). To identify the gene PAV under selection during the adaptation process, the frequency of each gene in the NCLs was plotted against its frequency in the CLRs. As shown in Figure 4A, 809 genes showed significant differences in frequency between the two groups (Fisher's exact test, false discovery rate < 0.05). Among them, 215 genes were considered favorable, as their frequencies increased from NCLs to CLRs, and 412 genes were considered unfavorable. Genes that showed opposite significant changes in frequency from NCLs to CLRs and from CLRs to CBLs were not considered further. Genes related to flowering regulation and programmed cell death were significantly enriched in both favorable and unfavorable genes during adaptation (Figure 4B).

Nine PAV events for genes related to flowering regulation were identified, three of which (jg13350, jg13746, and Pang80812) were present in most CLRs and CBLs. The phenotypic data revealed that the presence of these genes could promote early flowering in spring but was associated with late flowering in summer (Figure 4C). The other six genes (jg1521, jg5273, jg5274, jg5281, jg5284, and Pang68295) were absent from most CLRs and CBLs and present in most NCLs. Their association with flowering phenotype was exactly the opposite of that in the previous case. The absence, instead of the presence, of these genes was associated with early flowering in spring but late flowering in summer (Figure 4D). Data from different environments largely supported the above observations, the signatures of increased linkage disequilibrium (LD), and the loss of genetic diversity in the regions surrounding the PAVs (Supplemental Figure 14), demonstrating that genes that promote early flowering in spring were selected during the adaptation process. No functional groups were found to be significantly enriched for the PAV genes during the improvement process (from CLRs to CBLs), although several genes encoding glucosidases could be candidates for selection.

SNP- and gene PAV-based GWAS of agronomic traits

Phenotypic data for 33 agronomic traits in 217 mung bean accessions were collected in Shijiazhuang (SJZ) for 2 seasons (spring and summer) over 2 years (2017 and 2018) and in Zhangjiakou

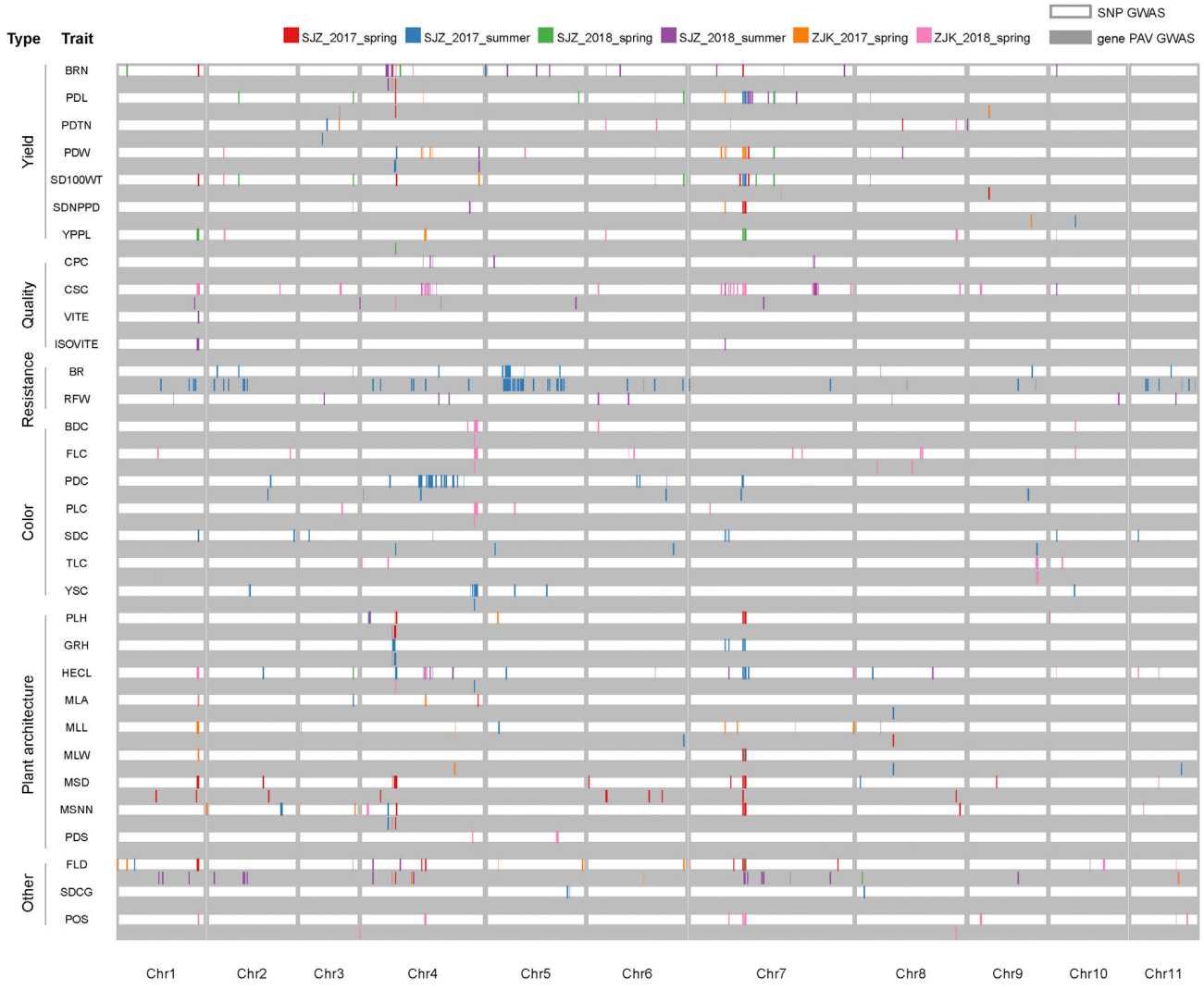

Figure 5. Summary of SNP GWAS and gene PAV GWAS results.
Distribution of all STAs and GPTA events identified by a SNP GWAS (white box) and a gene PAV GWAS (gray box) within the Vrad_JL7 genome for 32 traits of 6 types under 6 different environments. The abbreviations of the traits are shown in Supplemental Table 8.

(ZJK) for 1 season (spring) over 2 years (2017 and 2018). Overall, data for the same trait were strongly correlated across different environments and years. The average broad-sense heritability for traits observed multiple times was 0.72, ranging from 0.3 to 0.98 (Supplemental Table 8). Interestingly, some traits had strong correlations with others. For instance, the average Pearson correlation coefficient was 0.94 for VITE and ISOVITE; 0.89 for maximum leaf length, maximum leaf width, and maximum leaf area; and 0.78 for pod length (PDL), pod width, and 100-seed weight; and the average Spearman correlation coefficient was 0.85 for bud color (BDC), flower color (FLC), petiole color (PLC), and young stem color (YSC) (Supplemental Figure 15). It would not be surprising to discover genes involved in these traits showing pleiotropic effects.

SNP-trait association sites (STAs) were identified for all but one trait (trilobal leaf shape) in at least one environment via GWAS (Figure 5). A total of 2912 STAs were identified for each individual environment, and 1790, 43, and 23 of them were classified as robust, consistent, and stable, respectively (Supplemental Table 9; Supplemental Data 6). Although most STAs were located in intergenic regions, 35% of them were located in the transcribed regions of 248 genes (Supplemental text). Some STAs were shared among highly correlated traits, as was the case for 17 common STAs identified for maximum leaf length and maximum leaf width and 14 STAs for pod width and PDL. Notably, some genomic regions appeared to be hot spots for STAs, as they were associated with multiple traits that were not highly correlated. For example, the terminal region of chromosome 1 (35.499–35.986 Mb) contained 285 STAs for 14 traits, and the middle region of chromosome 7 (16.915–16.993 Mb) contained 132 STAs for 8 traits, including yield-, quality-, and plant architecture–related traits (Figure 5; Supplemental Data 6). These hot-spot genomic regions significantly associated with multiple traits, called "agro-islands," were also found on multiple chromosomes of chickpea (Plekhanova et al., 2017; Varshney et al., 2019a; Sokolkova et al., 2021) and pigeon pea (Varshney et al., 2017); these regions may be closely linked to

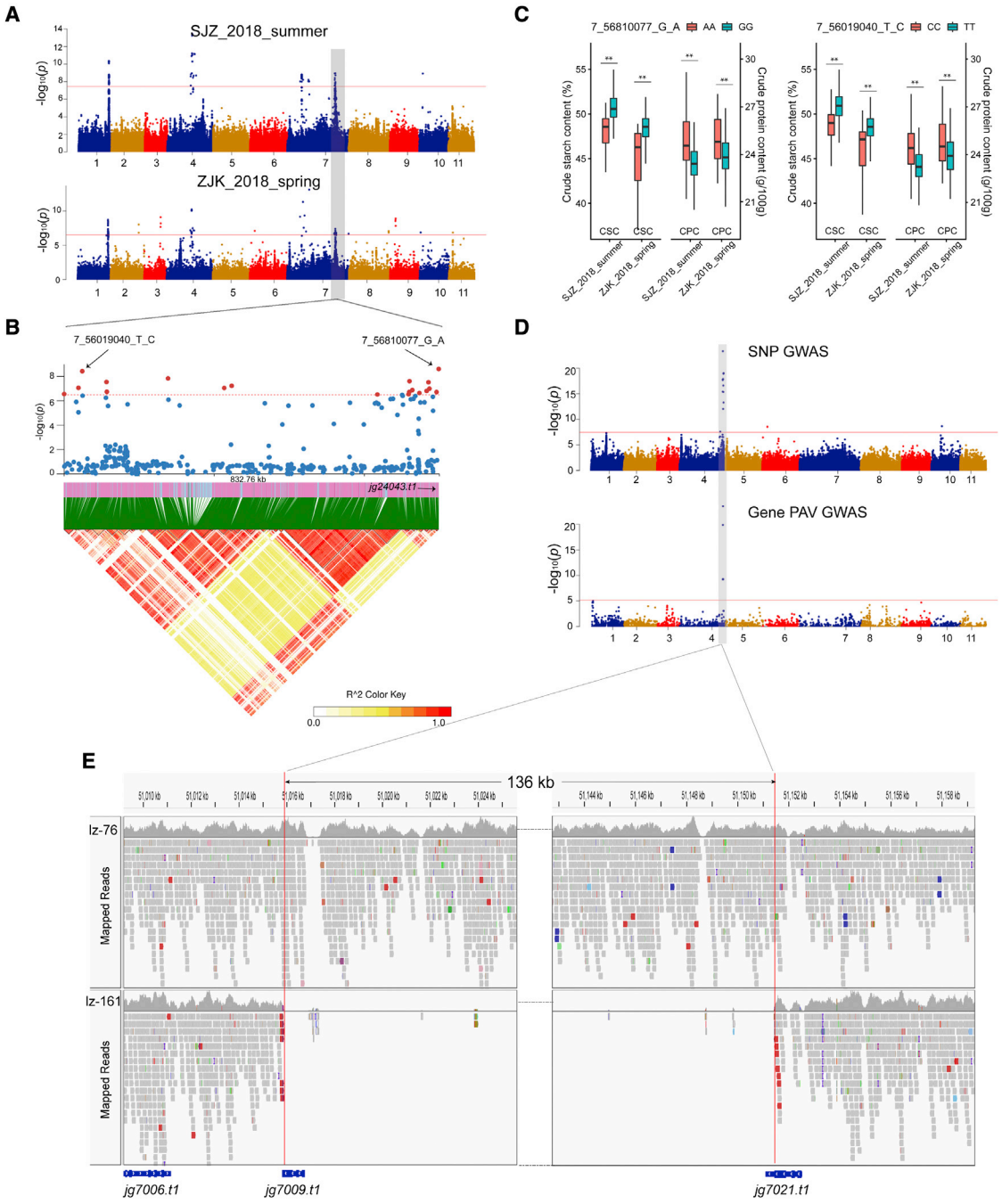

Figure 6. GWAS of CSC and color-related traits.

(A) Manhattan plot of GWAS of CSC traits in 2 different environments (SJZ_2018_summer and ZJK_2018_spring).

(B) LD heatmap showing the regions surrounding the strong peaks (chromosome 7: 55.979–56.812 Mb) identified by a SNP GWAS.

(C) Relationship between the alleles of the 2 STAs (7_56810077_G_A and 7_56019040_T_C) and the CSC and CPC in 2 environments. ** indicates a statistical significance level of $P < 0.001$.

(D) Manhattan plots of SNP-based GWAS and gene PAV-based GWAS for color-related traits, including BDC, FLC, PLC, and YSC.

(E) Presence/absence variants of 136 kb on chromosome 4 according to Integrative Genomics Viewer (IGV). Sample lz-76 (variety: Jilv7) contains this segment, and the color is purple. Sample lz-161 (variety: VC973A) is missing this segment, and the color is green or yellow.

domestication selection. The availability of "islands" in the mung bean genome (e.g., chromosomes 1, 4, and 7) facilitates the identification of breeding targets to reintroduce genetic diversity lost in modern breeding programs owing to domestication and crop improvement.

In addition to SNP GWAS, gene PAV GWAS was conducted by treating gene PAVs as genotyping data. A total of 391 gene PAV-trait association (GPTA) events were identified for 29 of 33 traits in multiple environments, corresponding to 259 unique genes (Figure 5; Supplemental Table 10; Supplemental Data 7).

Although most GPTA genes were associated with only a single trait, 52 genes were related to multiple traits. Approximately 60 GPTA genes were in close proximity (less than 100 kb) to STAs, half of which actually had a distance of less than 10 kb. As SVs would normally be coupled with SNPs in surrounding regions, consistent GPTA and STA events would be very helpful for identifying genes that regulate associated traits. However, reliable GPTA events not related to STA would be complementary to the SNP method for identifying candidate genes, as was reported in previous studies (Song et al., 2020).

Candidate genes associated with important agronomic traits

The contents of VITE and ISOVITE are important grain quality traits. The two traits were highly correlated (r = 0.94); thus, it was not surprising to identify common STAs for these two traits. Two regions were identified for regulation of ISOVITE, one on the terminus of chromosome 1 (35 556 181–35 673 741, 117 kb) and another on chromosome 7. The region on chromosome 1 was also identified for VITE. Twelve candidate genes were annotated in this region. None of them appeared to participate directly in the flavonoid biosynthesis pathway (KEGG map00941). Two genes (jg15859 and jg15860) were annotated as encoding glutamine synthetase, which could be used as a substrate for VITE/ISOVITE synthesis.

For other quality-related traits, a stable STA (7_56810077_G_A) on chromosome 7 was identified for crude starch content (CSC) (Figure 6A). The SNPs were located in the coding region of jg24043, a homolog of the soybean SWEET10 gene. This soybean homolog was reported to be a key gene that regulates seed size, oil content, and protein content (Wang et al., 2020). Interestingly, the most significant SNP (7_56019040_T_C) associated with crude protein content (CPC) was found to have strong LD with jg24043 (Figures 6B and 6C), indicating that the same gene may regulate CSC and CPC concurrently. Another STA identified for CSC was located on chromosome 3 between the jg9561 and jg9562 genes, and both were annotated as β-amylase, supporting their function in starch content.

Significant STA and GPTA events from the same region were identified for several color-related traits, including BDC, FLC, YSC, and PLC (Figure 6D). Eleven consecutive genes (jg7009–jg7020) and part of jg7021 exhibited PAV patterns among the mung bean population. It is very likely that the entire segment of 136 kb on chromosome 4 (51 015 805–51 151 503) containing these genes was missing in some accessions (Figure 6E). For 185 accessions that contained this segment, a purple color appeared in the buds, flowers, young stems, and petioles. For the other 32 accessions that lacked the segment, the corresponding tissues were either green or yellow. Several MYB90-like genes were annotated in the segment (Supplemental Figure 20). The soybean homolog R gene (Glyma.09G235100) plays an important regulatory role in the synthesis of anthocyanins during the process of seed coloring (Gao et al., 2021).

For yield-related traits, a reliable candidate region on chromosome 4 was identified for branch number by consistent STA and GPTA signals in multiple environments, narrowing candidate genes to one of five showing PAV events (Supplemental text). A homolog of NRT1/PTR FAMILY 2.13 could be a candidate gene for PDL, and a homolog of WUSCHEL-related homeobox 3 could be a candidate for yield per plant (YPPL) (Supplemental text). Many other candidate regions/genes were also identified, including a main region on chromosome 5 related to bruchid resistance (BR); two candidate genes (jg3587 and jg35209) for flowering date, which were homologous to the candidate genes Glyma03g01540 and Glyma09g33340/Glyma09g33350 identified via a soybean GWAS (Mao et al., 2017); and important regions for plant height, growth habit, pod shattering, and more. These candidate genes are valuable for further genetic discovery, and molecular markers could be developed to assist in the selection of target phenotypes for mung bean breeding.

DISCUSSION

Genome assembly at the chromosome scale is very important for crop genetics and breeding. In this study, we sequenced and assembled a high-quality mung bean genome (Vrad_JL7) through a combination of long-read and short-read sequencing technologies and annotated 40 125 protein-coding genes through the integration of *ab initio* prediction, RNA-seq, and homology data. We also assembled the first mung bean pan-genome from 217 accessions to understand the entire genome space and discovered genes related to important traits through SNP-based and PAV-based GWAS. The Vrad_JL7 assembly had improved completeness and continuity compared with those of the previously published VC1973A genome and the improved version (VC1973A_v2) (Supplemental Figure 9A; Table 1). It also displayed very good synteny with the genomes of adzuki bean and cowpea (Figure 1D). The annotated gene set was comprehensive, as indicated by the high protein BUSCO score (96.9%). Overall, among all of the available mung bean genome datasets, the Vrad_JL7 assembly and annotation had by far the highest quality.

The recent VC1973A genome (VC1973A_v2) assembled contigs/scaffolds under the guidance of a genetic map composed of 1100 molecular markers (Ha et al., 2021). Although the N50 of contigs and scaffolds seemed to be improved in the VC1973A_v2 assembly, the completeness was reduced by ∼5% (the BUSCO score decreased from 96.82% to 91.36%) (Table 1), and correctness was also compromised, as none of the VC1973A_v2 pseudochromosomes exhibited one-to-one collinearity with the adzuki bean and cowpea genomes (Supplemental Figure 9B). We noticed that some genomic regions of mung bean, cowpea, and adzuki bean seemed to be inverted in the comparison, which could be caused by misassembly in Vrad_JL7. For instance, the density distribution of genes and transposable elements suggested that some segments on chromosomes 1, 2, 3, 6, and 10 may need to be reversed in direction. These segments should be validated, and the assembly could be improved further by integrating genetic and/or optical physical maps. We also noticed that some gene models could be improved. Although 31 848 of 42 986 (74.09%) gene models had either RNA-seq or homology support, some were short (shorter than 200 nt), and others could be partial, fragmented, repetitive, or chimeric (Supplemental Data 1). More experimental evidence and manual curation are necessary to improve the genome annotation. Nonetheless, potential false-positive genes, such as

small open reading frames without homology evidence, did not have much influence on related conclusions in this study (functional enrichment analysis), as their functions could not be assigned.

Genetic variation was assessed in depth for the first time by resequencing 217 mung bean accessions mainly collected from China. Most Chinese accessions were clustered into two groups based on SNP and gene PAV data and were correlated with their geographic distribution (in terms of latitude). There were no clear genomic features distinguishing landraces from cultivars in the population structure analysis. This suggested that the improvement process in mung bean had been relatively slow, and few genomic regions were selected and fixed in modern breeding lines. Interestingly, we found that some landraces showed phenotypes similar to those of wild accessions, such as an indeterminate growth habit and pod shattering. These feralized landraces had similar levels of genetic diversity as other landraces. This phenomenon was probably caused by the de-domestication process, similar to that which occurred during the origin of weedy rice (Qiu et al., 2017). Because of the limited number of wild accessions collected in this study, we were not able to thoroughly analyze the genetics of mung bean domestication. As most of the 217 materials were from landraces and cultivars collected in China, which covered only a portion of global diversity, the assembled pan-genome could more appropriately be called the pan-genome of Chinese mung bean germplasm rather than the pan-genome of mung bean. Nevertheless, this study still provides new insights into the characteristics and genetic diversity of the mung bean genome.

Mung bean was proposed to have originated and been domesticated in India (Fuller, 2007; Kang et al., 2014), but there is no comprehensive population genetic evidence. For the first time, we compared the population diversity on a large scale by combining public GBS data from 533 accessions with resequencing data from 217 accessions. Although only a few thousand SNPs were common between the two datasets, nucleotide diversity ($\theta\pi$) (Figure 2E), LD decay (Supplemental Figure 11B), and other indicators supported the hypothesis that mung bean was introduced from south Asia to west Asia and then to east Asia, probably through the Silk Road. Although the diversity level of Southeast Asia accessions was also higher (3.18×10^{-4}) than that for east and west Asia, this was likely to be caused by collection of accessions from other regions by the Asian Vegetable R&D Center, located in Southeast Asia. After a long-term adaptation process, several distinct groups were formed among the mung bean population associated with their geographic regions, including two groups corresponding to south Asia and one group corresponding to east Asia. Accessions from west Asia and Southeast Asia were largely grouped together with accessions from south Asia. Within the east Asia group, the subgroup structure was very similar to the topology of trees constructed for the 217 accessions alone by the use of the large-scale SNP or PAV data. This demonstrated that the overall structure for all of the accessions based on the common SNP set was correct and that east Asia (mainly China) accessions had formed distinct population features during cultivation and improvement over the past 2000 years. It also strongly suggested that the improvement of mung bean could benefit greatly from the exchange of germplasm across different geographic regions.

Many genomic regions associated with agronomic traits were identified in this study, including a large region with important signals for BR on chromosome 5: 9 456 956–11 579 877 (~2.12 Mb) containing 156 genes. One gene, jg26964, was annotated as gibberellin-regulated protein 14, which corresponded to *Vradi05g03730* in the VC1973A genome and was located upstream of the BR marker W02a4 (Kang et al., 2014). Other previously studied candidate genes, such as genes encoding RD22 proteins and polygalacturonase-inhibiting proteins, were also included in this region (Liu and Fan, 2018; Kaewwongwal et al., 2020). Because neither VC1973A nor Vrad_JL7 shows resistance to bruchids, the two reference genomes may not contain functional resistance genes. We further used the PAV gene data to identify possible candidate genes that are likely to be missing in the reference genome. Of 217 accessions, 8 showed the bruchid resistance phenotype. Five pangenes (Pang34265, Pang44622, Pang57772, Pang58608, and Pang64254) were identified as possible candidate genes based on their frequencies and were also identified in our gene PAV-based GWAS. Among them, Pang58608 and Pang64254 were annotated as resistant specific protein-3, making them very good candidates for further verification. Taken together, the high-quality Vrad_JL7 genome, genetic variation map, and pan-genome constructed in this study lay a good foundation for gene discovery and breeding in mung bean.

METHODS

Sample collection and plant materials

In this study, 217 representative germplasms were selected from 589 accessions in the Chinese mung bean germplasm bank based on their phenotypes, the results of cluster analysis using simple sequence repeat markers, and their geographic distribution. The accessions were planted in SJZ (114.48E, 38.03N) and ZJK (114.88E, 40.82N), Hebei Province, China, for phenotypic observations. Detailed information for each accession is given in Supplemental Table 4. The seed coats of wild accessions were cut open with a knife on the opposite side of the hilum to promote germination. To reduce the environmental impact, we planted in SJZ in the spring and summer seasons of 2017 and 2018, respectively. However, ZJK was planted only in the spring season of the 2 years.

With respect to publicly available data, we downloaded three genotyping-by-sequencing datasets of the mung bean core germplasms from the NCBI BioProject database under accession numbers PRJNA645721, PRJNA609409, and PRJNA664607, corresponding to 296 (Sokolkova et al., 2020), 144 (Reddy et al., 2020), and 93 accessions (Wu et al., 2020b), respectively (Supplemental Data 4). In total, 750 mung bean accessions from 23 countries were used for the diversity analysis.

Genome sequencing and assembly

We selected the elite mung bean variety JL7 for whole-genome sequencing and assembly. Genomic DNA was extracted, and libraries were constructed and sequenced on the Illumina NovaSeq platform. The raw reads were preprocessed to remove adaptors and low-quality bases. The K-mer distribution was estimated using KMC (version 3.0) (Kokot et al., 2017) with the parameters "-k31 -t16 -m64 -ci1 -cs10000," and the genome size was estimated with GenomeScope (version 2.0) (Vurture et al., 2017).

For *de novo* assembly of the Vrad_JL7 genome, we used long-read sequencing based on the PacBio Sequel II platform. High-molecular-weight (HMW) DNA was used to construct a DNA library with a ~20-kb insert size, and the library was subsequently sequenced on the PacBio Sequel II sequencing platform at Novogene (Beijing, China). Flye (version

2.8.3-b1695) (Kolmogorov et al., 2019) was used for *de novo* assembly, and BWA-MEM (version 0.7.17) (Li and Durbin, 2009) was used to map Illumina PE reads to the assembled sequence. The genome was then polished using Pilon (version 1.24) (Walker et al., 2014). Hi-C sequencing library construction and sequencing (150-bp PE) were completed on the Illumina HiSeq platform. After removing low-quality raw reads, sequences were mapped to the assembled genome with Juicer (default parameters) (Durand et al., 2016a), and the 3D-DNA pipeline (https://github.com/aidenlab/3d-dna) was then used to obtain the chromosome-scale assembly. Finally, we used Juicebox Assembly Tools (version 1.11.9) (Durand et al., 2016b) to manually correct errors and visualize the results of the assembly.

For genome assessment, the Illumina PE reads were mapped to Vrad_JL7 using BWA-MEM, and the coverage of short reads on the genome was calculated using SAMtools (version 1.13) and Mosdepth (version 0.3.1) (Pedersen and Quinlan, 2018). Then, 2326 single-copy orthologs of dicot species (eudicots_odb10 database) were evaluated using BUSCO (version 3.1.0) (Simao et al., 2015). Finally, the LAI was calculated and evaluated with LTR_retriever (Ou et al., 2018b).

Genome annotation

We constructed a transposable element library by *ab initio* predictions of repeated sequences using RepeatModeler followed by RepeatMasker (version 4.0.9) (Tarailo-Graovac and Chen, 2009) to search for repeat sequence annotations. The BRAKER2 pipeline (Bruna et al., 2021), which integrates RNA-seq and protein homology-based methods, was executed to predict protein-coding genes. First, clean RNA-seq reads (~21.18 Gb from mung bean flower tissue) were mapped to the genome using HISAT2 (version 2.10.2) to obtain transcriptome mapping data. Second, all of the protein sequences from OrthoDB (version 10.0) (Kriventseva et al., 2019) were downloaded and mapped to the genome assembly with ProtHint (version 2.6.0). We then used GeneMark-EP+ (version 4.65) (Bruna et al., 2020) to integrate the two types of data. The final gene structure was predicted by Augustus (version 3.4.0), and the untranslated regions (UTRs) were predicted by GUSHR (version 1.0) (Stanke et al., 2008). BUSCO was used with eudicots_odb10 to evaluate the quality of the annotation. We also predicted rRNA using Barrnap (version 0.9) (https://github.com/tseemann/barrnap) and tRNA and other noncoding RNAs using Infernal (version 1.1.4) (Nawrocki and Eddy, 2013) by searching the Rfam (version 14.1) database (Kalvari et al., 2021). Functional annotations were assigned by homology searching against public databases, including the SwissProt, NR, and Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups databases, using DIAMOND ($E \leq 1e^{-6}$). The GO and KEGG annotations were assigned by transferring the annotation data from the Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups. Pfam and InterPro motif annotations were assigned using InterProScan (version 5.36).

Comparative genome analyses

Protein sequences from the following 12 eudicot genomes were downloaded: *Vigna angularis*, *Phaseolus vulgaris*, *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana* from Ensembl Plants (http://plants.ensembl.org/index.html); *Vigna unguiculata*, *Cajanus cajan*, *Cicer arietinum*, and *Arachis duranensis* from the NCBI Reference Sequence Database; *Pisum sativum* from https://urgi.versailles.inra.fr/download/pea/; *Phaseolus lunatus* from https://data.jgi.doe.gov/refine-download/phytozome?organism=Plunatus&expanded=563; and *Lotus japonicus* from https://drive.google.com/drive/folders/1yMR4fiRKt7fWZ6yTxlB7lsJoClQT9Q_0. Orthologous genes were identified using OrthoFinder (version 2.37) (Emms and Kelly, 2019). An ultrametric tree was constructed using r8s (version 1.81) (Sanderson, 2002), and TimeTree (http://www.timetree.org/) was used to calibrate divergence times. According to the evolutionary tree results with divergence times and gene family clustering by CAFE (version 4.2) (De Bie et al., 2006), the number of gene family members of each branch's ancestors was estimated by a birth mortality model to predict the contraction and expansion of gene families of species relative to their ancestors ($P < 0.05$). GO and KEGG enrichment analyses were performed using clusterProfiler (version 3.14.0) (Yu et al., 2012). Genome collinearity analysis was subsequently performed using MCScan (Python version, https://github.com/tanghaibao/jcvi/), and syntenic blocks were visualized using the dotplot script in the jcvi package. WGDdetector (version 1.00) (Yang et al., 2019) was used to identify whole-genome duplications (WGDs) with the synonymous mutation rate (Ks) method.

Pan-genome construction and analysis

The pan-genome was constructed following the method used for the tomato pan-genome (Gao et al., 2019). First, the genomes of 217 accessions were individually assembled using MaSuRCA (Zimin et al., 2013). Then, each assembled contig was aligned to Vrad_JL7 using QUAST (Gurevich et al., 2013), and contigs longer than 500 nt with an identity less than the threshold were extracted as non-reference sequences. We set the identity threshold to 90% after comparing the relationship between the length and identity of all contigs. Second, we aligned the non-reference sequences to the NT database and removed sequences ($E < 1e^{-5}$) that showed good homology to microorganisms, animals, and other non-Fabales plants. The clean non-reference sequences obtained were then clustered using CD-HIT (version 4.8.1) (Li and Godzik, 2006) to remove redundancy using an identity threshold of 90%. The non-redundant contigs were further aligned to the reference genome using BLAST to ensure that no single contig showed good identity to the reference. Finally, the non-redundant and non-reference contigs were merged with the reference mung bean genome as the mung bean pan-genome. The same annotation pipelines were used to annotate the gene structure and function of the pan-genome.

We used a map-to-pan strategy to identify core and variable genes. A total of 217 mung bean accessions were aligned to the reference genome using BWA-MEM and then Mosdepth (version 0.3.1) (Pedersen and Quinlan, 2018) to estimate CDS coverage. Genes with 2× coverage on at least 20% of the entire CDS region were considered present; otherwise, they were considered absent. By estimating the frequency of PAVs among all 217 accessions, we divided the genes into two categories: core (absence rate <0.05) and variable genes (absence rate ≥0.05). Four subcategories, hardcore, softcore, shell, and cloud, were defined as genes present in 100%, >99%, 1%–99%, and <1% of all accessions, respectively. To estimate the size of the pan-genome and core genome, samples were randomly picked (n = 1–217), and the process was iterated 100 times. Wilcoxon's test ($P < 0.05$) was used to denote the level of significance for differences in gene numbers in different subgroups.

Resequencing and variant calling

Genomic DNA was extracted from 217 mung bean accessions and sequenced on the Illumina NovaSeq platform. The clean reads were mapped to Vrad_JL7 using BWA-MEM with default parameters. Picard (version 2.18.17, http://broadinstitute.github.io/picard/) was used to remove PCR duplicates. Genetic variants, including SNPs and short indels (<15 bp), were detected using the Genome Analysis Toolkit (GATK) (version 3.8.1). SNPs were filtered with the following parameters: QD < 2.0, MQ < 40.0, FS > 60.0, SOR >3.0, MQRankSum < −12.5, and ReadPosRankSum < −8.0, and indels were removed with the parameters QD < 2.0, FS > 200.0, MQ < 40.0, SOR >10.0, and ReadPosRankSum < −20.0. Based on the filtered SNP set, we defined a core SNP/indel set by removing SNPs/indels with more than 2 alleles, a >10% missing rate, and a minor allele frequency of <5%. DELLY software (version 0.8.3) (Rausch et al., 2012) was used to identify SVs, including large indels (>15 bp), duplicate copy-number variations, inversions, and translocations. For public GBS data, the GATK best-practice pipeline described above and NGSEP (version 3.0.2) (Perea et al., 2016) with default parameters were used to call SNPs, and their results were consistent. A total of 5671 SNPs (40% missing rate filtered and imputed using Beagle [version 5.2]; Browning et al., 2018)

were ultimately merged from 750 accessions with GATK. According to the gene models of the Vrad_JL7 assembly, the genetic variants identified above were annotated using SnpEff (version 5.0e) (Cingolani et al., 2012), and the density along each chromosome was determined with VCFtools (version 0.1.17) using 500-kb sliding windows (Danecek et al., 2011).

Population genetic diversity and structure analysis

Based on the SNP data, population structure was determined using ADMIXTURE (version 1.3) (Alexander et al., 2009) with a block-relaxation algorithm. PCA was performed using the smartpca function in EIGENSOFT (version 6.1.4) (Price et al., 2006) with default parameters, and the first two eigenvectors were plotted. Neighbor-joining trees were constructed with the R package ape, and iTOL (https://itol.embl.de/) was used to visualize the trees. Population structure analysis of genic PAVs was performed using the hclust function in R. Weir and Cockerham's estimator of F_{ST} was used to measure genetic differentiation among multiple subpopulations. The values for F_{ST} and genome-wide diversity ($\theta\pi$) were calculated using VCFtools.

Phenotyping and GWAS

Thirty-three agronomic traits were observed for mung bean accessions planted in the field (2 seasons in SJZ and 1 season in ZJK for 2 consecutive years): 19 quantitative traits and 14 discrete traits. Fifteen traits (mostly quantitative traits) were recorded in all 6 fields, and another 15 traits were recorded only once. For traits collected more than three times, the average values were used in downstream analysis. Detailed information for each trait is given in Supplemental Table 8. GWASs based on SNP and gene PAV data were performed using a mixed linear model in the genome-wide efficient mixed-model association algorithm (GEMMA, version 0.98.1) (Zhou and Stephens, 2012). The first three principal components from PCA and a genomic relationship matrix were used to correct the population structure and random polygenic effect. Significant signals were identified using a $P < 0.05$ threshold and applying an adjusted Bonferroni test. STAs and GPTAs that passed the threshold were then evaluated for stability, consistency, and robustness based on standards used by Varshney et al. (2019a, 2019b). STAs/GPTAs with more than 15% phenotypic variation explained were considered robust; STAs/GPTAs identified for more than one location were considered stable; and STAs/GPTAs identified across >1 year/season were defined as consistent. The LD blocks surrounding GWAS signals were further evaluated using LDBlockShow (Dong et al., 2021).

ACCESSION NUMBERS

These sequence data have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA802060 and the National Genomics Data Center BioProject database (https://ngdc.cncb.ac.cn/bioproject/) under accession number PRJCA008996 (GSA: CRA006590). The genome assembly sequences, gene annotation, and supplemental data can be accessed from the FigShare database (https://doi.org/10.6084/m9.figshare.19583446).

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This research was supported by the National Key R&D Program of China (2019YFD1000700/2019YFD1000702), the China Agricultural Research System (CARS-08-G3), the Key Research and Development Program of Hebei (21326305D), the Hebei Agriculture Research System (HBCT2018070203), and the Hebei Talent Project.

AUTHOR CONTRIBUTIONS

J.T. and C.L. designed the experiments. C.L., Y.W., J.T., B.F., D.X., Z.C., Y.G., X.W., S.L., Q.S., Z.Z., S.W., Q.S., H.S., and Y.S. carried out the phenotyping. J.P., C.L., B.W., and X.W. carried out the sequencing and data analysis. J.T., C.L., J.P., B.W., J.W., and Y.W. wrote the manuscript. All of the authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jinfeng Chen at the Institute of Zoology, Chinese Academy of Sciences, and Dr. Jun Yang at the Shanghai Chenshan Botanical Garden for their useful comments on the data analysis and manuscript.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

Received: February 14, 2022
Revised: May 31, 2022
Accepted: June 22, 2022
Published: June 26, 2022

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**:1655–1664. https://doi.org/10.1101/gr.094052.109.

Breria, C.M., Hsieh, C.H., Yen, T.B., Yen, J.Y., Noble, T.J., and Schafleitner, R. (2020). A SNP-based genome-wide association study to mine genetic loci associated to salinity tolerance in mungbean (Vigna radiata L.). Genes **11**:759. https://doi.org/10.3390/genes11070759.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. **103**:338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform **2**:lqaa026. https://doi.org/10.1093/nargab/lqaa026.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform **3**:lqaa108. https://doi.org/10.1093/nargab/lqaa108.

Cao, D., Li, H., Yi, J., Zhang, J., Che, H., Cao, J., Yang, L., Zhu, C., and Jiang, W. (2011). Antioxidant properties of the mung bean flavonoids on alleviating heat stress. PLoS One **6**:e21071. https://doi.org/10.1371/journal.pone.0021071.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). 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 **6**:80–92. https://doi.org/10.4161/fly.19695.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics **27**:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics **22**:1269–1271. https://doi.org/10.1093/bioinformatics/btl097.

Dong, S.S., He, W.M., Ji, J.J., Zhang, C., Guo, Y., and Yang, T.L. (2021). LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief Bioinform. **22**:bbaa227. https://doi.org/10.1093/bib/bbaa227.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016a). Juicer provides a one-click System for analyzing loop-resolution Hi-C experiments. Cell Syst **3**:95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016b). Juicebox provides a visualization System for Hi-C contact maps with unlimited zoom. Cell Syst **3**:99–101. https://doi.org/10.1016/j.cels.2015.07.012.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. **20**:238. https://doi.org/10.1186/s13059-019-1832-y.

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. **18**:161. https://doi.org/10.1186/s13059-017-1289-9.

Fuller, D.Q. (2007). Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. Ann. Bot. **100**:903–924. https://doi.org/10.1093/aob/mcm048.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat. Genet. **51**:1044–1051. https://doi.org/10.1038/s41588-019-0410-2.

Gao, R., Han, T., Xun, H., Zeng, X., Li, P., Li, Y., Wang, Y., Shao, Y., Cheng, X., Feng, X., et al. (2021). MYB transcription factors GmMYBA2 and GmMYBR function in a feedback loop to control pigmentation of seed coat in soybean. J. Exp. Bot. **72**:4401–4418. https://doi.org/10.1093/jxb/erab152.

Garcia, T., Duitama, J., Zullo, S.S., Gil, J., Ariani, A., Dohle, S., Palkovic, A., Skeen, P., Bermudez-Santana, C.I., Debouck, D.G., et al. (2021). Comprehensive genomic resources related to domestication and crop improvement traits in Lima bean. Nat. Commun. **12**:702. https://doi.org/10.1038/s41467-021-20921-1.

Golicz, A.A., Batley, J., and Edwards, D. (2016). Towards plant pangenomics. Plant Biotechnol. J **14**:1099–1105. https://doi.org/10.1111/pbi.12499.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics **29**:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Ha, J., Satyawan, D., Jeong, H., Lee, E., Cho, K.H., Kim, M.Y., and Lee, S.H. (2021). A near-complete genome sequence of mungbean (Vigna radiata L.) provides key insights into the modern breeding program. Plant Genome **14**:e20121. https://doi.org/10.1002/tpg2.20121.

Hu, Z., Sun, C., Lu, K.C., Chu, X., Zhao, Y., Lu, J., Shi, J., and Wei, C. (2017). EUPAN enables pan-genome studies of a large number of eukaryotic genomes. Bioinformatics **33**:2408–2409. https://doi.org/10.1093/bioinformatics/btx170.

Hurgobin, B., Golicz, A.A., Bayer, P.E., Chan, C.K., Tirnaz, S., Dolatabadian, A., Schiessl, S.V., Samans, B., Montenegro, J.D., Parkin, I.A.P., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol. J **16**:1265–1274. https://doi.org/10.1111/pbi.12867.

Kaewwongwal, A., Liu, C., Somta, P., Chen, J., Tian, J., Yuan, X., and Chen, X. (2020). A second VrPGIP1 allele is associated with bruchid resistance (Callosobruchus spp.) in wild mungbean (Vigna radiata var. sublobata) accession ACC41. Mol. Genet. Genomics **295**:275–286. https://doi.org/10.1007/s00438-019-01619-y.

Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. **49**:D192–D200. https://doi.org/10.1093/nar/gkaa1047.

Kang, Y.J., Kim, S.K., Kim, M.Y., Lestari, P., Kim, K.H., Ha, B.K., Jun, T.H., Hwang, W.J., Lee, T., Lee, J., et al. (2014). Genome sequence of mungbean and insights into evolution within Vigna species. Nat. Commun. **5**:5443. https://doi.org/10.1038/ncomms6443.

Kokot, M., Długosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. Bioinformatics **33**:2759–2761. https://doi.org/10.1093/bioinformatics/btx304.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. **37**:540–546. https://doi.org/10.1038/s41587-019-0072-8.

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., and Zdobnov, E.M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. **47**:D807–D811. https://doi.org/10.1093/nar/gky1053.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**:1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

Liu, C.Y., Su, Q.Z., Fan, B.J., Cao, Z.M., Zhang, Z.X., Wu, J., Cheng, X.Z., and Tian, J. (2018). Genetic mapping of bruchid resistance gene in mungbean V1128. Acta Agron. Sin. **44**:1875. https://doi.org/10.3724/SP.J.1006.2018.01875.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al. (2020). Pan-genome of wild and cultivated soybeans. Cell **182**:162–176.e13. https://doi.org/10.1016/j.cell.2020.05.023.

Lonardi, S., Muñoz-Amatriaín, M., Liang, Q., Shu, S., Wanamaker, S.I., Lo, S., Tanskanen, J., Schulman, A.H., Zhu, T., Luo, M.C., et al. (2019). The genome of cowpea (Vigna unguiculata [L.] Walp.). Plant J. **98**:767–782. https://doi.org/10.1111/tpj.14349.

Mao, T., Li, J., Wen, Z., Wu, T., Wu, C., Sun, S., Jiang, B., Hou, W., Li, W., Song, Q., et al. (2017). Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. BMC Genomics **18**:415. https://doi.org/10.1186/s12864-017-3778-3.

Michael, T.P., and VanBuren, R. (2020). Building near-complete plant genomes. Curr. Opin. Plant Biol. **54**:26–33. https://doi.org/10.1016/j.pbi.2019.12.009.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29**:2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Noble, T.J., Tao, Y., Mace, E.S., Williams, B., Jordan, D.R., Douglas, C.A., and Mundree, S.G. (2017). Characterization of linkage disequilibrium and population structure in a mungbean diversity panel. Front. Plant Sci. **8**:2102. https://doi.org/10.3389/fpls.2017.02102.

Ou, L., Li, D., Lv, J., Chen, W., Zhang, Z., Li, X., Yang, B., Zhou, S., Yang, S., Li, W., et al. (2018a). Pan-genome of cultivated pepper (Capsicum) and its use in gene presence–absence variation analyses. New Phytol. **220**:360–363. https://doi.org/10.1111/nph.15413.

Ou, S., Chen, J., and Jiang, N. (2018b). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**:e126. https://doi.org/10.1093/nar/gky730.

Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics **34**:867–868. https://doi.org/10.1093/bioinformatics/btx699.

Perea, C., De La Hoz, J.F., Cruz, D.F., Lobaton, J.D., Izquierdo, P., Quintero, J.C., Raatz, B., and Duitama, J. (2016). Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. BMC Genomics **17** (*Suppl 5*):498. https://doi.org/10.1186/s12864-016-2827-7.

Plekhanova, E., Vishnyakova, M.A., Bulyntsev, S., Chang, P.L., Carrasquilla-Garcia, N., Negash, K., Wettberg, E.V., Noujdina, N., Cook, D.R., Samsonova, M.G., et al. (2017). Genomic and phenotypic analysis of Vavilov's historic landraces reveals the impact of environment and genomic islands of agronomic traits. Sci. Rep. **7**:4816. https://doi.org/10.1038/s41598-017-05087-5.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**:904–909. https://doi.org/10.1038/ng1847.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell **184**:3542–3558.e16. https://doi.org/10.1016/j.cell.2021.04.046.

Qiu, J., Zhou, Y., Mao, L., Ye, C., Wang, W., Zhang, J., Yu, Y., Fu, F., Wang, Y., Qian, F., et al. (2017). Genomic variation associated with local adaptation of weedy rice during de-domestication. Nat. Commun. **8**:15323. https://doi.org/10.1038/ncomms15323.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics **28**:i333–i339. https://doi.org/10.1093/bioinformatics/bts378.

Reddy, V.R.P., Das, S., Dikshit, H.K., Mishra, G.P., Aski, M., Meena, S.K., Singh, A., Pandey, R., Singh, M.P., Tripathi, K., et al. (2020). Genome-wide association analysis for phosphorus use efficiency traits in mungbean (Vigna radiata L. Wilczek) using genotyping by sequencing approach. Front. Plant Sci. **11**:537766. https://doi.org/10.3389/fpls.2020.537766.

Sanderson, M.J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**:301–302. https://doi.org/10.1093/bioinformatics/19.2.301.

Schafleitner, R., Nair, R.M., Rathore, A., Wang, Y.W., Lin, C.Y., Chu, S.H., Lin, P.Y., Chang, J.C., and Ebert, A.W. (2015). The AVRDC - the World Vegetable Center mungbean (Vigna radiata) core and mini core collections. BMC Genomics **16**:344. https://doi.org/10.1186/s12864-015-1556-7.

Shen, Y., Du, H., Liu, Y., Ni, L., Wang, Z., Liang, C., and Tian, Z. (2019). Update soybean Zhonghuang 13 genome to a golden reference. Sci. China Life Sci. **62**:1257–1260. https://doi.org/10.1007/s11427-019-9822-2.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Sokolkova, A., Burlyaeva, M., Valiannikova, T., Vishnyakova, M., Schafleitner, R., Lee, C.R., Ting, C.T., Nair, R.M., Nuzhdin, S., Samsonova, M., et al. (2020). Genome-wide association study in accessions of the mini-core collection of mungbean (Vigna radiata) from the World Vegetable Gene Bank (Taiwan). BMC Plant Biol. **20**:363. https://doi.org/10.1186/s12870-020-02579-x.

Sokolkova, A.B., Bulyntsev, S.V., Chang, P.L., Carrasquila-Garcia, N., Cook, D.R., von Wettberg, E., Vishnyakova, M.A., Nuzhdin, S.V., and Samsonova, M.G. (2021). The search for agroislands in the chickpea genome. Biophysics **66**:395–400. https://doi.org/10.1134/s0006350921030192.

Song, J.M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants **6**:34–45. https://doi.org/10.1038/s41477-019-0577-7.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**:637–644. https://doi.org/10.1093/bioinformatics/btn013.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics **Chapter 4**:Unit 4 10. https://doi.org/10.1002/0471250953.bi0410s25.

Tettelin, H., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae implications for the microbial pan-genome. Proc. Natl. Acad. Sci. USA **102**:0506758102.

Torkamaneh, D., Lemay, M.A., and Belzile, F. (2021). The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. Plant Biotechnol. J. **19**:1852–1862. https://doi.org/10.1111/pbi.13600.

Varshney, R.K., Saxena, R.K., Upadhyaya, H.D., Khan, A.W., Yu, Y., Kim, C., Rathore, A., Kim, D., Kim, J., An, S., et al. (2017). Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. Nat. Genet. **49**:1082–1088. https://doi.org/10.1038/ng.3872.

Varshney, R.K., Pandey, M.K., Bohra, A., Singh, V.K., Thudi, M., and Saxena, R.K. (2019a). Toward the sequence-based breeding in legumes in the post-genome sequencing era. Theor. Appl. Genet. **132**:797–816. https://doi.org/10.1007/s00122-018-3252-x.

Varshney, R.K., Thudi, M., Roorkiwal, M., He, W., Upadhyaya, H.D., Yang, W., Bajaj, P., Cubry, P., Rathore, A., Jian, J., et al. (2019b). Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. Nat. Genet. **51**:857–864. https://doi.org/10.1038/s41588-019-0401-3.

Varshney, R.K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M., Singh, N.P., Du, X., Upadhyaya, H.D., Khan, A.W., et al. (2021). A chickpea genetic variation map based on the sequencing of 3, 366 genomes. Nature **599**:622–627. https://doi.org/10.1038/s41586-021-04066-1.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., Schatz, M.C., and Berger, B. (2017). Genome Scope: fast reference-free genome profiling from short reads. Bioinformatics **33**:2202–2204. https://doi.org/10.1093/bioinformatics/btx153.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963. https://doi.org/10.1371/journal.pone.0112963.

Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y.C., Liu, Z., Frommer, W.B., Ma, J.F., Chen, L.Q., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. Natl. Sci. Rev. **7**:1776–1786. https://doi.org/10.1093/nsr/nwaa110.

Wu, J., Wang, L., Fu, J., Chen, J., Wei, S., Zhang, S., Zhang, J., Tang, Y., Chen, M., Zhu, J., et al. (2020a). Resequencing of 683 common bean genotypes identifies yield component trait associations across a north-south cline. Nat. Genet. **52**:118–125. https://doi.org/10.1038/s41588-019-0546-0.

Wu, X., Islam, A.S.M.F., Limpot, N., Mackasmiel, L., Mierzwa, J., Cortés, A.J., and Blair, M.W. (2020b). Genome-wide SNP identification and association mapping for seed mineral concentration in mung bean (Vigna radiata L.). Front. Genet. **11**:656. https://doi.org/10.3389/fgene.2020.00656.

Yang, Y., Li, Y., Chen, Q., Sun, Y., and Lu, Z. (2019). WGDdetector: a pipeline for detecting whole genome duplication events using the genome or transcriptome annotations. BMC Bioinf. **20**:75. https://doi.org/10.1186/s12859-019-2670-3.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS **16**:284–287. https://doi.org/10.1089/omi.2011.0118.

Zhao, J., Bayer, P.E., Ruperao, P., Saxena, R.K., Khan, A.W., Golicz, A.A., Nguyen, H.T., Batley, J., Edwards, D., and Varshney, R.K. (2020). Trait associations in the pangenome of pigeon pea (Cajanus cajan). Plant Biotechnol. J. **18**:1946–1954. https://doi.org/10.1111/pbi.13354.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. **44**:821–824. https://doi.org/10.1038/ng.2310.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. **33**:408–414. https://doi.org/10.1038/nbt.3096.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics **29**:2669–2677. https://doi.org/10.1093/bioinformatics/btt476.