The Plant Genome OPEN ACCESS

ORIGINAL RESEARCH

Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding

Philipp E. Bayer[1,*] | **Babu Valliyodan**[2,3,*] | **Haifei Hu**[1,*] | **Jacob I. Marsh**[1] |
Yuxuan Yuan[1,4] | **Tri D. Vuong**[3] | **Gunvant Patil**[3,5] | **Qijian Song**[6] |
Jacqueline Batley[1] | **Rajeev K. Varshney**[7,8] | **Hon-Ming Lam**[4] | **David Edwards**[1] |
Henry T. Nguyen[3]

[1] School of Biological Sciences and Inst. of Agriculture, The Univ. of Western Australia, Crawley, WA, Australia

[2] Dep. of Agriculture and Environmental Sciences, Lincoln Univ., Jefferson, City, MO 65101, USA

[3] Div. of Plant Sciences and National Ctr. for Soybean Biotechnology, Univ. of Missouri, Columbia, MO, USA

[4] Ctr. for Soybean Research of the State Key Lab. of Agrobiotechnology and School of Life Sciences, The Chinese Univ. of Hong Kong, Shatin, Hong Kong, China

[5] Dep. of Plant and Soil Science, Texas Tech Univ., Lubbock, TX, USA

[6] U.S. Dep. of Agriculture–Agricultural Research Service, Soybean Genomics and Improvement Lab., Beltsville, MD, USA

[7] Ctr. of Excellence in Genomics & Systems Biology, International Crops Research Inst. for the Semi-Arid Tropics (ICRISAT), Patancheru, India

[8] State Agricultural Biotechnology Ctr., Crop Research Innovation Ctr., Food Futures Inst., Murdoch Univ., Murdoch, WA, Australia

Correspondence
David Edwards, School of Biological Sciences and Inst. of Agriculture, The Univ. of Western Australia, Crawley, WA, Australia.
Henry T. Nguyen, Div. of Plant Sciences and National Ctr. for Soybean Biotechnology, Univ. of Missouri, Columbia, MO, USA.
Email: Dave.Edwards@uwa.edu.au, nguyen-henry@missouri.edu

*These authors contributed equally

Assigned to Associate Editor Agnieszka Golicz.

Abstract

The gene content of plants varies between individuals of the same species due to gene presence/absence variation, and selection can alter the frequency of specific genes in a population. Selection during domestication and breeding will modify the genomic landscape, though the nature of these modifications is only understood for specific genes or on a more general level (e.g., by a loss of genetic diversity). Here we have assembled and analyzed a soybean (*Glycine* spp.) pangenome representing more than 1,000 soybean accessions derived from the USDA Soybean Germplasm Collection, including both wild and cultivated lineages, to assess genomewide changes in gene and allele frequency during domestication and breeding. We identified 3,765 genes that are absent from the Lee reference genome assembly and assessed the presence/absence of all genes across this population. In addition to a loss of genetic diversity, we found a significant reduction in the average number of protein-coding genes per individual during domestication and subsequent breeding, though with some genes and allelic variants increasing in frequency associated with selection for agronomic traits. This analysis provides a genomic perspective of domestication and breeding in this important oilseed crop.

Abbreviations: BSR, brown stem rot; FST, fixation index; GO, gene ontology; LD, linkage disequilibrium; PAV, presence/absence variation; PCA, principal component analyses; QTL, quantitative trait loci; SNPs, single nucleotide polymorphisms; SVs, structural variants.

1 | INTRODUCTION

Cultivated soybean [*Glycine max* (L.) Merr.] is a staple crop that was domesticated 6,000–9,000 years ago in East Asia from wild soybean [*G. soja* (L.) Merr](Carter et al., 2004; Kim et al., 2010), a process that involved a 50% reduction in genetic diversity and the loss of 81% of rare alleles (Hyten et al., 2006; Z. Zhou et al., 2015). Since domestication, diverse cultivated lines have been produced, harboring improved agronomic traits; however, soybean yield is not increasing in pace with the growing demand for this crop (Ray et al., 2013). Soybean production needs to double by 2050 to keep track with a growing population, yet if current yield trends continue, soybean production will grow by only 55% by 2050 (Ray et al., 2013). At the same time, climate change is expected to reduce global soybean yields by 3.1% with each degree Celsius change (C. Zhao et al., 2017).

Intensive soybean breeding has been associated with further loss of diversity. Around 85% of genes present in North American lines may have been derived from only 19 landraces (Gizlice et al., 1996), and 79% of rare alleles present in diverse landraces have been lost during breeding (Hyten et al., 2006). Genomic analysis of these bottlenecks and the association of the lost diversity with agronomic traits can provide the foundation for increasing diversity in this crop and support the breeding of improved cultivars (Valliyodan et al., 2016).

The increasing availability of crop genome sequence data facilitates the study of genome composition changes during domestication and breeding. While many studies have examined the diversity of single nucleotide polymorphisms (SNPs) in populations, there has been increasing acknowledgment of the importance of gene presence/absence variation (PAV) in crop species (Alonge et al., 2020; Gao et al., 2019; Hurgobin & Edwards, 2017; Li et al., 2014; Liu et al., 2020; Lu et al., 2015), leading to the growth of pangenomics (Bayer et al., 2020; Danilevicz et al., 2020; Golicz et al., 2016; Golicz et al., 2020). Pangenomes have been constructed for several crop species, including maize (*Zea mays* L.; Hirsch et al., 2014), *Brassica oleracea* L. (Golicz et al., 2016), wheat (*Triticum aestivum* L.; Montenegro et al., 2017), canola (*Brassica napus* L.; Hurgobin et al., 2018), sesame (*Sesamum indicum* L.; Yu et al., 2019a), tomato (*Solanum lycopersicum* L.; Gao et al., 2019), rice (*Oryza sativa* L; Q. Zhao et al., 2018; Y. Zhou et al., 2020), and pigeon pea (*Cajanus cajan* L.; J. Zhao et al., 2020). These studies used whole-genome resequencing to assemble genomic regions not present in the reference genomes and to call gene PAV. They found extensive gene PAV ranging from 19% of genes being dispensable in *B. oleracea* (Golicz et al., 2016) to almost 40% of genes being dispensable in hexaploid bread wheat (*Triticum aestivum* L.; Montenegro et al., 2017), and in almost all of these studies, dispensable genes were enriched for biotic and abiotic stress-related annotations.

Core Ideas

- We assembled a soybean pangenome based on more than 1,000 lines from the USDA Soybean Germplasm Collection.
- We found 3,765 genes absent from the reference assembly.
- We found a reduction in the number of genes per individual during domestication and breeding.

Comparative genomics methods have been applied to soybean. A comparison of seven whole-genome assemblies of wild *G. soja* lines found loss-of-function frameshifts in domestication-related genes and PAV-regions reduced in frequency in *G. max* compared with *G. soja* (Li et al., 2014). A subsequent study examined 302 wild and cultivated soybean genomes to investigate the impact of domestication (Z. Zhou et al., 2015). This study identified 10 genomic regions under selection linked to nine domestication or breeding traits, mostly associated with oil content and fatty acid biosynthesis. A later study across 106 U.S. soybean lines identified 146 regions under selection (Valliyodan et al., 2016). They found that 43% of SNPs and 50% of PAV regions were not shared between wild *G. soja* and cultivated lines. Together these studies highlight the impact of domestication on the *Glycine* genome.

A recent soybean pangenome compares 26 de novo genome assemblies and data from an additional 2,872 wild, landrace, and cultivated lines (Liu et al., 2020). They identify 55,402 structural variants (SVs), with wild soybeans containing more SVs than landraces and cultivars. Genes were grouped into gene families, and only 35.88% of gene families were present in all lines. As with other pangenomes, dispensable genes were enriched with annotations for defense response, while core genes were associated with metabolic pathways. A genome-wide association study using these SVs as input identified a 10-kb deletion around a hydrophobic protein gene associated with seed luster, highlighting the importance of PAV in selection. This study also identified domestication-related SVs, including a 360-kb inversion on chromosome 7 that occurred approximately 4,700 years ago during soybean domestication.

Modern U.S. soybean breeding has led to a yield increase of 29 kg ha^{-1} yr^{-1} (Rincker et al., 2014), and understanding the genomic basis behind this improvement may provide indicators for further soybean improvement and adaptation. To investigate this, we assembled a pangenome and examined gene PAV as well as SNP diversity across 1,110 soybean lines (157 wild *G. soja*, 723 landraces, 228

cultivars, and two unclassified lines). These include 886 newly sequenced individuals, which represent the diversity present in the USDA Soybean Germplasm Collection. We demonstrate both a reduction in genetic diversity and a contraction in both gene number and estimated genome size during both during domestication and the subsequent breeding of modern U.S. soybean lines.

2 | MATERIALS AND METHODS

2.1 | Data availability

The sequence metadata for all 1,110 soybean accessions are summarized in Supplemental Table S2. Of these lines, 118 lines were previously published in PRJNA257011 (Fang et al., 2017; Z. Zhou et al., 2015) and 104 lines were previously published in PRJNA289660 (Valliyodan et al., 2016). All newly sequenced data are publicly available from the SRA project PRJNA639876. The assembled genomes and other data are available in Bayer et al. (Bayer et al., 2020). The constructed pangenome can be visualized using JBrowse (Buels et al., 2016) at http://appliedbioinformatics.com.au/soybean/. Pangenome annotation used available RNA-seq from NCBI (PRJNA238008,PRJNA246058 PRJNA246315,PRJNA246314,PRJNA246783,PRJNA197251, PRJNA254333,PRJNA280872,PRJNA149185,PRJNA350330, PRJNA304631,PRJNA389558,PRJNA182292,PRJNA197251). Protein sequences for *G. max, G. soja, Medicago truncatula* Gaertn.*, Vigna radiata* (L.) R. Wilczek var. *radiata*, and *Vigna angularis* (Willd.) Ohwi & H. Ohashi var. *angularis* were downloaded from Soybase (https://soybase.org/data/public/).

2.2 | Soybean germplasm, DNA isolation and sequencing

Diverse soybean germplasm were selected from the USDA Soybean Germplasm Collection (Song et al., 2015) and the seeds were germinated in the University of Missouri greenhouse for leaf sample collection and DNA extraction. Total DNA was extracted using the cetyltrimethylammonium bromidemethod (Murray & Thompson, 1980), and the sample heterogeneity was tested using Illumina Infinium SoySNP6K BeadChips BARCSoySNP6K beadchips containing SNPs that were selected from SoySNP50K (Song et al., 2013). All sequencing libraries were constructed using 5 µg of genomic DNA from each soybean germplasm following the Illumina sequencing protocols (Illumina Inc.). Paired-end sequencing libraries with an insert size of ~300 bp were sequenced on an Illumina HiSeq 2000 sequencer, at a minimum depth of 8.5× genome equivalent. Germplasm details, sequencing depth, and sequence identifiers are presented in Supplemental Table S1.

2.3 | Pangenome construction

The pangenome was assembled using a previously published pipeline (Golicz et al., 2016) using the chromosome-level Lee soybean assembly as the starting reference (Valliyodan et al., 2019). The pipeline consists of steps to assemble reads that do not align with the reference. The chloroplast and mitochondrial genomes (NC_020455.1; NC_007942.1) were first added to the reference. Adapters were removed from the sequence reads using Trimmomatic (Bolger et al., 2014) v0.36, and reads were aligned with the reference using Bowtie2 (Langmead & Salzberg, 2012) v2.3.3.1 (–end-to-end –sensitive -I 0 -X 1000). Unaligned reads were assembled using MaSuRCA (Zimin et al., 2013) v3.3.1, and contigs greater than 500 bp were retained.

2.4 | Annotation of the soybean pangenome

The pangenome was annotated using Augustus (Stanke et al., 2006) and SNAP (Korf, 2004). RNA-Seq data was trimmed to remove the low-quality sequences and adapters removed using Trimmomatic (Bolger, Lohse & Usadel, 2014) v0.36. The clean reads were mapped to the pangenome using Hisat2 (Kim et al., 2015) v2.1.0 and used to construct gene models using StringTie (Pertea et al., 2015) v2.0 and TransDecoder (Haas et al., 2013) v5.5.0. The Soybase protein sequences for *G. max, G. soja, M. truncatula, V. radiata* var. *radiata*, and *V. angularis* var. *angularis* were clustered and redundant sequences removed using CD-HIT (W. Li & Godzik, 2006) v4.6.8 with default settings. The de novo predicted and evidence models were used to annotate the pangenome using Maker 2 with the clustered Soybase proteins as external evidence (Holt & Yandell, 2011). Repeats were masked using RepeatMasker v4.0.4 (Smit & Hubley, 2008) using all repeats stored in Repbase 20150807 (Jurka et al., 2005). Predicted genes with protein length shorter than 33 amino acids were removed.

2.5 | PAV analysis

Genomic reads for each accession were aligned to the pangenome using Bowtie2 (Langmead & Salzberg, 2012) v2.3.3.1 (–end-to-end –sensitive -I 0 -X 1000). A gene is considered as missing when the horizontal coverage across exons is less than 5% and the vertical coverage less than two times as used in SGSGeneLoss (Golicz et al., 2016; Golicz et al., 2015) using Mosdepth v0.2.6 (Pedersen & Quinlan, 2018). A PAV matrix was generated showing the presence or absence of each gene for each accession. Statistical significance of gene frequency changes due to selection during domestication or breeding was calculated using Fisher's exact test. P-values

were adjusted for multiple comparisons using the Bonferroni method as implemented in p.adjust from R v3.5.0 (R Core Team, 2020). Genes with an adjusted p-value < .001 and difference frequency between groups ≥10% were identified.

Genome sizes were estimated using JELLYFISH v2.2.6 (settings: -h 1,000,000 for the upper limit of the histogram; Marcais & Kingsford, 2011) and GenomeScope (Vurture et al., 2017). Genome size estimates with model fits below 95% were removed, as were extreme outlier estimates (below 900 Mb, above 1,200 Mb).

2.6 | GO analysis

Functional annotation was performed using Blast2GO (Conesa et al., 2005) v2.5. Genes were aligned to the proteins in the Viridiplantae database using BLASTP (Camacho et al., 2009; E-values <1 × 10-5). Gene ontology (GO) analysis was conducted using topGO (Alexa & Rahnenführer, 2009) and Fisher's exact test with 'elim' used to correct for multiple comparisons.

2.7 | SNP discovery and population genetics analysis

Clean reads were mapped to the pangenome using BWA-MEM (H. Li, 2013) v0.7.17 with default settings and duplicates removed by Picard tools (http://broadinstitute.github.io/picard/). Reads were realigned by GATK (McKenna et al., 2010) v3.8-1-0 RealignerTargetCreator and IndelRealigner, followed by variant calling using GATK HaplotypeCaller. The resulting SNPs were filtered (QD < 2.0 || MQ < 40.0 || FS > 60.0 || QUAL < 60.0 || MQrankSum < −12.5 || ReadPosRankSum < −8.0) to remove low-quality SNPs.

High-confidence SNPs were identified by removing SNPs with minor allele frequency <0.05 and missing genotype rate <10% using VCFtools (Danecek et al., 2011). Neighbor-joining phylogenetic trees were constructed based on PAVs with 1,000 bootstraps using MEGA5 (Tamura et al., 2011). Principal component analysis was performed with the R package logisticPCA (Landgraf & Lee, 2015). Fixation index (FST) values and Tajima's D values were calculated using a 100-kb sliding window (with a 10-kb step for FST values calculation) using VCFftools (Danecek et al., 2011). Nucleotide diversity values (π) were calculated using pixy v1.0.4.beta1 using all invariant sites (Korunes & Samuk, 2021). Sliding windows with the top 1% of FST values were selected as significant windows and the overlapped significant windows were merged into the final nonredundant selective regions. The pairwise linkage disequilibrium (LD) between whole genomewide SNPs was calculated for each group based on allele frequency correlations (r^2) using PopLDdecay (Zhang

et al., 2019). Heterozygosity (F) was calculated using the –het option in vcftools v0.1.16 (Danecek et al., 2011).

3 | RESULTS AND DISCUSSION

We have assembled a soybean pangenome and examined the gene content for 1,110 public accessions (886 newly sequenced) from the USDA Soybean Germplasm Collection representing a wide distribution, from the soybean place of origin in East Asia to the current major soybean growing countries (Supplemental Table S1 and Supplemental Figure S1). Accessions were grouped into categories based on breeding history, including 157 wild soybean lineages (*G. soja*), 723 landraces, 228 cultivars, and two unclassified lines (Supplemental Table S1, Supplemental Figure S2 and S3). Cultivars were also split into two groups made up of 46 old cultivars, included cultivars developed during 1910s–1950s and 182 modern cultivars that were developed later. Modern cultivars show increased crop growth rate and produce enhanced yields and yield stability compared with old cultivars (Debruin & Pedersen, 2009). In addition to the Lee genome reference that was used as the basis for pangenome construction (Valliyodan et al., 2019), we assembled an additional 198.4 Mbp of sequence hosting 3,765 high confidence genes (Supplemental Table S2), to produce a pangenome of 1,213 Mbp and 51,414 predicted genes (Supplemental Table S3). Gene PAV was determined for all accessions, which revealed that 86.8% of genes are core (present in all accessions), and the remaining 13.2% are dispensable (absent in at least one accession). The percentage of dispensable genes is lower than previously observed in seven soybean species (~20% dispensable; Li et al., 2014), which is likely due to differences in gene comparison approaches, as the earlier publication used gene clustering approaches using OrthoMCL, while our study used more stringent read alignment methods. While the read alignment approach for calling PAVs using software such as SGSGeneLoss (Golicz et al., 2015) takes a strict approach in calling a gene as absent, this conservative approach avoids artificially inflating PAV numbers. The proportion of dispensable genes observed here is relatively low compared with some other crop studies that applied read mapping to call PAVs, such as bread wheat (36%; Montenegro et al., 2017), sesame (42%; Yu et al., 2019), or tomato (26%; Gao et al., 2019), although the proportion is similar to pigeon pea (13%; J. Zhao et al., 2020) and rice (11%; Schatz et al., 2014). A recent Chinese soybean pangenome reported 64% of gene families as being dispensable (Liu et al., 2020); however, they did not report the number of individual dispensable genes. The proportion of dispensable genes decreased slightly during domestication from 10.6% of genes in wild soybean to 9.8% in landraces (Supplemental Table S4); however, only
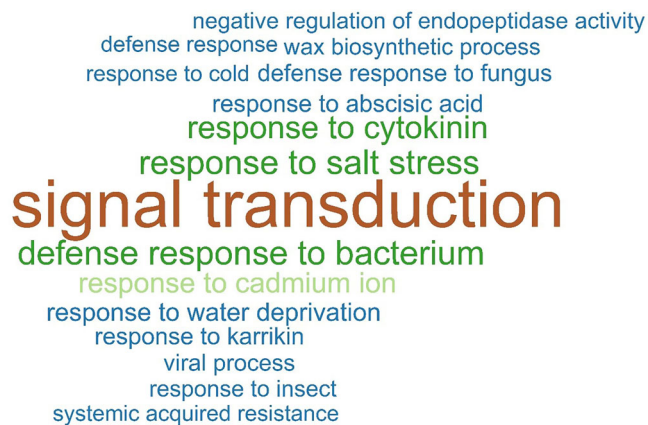
negative regulation of endopeptidase activity
defense response wax biosynthetic process
response to cold defense response to fungus
response to abscisic acid
response to cytokinin
response to salt stress
signal transduction
defense response to bacterium
response to cadmium ion
response to water deprivation
response to karrikin
viral process
response to insect
systemic acquired resistance

FIGURE 1 Significantly enriched gene ontology terms among dispensable genes. Font size and color scheme are proportional to –log (p)

5.9% of genes in modern cultivars are dispensable, reflecting the reduction in diversity during breeding and the fixation of genes.

Gene ontology analysis suggests that dispensable genes are enriched for terms associated with responses to biotic and abiotic stress, including defense response, response to abscisic acid, and response to salt stress (Figure 1 and Supplemental Table S5). These results are similar to findings in other crop pangenome studies. In soybean, Liu et al. (2020) found GO terms and Pfam domains associated with disease resistance and responses to biotic stimuli. Golicz et al. (2016), observed that *B. oleracea* dispensable genes are enriched for functions associated with disease resistance, response to salt stress, cold, and water deprivation, while (Montenegro et al. 2017) demonstrated that dispensable genes in wheat are enriched for functions associated with response to environmental stress and defense.

Phylogenetic and principal component analyses (PCA) based on gene PAV separated wild lineages and domesticated lines into major clusters, with only a few exceptions (Supplemental Figures S2a, S3). Interestingly, a PCA based on SNPs alone does not divide cultivated lines into subgroups, showing how gene PAV-based PCA can find patterns not contained in SNPs alone as observed in other plants (De Oliveira et al., 2020; Golicz et al., 2016; Mamidi et al., 2020; Tan et al., 2012; Supplemental Figure 2b). The gene–PAV distribution is similar to that observed by Han et al. (2016) using SNPs, with the domesticated lines forming two clusters, supporting the possibility of multiple domestication events. We find no geographic differences in these two clusters: Domesticated lines of both clusters were collected mostly in Korea (43% of individuals in Cluster 1 and 31% in Cluster 2), followed by Chinese individuals (22 and 29%), Japanese individuals (24 and 25%), and Russian individuals (11 and 14%).

This is contrary to the hypothesis (reviewed in Sedivy et al., 2017) based on domestication-specific alleles such as the pod shattering-resistant allele *SHAT1-5*, which appears in nearly all domesticated soybeans but not in wild soybeans (Dong et al., 2014), or a transposon insertion in a *FLOWERING LOCUS T (FT)* orthologue that appears only in domesticated lines (Wu et al., 2017). However, there is some evidence for multiple domestication events. For example, 302 chloroplast genomes revealed multiple maternal clades indicating that multiple maternal lines were selected in early soybean domestication stages (Fang et al., 2016). Similarly, resequencing of 302 soybean lines revealed a separate cluster of diversity unique to Japan and Korea indicating a separate domestication event (Z. Zhou et al., 2015), which is supported by domestication-associated SNPs detected only in Japanese lines (Jeong et al., 2019). The presence/absence of specific genes associated with the clusters presented here provide genic markers associated with this diversity (Supplemental Figure S4).

Domestication from wild soybean to cultivated soybean and subsequent selective breeding decreased nucleotide diversity, with the loss of the majority of the rare alleles and more than half of the genetic diversity (Hyten et al., 2006; Z. Zhou et al., 2015), and collectively only 17 landraces account for 86% of the North American genepool (Gizlice et al., 1993; Rincker et al., 2014). Soybean cultivars with a broad range of maturity and flowering time traits have been developed (Valliyodan et al., 2016; Z. Zhou et al., 2015), and an understanding of the genomic changes that occurred during domestication and breeding may assist in the identification of new alleles or genes to support future soybean breeding.

Analyzing gene content across this diverse population demonstrated a significant reduction in average gene number per individual following domestication and during subsequent breeding, similar to what was observed in a previous tomato pangenome study (Gao et al., 2019). Wild soybean contains the greatest average number of genes (48,785 $\pm$ 237), with a reduction in domesticated landraces (48,371 $\pm$ 139) and further declines in old cultivars (48,350 $\pm$ 232) and modern cultivars (48,165 $\pm$ 55) (Figure 2 and Supplemental Tables S6–S9). The loss of genes reflects in an overall reduction in genome size, with modern cultivars having an estimated average genome size of 877 Mbp compared with 898 Mbp for wild soybean (Supplemental Figure S5, Supplemental Table S10). On a country-by-country basis, the U.S. lines have a lower average gene number (48,286) than the other four major countries, for example, China (48,361), Korea (48,390), Japan (48,371), and Russia (48,344) (Supplemental Figure S6), mostly due to reduced average gene number in modern cultivars, suggesting that gene loss has accelerated in recent U.S. breeding programs. We also observed a greater average gene number in northern (48,332) compared with southern (48,204) adapted U.S. cultivars (Supplemental
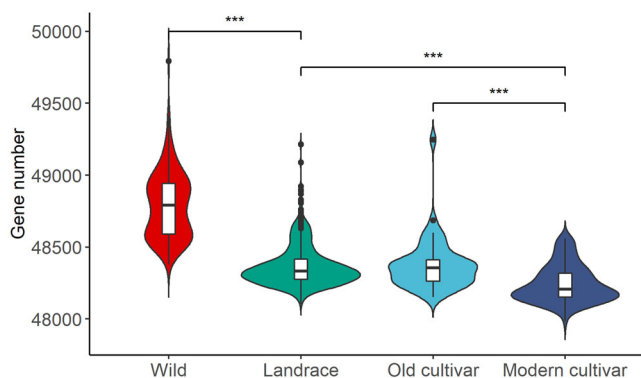
FIGURE 2 Violin plots showing gene abundance for the wild (*G. soja*), landraces, old and modern cultivars. Significance differences between groups is indicated (***$p < .005$)

FIGURE 3 Comparison of gene frequency during soybean (a) domestication, and (b) breeding. Colors indicate *p*-value, with purple ($p < $ 1e-20), blue ($p > $ 1e-10 - $\leq$ 1e-20), green ($p > $ .01 - $\leq$ 1e $-$10), and red ($p < = $ .01)

Figure S7). In heterozygous grapevine (*Vitis vinifera* ssp. *Sativa* L.), hemizygous genes are associated with PAVs accumulated during domestication and breeding (Y. Zhou et al., 2019). We expect hemizygosity to correlate with heterozygosity and therefore investigated patterns of heterozygosity across *G. soja*, landraces, as well as old and modern cultivars. There was a statistically significant difference in heterozygosity between *G. soja* and old cultivars, *G. soja* and modern cultivars, landraces and old cultivars, and landraces and modern cultivars ($p < .05$). However, the loss of genes following domestication does not mirror the decline of heterozygosity with no statistical difference in heterozygosity between old and modern cultivars (Supplemental Figure S8).

The reduction in average gene numbers hides a more complex pattern of increases and decreases in the frequency of specific genes across the population. To identify gene PAV changes during soybean domestication, we compared gene frequencies between wild soybean and landraces (Figure 3). A total of 1,478 genes decreased in frequency following domestication, while 261 genes increased in frequency (Figure 3a, Supplemental Table S6–S7). Among the annotated genes with decreased frequency, 98 were associated with defense response, 88 were associated with protein kinase activity, 44 with oxidation-reduction process, and 36 with response to salt process. Thirteen of the 98 defense response genes are colocated with disease resistance quantitative trait loci (QTL), including *Sclerotinia* resistance, Sclero 3-g31 and Sclero 3-g58 (Moellers et al., 2017), brown stem rot (BSR) resistance, BSR 1-g2 (Chang et al., 2016), and *Phytophthora* resistance, Phytoph 2-g1, Phytoph 2-g6 and Phytoph 2-g17 (Qin et al., 2017; Supplemental Table S11).

Genes associated with pubescence color were affected by domestication (Han et al., 2016), and the pubescence color gene GlymaLee.12G119700 shows a reduction in gene frequency from 79% in wild soybean to 38% in the landraces. Flowering time is also under strong selection during domesti-
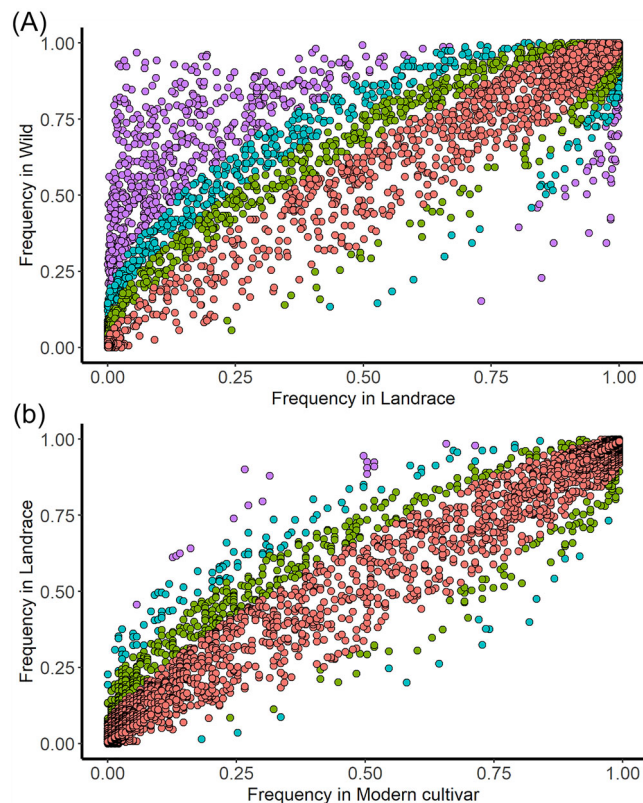
cation, breeding, and adaptation, and several flowering related genes, including FRIGIDA-like protein 4a, demonstrate a reduction in frequency during domestication (Supplemental Table S6). While fewer genes increase in frequency following domestication, they include 22 disease resistance genes and 10 salt stress tolerance genes suggesting selection for these traits (Supplemental Table S7).

Early breeding efforts developed cultivars suitable for North American production systems, and as soybean production increased, the breeding programs focused on yield improvement and disease resistance traits. Breeding for yield over the last 60 yr has had no major influence on seed protein composition, possibly because of limited genetic diversity among the parental lines (Mahmoud et al., 2006). The average number of genes per individual declined during breeding (Figure 2), and we observed a decrease in frequency for 483 genes, while 100 genes increased in frequency during the transition from landrace to modern cultivar (Figure 3b). Among the genes that reduce in frequency, 49 were associated with defense response, 36 with signal transduction, 15 with oxidation-reduction process, and 7 with response to auxin stimulus (Supplemental Table S8). Genes that reduce in frequency during breeding are associated with QTL for

plant architecture and seed composition traits, including three genes under the Shoot Fe 1-g20, one under the Leaf carotenoid content QTL 1-g13.4, and another 5 associated with seed composition and yield QTL (Supplemental Table S12). Genes that increase in frequency during breeding are mainly associated with flowering time, seed composition, and stress tolerance traits (both disease resistance and abiotic stress), though genes encoding several auxin responsive proteins that share maturity and seed composition functions also increased frequency during breeding (Supplemental Table S9).

Comparing cultivars from the five most represented countries (Russia, China, Japan, United States, Korea) identified 16 genes that increased in frequency and 64 genes that decreased in frequency in U.S. cultivars compared with each of the four other countries (Supplemental Table S13–S14). Several of the genes that reduce in frequency encode disease resistance genes, including UWASoyPan03234, a Leaf Rust 10 disease-resistance locus receptor-like gene; UWASoyPan03449, a TMV resistance protein N isoform X3; and UWASoyPan05034, a disease resistance RML1A-like gene. More detailed comparison of northern and southern U.S. lines identified 27 genes that have a lower frequency and 8 genes that have a higher frequency in southern cultivars (Supplemental Table S15). Of the 27 genes, 3 show similarity to transcription factors, while 5 show similarity to disease resistance genes. The eight genes that increased in frequency in southern adapted cultivars include *ZPR1*, which encodes a clock-associated zinc finger protein required for circadian-regulated gene expression in plants (Kiełbowicz-Matuk et al., 2017; J. Li et al., 2013), and so may play a role in adaptation. While we have sequenced a large and diverse collection of germplasm, we only have a limited insight into local diversity and selection, and the sequencing of additional lines may reveal a more complete picture of genome variation due to local soybean breeding efforts.

Studies have shown that domestication from wild soybean to landraces resulted in a reduction in genetic diversity and the loss of more than 81% of rare alleles (Hyten et al., 2006; Z. Zhou et al., 2015). Early North American landraces have low genetic diversity compared to Chinese lines (Y. Li et al., 2008), and southern elite cultivars are less diverse compared to the ancestral U.S. cultivar pool (Kisha et al., 1998; Thompson & Nelson, 1998). Here, we annotated 13,039,091 high-quality SNP loci across the 1,110 individuals and called 14,285,049,178 genotypes. The nucleotide diversity (π) of wild soybeans (3.75×10^{-3}) was higher than landraces (2.12×10^{-3}), old cultivars (2.11×10^{-3}), and modern cultivars (1.48×10^{-3}), reflecting the loss of diversity during domestication and breeding. These values are similar to previous observations in U.S. (Hyten et al., 2006; Valliyodan et al., 2016) and Chinese soybean lines (Z. Zhou et al., 2015), suggesting that U.S.
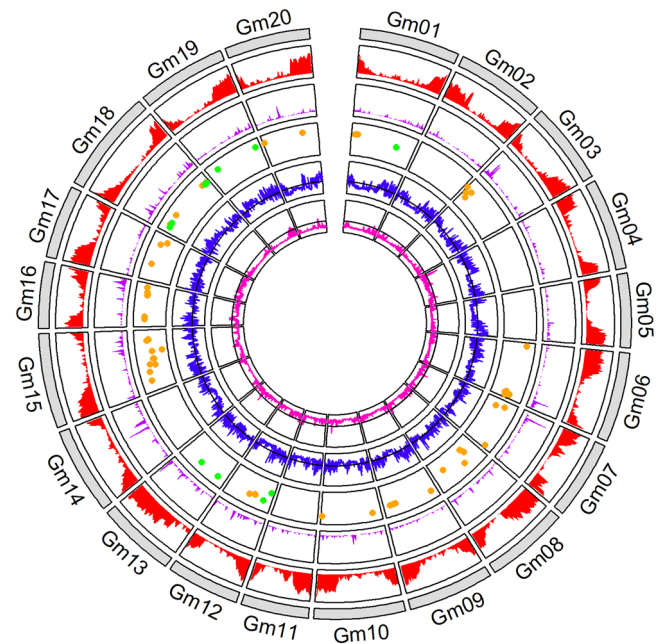

FIGURE 4 Circos plot showing genetic diversity and signals of selection between landraces and modern cultivars. From outer-most track to innermost track: (a) gene density; (b) dispensable gene density; (c) genes increased and decreased between landraces and modern cultivars (false discovery rate (FDR)-adjusted $p < 0.05$, orange: genes increased in frequency in modern cultivars, green: genes decreased in frequency in modern cultivars) (y-position assigned to avoid overlapping points); (d) Tajima's D between landraces and modern cultivars (black line: D = 0); and (e) Fixation Index (FST) between landraces and modern cultivars (black line: fixation index [FST] = 0.15)

and Chinese lines show similar nucleotide diversity and similar loss of diversity during domestication and subsequent breeding.

The average distance over which LD decays to half of its maximum value was substantially shorter in wild soybean than landraces and old and modern cultivars, which shows similar trend to previous studies (Hyten et al., 2006; Valliyodan et al., 2016; Z. Zhou et al., 2015; Supplemental Figure S9 and Supplemental Table S16). We searched for selective sweeps during domestication and breeding and identified 110 genomic regions with signatures of domestication-selective sweeps harboring 1,266 protein-coding genes. We also identified 86 genomic regions with signatures of breeding-selective sweeps harboring 1,434 protein-coding genes (Supplemental Table S17–S18, Figure 4). Among the genes located within the domestication-selective sweeps, 51 genes are dispensable, with a probable receptor-like protein kinase GlymaLee.05G082900 and a L-10 interacting MYB domain-containing protein GlymaLee.05G083000 showing a significant decrease in frequency during domestication. In total, 55 genes are dispensable among the genes located within selective sweep regions during

breeding. Four of these dispensable genes—including omega-6 fatty acid desaturase GlymaLee.15G174100, a phosphate transporter GlymaLee.15G174200, a nonannotated gene GlymaLee.17G150300, and a GEM-like protein GlymaLee.20G073500—significantly increased in frequency during breeding, and only one gene GlymaLee.19G173000, encoding a transmembrane protein, significantly decreased in frequency.

Domestication selection sweeps on chromosome Gm20 (6.9–12 Mb), overlapping with the seed protein QTL were previously detected in other reported domestication-related QTL regions (Grant et al., 2010; Nichols et al., 2006). We also found a breeding-related selective sweep region on Gm20 (36.1–37.2 Mb), which overlapped with the reported Seed yield 31–38, seed oil and seed protein QTL region (Grant et al., 2010; Hacisalihoglu et al., 2018). These results show that selective sweeps acted on different QTL regions during domestication and breeding.

Calculation of the divergence index value (FST) between different groups identified genomic regions associated with domestication and subsequent breeding. The largest differences of FST were observed during domestication, with a mean weighted value of 0.215 (Supplemental Table S16), compared with 0.06 between landraces and modern cultivars. These results are consistent with previous studies showing that wild soybean contains the most diverse gene pools and that selective sweeps are stronger during domestication than during breeding (Hyten et al., 2006; Song et al., 2020; Valliyodan et al., 2016 ; Z. Zhou et al., 2015).

In this study, we have examined changes in the frequency of dispensable genes during domestication and breeding, providing information that will assist the production of improved cultivars. The reduction in average gene number and genome size during breeding was unexpected and raises several questions. If breeders are selecting for gene absence, then selection can only occur for the relatively small proportion of genes that show PAV. Further analysis may identify candidate core genes that, if deleted using tools such as genome editing, could further improve this important crop. Moreover, this pangenome, along with the publicly available USDA Soybean Germplasm Collection, provides a valuable resource to design more efficient and targeted molecular breeding strategies for soybean improvement.

AUTHOR CONTRIBUTIONS
Philipp E. Bayer, Babu Valliyodan, and Haifei Hu contributed equally.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ORCID
Philipp E. Bayer https://orcid.org/0000-0001-8530-3067
Babu Valliyodan https://orcid.org/0000-0001-9457-9508
Haifei Hu https://orcid.org/0000-0003-1070-213X
Yuxuan Yuan https://orcid.org/0000-0003-0741-4196
Tri D. Vuong https://orcid.org/0000-0002-4782-4061
Rajeev K. Varshney https://orcid.org/0000-0002-4562-9131
Hon-Ming Lam https://orcid.org/0000-0002-6673-8740
David Edwards https://orcid.org/0000-0001-7599-6760
Henry T. Nguyen https://orcid.org/0000-0002-7597-1800

REFERENCES
Alexa, A., & Rahnenführer, J. (2009). Gene set enrichment analysis with topGO. *Bioconductor Improv*, 27, 1–26.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., & Ciren, D. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, *182*, 145–161.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K. W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., MacLachlan, R. P., Sharpe, A. G., Fritz, A., ... Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, *357*(6346), 93–97. https://doi.org/10.1126/science.aan0032/J

Bayer, P., Hu, R., Valliyodan, B., Marsh, J., Yuan, A., Vuong, T., Patil, G., Song, Q., Batley, J., Varshney, R., Lam, H. M., Edwards, D., &

Nguyen, H. (2020). Soybean pangenome SNPs, assembly, annotation, PAV *[Data set]*. The University of Western Australia. https://doi.org/10.26182/5f34ac3377313

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., & Stein, L. (2016). JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biology*, *17*, 66.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Carter, T., Hymowitz, T., & Nelson, R. (2004). Biogeography, local adaptation, Vavilov, and genetic diversity in soybean. In D. Werner (Ed.), *Biological resources and migration* (pp. 47—59). Springer.

Chang, H. X., Lipka, A. E., Domier, L. L., & Hartman, G. L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology*, *106*(10), 1139–1151. https://doi.org/10.1094/phyto-01-16-0042-fi

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

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., McVean, G., & Durbin, R., 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danilevicz, M. F., Fernandez, C. G. T., Marsh, J. I., Bayer, P. E., & Edwards, D. (2020). Plant pangenomics: approaches, applications and advancements. *Current Opinion in Plant Biology*, *54*, 18–25.

De Oliveira, R., Rimbert, H., Balfourier, F., Kitt, J., Dynomant, E., Vrána, J., Doležel, J., Cattonaro, F., Paux, E., & Choulet, F. (2020). Structural variations affecting genes and transposable elements of chromosome 3B in wheats. *Frontiers in Genetics*, *11*, 891. https://doi.org/10.3389/fgene.2020.00891

Debruin, J., & Pedersen, P. (2009). Growth, Yield, and Yield Component Changes among Old and New Soybean Cultivars. *Agronomy Journal*, *101*. https://doi.org/10.2134/agronj2008.0187

Dong, Y., Yang, X., Liu, J., Wang, B. - H., Liu, B.- L., & Wang, Y.-Z. (2014). Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. *Nature Communications*, *5*(1), 1–11. https://doi.org/10.1038/ncomms4352

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., & Zhang, M. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology*, *18*, 1–14.

Fang, C., Ma, Y., Yuan, L., Wang, Z., Yang, R., Zhou, Z., Liu, T., & Tian, Z. (2016). Chloroplast DNA underwent independent selection from nuclear genes during soybean domestication and improvement. *Journal of Genetics and Genomics*, *43*(4), 217. https://doi.org/10.1016/j.jgg.2016.01.005

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.,

Thannhauser, T. W., Foolad, M. R., Diez, M. J., Blanca, J., Canizares, J., Xu, Y., Knaap, E., Huang, S., Klee, H. J., ... Fei, Z. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics*, *51*(6), 1044–1051. https://doi.org/10.1038/41588-019-0410-2

Gizlice, Z., Carter, T. E., Jr., & Burton, J. W. (1993). Genetic Diversity in North American Soybean: I. Multivariate analysis of founding stock and relation to coefficient of parentage. *Crop Science*, *33*(3), 614–620. https://doi.org/10.2135/cropsci1993.0011183X003300030038x

Gizlice, Z., Carter, T. E., Jr., & Burton, J. W. (1996). Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. *Crop Science*, *36*, 753–765. https://doi.org/10.2135/cropsci1996.0011183X003600030038x

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*, 13390. https://doi.org/10.1038/ncomms13390

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., & Edwards, D. (2020). Pangenomics comes of age: From bacteria to plant and animal applications. *Trends in Genetics*, *36*, 132–145.

Golicz, A. A., Martinez, P. A., Zander, M., Patel, D. A., Van De Wouw, A. P., Visendi, P., Fitzgerald, T. L., Edwards, D., & Batley, J. (2015). Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Functional & Integrative Genomics*, *15*(2), 189–196.

Grant, D., Nelson, R. T., Cannon, S. B., & Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research*, *38*(Database issue), D843–D846. https://doi.org/10.1093/nar/gkp798

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hacisalihoglu, G., Burton, A. L., Gustin, J. L., Eker, S., Asikli, S., Heybet, E. H., Ozturk, L., Cakmak, I., Yazici, A., Burkey, K. O., Orf, J., & Settles, A. M. (2018). Quantitative trait loci associated with soybean seed weight and composition under different phosphorus levels. *Journal of Integrative Plant Biology*, *60*(3), 232–241. https://doi.org/10.1111/jipb.12612

Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D. A., Yang, Z., Zhao, L., Zhou, G., Wang, Z., Huang, L., Zhang, Z., Qiu, L., Zheng, H., & Li, W. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytologist*, *209*(2), 871–884. https://doi.org/10.1111/nph.13626

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., Leon, N., Kaeppler, S. M., & Barry, K. (2014). Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell*, *26*(1), 121–135. https://doi.org/10.1105/tpc.113.119982

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491. https://doi.org/10.1186/1471-2105-12-491

Hurgobin, B., & Edwards, D. (2017). SNP discovery using a pangenome: has the single reference approach become obsolete? *Biology*, *6*, 21.

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., & Parkin, I. A. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid*Brassica napus*. *Plant Biotechnology Journal*, *16*, 1265–1274.

Hyten, D. L., Song, Q., Zhu, Y., Choi, I. Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., & Cregan, P. B. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences*, *103*(45), 16666–16671. https://doi.org/10.1073/pnas.0604379103

Jeong, S.-C., Moon, J.-K., Park, S.-K., Kim, M.-S., Lee, K., Lee, S. R., Jeong, N., Choi, M. S., Kim, N., & Kang, S.-T. (2019). Genetic diversity patterns and domestication origin of soybean. *Theoretical and Applied Genetics*, *132*(4), 1179–1193. https://doi.org/10.1007/s00122-018-3271-7

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1–4), 462–467. https://doi.org/10.1159/000084979

Kiełbowicz-Matuk, A., Czarnecka, J., Banachowicz, E., Rey, P., & Rorat, T. (2017). *Solanum tuberosum ZPR1* encodes a light-regulated nuclear DNA-binding protein adjusting the circadian expression of *StBBX24* to light cycle. *Plant, Cell and Environment*, *40*(3), 424–440. https://doi.org/10.1111/pce.12875

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, M.Y., Lee, S., Van, K., Kim, T.-H., Jeong, S.-C., Choi, I.-Y., Kim, D.-S., Lee, Y.-S., Park, D., & Ma, J. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proceedings of the National Academy of Sciences*, *107*, 22032–22037.

Kisha, T., Diers, B. W., Hoyt, J., & Sneller, C. (1998). Genetic diversity among soybean plant introductions and North American germplasm. *Crop Science*, *38*(6), 1669–1680. https://doi.org/10.2135/cropsci1998.0011183X003800060042x

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Korunes, K. L., & Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, *21*(4), 1359–1368. https://doi.org/10.1111/1755-0998.13326

Landgraf, A. J., & Lee, Y. (2015). *Dimensionality reduction for binary data through the projection of natural parameters*. arXiv. https://arxiv.org.pdf/1510.06112.pdf

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. https://arxiv.org/pdf/1303.3997

Li, J., Sima, W., Ouyang, B., Luo, Z., Yang, C., Ye, Z., & Li, H. (2013). Identification and expression pattern of a *ZPR1* gene in wild tomato (*Solanum pennellii*). *Plant molecular biology reporter*, *31*(2), 409–417. https://doi.org/10.1007/s11105-012-0509-4

Li, W., & Godzik, A. (2006). CD-HIT: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinfor-

matics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., Zhang, S. S., Zuo, Q., Shi, X. H., Li, Y. F., Zhang, W. K., Hu, Y., Kong, G., Hong, H. L., Tan, B., Song, J., … Qiu, L. J. (2014). *De novo* assembly of soybean wild relatives for pangenome analysis of diversity and agronomic traits. *Nature Biotechnology*, *32*, 1045–1052.

Li, Y., Guan, R., Liu, Z., Ma, Y., Wang, L., Li, L., Lin, F., Luan, W., Chen, P., Yan, Z., Guan, Y., Zhu, L., Ning, X., Smulders, M. J. M., Li, W., Piao, R., Cui, Y., Yu, Z., Guan, M., …, & Qiu, L. (2008). Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. *Theoretical and applied genetics*, *117*(6), 857–871. https://doi.org/10.1007/s00122-008-0825-0

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., & Tian, Z. (2020). Pan-genome of wild and cultivated soybeans. *Cell*, *182*(1), 162–176.e113. https://doi.org/10.1016/j.cell.2020.05.023

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., Li, Y., Li, Y., Semagn, K., Zhang, X., Hernandez, A. G., Mikel, M. A., Soifer, I., Barad, O., & Buckler, E. S. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, *6*, 6914.

Mahmoud, A. A., Natarajan, S. S., Bennett, J. O., Mawhinney, T. P., Wiebold, W. J., & Krishnan, H. B. (2006). Effect of six decades of selective breeding on soybean protein composition and quality: A biochemical and molecular analysis. *Journal of Agricultural and Food Chemistry*, *54*(11), 3916–3922. https://doi.org/10.1021/jf060391m

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., Sreedasyam, A., Shu, S., Lovell, J. T., Feldman, M., Wu, J., Yu, Y., Chen, C., Johnson, J., Sakakibara, H., Kiba, T., Sakurai, T., Tavares, R., Nusinow, D. A., … Feldman, M. (2020). A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. *Nature Biotechnology*, *38*(10), 1203–1210. https://doi.org/10.1038/s41587-020-0681-2

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Moellers, T. C., Singh, A., Zhang, J., Brungardt, J., Kabbage, M., Mueller, D. S., Grau, C. R., Ranjan, A., Smith, D. L., Chowda-Reddy, R. V., & Singh, A. K. (2017). Main and epistatic loci studies in soybean for *Sclerotinia sclerotiorum* resistance reveal multiple modes of resistance in multi-environments. *Scientific Reports*, *7*(1), 3554. https://doi.org/10.1038/s41598-017-03695-9

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Batley, J. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*(5), 1007–1013. https://doi.org/10.1111/tpj.13515

Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, *8*, 4321–4325.

Nichols, D. M., Glover, K. D., Carlson, S. R., Specht, J. E., & Diers, B. W. (2006). Fine mapping of a seed protein QTL on soybean linkage

group I and its correlated effects on agronomic traits. *Crop Science*, *46*(2), 834–839. https://doi.org/10.2135/cropsci2005.05-0168

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Qin, J., Song, Q., Shi, A., Li, S., Zhang, M., & Zhang, B. (2017). Genome-wide association mapping of resistance to Phytophthora sojae in a soybean [*Glycine max* (L.) Merr.] germplasm panel from maturity groups IV and V. *PLOS ONE*, *12*(9), e0184613. https://doi.org/10.1371/journal.pone.0184613

Ray, D. K., Mueller, N. D., West, P. C., & Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLOS ONE*, *8*(6), e66428. https://doi.org/10.1371/journal.pone.0066428

Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S. R., Casteel, S., Conley, S., Chen, P., Davis, V., Fox, C., Graef, G., Godsey, C., Holshouser, D., Jiang, G., Kantartzi, S. K., Kenworthy, W., Lee, C., Mian, R., … Diers, B. (2014). Genetic improvement of U.S. soybean in Maturity Groups II, III, and IV. *Crop Science*, *54*(4), 1419–1432. https://doi.org/10.2135/cropsci2013.10.0665

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Schatz, M. C., Maron, L. G., Stein, J. C., Hernandez Wences, A., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., Wright, M. H., Chia, J., Ware, D., McCouch, S. R., & McCombie, W. R. (2014). Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biology*, *15*(11), 506. https://doi.org/10.1186/preaccept-2784872521277375

Sedivy, E. J., Wu, F., & Hanzawa, Y. (2017). Soybean domestication: The origin, genetic architecture and molecular bases. *New Phytologist*, *214*(2), 539–553. https://doi.org/10.1111/nph.14418

Smit, A. F., & Hubley, R. (2008). RepeatModeler Open-1.0. http://www.repeatmasker.org

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W., Cheng, Y., Zhang, Y., Liu, K., Yang, Q., Chen, L., & Zhou, R. (2020). Eight high-quality genomes reveal pangenome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants*, *6*(1), 34–45.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLOS ONE*, *8*, e54985.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2015). Fingerprinting soybean germplasm and its utility in genomic research. *G3 (Bethesda)*, *5*, 1999–2006.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435–W439. https://doi.org/10.1093/nar/gkl200

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*(10), 2731–2739. https://doi.org/10.1093/molbev/msr121

Tan, S., Zhong, Y., Hou, H., Yang, S., & Tian, D. (2012). Variation of presence/absence genes among Arabidopsis populations. *BMC Evolutionary Biology*, *12*(1), 86. https://doi.org/10.1186/1471-2148-12-86

Thompson, J. A., & Nelson, R. L. (1998). Utilization of diverse germplasm for soybean yield improvement. *Crop Science*, *38*(5), 1362–1368. https://doi.org/10.2135/cropsci1998.0011183X003800050035x

Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S., Brown, A. V., Ren, L., Jenkins, J., Chung, C. Y. L., Chan, T., Daum, C. G., Plott, C., Hastie, A., Baruch, K., Barry, K. W., Huang, W., Patil, G., Varshney, R. K., Hu, H., Batley, J., … Nguyen, H. T. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. *Plant Journal*, *100*(5), 1066–1082. https://doi.org/10.1111/tpj.14500

Valliyodan, B., Qiu, D., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., Song, L., Vuong, T. D., Musket, T. A., Xu, D., Shannon, J. G., Shifeng, C., Liu, X., & Song, L. (2016). Landscape of genomic diversity and trait discovery in soybean. *Scientific Reports*, *6*, 23598. https://doi.org/10.1038/srep23598

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wu, F., Sedivy, E. J., Price, W. B., Haider, W., & Hanzawa, Y. (2017). Evolutionary trajectories of duplicated FT homologues and their roles in soybean domestication. *The Plant Journal*, *90*(5), 941–953. https://doi.org/10.1111/tpj.13521

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., Wang, L., You, J., Fan, D., Edwards, D., & Zhang, X. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnology Journal*, *17*(5), 881–892. https://doi.org/10.1111/pbi.13022

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., & Yang, T. L. (2019). PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, *35*(10), 1786–1788. https://doi.org/10.1093/bioinformatics/bty875

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Yao, Y., Bassu, S., Ciais, P., Durand, J., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., … Ciais, P. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences*, *114*(35), 9326–9331. https://doi.org/10.1073/pnas.1701762114

Zhao, J., Bayer, P., Ruperao, P., Saxena, R., Khan, A., Golicz, A., Nguyen, H. T., Batley, J., Edwards, D., & Varshney, R. (2020). Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant biotechnology journal*, *18*(9), 1946–1954. https://doi.org/10.1111/pbi.13354

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., Xu, Q., … Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*, 278–284.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., & Parakkal, P. (2020) A platinum standard pan-genome resource that represents the population structure of Asian rice. *Scientific Data 7*, 1–11.

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., & Gaut, B. S. (2019). The population genetics of structural

variants in grapevine domestication. *Nature Plants*, 5(9), 965–979. https://doi.org/10.1038/s41477-019-0507-8

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., … Tian, Z. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, 33(4), 408–414. https://doi.org/10.1038/nbt.3096

Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29, 2669–2677.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.