Identification and validation of candidate genes associated with domesticated and improved traits in soybean

Published July 7, 2016

ORIGINAL RESEARCH

Identification and Validation of Candidate Genes Associated with Domesticated and Improved Traits in Soybean

Ling Zhou,† Longhai Luo,† Jian-Fang Zuo, Linfeng Yang, Li Zhang, Xuanmin Guang, Yuan Niu, Jianbo Jian, Qing-Chun Geng, Liping Liang, Qijian Song, Jim M. Dunwell, Zhenzhen Wu, Jia Wen, Yu-Qin Liu, and Yuan-Ming Zhang*

Abstract

Soybean, an important source of vegetable oils and proteins for humans, has undergone significant phenotypic changes during domestication and improvement. However, there is limited knowledge about genes related to these domesticated and improved traits, such as flowering time, seed development, alkaline-salt tolerance, and seed oil content (SOC). In this study, more than 106,000 single nucleotide polymorphisms (SNPs) were identified by restriction site associated DNA sequencing of 14 wild, 153 landrace, and 119 bred soybean accessions, and 198 candidate domestication regions (CDRs) were identified via multiple genetic diversity analyses. Of the 1489 candidate domestication genes (CDGs) within these CDRs, a total of 330 CDGs were related to the above four traits in the domestication, gene ontology (GO) enrichment, gene expression, and pathway analyses. Eighteen, 60, 66, and 10 of the 330 CDGs were significantly associated with the above four traits, respectively. Of 134 trait-associated CDGs, 29 overlapped with previous CDGs, 11 were consistent with candidate genes in previous trait association studies, and 66 were covered by the domesticated and improved quantitative trait loci or their adjacent regions, having six common CDGs, such as one functionally characterized gene *Glyma15g17480* (*GmZTL3*). Of the 68 seed size (SS) and SOC CDGs, 37 were further confirmed by gene expression analysis. In addition, eight genes were found to be related to artificial selection during modern breeding. Therefore, this study provides an integrated method for efficiently identifying CDGs and valuable information for domestication and genetic research.

During the approximate 10,000-yr period of domestication, many morphological and physiological traits in wild species of plants and animals have undergone dramatic modification to meet human needs. In plants, these traits include more or larger seeds, reduced seed dispersal, more robust plants, and decreased chemical and morphological defenses. Studying these domestication traits (DTs) is a useful way to identify key genes in their wild ancestors. Extensive studies have been

L. Zhou, J.-F. Zuo, and Y.-M. Zhang, Statistical Genomics Lab, College of Plant Science and Technology, Huazhong Agric. Univ., Wuhan 430070, China; L. Zhou, Inst. of Agro-biotechnology, Jiangsu Academy of Agric. Sci., Nanjing 210014, China; and L. Zhou, L. Zhang, Y. Niu, Q.-C. Geng, and J. Wen, State Key Laboratory of Crop Genetics and Germplasm Enhancement/National Center for Soybean Improvement, Nanjing Agric. Univ., Nanjing 210095, China; L. Luo, L. Yang, X. Guang, J. Jian, L. Liang, and Z. Wu, BGI-Shenzhen, Shenzhen 518083, China; Q. Song, Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705; J. M. Dunwell, School of Agriculture, Policy and Development, Univ. of Reading, Reading RG6 6AR, UK; Y.-Q. Liu, Crop Research Inst., Linyi Academy of Agricultural Sciences, Linyi 276012, China. † These authors have contributed equally to this work. Received 27 June 2015. Accepted 23 Jan. 2016. *Corresponding author (soyzhang@mail.hzau.edu.cn).

Abbreviations: 100SW, 100-seed weight; ABC, adenosine triphosphate-binding cassette; AST, alkaline-salt tolerance; CDG, candidate domestication gene; CDR, candidate domestication region; CDS, coding DNA sequence; DAF, days after flowering; DEG, differential expression gene; DG, domestication gene; DT, domestication trait; F_{ST}, fixation index; GATK, Genome Analysis Toolkit; GO, gene ontology; LD, linkage disequilibrium; LH, length of hypocotyls; LR, length of main root; MAF, minor allele frequency; MRP, multidrug resistance associated protein; PCA, principle component analysis; PDR, pleiotropic drug resistance; QTL, quantitative trait loci; RAD, restriction site associated DNA; ROD, reduction of diversity; seq, sequenced; SL, seed length; SNP, single nucleotide polymorphism; SOC, seed oil content; SS, seed size; ST, seed thickness; SW, seed width.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2015.09.0090

conducted in wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), rice (*Oryza sativa* L.), and sunflower (*Helianthus annuus* L.), and domestication genes (DGs) and their associated DTs were summarized by Doebley et al. (2006), Gross and Olsen (2010), and Olsen and Wendel (2013). These studies have provided useful resources for understanding the domestication process and artificial selection during domestication. However, for a considerable proportion of plants, including soybean [*Glycine max* (L.) Merr.], there is still little understanding of the DGs responsible for DTs during domestication.

Soybean was first domesticated by Chinese farmers between 6000 and 9000 yr ago (Carter et al., 2004). However, cultivated soybeans went through a genetic bottleneck of reduced genetic diversity during domestication (Guo et al., 2010; Tang et al., 2010). For example, soybean landraces exhibit only 41.9% of the allelic diversity found in wild soybean (Guo et al., 2010); with 50% of the genetic diversity and 81% of the rare alleles having been lost during domestication (Hyten et al., 2006), nucleotide diversity in cultivated soybean is lower than that in wild soybean (Lam et al., 2010). Moreover, many morphological and physiological traits have undergone significant changes to meet the specific needs of humans during domestication. These needs include alterations in flowering time, SS, tolerance to biotic and abiotic stresses, SOC, and shattering (Broich and Palmer, 1980), resulting in a loss of important traits that are still preserved in wild relatives. Therefore, identification of major novel genes responsible for DTs would facilitate genetic improvement in this crop.

Linkage and association analyses in populations derived from *G. max* (cultivated) and *G. soja Siebold & Zucc.* (wild) has previously identified many quantitative trait loci (QTL) for DTs, such as determinate habit (Liu et al., 2010; Tian et al., 2010), pod dehiscence (Bailey et al., 1997; Funatsuki et al., 2006; Liu et al., 2007; Kang et al., 2009; Dong et al., 2014; Van et al., 2014), flowering time (Keim et al., 1990; Liu et al., 2011), seed yield (Concibido et al., 2003; Liu et al., 2007; Li et al., 2008), and stem diameter (Keim et al., 1990). Some loci, such as *Dt1* for determinate growth habit (Tian et al., 2010), *E1-E4* for flowering time (Liu et al., 2008; Watanabe et al., 2009, 2011; Xia et al., 2012), and *SHAT1-5* gene for pod shattering (Dong et al., 2014) have been fine mapped and functionally characterized. In addition, evidence has shown that many DTs in soybean were controlled by one or two major QTL or genes (Liu et al., 2007), indicating that useful genes from wild soybean can be easily introgressed into cultivated soybean.

To date, whole genome sequences of >10 soybean accessions have been released, including Williams 82 (*G. max*; Schmutz et al., 2010), IT182932 (*G. soja*; Kim et al., 2010), and seven other wild accessions (Li et al., 2014). The nucleotide sequence of the *G. soja* genome (IT182932) is ~0.31% different from that of *G. max*, and the percentage increased to 3.76% when the 32.4 Mb of *G. max*-specific sequences in 712 genes, partially or completely absent in IT182932, were considered (Kim et al., 2010). Approximately 80% of the pan-genome was present in all seven wild accessions, whereas the rest was dispensable, and some genes with *soja*-specific presence-absence variations may contribute to the variation in defense response, cell growth, and photosynthesis (Li et al., 2014). To conduct a more in-depth study of soybean domestication, additional resequencing and microarray studies have been reported. Lam et al. (2010) resequenced a total of 17 wild and 14 cultivated soybean genomes and found higher allelic diversity in wild soybean, and a high level of linkage disequilibrium (LD) in the soybean genome; and some genes identified in the 470 domestication regions (DRs) were associated with stem elongation-related traits, disease resistance, and metabolism. Song et al. (2013) identified a total of 620 DRs by genotyping 96 landraces and 96 wild soybeans with the SoySNP50K Illumina Infinium BeadChip. Li et al. (2013) found that 60 genes from 21 CDRs and 106 genes from 20 regions associated with soybean improvement were related to several important agronomic traits, such as yield, plant height, lodging, maturity time, seed weight, seed hardness, seed-coat color, and flower color, by analyzing the whole genome sequence from 25 accessions and sequence data of the 30 accessions deposited at NCBI website. Chung et al. (2014) identified 3068 DGs in 206 DRs by analyzing sequences from 10 cultivated and six wild soybeans. However, the above studies were based on limited sample size, although Zhou et al. (2015b) identified eight genes by associating 48 domestication-related loci with flowering time and SS in 286 soybean accessions.

In this study, a total of 286 soybean accessions were resequenced by restriction site associated DNA (RAD) sequencing technology to investigate genetic diversity, and to mine CDGs during domestication. An effort was made to relate to all the CDGs to DTs by the GO enrichment, gene expression, and pathway analyses. These CDGs were further verified by the trait-gene association studies and RNA-seq (sequenced) analysis.

Materials and Methods

Germplasm for RAD-Sequencing and Their Phenotypic Measurements

A total of 286 soybean accessions, including 14 wild, 153 landrace, and 119 cultivars from six geographic regions in China, were grown in Nanjing, China, with a complete randomized design from 2008 to 2012. The plots were 1.5 m wide and 2 m long, and a list of the accessions was described by Zhou et al. (2015b).

Seeds from five plants in the middle row of each plot were measured for SS. Seed length (SL), seed width (SW), and seed thickness (ST) were measured using digital vernier calipers, and 100 seeds from dried samples were weighted (100-seed weight, 100SW) from 2008 to 2012. The SL, SW, and ST for each accession were averaged based on 20 seeds and 100SW for each accession was averaged based on three replicates.

The first and full flowering times were the days from the date of emergency to the date of the first flower and 75% plant flowering, respectively, and both traits were observed in the field from 2010 to 2012.

A salt-water flooding method was used to evaluate the alkaline and salt tolerances of 286 soybeans (Sobhanian et al., 2010). In brief, 12 seeds for each accession were sown in a 30 × 20 × 15 cm plastic container with 3.5-cm-deep sand, and then treated with control (CK; pH 7.0), 100 mM NaCl (pH 7.0), and 10 mM Na_2CO_3 (pH 11.1), each with two 7-d replications. A 350-mL aliquot of the appropriate solution for each treatment was applied to each plastic container filled with sand. Twelve soybean seeds for each treatment were grown in a growth chamber under white fluorescent light (600 μmol m^{-2} s^{-1}; 14 h light/10 h dark) at 25 ± 1°C. Two alkaline-salt tolerance traits, length of main root (LR) and length of hypocotyls (LH) for healthy seedlings, were measured from 5 plants 7 d after sowing from 2009 to 2010. To measure the degree of salt-alkaline tolerance, original trait observations were transferred into salt-alkaline tolerance index using the following equations

$$\mathrm{STI} = (x_{\mathrm{CK}} - x_{\mathrm{NaCl}})/x_{\mathrm{CK}} \times 100\%$$

$$\mathrm{ATI} = \left(x_{\mathrm{CK}} - x_{\mathrm{Na_2CO_3}}\right)\Big/x_{\mathrm{CK}} \times 100\%$$

where STI is salt tolerance index, ATI is alkaline tolerance index, and x_{CK}, x_{NaCl}, and $x_{\mathrm{Na_2CO_3}}$ stand for phenotypic values in control, saline, and alkaline treatments, respectively.

Approximate 10 g of seeds was collected from five plants per accession. Based on the method of Baydar and Akkurt (2001), five fatty acids for each accession were measured by gas chromatography with a flame ionization detector and a Permabond FFAP stainless steel column (50 m × 0.2 mm × 0.33 μm, ThermoFisher Scientific, Waltham, MA) at the Wuhan Research Branch of the National Rapeseed Genetic Improvement Center in 2015. Using methyl heptadecanoate (C17) as internal standard, SOC was calculated by

$$\mathrm{SOC}\ (\%) = \frac{\text{TPA for 5 FA}}{\text{ISPA}} \times \frac{2\ \text{mg/mL} \times 0.5\ \text{mL}}{30\ \text{mg}} \times 100\%$$

where TPA is total peak area, FA is fatty acids, and ISPA is internal standard peak area.

RAD-Sequencing and Sequence Alignment

DNA for the 286 accessions were extracted from fresh leaves of multiple plants per accession and digested with the *EcoRI* restriction enzyme. The sequence library was prepared by Baird et al. (2008), and 50 bp at each end of each fragment was sequenced. The reads were aligned to the Glyma1.1 of Williams 82 (Schmutz et al., 2010; http://www.jgi.doe.gov, verified 3 Mar. 2016) using the Burrows-Wheeler Alignment Tool (http://bio-bwa.sourceforge.net/bwa.shtml, verified 3 Mar. 2016) after elimination of low quality reads and trimming of adaptors. The SNP alleles were called using the Genome Analysis Toolkit (GATK; http://www.broadinstitute.org/gatk/index.php, verified 3 Mar. 2016) as described by Zhou et al. (2015b).

Population Structure Analyses

Population Structure Determination

Based on the 106,013 SNPs, the population structure of the 286 soybean accessions was determined using the STRUCTURE 2.2 software (Pritchard et al., 2000; Falush et al., 2003, 2007). The number of subgroups (K) was set from 2 to 7. In the Markov chain Monte Carlo Bayesian analysis for each K, the length of a Markov chain consisted of 100,000 sweeps. The first 10,000 sweeps (the burn-in period) were deleted, and thereafter, the chain was used to calculate the mean of log-likelihood. This process was repeated five times, and the total average of log-likelihood at fixed K was used. The ad hoc statistic ΔL, based on the rate of change in the log-likelihood of data between successive K values, was used to determine the suitable value of K (Evanno et al., 2005). The **Q** matrix at the estimated K was calculated based on all SNPs.

A SNP was described as rare if minor allele frequency (MAF) was <5%, or as specific if an allelic variant was identified in one subgroup but not in the others. In contrast, a SNP was defined as common if allelic variations were observed in all the subgroups.

Principle Component Analysis

As stated by Patterson et al. (2006), principle component analysis (PCA) was conducted using sample covariance matrix $\mathbf{X} = \mathbf{M}\mathbf{M}^{\mathrm{T}}/_{\mathrm{S}}$ of the accessions, where no. of SNPs (S) = 106,013, T is matrix transposition, and $\mathbf{M} = \left(d'_{ik}\right)_{n\times s}$ was the normalized genotypic information matrix (i = 1, …, n; k = 1, …, S). d'_{ik} was defined by

$$d'_{ik} = \frac{d_{ik} - E(d_k)}{\sqrt{E(d_k)\left[1 - E(d_k)\right]/2}}$$

where $E(d_k) = \frac{1}{n}\sum_{i=1}^{n} d_{ik}$, and d_{ik} for the kth SNP genotype of the ith accession was defined as d_{ik} = 0 for homozygous of the reference allele, d_{ik} = 1 for heterozygous, and d_{ik} = 2 for homozygous of the nonreference allele. The eigen-vector decomposition was performed in the *R* function eigen, and the significance of the eigenvectors was determined with a Tracey-Widom test implemented in the program twstats that was provided with the EIGENSOFT software (Patterson et al., 2006).

Phylogeny Tree Construction

Pair-wise distance among all accessions was calculated based on the p-distance (porportion [p] of amino acid sites at which the two sequences to be compared are different) model, and a neighbor-joining method (Saitou and

Nei, 1987) was used to construct the phylogenetic tree. The software PHYLIP 3.68 (Retief, 2000) was used for the above analysis. The p-distance D_{ij} between two individuals i and j was defined as $D_{ij} = \sum_{l=1}^{L} c_{ij}^{(l)} \Big/ L$, where L is the length of regions where high quality SNPs were identified, and c_{ij} was defined as $c_{ij}^{(l)} = 0$ if the genotypes at position l for the two accessions were AA and AA, $c_{ij}^{(l)} = 0.5$ if the genotypes at position l were AC and AC (or AA and AC), and $c_{ij}^{(l)} = 1$ if the genotypes at position l were AA and CC.

Linkage Disequilibrium Analysis

To measure LD levels in wild, landrace, and bred soybeans, the software Haploview (Barrett et al., 2005) was used to calculate the correlation coefficient (r^2) of alleles at any two SNP loci by setting maxdistance to 200, dprime-minMAF to 0.1, and hwcutoff at 0. The curves and plot of average r^2 against pair-wise marker distances were drawn with R scripts.

Genetic Diversity, and Candidate Domestication-Related Regions and Genes

π, θ, and Tajima's D Estimation

Tajima's D was calculated by the difference between the average pairwise nucleotide diversity (π), and the number of segregating sites as measured by θ, while π and θ were calculated using the definition of Xu et al. (2012). The π, θ, and Tajima's D were calculated for nonoverlapping 20-kb sliding windows across the genome. The size of the window was selected from 10, 20, 50, and 100 kb. Because genomic regions with artificial selection are nonneutral in the D test of Tajima (1989), a criterion of |D| > 2 was adopted in the determination of CDRs (Varshney et al., 2013). The Tajima's D statistic is expected to be zero under the neutral equilibrium model and the departure from the standard neutral model will lead to nonzero values (Tajima, 1989).

Fixation Index Calculation

The fixation index (F_{ST}) is a measure of population differentiation due to genetic structure. The F_{ST} value was calculated by the software Genepop v.4.2 (http://kimura.univ-montp2.fr/~rousset/Genepop.htm, verified 3 Mar. 2016) in 20-kb nonoverlapping sliding windows along the entire genome among 119 bred, 153 landrace, and 14 wild soybeans (Xu et al., 2012). On the basis of the threshold of $F_{ST} > 0.45$ (Lam et al., 2010), the CDRs were identified.

Reduction of Diversity Estimation

The reduction of diversity (ROD) was defined as ROD = $1 - \pi_{\text{bred or landrace}}/\pi_{\text{wild}}$, based on the ratio of diversity in bred (or landrace) soybean vs. wild soybean ($\pi_{\text{bred}}/\pi_{\text{wild}}$, $\pi_{\text{landrace}}/\pi_{\text{wild}}$) in 20-kb nonoverlapping sliding windows along the entire genome (Xu et al., 2012). The value of ROD > 0.98 (Chung et al., 2014) was adopted in the identification of CDRs.

The genes in the above CDRs were defined as CDGs.

Gene Ontology Enrichment Analysis

The GO annotations of the CDGs, including molecular function, molecular location and biological process, were conducted using the online tool Goanna (McCarthy et al., 2006). The GO enrichment analysis was performed using GOstats at the 0.05 significance level (Falcon and Gentleman, 2007).

Gene Expression Data and Trait-associated Analysis

Two transcriptome datasets of soybean (Jones and Vodkin, 2013) were obtained from the Gene Expression Omnibus database at http://www.ncbi.nlm.nih.gov/geo (verified 3 Mar. 2016). Dataset I (accession number GSE29163) included transcriptome sequences from six tissues: leaf (SRX062330), root (SRX062331), floral bud (SRX062333), seedling (SRX062334), globular stage whole seed (SRX062325), heart stage whole seed (SRX062326), cotyledon stage whole seed (SRX062327), early-maturation stage whole seed (SRX062328), and dry whole seed (SRX062329). Dataset II (accession number GSE42871) included transcriptome sequences from seven stages of seed development: 4 d after flowering (DAF) whole seed (SRX212244), 12 to 14 DAF whole seed (SRX212246), 22 to 24 DAF whole seed (SRX212248), 5 to 6 mg whole seed (SRX212250), 100 to 200 mg cotyledon (SRX212252), 400 to 500 mg cotyledon (SRX212254), and dry whole seed (SRX212256). Seed lipids are generated during the developmental stages of R4 to R7 (Pedersen, 2009; Jones and Vodkin, 2013). Using the quite stringent threshold of "> twofold" above the median, most genes significantly expressed in floral bud, all the seed development stages, or in the R4-R7 development stages can be found (Barnes and Derwent, 2007).

Coexpression Analysis and Identification of Important Genes in Networks

Using the Pearson correlation coefficient, coexpression of each flowering time candidate gene with the others in soybean was examined, and the significant *P*-value threshold was set at 1E-09.

Based on the gene expression profile, two genes were considered to be linked in a gene network if their correlation coefficient was significant at the 0.01 level. If the number of links with a gene in the network was more than one constant, the gene was considered as important. The analysis was conducted among candidate flowering-time genes identified in this study, and between the candidate flowering-time genes identified in this study and the genes reported by Jung et al. (2012). SAS 9.3 software was used for the analysis. The networks were visualized using the BioLayout Expression^{3D} v.3.1 (www.biolayout.org, verified 3 Mar. 2016).

Candidate Gene Association Study

All the candidate genes for each trait in the domestication, GO enrichment, gene expression, and pathway

analyses were used to be associated with the corresponding trait using the mixed linear model (MLM) method in TASSEL v.5.0 (www.maizegenetics.net/tassel, verified 3 Mar. 2016) with the population average and population structure (**Q** matrix) as covariates. Using the same TASSEL v.5.0, the kinship coefficient matrix for polygenic background control was calculated from the 106,013 SNPs. The *P*-value was used to determine the association significance in the candidate gene association study, and the critical value for significance was set at the 0.01 level.

Differential Expression Genes Analysis Based on RNA-Sequenced Data

One cultivated soybean (No. 101) and two wild soybean accessions (No. 265 and 272) were chosen for RNA-seq analysis. The No. 101 is a cultivar with early flowering (days to first flower: 44), large seed (100SW: 11.81 g), middle tolerance to salt (tolerance index for LR: 0.50) and high SOC (20.9%), while No. 265 and 272 are of late flowering (67 and 54 d, respectively), small seed (100SW: 2.35 and 2.62 g, respectively), high tolerance to salt (tolerance index for LR: 0.22 and 0.31, respectively), and low SOC (11.9 and 12.5%, respectively). Two plants from each accession were planted in a soil-filled plastic pot (35 × 28 cm) and grown under controlled environment conditions in 2014, and seed were collected at five seed development stages (15, 25, 35, 45, and 55 d after first flower) for RNA extraction.

Total RNA was extracted using *TRIzol* reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The RNA was analyzed in an Illumina Hiseq 2500 Sequencer. Sequence reads were aligned using SAM format (Li et al., 2009), a generic alignment format for storing read alignments against reference sequence, supporting short and long reads produced by different sequencing platforms. The obtained sequence reads were subsequently aligned to the soybean reference genome (Glyma2.0, http://www.phytozome.net/soybean, verified 3 Mar. 2016; Schmutz et al., 2010). TopHat (http://tophat.cbcb.umd.edu, verified 3 Mar. 2016) was used to identify splice junctions (Trapnell et al., 2009).

Differential expression genes (DEGs) between the cultivated (No. 101) and two wild soybean accessions (No. 265 and 272) at the five development stages after first flower were identified by the statistical R package DEGseq (https://www.bioconductor.org/packages/release/bioc/html/DEGseq.html; verified 8 Mar. 2016) with a random sampling model (Wang et al., 2010), and at the adjusted significance level of 0.001. The purpose was to further confirm the CDGs obtained from the above analyses.

Results and Discussion

Identification and Distribution of SNPs

Using the RAD-sequencing approach, at least 400 Mb (×0.35) of sequence were generated for each of the 286 soybean accessions except for No.70, where the average amount of sequence was >1 G (×0.85).These sequences were aligned to the whole genome sequence of Williams 82 (Glyma1.1, Schmutz et al., 2010), and approximately 0.24 million candidate SNPs were identified among the 286 accessions. The GATK was used to detect the missing genotypes for each SNP, the frequent distribution for the missing proportion was presented in Supplemental Fig. S1, and the average plus standard deviation were 2.8% ± 2.7%. Using the quality control criterion described in Zhou et al. (2015b), a total of 106,013 high-quality SNPs were identified.

Table 1. Distribution of single nucleotide polymorphisms in the soybean genome.

	CDS†							
Chr	CDSStart	CDS	CDSEnd	Intron	5′UTR	3′UTR	Nongenic sequence	Total
1	5	340	2	600	77	140	4,227	5,391
2	15	552	3	825	147	188	3,703	5,433
3	3	448	2	701	146	206	3,496	5,002
4	4	430	2	755	162	212	3,743	5,308
5	1	326	6	658	137	165	3,047	4,340
6	5	500	5	979	130	228	4,063	5,910
7	7	529	7	867	122	209	3,576	5,317
8	9	594	7	1,147	180	274	3,683	5,894
9	10	480	5	767	124	141	3,449	4,976
10	6	463	4	812	137	198	3,405	5,025
11	7	446	3	697	99	160	2,913	4,325
12	4	385	2	600	151	180	3,122	4,444
13	3	694	7	1,058	225	270	3,764	6,021
14	8	366	6	678	80	124	3,902	5,164
15	2	458	7	838	89	206	4,522	6,122
16	1	503	4	780	99	249	3,415	5,051
17	0	337	6	688	102	134	3,014	4,281
18	1	587	9	1,058	180	256	5,474	7,565
19	2	346	2	759	97	159	3,843	5,208
20	2	465	4	763	132	163	3,707	5,236
Total	95	9,249	93	16,030	2,616	3,862	74,068	106,013

† CDS, coding DNA sequence.

Of the 106,013 SNPs, a total of 31,945, 9249, 16,030, 95, 93, 2616, and 3862 were located in genes, coding DNA sequence (CDS), introns, CDSStart, CDSEnd, 5′UTR, and 3′UTR, respectively (Table 1); and a total of 16,642 and 16,613 were within 2 kb upstream and downstream of genes, respectively. Overall, 65,200 (61.50%) SNPs were located within genes or within 2 kb upstream or downstream of genes.

Of the 25,100 SNPs with MAF < 5%, 4389 (17.49%), 1732 (6.90%), and 702 (2.80%) were specific to wild, landrace, and bred soybeans, respectively, while 14,995 (59.74%) were common in the three subpopulations (Fig. 1A). A large number of specific SNPs, were gradually lost during the processes of artificial selection from the wild to bred soybeans (domestication), indicating the reduction of genetic variation from wild to elite soybeans. A similar result was reported by Hyten et al. (2006), who showed that 50% genetic diversity and 81% of rare alleles have been lost during domestication. In addition, new alleles were observed in bred soybean, although the proportion is only 2.80%. The possible reason is strong

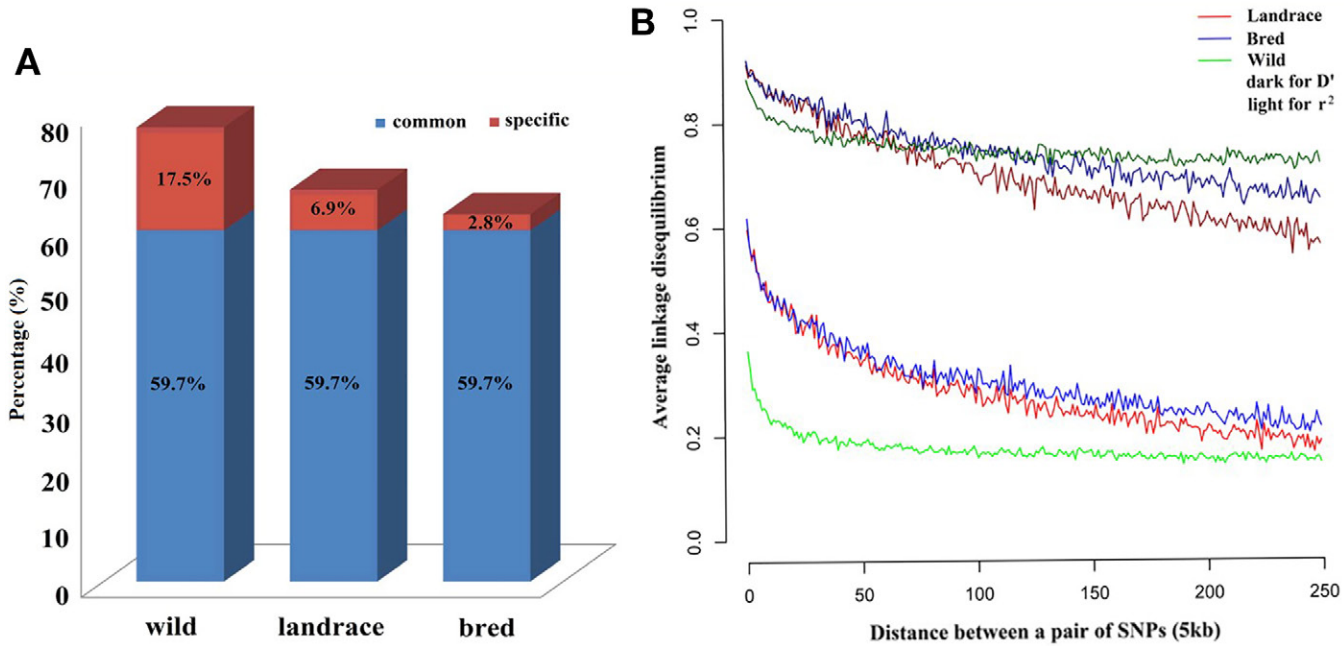

Fig. 1. (A) Distribution and (B) linkage disequilibrium of single nucleotide polymorphisms (SNPs) in wild, landrace, and bred soybeans. (A) Distribution of common and specific SNPs in wild, landrace, and bred soybeans; and (B) Plot of r^2 and D′ against distance (5 kb) between a pair of SNPs in wild, landrace, and bred soybeans.

selection for a novel allele of an important trait to maintain the mutation of its adjacent loci.

Within a 50 kb window, the rate of LD decay in wild soybean was significantly faster than those in landrace and bred soybeans (Fig. 1B). This result was similar to that reported by Lam et al. (2010). The slower decay rate in bred soybean may be caused by a limited number of elite bred cultivars or by other factors such as a bottleneck (Barecki and Suarez, 2001). The r^2 values in wild soybean were significantly lower than those in cultivated soybean, which may result from the loss of genetic diversity during domestication. The D′ values were larger than the r^2 values in each of the wild, landrace, and bred soybeans, larger D′ values in the >250 kb window in wild soybean relative to landrace or bred soybean were partly due to the smaller sample size in wild soybean (Xu, 2010).

Population Structure

Analysis of population structure of the 286 accessions based on the 106,013 SNPs showed that the maximum peak value of Δk appeared at $K = 4$ when the number of subpopulations increases from 2 to 7. When $K = 2$, wild and cultivated soybeans were separated, whereas when $K = 3$, the cultivated soybean was further divided into two groups: landrace and bred soybeans (Fig. 2A). When $K = 4$, 106 accessions were assigned to Subgroup 1 (bred), 134 to Subgroup 2 (mainly landrace), 12 to Subgroup 3 (mainly wild), and 34 to Subgroup 4 (mixture; Fig. 2A).

In the PCA, four principle components explained 41.11% of the total SNP variance, while the first and second principle components explained 13.63 and 12.58% of the total SNP variance, respectively. Based on the analysis of the first and second principle components, wild, landrace, and bred soybeans were clustered into different groups, except a few cultivated soybeans were assigned to the subgroup with mixture (Fig. 2B). This finding was also consistent with the Structure result.

Neighbor-joining analysis showed similar results (Fig. 2C), with the largest genetic distance between wild and bred soybeans, and the intermediate genetic distance between landrace and wild (or bred) soybeans.

Genetic Diversity and Candidate Domestication Regions (Genes)

π, θ, and Tajima's D

To measure the genetic diversities in wild, landrace, and bred soybeans, two commonly used metrics of genetic diversity, namely the average pairwise nucleotide diversity (π) and the number of segregating sites as measured by θ, were calculated based on the 106,013 SNPs. The estimates of π and θ in bred, landrace, and wild soybeans were 0.1030 and 0.1044, 0.1049 and 0.1046, and 0.2147 and 0.2377, respectively. The genetic diversity was two-fold lower in cultivated soybean than in wild soybean. This result is similar to those of Chung et al. (2014) and Lam et al. (2010).

To further dissect genome-wide pattern of the nucleotide diversity among wild, landrace, and bred soybeans, the Gaussian kernel smoothing function with a step length of 20 kb across each chromosome was applied. As shown in Fig. 3, the ranges for estimates of π in wild, landrace, and bred soybeans were 0.0344 to 0.3996, 0 to 0.3724, and 0 to 0.3788, respectively; and the averages plus standard deviations were 0.2160 ± 0.0233, 0.1055 ± 0.1058, and 0.1035 ± 0.1078, respectively. The ranges for estimates of θ in the above three soybeans were 0.0786

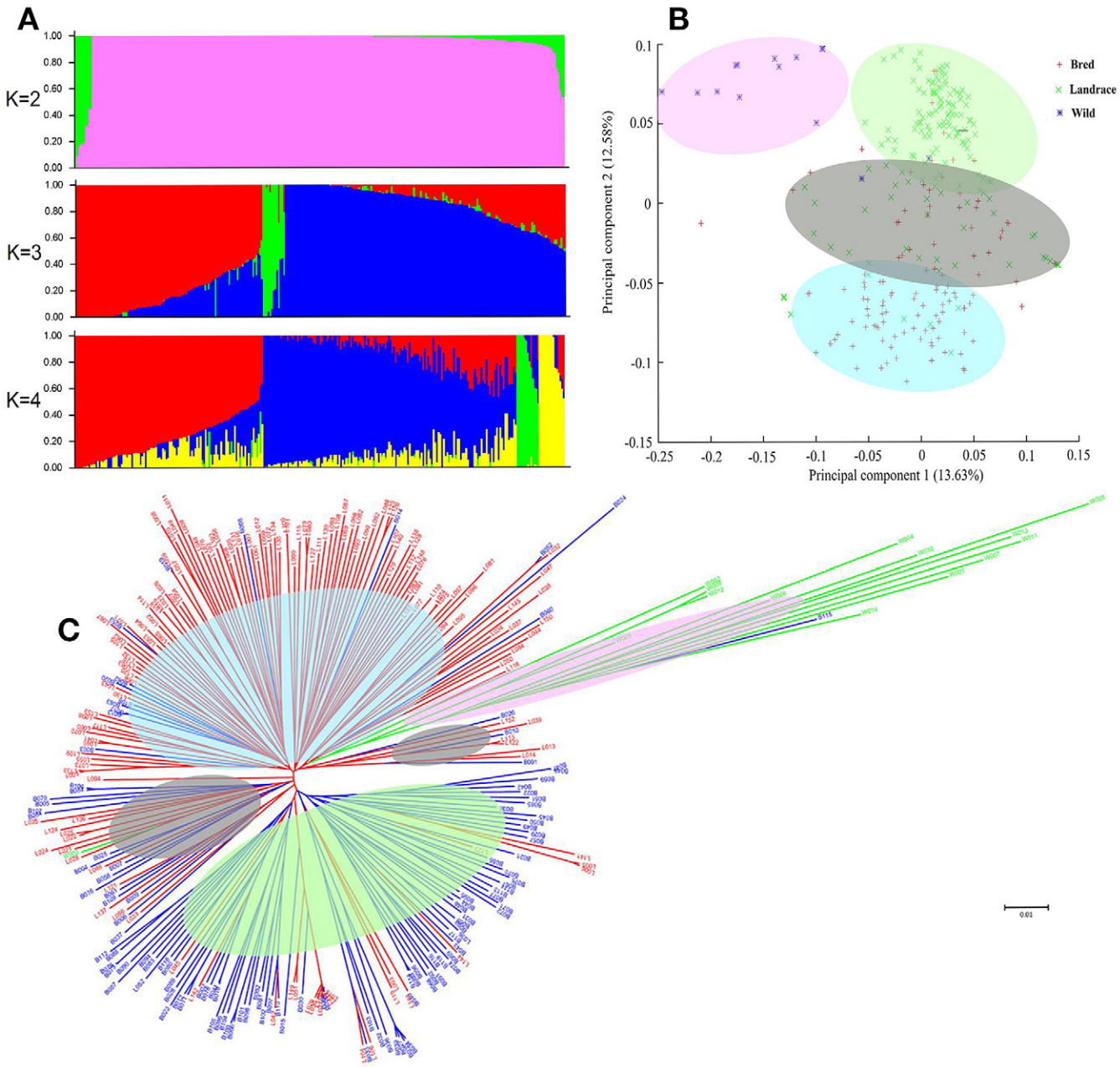

Fig. 2. Population structure of 286 soybean accessions. (A) Structure analysis. The wild (green) and cultivated (pink) soybeans were separated (number of subgroups = *K* = 2), cultivated soybean was further split into landrace (blue) and bred (red) soybeans (*K* = 3), and some cultivated soybeans were mixture (yellow; *K* = 4). (B) Principle component analysis. (C) Neighbor-joining tree analysis. In the phylogenetic tree, wild, landrace, and bred soybeans were represented by pink, green, and blue shadows, respectively.

to 0.3311, 0 to 0.1818, and 0 to 0.1887, respectively; and the averages plus standard deviations were 0.2393 ± 0.0029, 0.1055 ± 0.0273, and 0.1044 ± 0.0280, respectively. The above estimates for π and θ were used to calculate Tajima's D. The ranges for D in wild, landrace, and bred soybeans were –2.6324 to 1.9218, –2.5195 to 4.4259, and –2.6500 to 3.7322, respectively; and the averages plus standard deviations were –0.4007, 0.1185, and –0.0746, respectively. Based on a threshold of Tajima's $|D| > 2$, which are generally considered to be CDRs (Varshney et al., 2013), 487 and 616 CDRs were identified in landrace and bred soybeans, respectively, indicating artificial selection in cultivated soybean. A total of 166 (3.1%) common CDRs between landrace and bred soybeans were identified genome-wide. In these common CDRs, 1344 CDGs were identified (Supplemental Table S1).

Differentiation Coefficient F_{ST} and ROD

Average F_{ST} estimates between landrace and wild soybeans, between bred and wild soybeans, and between landrace and bred soybeans were 0.155, 0.1908, and 0.0267, respectively, indicating high differentiation between wild and cultivated soybeans and low differentiation between landrace and bred soybeans.

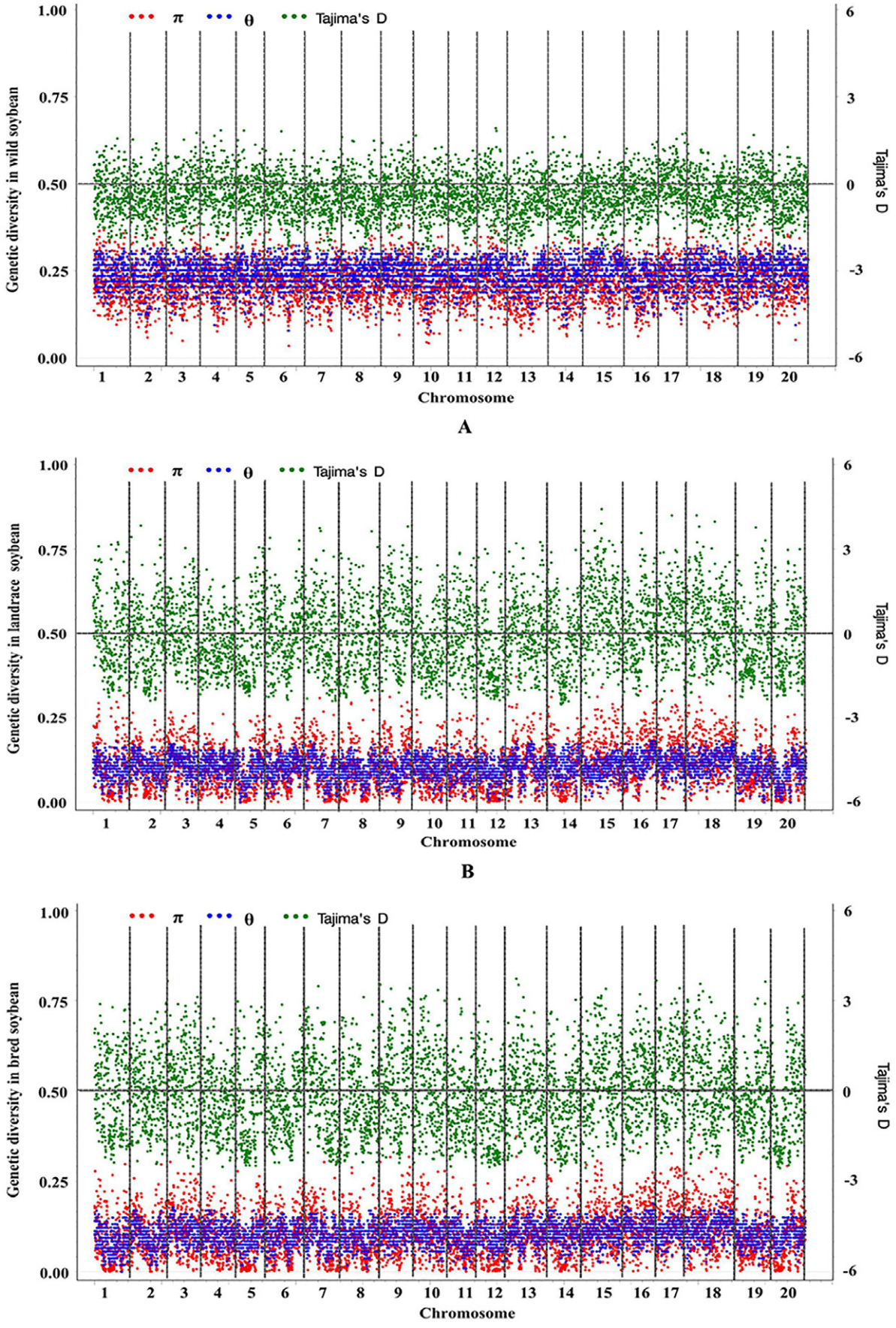

Fig. 3. Genetic diversity parameters π, θ, and Tajima's D in the nonoverlapping 20-kb sliding windows across the genome in (A) wild, (B) landrace, and (C) bred soybeans.

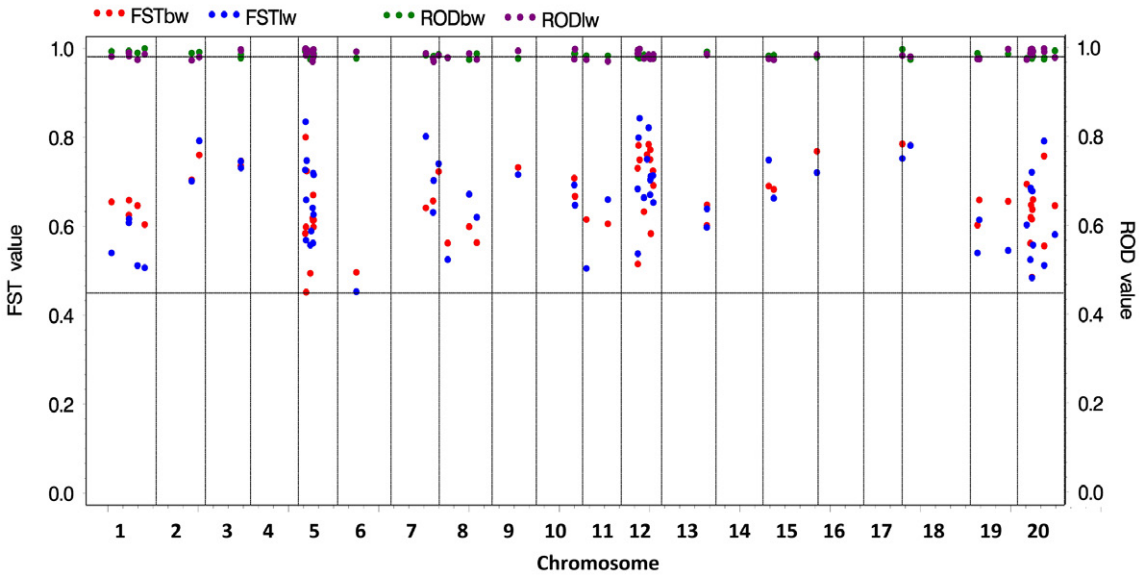

Fig. 4. Candidate domestication regions in landrace (l) and bred (b) soybeans compared with wild (w) soybean based on reduction of diversity (ROD) and fixation index (F_{ST}) analyses.

For the purpose of detecting selective sweeps from domestication, regions with high differentiation between wild and cultivated soybeans were identified by calculating the differentiation index [$F_{ST(bred-wild)}$ and $F_{ST(landrace-wild)}$] and regions with significantly lower levels of polymorphisms were identified by calculating the ROD (Xu et al., 2012):

$$ROD = 1 - \pi_{bred}/\pi_{wild} \text{ or } 1 - \pi_{landrace}/\pi_{wild}.$$

As a result, 45 common CDRs were identified (Fig. 4) by the F_{ST} and ROD approaches. In these regions, 241 genes were found and considered to be CDGs. Among these genes, 61 genes were common to those identified by Chung et al. (2014).

Based on the results from the Tajima' D, F_{ST} and ROD analyses, 13 common CDRs were determined and these CDRs contained 96 genes. A total of 1489 candidate genes in the 198 CDRs which were determined by at least one of the three analysis methods were identified and subjected to further analysis.

Based on GO enrichment analysis, we identified 252 significantly overrepresented GO categories, which were related to flowering time, seed development, tolerance/resistance, and oil biosynthesis (Supplemental Table S2).

Flowering Time

Through GO enrichment analyses, nine significantly enriched terms ($P < 0.05$) for the biological processes related to flowering time were identified (Supplemental Table S3), including GO:0009909 (regulation of flower development), GO:0009911 (positive regulation of flower development), GO:0009640 (photomorphogenesis), GO:0009649 (entrainment of circadian clock), GO:0048573 (photoperiodism, flowering), and GO:0009910 (negative regulation of flower development). Among the 96 candidate genes involved in these terms, 54 genes showed high expression in flower tissue and were viewed as flowering-time-related genes (Supplemental Table S4). In these highly expressed genes, *Glyma05 g05620*, *Glyma09 g00346*, *Glyma09 g11600*, *Glyma10 g36720*, *Glyma15 g17480*, and *Glyma17 g06950* were homologous to the *Arabidopsis thaliana* (L.) Heynh. flowering-time genes *CKB1*, *FAR1*, *PHYB* (*PHYE*), *ASHH1*, *FKF1*, and *ZTL*, respectively. These genes are different from those in Zhou et al. (2015b), because only 48 domestication-related loci were used to perform association studies in Zhou et al. (2015b). In Arabidopsis, the gene *CO* is associated with flowering time (Putterill et al., 1995) and regulated by six genes, five of which, *CKB1* (Sugano et al., 1998; Lu et al., 2012), *FAR1* (Hudson et al., 1999; Yanovsky and Kay, 2002), *PHYB* (*PHYE*; Iñigo et al., 2012), *FKF1* (Boss et al., 2004; Putterill et al., 2004; Song et al., 2012), and *ZTL* (Somers et al., 2004; Kiba et al., 2007; Ito et al., 2008, 2012b; Kim et al., 2013) were in the light-signaling pathway and one gene *ASHH1* was in the vernalization or autonomous pathway (Michaels and Amasino, 1999; Hepworth et al., 2002; He and Amasino, 2005; Xu et al., 2008; Fig. 5). Co-expression analysis of these six genes with all the other genes in soybean showed that four of the six genes were co-expressed with 12 genes at the 1E-9 level: *Glyma09 g00346* was co-expressed with *Glyma02 g46570*, *Glyma03 g31560*, *Glyma06 g19190*, *Glyma10 g28950*, *Glyma14 g24910*, and *Glyma17 g37470*, *Glyma10 g36720* with *Glyma01 g30320* and *Glyma03 g07790*, *Glyma15 g17480* with *Glyma08 g39830* and *Glyma08 g25750*, and *Glyma17 g06950* with *Glyma17 g12590* and *Glyma19 g43700* (Supplemental Fig. S2A, Supplemental Table S5).

A co-expression network analysis showed that six genes (*Glyma02 g40400*, *Glyma04 g01140*, *Glyma06 g34850*, *Glyma08 g01100*, *Glyma08 g03980*, and *Glyma09 g00346*) among the above 54 genes had 20 or more

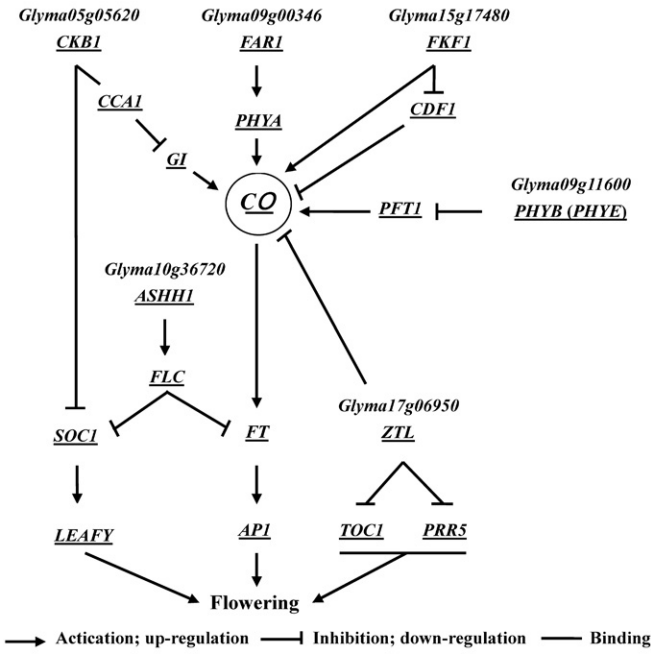

Fig. 5. Simplified pathway model of the hypothesized connections in flowering pathways. An underlined gene denotes an Arabidopsis gene, but not a soybean gene.

significant expression links at the 0.01 level (Supplemental Fig. S2B, Supplemental Table S4); and 14 genes (*Glyma02 g40400*, *Glyma04 g01140*, *Glyma06 g34850*, *Glyma08 g01100*, *Glyma08 g03980*, *Glyma09 g00346*, *Glyma10 g36720*, *Glyma14 g16040*, *Glyma14 g36100*, *Glyma15 g17030*, *Glyma15 g17710*, *Glyma15 g19450*, *Glyma18 g07830*, and *Glyma18 g13250*) had 50 or more significant expression links with 185 Arabidopsis homologous flowering-time genes at the 0.01 level (Jung et al., 2012; Supplemental Fig. S2C, Supplemental Table S6).

Seed Development

GO enrichment analysis identified 12 enriched terms, with 101 candidate genes, in the biological process that were related to seed development, such as GO: 0048580 (regulation of postembryonic development) and GO: 2000241 (positive regulation of reproductive process; Supplemental Table S3). Using comparative genome analysis, a total of 73 CDGs in this study were found to be homologous to the seed development genes in Arabidopsis, including seed coat development (2 genes), embryo development ending in seed dormancy (59), seed germination (2), ovule development (3), endosperm development (1), and other seed development processes (6; Supplemental Table S7). Among the 59 genes of embryo development ending in seed dormancy, 25 embryo-defective genes were likely to encode chloroplast-localized proteins that result in embryo lethality (Tzafrir et al., 2003; www.seedgenes.org, verified 3 Mar. 2016). Between the above 101 and 73 genes, 15 common genes were identified. Among all the 159 genes, 112 genes were highly expressed during the seed development process (Supplemental Table S7). These genes are different from those in Zhou et al. (2015b). The reason is also because only 48 domestication-related loci were used to perform association studies in Zhou et al. (2015b).

In Arabidopsis, *AUXIN RESPONSE FACTOR2* (*ARF2*) has been identified to be associated with SS (Okushima et al., 2005) and was homologous to *Glyma08 g01100*, which was highly expressed during seed development. The gene *Glyma08 g01100* may regulate SS in soybean. Although *HAIKU2* (*IKU2*) has been identified to be associated with SS in Arabidopsis (Garcia et al., 2003; Luo et al., 2005), its homolog *Glyma14 g21901* was not expressed during seed development in soybean.

The soybean gene *Glyma15 g09500*, which was highly expressed during seed development, was homologous to the three genes, *SHATTERPROOF1* (*SHP1*), *SHP2* (Liljegren et al., 2000), and *SEEDSTICK* (*STK*; Pinyopich et al., 2003), which have been identified to be associated with seed dispersal and shattering in Arabidopsis (Martínez-Andújar et al., 2012). The *Glyma15 g09500* gene may regulate seed shattering in soybean but be different from the *SHAT1–5* gene controlling soybean shattering (Dong et al., 2014).

Tolerance and Resistance

GO enrichment analysis showed that 21 enriched terms, with 169 candidate genes, in the biological process were associated with tolerance or resistance in soybean, such as GO:0071214 (cellular response to abiotic stimulus) and GO:0009870 (defense response signaling pathway, resistance gene-dependent; Supplemental Table S3). Based on functional annotation and domain conservation, 37 candidate genes were identified to be associated with tolerance or resistance in soybean, including the adenosine triphosphate-binding cassette (ABC) transporters, abiotic tolerance genes, and disease resistance genes (Supplemental Table S8). Between the 169 and 37 genes mentioned above, 10 common genes were identified. Among all the 196 CDGs, 7, 3, and 27 belonged to ABC transporters, abiotic tolerance genes, and disease resistance genes, respectively (Supplemental Table S8).

In Arabidopsis, ABC transporters, including pleiotropic drug resistance (PDR), multidrug resistance, and multidrug resistance associated protein (MRP) subfamilies, are important for defense and normal growth and development (Kang et al., 2011). *AtPDR1* (Alejandro et al., 2012; Sibout and Höfte, 2012), *AtPDR11* (Xi et al., 2012) and *AtPDR12* (Lee et al., 2005; Kang et al., 2010) are associated with PDR and have been identified to be associated with tolerance in Arabidopsis. *Glyma13 g19920* was homologous to *AtPDR1/AtABCG29*; *Glyma02 g35841* and *Glyma03 g32530* to *AtPDR11*; *Glyma03 g32520* and *Glyma03 g32540* to *AtPDR12/AtABCG40*. *Glyma10 g17300* and *Glyma03 g32500* were homologous to *AtMRP3* and *AtMRP5*. *AtMRP3* (Bovet et al., 2003, 2005; Zientara et al., 2009) and *AtMRP5* (Klein et al., 2003; Lee et al., 2004) have been identified to be associated with tolerance in Arabidopsis.

Abiotic stresses, such as high salinity, drought, and low temperature, substantially affect plant growth. Based

on functional annotation, the CDGs *Glyma04 g01120*, *Glyma05 g16860*, and *Glyma19 g40150* were associated with abiotic stresses in soybean (Supplemental Table S8). The *Glyma04 g01120* is a homolog of *SALT TOLERANCE HOMOLOG 2* (*STH2*) which controls de-etiolation and regulates shade avoidance in Arabidopsis (Datta et al., 2007; Crocco et al., 2010), and may response to abiotic stresses in soybean.

R genes have conserved domains like TIR-NBS-LRR, CC-NBS-LRR (Hulbert et al., 2001), LRR (Ellis et al., 2000), TIR (Bernoux et al., 2011), and NB-ARC (van der Biezen and Jones, 1998). A total of 27 genes, such as *Glyma15 g17310*, had one of the above five domains (Supplemental Table S8). Three of the five R-related genes have been reported previously. The *Glyma20 g12720* is a R-gene with copy number variation between seven *G. soja* and *G. max* Williams 82 (Li et al., 2014), the *Glyma10 g37230* and *Glyma10 g37250* were deleted in *G. soja* but present in *G. max* (Joshi et al., 2013), and were homologous to *At2 g34930*. The loss-of-function mutation for *At2 g34930* showed increased susceptibility to the powdery mildew pathogen *Erysiphe cichoracearum* in Arabidopsis (Ramonell et al., 2005).

Oil Biosynthesis

Using GO enrichment analysis, five enriched terms with 63 candidate genes in the biological process were found to be associated with oil synthesis; these terms included GO:0042304 (regulation of fatty acid biosynthetic process), GO:0001676 (long-chain fatty acid metabolic process), GO:0006631 (fatty acid metabolic process), GO:0019682 (glyceraldehyde-3-phosphate metabolic process), and GO:0006090 (pyruvate metabolic process; Supplemental Table S3). Among the 1483 CDGs identified in this study, 20 were common to the oil-related genes described by Schmutz et al. (2010). These 20 genes belonged to seven membrane lipid classes, including the synthesis of plastid fatty acids, plastid membrane lipids, and endomembrane system membrane lipids, metabolism of mitochondria acyl lipids, degradation of storage lipids and straight fatty acids, and lipid signaling (Supplemental Table S9). There were eight common genes between the above 63 and 20 genes. Among all the 75 genes, 48 were highly expressed in oil biosynthesis processes (Supplemental Table S9).

In Arabidopsis, *CDS3* (*At4 g26770*; Hu et al., 2012), *KASI* (*At5 g46290*; Moche et al., 2001) and *LACS9* (*At1 g77590*; Schnurr et al., 2002) were associated with seed oil synthesis. The gene *Glyma08 g08910* was homologous to *KASI*, and *Glyma13 g03280* was homologous to *LACS9*. The two soybean genes were highly expressed in the process of oil biosynthesis, and may regulate seed oil biosynthesis in soybean. Although *Glyma14 g02210* and *Glyma05 g36690* were homologous to *CDS3* and *KASI*, respectively, they had low expression.

In summary, a total of 330 (54 + 112 + 196 + 48 – 80 common) CDGs that might be related to DTs was identified in this study.

Association Studies of Domesticated and Improved Traits with Candidate Genes

All the SNPs within or adjacent to 330 CDGs were selected for the association studies with DTs (flowering times [first and full] and SS [SL, SW, ST, and 100SW]) and improvement traits (alkaline-salt tolerances [LR and LH] and SOC) in this study (Supplemental Table S10). As a result, 18 of the 54 flowering time CDGs, 60 of the 112 seed development CDGs, 66 of the 196 biotic and abiotic tolerance or resistance CDGs, and 10 of the 48 oil biosynthesis CDGs were found to be associated with the corresponding traits. Among these trait-associated CDGs, 17 were common across various traits, for example, *Glyma05 g06070* was associated with tolerance and SS, *Glyma20 g10240* with SOC and SS, *Glyma15 g17480* with flowering time and tolerance, and *Glyma14 g16040* with flowering time, SS, and tolerance (Supplemental Table S11). Therefore, a total of 134 trait-associated CDGs were identified.

Common CDGs for Domesticated and Improved Traits

Of the 134 trait-associated CDGs, 29 overlapped with previous CDGs, 11 were exactly the same as candidate genes in previous trait association studies, and 66 were covered by previous domesticated and improved quantitative trait loci or their adjacent regions (Table 2 and Supplemental Table S12). Among the 29, 11, and 66 CDGs, six were common (Table 2), including the SOC gene *Glyma20 g10240* (Zhou et al., 2015a), the SS gene *Glyma08 g09310* (Bolon et al., 2014), the alkaline-salt tolerance (AST) gene *Glyma05 g06070* (Mochida et al., 2010), and the FT genes *Glyma14 g16040* and *Glyma15 g17480* (Jung et al., 2012; Kim et al., 2012b; Xue et al., 2012; Chung et al., 2014; Zhou et al., 2015b) (Table 2). More importantly, *Glyma15 g17480* (*GmZTL3*) has been functionally characterized in soybean (Xue et al., 2012).

Of the above 29 common CDGs, 15 were found to be associated with the above domesticated and improved traits. For example, *Glyma14 g02210* (Joshi et al., 2013) and *Glyma20 g10240* (Zhou et al., 2015a) with SOC, *Glyma02 g41700* (Li et al., 2013), *Glyma02 g41710* (Li et al., 2013; Zhao et al., 2015), *Glyma07 g39770*, *Glyma20 g10600* (Zhou et al., 2015a), *Glyma13 g10260* (Chung et al., 2014), *Glyma20 g12720* (Li et al., 2014) and *Glyma19 g39960* (Han et al., 2015) with AST, *Glyma08 g05550* (Li et al., 2013), *Glyma07 g34530*, *Glyma09 g03490*, *Glyma20 g10360* and *Glyma20 g10380* (Zhou et al., 2015a) with SS, and *Glyma14 g16040* (Chung et al., 2014) with FT, SS, and AST, although some traits associated are different from those previously reported.

Differential Expression Analysis

With differential expression analysis, the RAN-seq data were used to identify DEGs between wild and cultivated soybeans in this study. Among 56,044 mapped soybean genes, a total of 22,659 and 18,024 DEGs were differentially expressed between No. 101 and 265 (DE101–265), and between No. 101 and 272 (DE101–272), respectively, at the 0.001 significance level (Supplemental Table S11). These

Table 2. Comparison of candidate domestication genes (DG) detected in this study with domestication quantitative trait loci (QTL) or genes in previous studies.

DG associated with trait in this study		Trait associated with the gene in previous studies		References with consistent DG/region and related-trait in this study		QTL near DG in previous reports			
Gene	Trait associated†	Trait associated	Reference	Domestication detected	Traits	Marker associated	Position (bp)	Trait	Reference
Glyma05g06070	FT, SS, AST	Stress-related	Mochida et al., 2010	Chung et al., 2014, Zhou et al., 2015a		Satt454-Satt572	Gm05:25287781–3403500	Drought tolerance	Carpentieri-Pipolo et al., 2012
						Satt276	Gm05:3442439–3442495	Seed wt. per plant	Chen et al., 2007
Glyma08g05550	SS			Li et al., 2013	seed coat color	Satt207-Satt315	Gm08:4000325–6751725	Seed length	Salas et al., 2006
						Satt424-Satt390	Gm08:1306479–10721881	Seed wt.	Han et al., 2012
Glyma08g09310	SS	Seed-specific gene	Bolon et al., 2014	Chung et al., 2014		Sat_215-Sat_409	Gm08:5715407–9211886	Pod no.	Zhang et al., 2010
						Satt207-Satt315	Gm08:4000325–6751725	Seed length	Salas et al., 2006
						Satt424-Satt390	Gm08:1306479–10721881	Seed wt.	Han et al., 2012
						Sat_181	Gm08:5770755–5770812	Seed wt.	Han et al., 2012
Glyma14g16040	FT, SS, AST			Chung et al., 2014; Zhao et al., 2015	flower development	Satt304	Gm14:13284499–13284588	Seed wt.	Hoeck et al., 2003
Glyma15g17480	FT, AST	FT gene	Jung et al., 2012; Kim et al., 2012b; Xue et al., 2012; Zhou et al., 2015b	Kim et al., 2012b		Sat_124-Satt598	Gm15:11099721–13638395	Flower no.	Zhang et al., 2010
Glyma20g10240	SS, AST, SOC	Oil-related	Zhou et al., 2015a	Kim et al., 2012a; Zhou et al., 2015a		Satt127-Satt239	Gm20:12169135–24129775	Seed oil	Tajuddin et al., 2003

† AST, alkaline-salt tolerance; FT, flowering time; SOC, seed oil content; SS, seed size.

DEGs were compared with the above trait-associated CDGs for SS and SOC. As a result, 36 of the 60 seed development genes and 2 of the 10 SOC-related genes were found to be differentially expressed both in DE101–265 and DE101–272, having one common gene *Glyma20 g10240*. Thus, 37 trait-associated CDGs were further confirmed (Table 3).

Candidate Artificial Selection Genes (CASGs) in Soybean

We detected the selective sweeps of modern breeding practice by calculating ROD = $1 - \pi_{bred}/\pi_{landrace}$ (Xu et al., 2012). Although no genomic regions with significant difference were found at the *ROD* critical value of 0.98, 47 candidate artificial selection genes in nine regions were identified at the value of 0.90, and were close to QTL responsive for important agronomic traits such as seed oil, flood tolerance, drought index, first flower, and plant height (Supplemental Table S13).

Among these candidate genes, *Glyma02 g29830* was homologous to flowering time-related gene *FBH* in rice and Arabidopsis (Ito et al., 2012a), *Glyma02 g30885* was homologous to *LOC_Os11 g31450* strongly expressed in a susceptible cultivar for rice stripe virus (Kwon et al., 2012), and *Glyma20 g08560* was homologous to *Medtr6 g086560* regulating drought in *Medicago* (Zhang et al., 2014; Supplemental Table S13).

Enrichment analysis identified 39 significantly over-represented GO categories (Supplemental Fig. S3, Supplemental Table S14). Among these GO categories, three, one, three, and two significantly enriched terms ($P < 0.05$) for the biological processes were related to SOC (four genes), flowering time (six genes), tolerance (five genes), and seed development (four genes), respectively (Supplemental Table S13). Further association analysis showed that two of the four oil biosynthesis candidate genes, six of the six flowering time candidate genes, three of the five tolerance candidate genes, and four of the four seed development candidate genes were associated with the corresponding trait. Interestingly, the *Glyma20 g05170*, *Glyma20 g05360*, and *Glyma20 g05420* genes were associated with all the above four traits (Supplemental Table S14). Therefore, eight candidate artificial selection genes were confirmed by association studies.

Table 3. Validation of candidate domestication genes for two domestication traits in soybean using association studies and RNA-seq (sequenced) differential expression analyses.

		Association studies of candidate genes				RNA-seq expression analysis between Accessions 101 and 265		RNA-seq expression analysis between Accessions 101 and 272		Homologs in Arabidopsis	
Gene	Chr	No. of SNPs	Trait†	*P*-value	r^2	log_2Fold	*P*-value	log_2Fold	*P*-value	Gene	Reference
					%						
Glyma01g05820	1	1	SS	1.33E-03 to 8.80E-03	3.71 to 4.23	2.29	1.22E-03	2.35	5.48E-04	*At5G39785*	
Glyma02g40310	2	1	SS	5.84E-05 to 2.92E-03	3.79 to 6.67	1.89	1.08E-15	1.16	4.52E-08	*EMB1401*	
Glyma02g40320	2	1	SS	2.43E-03	4.04	2.73	5.94E-16	1.54	3.99E-08	*CRF4*	
Glyma03g41460	3	1	SS	2.94E-03	3.91	−1.39	3.93E-30	−1	2.05E-14	*PP2–B1*	Pagnussat et al., 2005
Glyma04g01130	4	2	SS	1.34E-03 to 9.68E-03	4.63 to 5.22	2.08	4.66E-192	1.36	8.56E-101	*COR47*	
Glyma04g01140	4	1	SS	1.72E-05	8.4	1.78	4.51E-07	1.43	2.03E-05	*SHL1*	
Glyma05g06070	5	2	SS	2.46E-03 to 9.68E-03	3.77 to 5.72	2.82	5.98E-11	1.49	3.05E-05	*APRR2*	
Glyma05g17900	5	10	SS	9.26E-07 to 8.22E-03	3.91 to 13.05	1.73	2.60E-10	0.86	3.34E-04	*SCE1*	Meinke et al., 2008; Saracco et al., 2007
Glyma06g34850	6	7	SS	1.81E-05 to 7.17E-03	2.99 to 7.71	1.73	8.88E-15	1.39	9.60E-11	*EBS*	
Glyma07g02150	7	1	SS	1.92E-03 to 5.43E-03	3.78 to 4.33	2.71	1.64E-05	1.9	5.79E-04	*At1G52190*	Almagro et al., 2008
Glyma08g08910	8	3	SS	1.13E-03 to 5.70E-03	3.48 to 5.98	1.94	2.21E-32	1.45	1.80E-21	*KAS I*	Hakozaki et al., 2008
Glyma08g09310	8	1	SS	2.27E-04 to 7.69E-03	3.72 to 6.45	1.71	4.42E-149	1.99	1.46E-177	*AtPER1*	Haslekas et al., 1998
Glyma08g28320	8	2	SS	1.22E-03 to 5.15E-03	3.25 to 4.41	2.35	5.59E-09	2.07	1.45E-07	*At3G15460*	
Glyma08g45380	8	3	SS	5.90E-05 to 9.29E-03	3.85 to 6.51	2.51	4.53E-11	1.53	4.37E-06	*ADL1E*	Kang et al., 2003
Glyma09g00310	9	3	SS	1.73E-04 to 8.88E-03	2.87 to 6.80	2.18	3.61E-15	1.05	4.94E-06	*emb2444*	Meinke et al., 2008
Glyma09g03490	9	1	SS	5.48E-03	3.5	2.52	4.35E-21	1.82	6.61E-14	*HSR8*	Pagnussat et al., 2005
Glyma09g10560	9	1	SS	2.50E-04 to 4.91E-03	7.94 to 9.72	2.02	8.00E-10	1.34	1.09E-05	*APO2*	Meinke et al., 2008
Glyma09g11740	9	4	SS	3.99E-04 to 9.93E-03	3.31 to 5.11	1.53	3.59E-08	2.01	3.40E-11	*At2G34620*	Errampalli et al., 1991
Glyma09g12420	9	3	SS	6.98E-03	4.92	1.44	9.07E-07	1.38	2.69E-06	*RPS11–BETA*	Meinke et al., 2008
Glyma10g36720	10	4	SS	1.77E-03 to 9.17E-03	3.75 to 5.01	2.25	3.31E-04	1.57	5.60E-03	*ASHH1, SDG26*	Michaels and Amasino, 1999; Hepworth et al., 2002; He and Amasino, 2005; Xu et al., 2008
Glyma12g26980	12	4	SOC	3.81E-03 to 8.87E-03	2.08 to 4.81	1.51	5.76E-06	1.41	3.81E-05	*At5G53940*	
Glyma13g36370	13	1	SS	6.85E-04 to 9.85E-03	3.84 to 5.87	3.76	2.37E-10	2.51	6.83E-07	*AtMYB106*	
Glyma14g16040	14	17	SS	2.98E-06 to 9.44E-03	3.52 to 6.43	2.28	5.08E-09	1.71	2.32E-06	*LUG*	
Glyma14g36070	14	3	SS	8.44E-03	3.25	3.18	8.25E-48	2.98	3.06E-43	*ATR1*	
Glyma15g19450	15	1	SS	4.28E-06 to 1.35E-03	5.48 to 8.44	2.17	3.29E-07	1.93	6.15E-06	*ELF8*	
Glyma15g19460	15	1	SS	6.64E-03	5.07	1.93	2.91E-05	1.58	4.48E-04	*AtGRF5*	
Glyma15g23780	15	8	SS	1.09E-04 to 9.46E-03	3.20 to 7.58	1.95	9.03E-06	1.63	1.10E-04	*At4G39630*	
Glyma17g18800	17	3	SS	3.05E-04 to 9.49E-03	3.19 to 5.51	2.17	1.84E-31	1.92	2.96E-26	*CCoAOMT1*	Fellenberg et al., 2008
Glyma18g07830	18	1	SS	6.13E-03	2.5	2.04	2.98E-06	1.67	4.23E-05	*At5G67630*	
Glyma18g12100	18	3	SS	9.54E-04 to 7.02E-03	4.05 to 6.70	1.42	3.35E-05	1.1	9.07E-04	*DDL*	
Glyma18g13250	18	2	SS	2.21E-03 to 3.84E-03	3.71 to 5.97	2.31	1.35E-06	1.43	6.89E-04	*VCL1*	
Glyma18g29560	18	1	SS	1.45E-03 to 1.95E-03	4.05 to 6.02	2.61	2.16E-04	1.8	6.24E-03	*KAC1*	
Glyma20g09810	20	1	SS	5.81E-03	3.39	2.13	3.04E-08	1.41	7.02E-05	*AtPPC3*	
Glyma20g10240	20	10	SS, SOC	4.28E-06 to 6.82E-03	2.78 to 8.48	1.84	5.66E-04	1.52	2.37E-03	*At5G42250*	Lee et al., 2008
Glyma20g10360	20	3	SS	1.18E-03 to 8.86E-03	3.88 to 5.46	2.76	3.93E-11	1.86	4.58E-07	*At4G12640*	
Glyma20g10380	20	14	SS	4.56E-05 to 9.43E-03	3.40 to 7.79	1.87	1.02E-06	1.47	5.48E-05	*At4G12640*	
Glyma20g22230	20	8	SS	5.28E-05 to 9.61E-03	3.18 to 7.87	2.98	3.28E-13	2.5	9.91E-11	*AtMYB61*	Penfield et al., 2001

† SOC, seed oil content; SS, seed size.

Conclusion

A total of 106,013 SNPs were identified by RAD-sequencing of 286 soybean accessions. The analyses of these SNPs revealed a high degree of differentiation between wild and cultivated soybeans; for example, the number of rare SNPs reduced gradually from wild to bred soybeans during domestication and genetic diversity was twofold larger in wild soybean than in cultivated soybean. A total of 330 CDGs were identified by the genetic diversity, domestication, GO enrichment, and gene expression analyses, and were homologs controlling flowering time, SS, and shattering, tolerance or resistance, and oil biosynthesis in Arabidopsis. Of 134 trait-associated CDGs, 29 and 11 were consistent with previous CDGs

and candidate genes in previous trait association studies, respectively, and 66 were close to previously identified domesticated and improved QTL, having six common CDGs, such as one functionally characterized gene *Glyma15 g17480* (*GmZTL3*). Of the 68 SS and SOC CDGs, 37 were further confirmed by gene expression analysis. In addition, eight genes were found to be related to artificial selection during modern breeding. This study provides an integrated approach for efficiently identifying CDGs and valuable information for evolution, molecular biology, and breeding in soybean.

Supplemental Information Available

Supplemental information is included with this article.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (31571268), Fundamental Research Funds for the Central Universities (KYT201002), and Huazhong Agricultural University Scientific and Technological Self-innovation Foundation (2014RC020).

References

Alejandro, S., Y. Lee, T. Tohge, D. Sudre, S. Osorio, J. Park, et al. 2012. AtABCG29 is a monolignol transporter involved in lignin biosynthesis. Curr. Biol. 22:1207–1212. doi:10.1016/j.cub.2012.04.064

Almagro, A., S.H. Lin, and Y.F. Tsay. 2008. Characterization of the *Arabidopsis* nitrate transporter *NRT1.6* reveals a role of nitrate in early embryo development. Plant Cell 20:3289–3299. doi:10.1105/tpc.107.056788

Bailey, M.A., M.A.R. Mian, T.E. Carter, D.A. Ashley, and H.R. Boerma. 1997. Pod dehiscence of soybean: Identification of quantitative trait loci. J. Hered. 88:152–154. doi:10.1093/oxfordjournals.jhered.a023075

Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376. doi:10.1371/journal.pone.0003376

Barecki, I.B., and B.K. Suarez. 2001. Linkage and association: Basic concepts. Adv. Genet. 42:45–66. doi:10.1016/S0065-2660(01)42014-1

Barnes, M.R., and P.S. Derwent. 2007. Needle in a haystack? Dealing with 500,000 SNP genome scans. In: M.R. Barnes, editor, Bioinformatics for geneticists. A bioinformatics primer for the analysis of genetic data. 2nd ed. John Wiley and Sons, New York. p. 447–493. doi:10.1002/9780470059180.ch18

Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265. doi:10.1093/bioinformatics/bth457

Baydar, N.G., and M. Akkurt. 2001. Oil content and oil quality properties of some grape seeds. Turk. J. Agric. For. 25:163–168.

Bernoux, M., T. Ve, S. Williams, C. Warren, D. Hatters, E. Valkov, et al. 2011. Structural and functional analysis of a plant resistance protein TIR domain reveals interfaces for self-association, signaling, and autoregulation. Cell Host Microbe 9:200–211. doi:10.1016/j.chom.2011.02.009

Bolon, Y.T., D.L. Hyten, J.H. Orf, C.P. Vance, and G.J. Muehlbauer. 2014. eQTL networks reveal complex genetic architecture in the immature soybean seed. Plant Gen. 7:1–14. doi:10.3835/plantgenome2013.08.0027

Boss, P.K., R.M. Bastow, J.S. Mylne, and C. Dean. 2004. Multiple pathways in the decision to flower: Enabling, promoting, and resetting. Plant Cell 16:S18–S31. doi:10.1105/tpc.015958

Broich, S., and R. Palmer. 1980. A cluster analysis of wild and domesticated soybean phenotypes. Euphytica 29:23–32. doi:10.1007/BF00037246

Bovet, L., T. Eggmann, M. Meylan-Bettex, J. Polier, P. Kammer, E. Marin, et al. 2003. Transcript levels of *AtMRP*s after cadmium treatment: Induction of *AtMRP3*. Plant Cell Environ. 26:371–381. doi:10.1046/j.1365-3040.2003.00968.x

Bovet, L., U. Feller, and E. Martinoia. 2005. Possible involvement of plant ABC transporters in cadmium detoxification: A cDNA sub-microarray approach. Environ. Int. 31:263–267. doi:10.1016/j.envint.2004.10.011

Carpentieri-Pipolo, V., A.E. Pipolo, H. Abdel-Haleem, H.R. Boerma, and T.R. Sinclair. 2012. Identification of QTLs associated with limited leaf hydraulic conductance in soybean. Euphytica 186:679–686. doi:10.1007/s10681-011-0535-6

Carter, T.E., J.R. Nelson, C.H. Sneller, and Z. Cui. 2004. Genetic diversity in soybean. In: H.R. Boerma and J.E. Specht, editors, Soybeans: Improvement, production and uses. ASA, Madison, WI. p. 303–416.

Chen, Q.S., Z.C. Zhang, C.Y. Liu, D.W. Xin, H.M. Qiu, and D.P. Shan. 2007. QTL analysis of major agronomic traits in soybean. Agric. Sci. China 6:399–405. doi:10.1016/S1671-2927(07)60062-5

Chung, W.H., N. Jeong, J. Kim, W.K. Lee, Y.G. Lee, S.H. Lee, et al. 2014. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. 21:153–167. doi:10.1093/dnares/dst047

Concibido, V.C., V.B. La, P. Mclaird, N. Pineda, J. Meyer, L. Hummel, et al. 2003. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. Theor. Appl. Genet. 106:575–582. doi:10.1007/s00122-002-1071-5

Crocco, C.D., M. Holm, M.J. Yanovsky, and J.F. Botto. 2010. *AtBBX21* and *COP1* genetically interact in the regulation of shade avoidance. Plant J. 64:551–562. doi:10.1111/j.1365-313X.2010.04360.x

Datta, S., C. Hettiarachchi, H. Johansson, and M. Holm. 2007. SALT TOLERANCE HOMOLOG2, a B-box protein in *Arabidopsis* that activates transcription and positively regulates light-mediated development. Plant Cell 19:3242–3255. doi:10.1105/tpc.107.054791

Doebley, J.F., B.S. Gaut, and B.D. Smith. 2006. The molecular genetics of crop domestication. Cell 127:1309–1321. doi:10.1016/j.cell.2006.12.006

Dong, Y., X. Yang, J. Liu, B.H. Wang, B.L. Liu, and Y.Z. Wang. 2014. Pod shattering resistance associated with domestication is mediated by a *NAC* gene in soybean. Nat. Commun. 5:3352. doi:10.1038/ncomms4352

Ellis, J., P. Dodds, and T. Pryor. 2000. Structure, function and evolution of plant resistance genes. Curr. Opin. Plant Biol. 3:278–284. doi:10.1016/S1369-5266(00)00080-7

Errampalli, D., D. Patton, L. Castle, L. Mickelson, K. Hansen, J. Schnall, et al. 1991. Embryonic lethals and T-DNA insertional mutagenesis in *Arabidopsis*. Plant Cell 3:149–157. doi:http://dx.doi.org/10.1105/tpc.3.2.149

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14:2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Falcon, S., and R. Gentleman. 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23:257–258. doi:10.1093/bioinformatics/btl567

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Falush, D., M. Stephens, and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol. Ecol. Notes 7:574–578. doi:10.1111/j.1471-8286.2007.01758.x

Fellenberg, C., C. Milkowski, B. Hause, P.R. Lange, C. Böttcher, J. Schmidt, et al. 2008. Tapetum-specific location of a cation-dependent O-methyltransferase in *Arabidopsis thaliana*. Plant J. 56:132–145. doi:10.1111/j.1365-313x.2008.03576.x

Funatsuki, H., M. Ishimoto, H. Tsuji, K. Kawaguchi, M. Hajika, and K. Fujino. 2006. Simple sequence repeat markers linked to a major QTL controlling pod shattering in soybean. Plant Breed. 125:195–197. doi:10.1111/j.1439-0523.2006.01199.x

Garcia, D., V. Saingery, P. Chambrier, U. Mayer, G. Jurgens, and F. Berger. 2003. *Arabidopsis haiku* mutants reveal new controls of seed size by endosperm. Plant Physiol. 131:1661–1670. doi:10.1104/pp.102.018762

Gross, B.L., and K.M. Olsen. 2010. Genetic perspectives on crop domestication. Trends Plant Sci. 15:529–537. doi:10.1016/j.tplants.2010.05.008

Guo, J., Y. Wang, C. Song, J. Zhou, L. Qiu, H. Huang, et al. 2010. A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): Implications from microsatellites and nucleotide sequences. Ann. Bot. 106(3):505–514. doi:10.1093/aob/mcq125

Han, Y., D. Li, D. Zhu, H. Li, X. Li, W. Teng, et al. 2012. QTL analysis of soybean seed weight across multi-genetic backgrounds and environments. Theor. Appl. Genet. 125:671–683. doi:10.1007/s00122-012-1859-x

Han, Y., X. Zhao, D. Liu, Y. Li, D.A. Lightfoot, Z. Yang, et al. 2015. Domestication footprints anchor genomic regions of agronomic importance in soybeans. New Phytol. doi:10.1111/nph.13626.

Hakozaki, H., J.I. Park, M. Endo, Y. Takada, T. Kazama, Y. Takeda, et al. 2008. Expression and developmental function of the 3-ketoacyl-ACP synthase2 gene in *Arabidopsis thaliana*. Genes Genet. Syst. 83:143–152. doi:org/10.1266/ggs.83.143

Haslekas, C., R.A. Stacy, V. Nygaard, F.A. Culiáñez-Macià, and R.B. Aalen. 1998. The expression of a peroxiredoxin antioxidant gene, *AtPer1*, in *Arabidopsis thaliana* is seed-specific and related to dormancy. Plant Mol. Biol. 36:833–845. doi:10.1023/A:1005900832440

He, Y.H., and R.M. Amasino. 2005. Role of chromatin modification in flowering-time control. Trends Plant Sci. 10:30–35. doi:10.1016/j.tplants.2004.11.003

Hepworth, S.R., F. Valverde, D. Ravenseroft, A. Mouradov, and G. Coupland. 2002. Antagonistic regulation of flowering-time gene *SOC1* by *CONSTANS* and *FLC* via separate promoter motifs. EMBO J. 21:4327–4337. doi:10.1093/emboj/cdf432

Hoeck, J.A., W.R. Fehr, R.C. Shoemaker, G.A. Welke, S.L. Johnson, and S.R. Cianzio. 2003. Molecular marker analysis of seed size in soybean. Crop Sci. 43:68–74. doi:10.2135/cropsci2003.6800

Hu, Z., Z. Ren, and C. Lu. 2012. The phosphatidylcholine diacylglycerol cholinephosphotransferase is required for efficient hydroxy fatty acid accumulation in transgenic *Arabidopsis*. Plant Physiol. 158:1944–1954. doi:10.1104/pp.111.192153

Hudson, M., C. Ringli, M.T. Boylan, and P.H. Quail. 1999. The *FAR1* locus encodes a novel nuclear protein specific to phytochrome A signaling. Genes Dev. 13:2017–2027. doi:10.1101/gad.13.15.2017

Hulbert, S.H., C.A. Webb, S.M. Smith, and Q. Sun. 2001. Resistance gene complexes: Evolution and utilization. Annu. Rev. Phytopathol. 39:285–312. doi:10.1146/annurev.phyto.39.1.285

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, R.L. Nelson, J.M. Costa, et al. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA 103:16,666–16,671. doi:10.1073/pnas.0604379103

Iñigo, S., M.J. Alvarez, B. Strasser, A. Califano, and P.D. Cerdán. 2012. PFT1, the MED25 subunit of the plant Mediator complex, promotes flowering through *CONSTANS* dependent and independent mechanisms in *Arabidopsis*. Plant J. 69:601–612. doi:10.1111/j.1365-313X.2011.04815.x

Ito, S., Y. Niwa, N. Nakamichi, H. Kawamura, T. Yamashino, and T. Mizuno. 2008. Insight into missing genetic links between two evening-expressed pseudo-response regulator genes *TOC1* and *PRR5* in the circadian clock-controlled circuitry in *Arabidopsis thaliana*. Plant Cell Physiol. 49:201–213. doi:10.1093/pcp/pcm178

Ito, S., Y.H. Song, and T. Imaizumi. 2012a. LOV domain-containing F-box proteins: Light-dependent protein degradation modules in *Arabidopsis*. Mol. Plant 5:573–582. doi:10.1093/mp/sss013

Ito, S., Y.H. Song, A.R. Josephson-Day, R.J. Miller, G. Breton, R.G. Olmstead, et al. 2012b. FLOWERING BHLH transcriptional activators control expression of the photoperiodic flowering regulator CONSTANS in *Arabidopsis*. Proc. Natl. Acad. Sci. USA 109(9):3582–3587. doi:10.1073/pnas.1118876109

Jones, S.I., and L.O. Vodkin. 2013. Using RNA-seq to profile soybean seed development from fertilization to maturity. PLoS ONE 8:e59270. doi:10.1371/journal.pone.0059270

Joshi, T., B. Valliyodan, J.H. Wu, S.H. Lee, D. Xu, and H.T. Nguyen. 2013. Genomic differences between cultivated soybean, *G. max* and its wild relative *G. soja*. BMC Genomics 14:S5. doi:10.1186/1471-2164-14-S1-S5

Jung, C.H., C.E. Wong, and M.B. Singh. 2012. Comparative genomic analysis of soybean flowering genes. PLoS ONE 7:e38250. doi:10.1371/journal.pone.0038250

Kang, J., J.U. Hwang, M. Lee, Y.Y. Kim, S.M. Assmann, E. Martinoia, et al. 2010. PDR-type ABC transporter mediates cellular uptake of the phytohormone abscisic acid. Proc. Natl. Acad. Sci. USA 107:2355–2360. doi:10.1073/pnas.0909222107

Kang, S.T., M. Kwak, H.K. Kim, M. Suzuki, S. Hagihara, Y. Tanaka, et al. 2009. [*Glycine max* (L.) Merr.] Population-specific QTLs and their different epistatic interactions for pod dehiscence in soybean. Euphytica 166:15–24. doi:10.1007/s10681-008-9810-6

Kang, J., J. Park, H. Choi, B. Burla, T. Kretzschmar, Y. Lee, et al. 2011. Plant ABC transporters. Arabidopsis Book 9:e0153. doi:10.1199/tab.0153

Kang, B.H., J.S. Busse and S.Y. Bednarek. 2003. Members of the *Arabidopsis* dynamin-like gene family, *ADL1*, are essential for plant cytokinesis and polarized cell growth. Plant Cell 15: 899–913. doi:10.1105/tpc.009670

Keim, P., B.W. Diers, T.C. Olson, and R.C. Shoemaker. 1990. RFLP mapping in soybean: Association between marker loci and variation in quantitative traits. Genetics 126:735–742.

Kiba, T., R. Henriques, H. Sakakibara, and N.H. Chua. 2007. Targeted degradation of PSEUDO-RESPONSE REGULATOR5 by an SCF^{ZTL} complex regulates clock function and photomorphogenesis in *Arabidopsis thaliana*. Plant Cell 19:2516–2530. doi:10.1105/tpc.107.053033

Kim, M.Y., S. Lee, K. Van, T.H. Kim, S.C. Jeong, I.Y. Choi, et al. 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. USA 107:22032–22037. doi:10.1073/pnas.1009526107

Kim, Y., J. Lim, M. Yeom, H. Kim, J. Kim, L. Wang, et al. 2013. ELF4 regulates GIGANTEA chromatin access through subnuclear sequestration. Cell Reports 3:671–677. doi:10.1016/j.celrep.2013.02.021

Kim, M.Y., J.H. Shin, Y.J. Kang, S.R. Shim, and S.H. Lee. 2012b. Divergence of flowering genes in soybean. J. Biosci. 37:857–870. doi:10.1007/s12038-012-9252-0

Kim, M.Y., K. Van, Y.J. Kang, K.H. Kim, and S.H. Lee. 2012a. Tracing soybean domestication history: From nucleotide to genome. Breed. Sci. 61:445–452. doi:10.1270/jsbbs.61.445

Klein, M., L. Perfus-Barbeoch, A. Frelet, N. Gaedeke, D. Reinhardt, B. Mueller-Roeber, et al. 2003. The plant multidrug resistance ABC transporter AtMRP5 is involved in guard cell hormonal signalling and water use. Plant J. 33:119–129. doi:10.1046/j.1365-313X.2003.016012.x

Kwon, T., J.H. Lee, S.K. Park, U.H. Hwang, J.H. Cho, D.Y. Kwak, et al. 2012. Fine mapping and identification of candidate rice genes associated with $qSTV11^{SG}$, a major QTL for rice stripe disease resistance. Theor. Appl. Genet. 125(5):1033–1046. doi:10.1007/s00122-012-1893-8

Lam, H.M., X. Xu, X. Liu, W. Chen, G. Yang, F.L. Wong, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42:1053–1059. doi:10.1038/ng.715

Lee, E.K., M. Kwon, J.H. Ko, H. Yi, M.G. Hwang, S. Chang, et al. 2004. Binding of sulfonylurea by AtMRP5, an *Arabidopsis* multidrug resistance-related protein that functions in salt tolerance. Plant Physiol. 134:528–538. doi:10.1104/pp.103.027045

Lee, M., K. Lee, J. Lee, E.W. Noh, and Y. Lee. 2005. AtPDR12 contributes to lead resistance in *Arabidopsis*. Plant Physiol. 138:827–836. doi:10.1104/pp.104.058107

Lee, U., C. Wie, B.O. Fernandez, M. Feelisch, and E. Vierling. 2008. Modulation of nitrosative stress by S-nitrosoglutathione reductase is critical for thermotolerance and plant growth in *Arabidopsis*. Plant Cell 20:786–802. doi:10.1105/tpc.107.052647

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Home, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Li, D.D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci. 48:571–581. doi:10.2135/cropsci2007.06.0361

Li, Y.H., S.C. Zhao, J.X. Ma, D. Li, L. Yan, J. Li, et al. 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14:579. doi:10.1186/1471-2164-14-579

Li, Y.H., G. Zhou, J. Ma, W. Jiang, L.G. Jin, Z. Zhang, et al. 2014. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32:1045–1052. doi:10.1038/nbt.2979

Liljegren, S.J., G.S. Ditta, Y. Eshed, B. Savidge, J.L. Bowman, and M.F. Yanofsky. 2000. *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. Nature 404:766–770. doi:10.1038/35008089

Liu, B., T. Fujita, Z.H. Yan, S. Sakamoto, D. Xu, and J. Abe. 2007. QTL mapping of domestication-related traits in soybean (*Glycine max*). Ann. Bot. 100(5):1027–1038. doi:10.1093/aob/mcm149

Liu, B., A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, and J. Abe. 2008. Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics 180:995–1007. doi:10.1534/genetics.108.092742

Liu, W.X., M.Y. Kim, K. Van, Y.H. Lee, H.L. Li, X.H. Liu, et al. 2011. QTL identification of yield-related traits and their association with flowering and maturity in soybean. J. Crop Sci. Biotech. 14:65–70. doi:10.1007/s12892-010-0115-7

Liu, B., S. Watanabe, T. Uchiyama, F. Kong, A. Kanazawa, Z. Xia, et al. 2010. The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis TERMINAL FLOWER1*. Plant Physiol. 153:198–210. doi:10.1104/pp.109.150607

Lu, S.X., C.J. Webb, S.M. Knowles, S.H. Kim, Z. Wang, and E.M. Tobin. 2012. CCA1 and ELF3 interact in the control of hypocotyl length and flowering time in *Arabidopsis*. Plant Physiol. 158:1079–1088. doi:10.1104/pp.111.189670

Luo, M., E.S. Dennis, F. Berger, W.J. Peacock, and A. Chaudhury. 2005. *MINISEED3* (*MINI3*), a *WRKY* family gene, and *HAIKU2* (*IKU2*), a leucine-rich repeat (*LRR*) *KINASE* gene, are regulators of seed size in *Arabidopsis*. Proc. Natl. Acad. Sci. USA 102:17,531–17,536. doi:10.1073/pnas.0508418102

Martínez-Andújar, C., R.C. Martin, and H. Nonogaki. 2012. Seed traits and genes important for translational biology—Highlights from recent discoveries. Plant Cell Physiol. 53:5–15. doi:10.1093/pcp/pcr112

McCarthy, F.M., N. Wang, G.B. Magee, B. Nanduri, M.L. Lawrence, E.B. Camon, et al. 2006. AgBase: A functional genomics resource for agriculture. BMC Genomics 7:229. doi:10.1186/1471-2164-7-229

Meinke, D., R. Muralla, C. Sweeney, and A. Dickerman. 2008. Identifying essential genes in *Arabidopsis thaliana*. Trends Plant Sci.13:483–491. doi:10.1016/j.tplants.2008.06.003

Michaels, S.D., and R.M. Amasino. 1999. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11:949–956. doi:10.1105/tpc.11.5.949

Moche, M., K. Dehesh, P. Edwards, and Y. Lindqvist. 2001. The crystal structure of β-ketoacyl-acyl carrier protein synthase II from *Synechocystis sp.* at 1.54 Å resolution and its relationship to other condensing enzymes. J. Mol. Biol. 305:491–503. doi:10.1006/jmbi.2000.4272

Mochida, K., T. Yoshida, T. Sakurai, K. Yamaguchi-Shinozaki, K. Shinozaki, and L.S. Tran. 2010. Genome-wide analysis of two-component systems and prediction of stress-responsive two-component system members in soybean. DNA Res. 17:303–324. doi:10.1093/dnares/dsq021

Okushima, Y., I. Mitina, H.L. Quach, and A. Theologis. 2005. AUXIN RESPONSE FACTOR 2 (ARF2): A pleiotropic developmental regulator. Plant J. 43:29–46. doi:10.1111/j.1365-313X.2005.02426.x

Olsen, K.M., and J.F. Wendel. 2013. Crop plants as models for understanding plant adaptation and diversification. Front. Plant Sci. 4:290. doi:10.3389/fpls.2013.00290

Pagnussat, G.C., H.J. Yu, Q.A. Ngo, S. Rajani, S. Mayalagu, C.S. Johnson, et al. 2005. Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. Development 132:603–614. doi:10.1242/dev.01595

Patterson, N., A.L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190. doi:10.1371/journal.pgen.0020190

Pedersen, P. 2009. Soybean growth and development. Extension Publ. PM1945. Iowa State Univ., Ames.

Penfield, S., R.C. Meissner, D.A. Shoue, N.C. Carpita, and M.W. Bevan. 2001. *MYB61* is required for mucilage deposition and extrusion in the *Arabidopsis* seed coat. Plant Cell 13:2777–2791. doi:10.1105/tpc.010265

Pinyopich, A., G.S. Ditta, B. Savidge, S.J. Liljegren, E. Baumann, E. Wisman, et al. 2003. Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature 424:85–88. doi:10.1038/nature01741

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Putterill, J., R. Laurie, and R. Macknight. 2004. It's time to flower: The genetic control of flowering time. BioEssays 26:363–373. doi:10.1002/bies.20021

Putterill, J., F. Robson, K. Lee, R. Simon, and G. Coupland. 1995. The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. Cell 80:847–857. doi:10.1016/0092-8674(95)90288-0

Ramonell, K., M. Berrocal-Lobo, S. Koh, J. Wan, H. Edwards, G. Stacey, et al. 2005. Loss-of-function mutations in chitin responsive genes show increased susceptibility to the powdery mildew pathogen *Erysiphe cichoracearum*. Plant Physiol. 138:1027–1036. doi:10.1104/pp.105.060947

Retief, J.D. 2000. Phylogenetic analysis using PHYLIP. Methods Mol. Biol. 132:243–258.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Salas, P., J.C. Oyarzo-Llaipen, D. Wang, K. Chase, and L. Mansur. 2006. Genetic mapping of seed shape in three populations of recombinant inbred lines of soybean (*Glycine max* L. Merr.). Theor. Appl. Genet. 113:1459–1466. doi:10.1007/s00122-006-0392-1

Saracco, S.A., M.J. Miller, J. Kurepa, and R.D. Vierstra. 2007. Genetic analysis of SUMOylation in *Arabidopsis*: conjugation of *SUMO1* and *SUMO2* to nuclear proteins is essential. Plant Physiol. 145:119–134. doi:10.1104/pp.107.102285

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi:10.1038/nature08670

Schnurr, J.A., J.M. Shockey, G.J. de Boer, and J.A. Browse. 2002. Fatty acid export from the chloroplast. Molecular characterization of a major plastidial acyl-coenzyme A synthetase from *Arabidopsis*. Plant Physiol. 129:1700–1709. doi:10.1104/pp.003251

Sibout, R., and H. Höfte. 2012. Plant cell biology: The ABC of monolignol transport. Curr. Biol. 22:R533–R535. doi:10.1016/j.cub.2012.05.005

Sobhanian, H., R. Razavizadeh, Y. Nanjo, A.A. Ehsanpour, F.R. Jazii, N. Motamed, et al. 2010. Proteome analysis of soybean leaves, hypocotyls and roots under salt stress. Proteome Sci. 8:19. doi:10.1186/1477-5956-8-19

Somers, D.E., W.Y. Kim, and R. Geng. 2004. The F-box protein ZEITLUPE confers dosage-dependent control on the circadian clock, photomorphogenesis, and flowering time. Plant Cell 16:769–782. doi:10.1105/tpc.016808

Song, Q., D.L. Hyten, G.F. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2013. Development and evaluation of soySNP50k, a high-density genotyping array for soybean. PLoS ONE 8:E54985. doi:10.1371/journal.pone.0054985

Song, Y.H., R.W. Smith, B.J. To, A.J. Millar, and T. Imaizumi. 2012. FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. Science 336:1045–1049. doi:10.1126/science.1219644

Sugano, S., C. Andronis, R.M. Green, Z.Y. Wang, and E.M. Tobin. 1998. Protein kinase CK2 interacts with and phosphorylates the *Arabidopsis* circadian clock-associated 1 protein. Proc. Natl. Acad. Sci. USA 95:11,020–11,025. doi:10.1073/pnas.95.18.11020

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tajuddin, T., S. Watanabe, N. Yamanaka, and K. Harada. 2003. Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. Breed. Sci. 53:133–140. doi:10.1270/jsbbs.53.133

Tang, H., U. Sezen, and A.H. Paterson. 2010. Domestication and plant genome. Curr. Opin. Plant Biol. 13:160–166. doi:10.1016/j.pbi.2009.10.008

Tian, Z., X. Wang, R. Lee, Y. Li, J.E. Specht, R.L. Nelson, et al. 2010. Artificial selection for determinate growth habit in soybean. Proc. Natl. Acad. Sci. USA 107:8563–8568. doi:10.1073/pnas.1000088107

Trapnell, C., L. Pachter, and S.L. Salzberg. 2009. TopHat: Discovering splice junctions with RNA-seq. Bioinformatics 25:1105–1111. doi:10.1093/bioinformatics/btp120

Tzafrir, I., A. Dickerman, O. Brazhnik, Q. Nguyen, J. McElver, C. Frye, et al. 2003. The *Arabidopsis* seed genes project. Nucleic Acids Res. 31:90–93. doi:10.1093/nar/gkg028

Van, K., M.Y. Kim, J.H. Shin, K.D. Kim, Y.H. Lee, and S.H. Lee. 2014. Molecular evidence for soybean domestication. In: R. Tuberosa, A. Graner, and E. Frison, editors, Genomics of plant genetic resources. Vol. 1, Managing, sequencing and mining genetic resources. p. 465–481.

van der Biezen, E.A., and J.D.G. Jones. 1998. The NB-ARC domain: A novel signaling motif shared by plant resistance gene products and regulators of cell death in animals. Curr. Biol. 8:R226–R227. doi:10.1016/S0960-9822(98)70145-9

Varshney, R.K., C. Song, R.K. Saxena, S. Azam, S. Yu, A.G. Sharpe, et al. 2013. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat. Biotechnol. 31:240–246. doi:10.1038/nbt.2491

Wang, L.K., Z.X. Feng, X. Wang, X. Wang, and X. Zhang. 2010. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26:136–138. doi:10.1093/bioinformatics/btp612

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, et al. 2009. Map-based cloning of the gene associated with the soybean maturity locus *E3*. Genetics 182:1251–1262. doi:10.1534/genetics.108.098772

Watanabe, S., Z. Xia, R. Hideshima, Y. Tsubokura, S. Sato, N. Yamanaka, et al. 2011. A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering. Genetics 188:395–407. doi:10.1534/genetics.110.125062

Xi, J., P. Xu, and C.B. Xiang. 2012. Loss of AtPDR11, a plasma membrane-localized ABC transporter, confers paraquat tolerance in *Arabidopsis thaliana*. Plant J. 69:782–791. doi:10.1111/j.1365-313X.2011.04830.x

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. USA 109:2155–2164. doi:10.1073/pnas.1117982109

Xu, Y. 2010. Molecular plant breeding. CABI, Wallingford, UK.

Xu, X., X. Liu, S. Ge, J.D. Jensen, F. Hu, X. Li, et al. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. 30:105–111. doi:10.1038/nbt.2050

Xu, L., Z. Zhao, A. Dong, L. Soubigou-Taconnat, J.P. Renou, A. Steinmetz, et al. 2008. Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in *Arabidopsis thaliana*. Mol. Cell. Biol. 28:1348–1360. doi:10.1128/MCB.01607-07

Xue, Z.G., X.M. Zhang, C.F. Lei, X.J. Chen, and Y.F. Fu. 2012. Molecular cloning and functional analysis of one *ZEITLUPE* homolog *GmZTL3* in soybean. Mol. Biol. Rep. 39:1411–1418. doi:10.1007/s11033-011-0875-2

Yanovsky, M.J., and S.A. Kay. 2002. Molecular basis of seasonal time measurement in *Arabidopsis*. Nature 419:308–312. doi:10.1038/nature00996

Zhang, D., H. Cheng, H. Wang, H. Zhang, C. Liu, and D. Yu. 2010. Identification of genomic regions determining flower and pod numbers development in soybean (*Glycine max* L). J. Genet. Genom. 37:545–556. doi:10.1016/S1673-8527(09)60074-6

Zhang, J.Y., M.H.C. de Carvalho, I. Torres-Jerez, Y. Kang, S.N. Allen, D.V. Huhman, et al. 2014. Global reprogramming of transcription and metabolism in *Medicago truncatula* during progressive drought and after rewatering. Plant Cell Environ. 37:2553–2576. doi:10.1111/pce.12328

Zhao, S., F. Zheng, W. He, H. Wu, S. Pan, and H.M. Lam. 2015. Impacts of nucleotide fixation during soybean domestication and improvement. BMC Plant Biol. 15:81. doi:10.1186/s12870-015-0463-z

Zhou, Z., Y. Jiang, Z. Wang, Z.H. Gou, J. Lyu, W.Y. Li, et al. 2015a. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33:408–414. doi:10.1038/nbt.3096

Zhou, L., S.B. Wang, J. Jian, Q.C. Geng, J. Wen, Q.J. Song, et al. 2015b. Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. Sci. Rep. 5:9350. doi:10.1038/srep09350

Zientara, K., A. Wawrzyńska, J. Lukomska, J.R. López-Moya, F. Liszewska, A.G. Assunção, et al. 2009. Activity of the *AtMRP3* promoter in transgenic *Arabidopsis thaliana* and *Nicotiana tabacum* plants is increased by cadmium, nickel, arsenic, cobalt and lead but not by zinc and iron. J. Biotechnol. 139:258–263. doi:10.1016/j.jbiotec.2008.12.001