Research review

Soybean domestication: the origin, genetic architecture and molecular bases

Author for correspondence:
Yoshie Hanzawa
Tel: +1 217 333 4685
Email: yhanzawa@illinois.edu

Received: *16 July 2016*
Accepted: *28 November 2016*

Eric J. Sedivy*, Faqiang Wu* and Yoshie Hanzawa

Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

New Phytologist (2017) **214:** 539–553
doi: 10.1111/nph.14418

Key words: artificial selection, convergent evolution, domestication, genome resequencing, molecular evolution, soybean.

Summary

Domestication provides an important model for the study of evolution, and information learned from domestication research aids in the continued improvement of crop species. Recent progress in *de novo* assembly and whole-genome resequencing of wild and cultivated soybean genomes, in addition to new archeological discoveries, sheds light on the origin of this important crop and provides a clearer view on the modes of artificial selection that drove soybean domestication and diversification. This novel genomic information enables the search for polymorphisms that underlie variation in agronomic traits and highlights genes that exhibit a signature of selection, leading to the identification of a number of candidate genes that may have played important roles in soybean domestication, diversification and improvement. These discoveries provide a novel point of comparison on the evolutionary bases of important agronomic traits among different crop species.

The origin of soybean domestication

Soybean (*Glycine max* (L.) Merr.) is an important legume crop that is a leading source of dietary protein and oil in animal feed, as well as a staple for human consumption (Hartman *et al.*, 2011). It is widely believed that modern cultivated soybean was domesticated from wild soybean (*Glycine soja* Sieb. & Zucc.) in East Asia 6000–9000 yr ago (Carter *et al.*, 2004; Kim *et al.*, 2012b). The origin of soybean domestication has been mysterious partly due to a lack of molecular-based studies and archeological information. However, recent progress in whole-genome sequencing of cultivated and wild soybeans as well as new archeological discoveries have shed light on the history of this important crop.

Despite historical evidence suggesting that soybean was introduced from north-eastern China *c.* 2510 BP, leading to the agricultural revolution in the Eastern Zhou Dynasty (Ho, 1975), soybean landraces with the highest genetic diversity are found in the Huanghe region around the Huanghe River (Yellow River) (Dong *et al.*, 2004; Li *et al.*, 2010). Centered around this region, an abundance of archeological, charred soybean specimens (Zhao, 2004; Lee *et al.*, 2011) place the Yellow River basin as a prime candidate for the origin of soybean domestication. Alternatively, the Yangtze basin (Southern China) has also been proposed as a birthplace of soybean based on phylogenetic and clustering analyses using microsatellites and nucleotide diversity (Guo *et al.*, 2010). Supporting the Southern origin of soybean, a previous study of polymorphisms using chloroplast and mitochondrial DNA identifies the Yangtze region as the most genetically diverse (Shimamoto *et al.*, 2000). However, there is currently no archeological evidence that supports Southern China as the origin of soybean domestication (Lee *et al.*, 2011). The long-standing intense debate concerning the origin of soybean appears to have come to a settlement at last. Han *et al.* (2016) sequenced > 50 000 targeted genomic regions using 404 accessions of *G. max* and 72 *G. soja*, as well as 36 accessions of *Glycine gracilis*, semi-wild soybeans that have been classified as landraces or wild soybeans (Han *et al.*, 2016). *G. gracilis* has been considered either a preliminary evolutionary product toward domesticated soybean (Fukuda, 1933) or a result of hybridization between *G. max* and *G. soja* (Hymowitz, 1970). Previous studies of chloroplast DNA (Xu *et al.*, 2002) and gene flow (Wang & Li, 2011), as well as the genome resequencing of 10 semi-wild soybeans (Qiu *et al.*, 2014) have suggested the latter possibility of

*These authors contributed equally to this work.

hybridization. However, recent analyses of the genetic differentiation and gene flow among *G. max*, *G. gracilis* and *G. soja* accessions reveal two novel insights about soybean domestication (Han *et al.*, 2016). First, *G. gracilis* is a transitional species derived from the evolutionary process of domesticated soybean. While no gene flow was observed from *G. max* to *G. gracilis* or *G. soja*, significant gene flow was observed from *G. soja* to *G. gracilis* and from *G. gracilis* to *G. max*, as well as moderate gene flow from *G. soja* to *G. max*, supporting *G. soja* as the progenitor of both *G. gracilis* and *G. max*. Second, the Huang-Huai Valley in Central China, the region between the Yellow River and Huai River, is the most likely location of soybean domestication. Among the *G. max* accessions examined, the accessions from the Huang-Huai Valley showed greater genetic introgressions from *G. soja* than those from other geographic regions. A Bayesian-based migration analysis also suggested gene flow from the Huang-Huai Valley to North-eastern and Southern China. Furthermore, the accessions from the Huang-Huai Valley possessed higher genetic diversity than those from North-eastern and Southern China. These recent observations agree with a previously proposed hypothesis in which the transition to domesticated soybean occurred as a gradual process. Before the estimated domestication time of soybean, divergence studies of *G. max* and *G. soja* genomes suggested that the ancestor of domesticated soybean diverged from *G. soja* 0.27 or 0.8 Ma, creating a *G. soja*–*G. max* complex (Kim *et al.*, 2010b; Li *et al.*, 2014a). It is therefore possible that the evolutionary intermediate species *G. gracilis* represents such a *G. soja*–*G. max* complex that humans had interacted with long before the domestication event of soybean.

Despite being another point of debate, a single origin of domesticated soybean appears widely accepted (the single origin hypothesis; Fig. 1a), supported by recent genome resequencing studies (Zhou *et al.*, 2015b; Han *et al.*, 2016) as well as a previous analysis of a relatively small number of microsatellite markers (Guo *et al.*, 2010). Resequencing of 302 wild, landrace or improved soybeans suggests that all domesticated soybeans derived from a single cluster of *G. soja* wild soybeans, supporting the single origin hypothesis that all currently grown domesticated soybeans originated from a single domestication event (Zhou *et al.*, 2015b). Similarly, all 79 soybean landraces used in Guo *et al.* (2010) clustered together, suggesting a monomorphic origin of domesticated soybeans. Providing additional support, the domesticated alleles of the soybean domestication genes *SHATTERING1-5* (*SHAT1-5*) have derived from a single domestication event as described later in this review (Dong *et al.*, 2014). Contrary to the prevailing single domestication hypothesis, comparative phylogenetic studies using chloroplast or nuclear microsatellite markers in wild and cultivated soybeans have suggested multiple origins of soybean in East Asia (Xu *et al.*, 2002; Abe *et al.*, 2003) (the multiple origin hypothesis; Fig. 1b). Recent analysis of chloroplast genomes from the earlier mentioned 302 wild, landrace and improved soybeans also suggests that multiple maternal lines account for domesticated soybeans (Fang *et al.*, 2016). Korean and Japanese soybeans possess significantly different gene pools in their chloroplast and nuclear genomes (Xu *et al.*, 2002; Abe *et al.*, 2003; Zhou *et al.*, 2015b), supporting the idea that independent domestication events may have taken place in these regions. High genetic diversity of soybeans was also reported in the Korean peninsula (Lee & Park, 2006). Moreover, archeological records show larger soybean seeds in Korea and Japan compared with seeds found in the Yellow River basin in China during the period 5000–3000 BP (Lee *et al.*, 2011). In particular, soybean seed samples from Central Japan are reported to be the largest during this time frame. These findings, together with the long divergence time between *G. max* and *G. soja* as described earlier, indicate that there may well have been multiple independent efforts to domesticate wild soybeans, either *G. soja* or the *G. soja*–*G. max* complex, at different locations in East Asia. Indeed, the presence of wild soybeans in grain impressions on pottery appears as early as 7000–5000 BP in Japan (Obata, 2011; Obata & Manabe, 2011). Taken together, these observations project a novel view on soybean domestication in which a prolonged period of low-intensity management or semi-cultivation of wild soybeans at multiple locations preceded the domestication event (the complex hypothesis; Fig. 1c). These potential early domesticates may have either disappeared among wild soybeans or been integrated into the domesticated soybean from China, which may have possessed more advantageous traits for cultivation, during its spread throughout the Korean peninsula and Japan, resulting in the continuous yet distinct subpopulations in these regions. Future genomic studies of a larger set of semi-wild soybeans from diverse geographic areas, as well as efforts to extract genome sequence information from archeological materials, will further clarify the early history of soybeans in the pre-domestication era.

In the following sections, we refer to *G. soja* as wild soybean, as the genetic and genomic information of *G. gracilis* relevant to the molecular bases of soybean domestication has yet to be clarified.

Genetic architecture of wild and domesticated soybeans

Several severe genetic bottlenecks occurred during soybean domestication and diversification, notably in the domestication of Asian landraces and in the introduction of relatively few of those landraces to North America (Hyten *et al.*, 2006). Approximately half of the genetic diversity (Zhou *et al.*, 2015b) and 81% of rare alleles (Hyten *et al.*, 2006) were lost during soybean domestication from *G. soja* to landraces. Only 19 landraces are thought to have contributed as much as 85% of the genes to the North American breeding pools (Gizlice *et al.*, 1996). Accordingly, nucleotide diversity (π) decreased sharply from 2.17×10^{-3} in *G. soja* to 1.47×10^{-3} in landraces and moderately to 1.14×10^{-3} in North American ancestors and 1.11×10^{-3} in elite cultivars, indicating the extent of the bottleneck effects in soybean domestication and the introduction to North America (Hyten *et al.*, 2006). Similar nucleotide diversity levels are reported in separate studies (Li *et al.*, 2013; Valliyodan *et al.*, 2016). Although modern plant breeding is known to reduce genetic diversity in elite cultivars in many domesticated crops, soybeans appear to show a different pattern. Despite its obvious severity, the moderate level of nucleotide diversity in North American ancestors and elite cultivars suggests that the extent of the bottleneck in soybean's North American introduction is surprisingly weak compared with the domestication

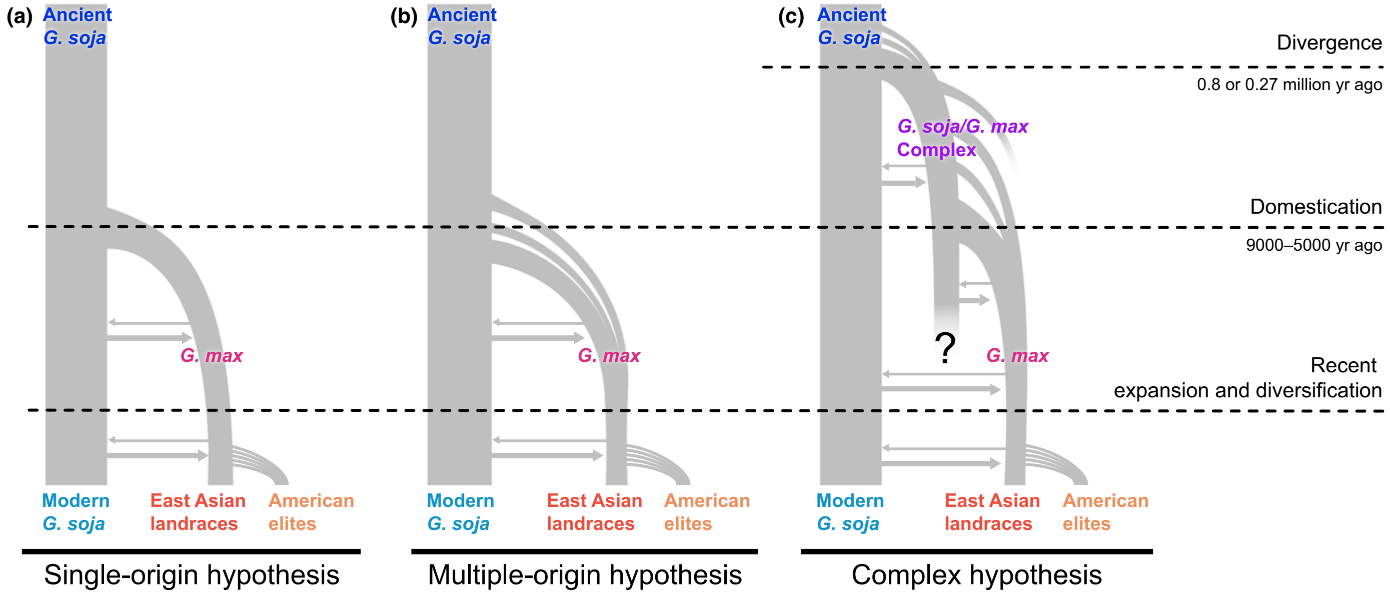

Fig. 1 Hypotheses of the origin of domesticated soybean (*Glycine max*). (a) The single founder model. (b) The multiple founder model. (c) The complex model. The width of lines indicates genetic diversity. Horizontal gray arrows represent gene flow between populations.

bottleneck (Hyten *et al.*, 2006; Valliyodan *et al.*, 2016). One potential explanation for this is that selection during the modern soybean breeding programs in North America may have pinpointed only small genomic regions that contain favorable traits, minimally affecting overall genomic diversity. Conversely, it is also possible that subsequent introgression of *G. soja* or landrace genomes in the modern breeding efforts may have increased diversity among elite cultivars. Additionally, balancing selection for adaptation to a photoperiodic gradient across a wide latitudinal range of North America would have maintained a moderate level of diversity, mitigating the founder effect of the North American introduction.

Owing to the recently accomplished *de novo* assembly of *G. soja* genomes (Li *et al.*, 2014a) and the resequencing of diverse soybean accessions (Table 1), we now have a much deeper understanding of the consequences of soybean domestication at the genomic level. The genomes of wild and cultivated soybeans possess a vast number of single nucleotide polymorphisms (SNPs) and insertion/deletion (indel) variants that may affect gene function in a lineage-specific manner, providing a reservoir of novel genes and genetic variation for the future of soybean improvement. Based on the comparison of seven *G. soja* accessions and the *G. max* reference accession Williams 82 (Li *et al.*, 2014a), 1764 loci were found to possess a stop codon in a coding region of *G. soja* accessions but not in the corresponding gene in *G. max*, and 2285 loci possessed a stop codon in *G. max* but not in *G. soja*. *G. soja* possesses 2989–4181 indels that result in frameshifts compared with *G. max*. Copy-number variations (CNVs) in gene coding regions also differ greatly between wild and cultivated soybeans. Compared with *G. max*, the *G. soja* accessions possess 1978 genes carrying CNVs, many of which are genes involved in biotic and abiotic stress responses, suggesting that these genes may have roles in environmental adaptation. One such example is the *Resistance to Heterodera glycines* (*Rhg1*) locus that confers soybean cyst nematode resistance (Kim *et al.*, 2010b) as detailed in the next section. In addition to CNVs, presence–absence variations (PAVs) are prevalent between wild and cultivated soybeans. Unlike CNVs, PAVs are shown to occur throughout each chromosome's euchromatic regions (Wang *et al.*, 2014). *G. soja*-specific PAVs that are present in *G. soja* but absent in *G. max* were reported to span 2.3–3.9 Mbp in the *G. soja* genome, while *G. max*-specific PAVs span 1.8 Mbp (Li *et al.*, 2014a). These PAVs contain at least 338 *G. soja*-specific genes and 16 *G. max*-specific genes in a variety of functional categories. In cultivated soybeans, CNVs tend to cluster in gene-rich regions on chromosomes 3, 6, 7, 16 and 18 (McHale *et al.*, 2012), in contrast to maize in which CNVs are distributed throughout the genome (Swanson-Wagner *et al.*, 2010). Variation between wild and cultivated soybeans, however, appears to centralize significantly more in pericentromeric regions than in the chromosome arms (Lee *et al.*, 2015).

Despite soybean's inbreeding habits and stringent cleistogamy, the linkage disequilibrium (LD) in soybean landraces and improved cultivars is moderate (83 and 133 kb, respectively) based on one of the most recent and comprehensive resequencing works (Zhou *et al.*, 2015b). The extent of LD in wild soybeans is *c.* 27 kb, similar to that of wild rice (Huang *et al.*, 2012b) and wild maize (Hufford *et al.*, 2012), making genome-wide association studies (GWAS) feasible in soybean. Additionally, genome sequencing information provides us with an accurate estimate of genome-wide population genetic statistics and identification of loci that are potentially under selection. Genomic regions associated with soybean domestication or subsequent diversification/improvement are sought based on test statistics using the levels and pattern of nucleotide diversity, and the extent of LD and haplotype extension for complete or partial selective sweeps. Population differentiation analysis (F_{st}) has also been popularly employed, most successfully in

Table 1 Recent genome sequencing studies in soybean in the past 5 years

Reference	No. of accessions studied	Accessions sequenced	Additional materials	Sequencing method	Analysis performed
Lee *et al.* (2016)	2	Landrace (1)/Landrace mutant (1)		Whole-genome resequencing	SNP/indel identification, GO analysis, nucleotide diversity π, qRT-PCR
Valliyodan *et al.* (2016)	106	Wild (7)/Landrace (43)/Elite (56)		Whole-genome resequencing	Genetic diversity (θ_π), Watterson's estimator (θ_w), phylogenetic tree, PCA, population structure, Bayesian clustering, Tajima's *D*, F_{st}, PAV, CNV, LD decay, selective sweeps (π_{wild}/$\pi_{cultivated}$)
Wang *et al.* (2016a)	367	Wild (105)/Cultivated (262)		Affymetrix Axiom Genome-Wide BOS 1 Array (designed using 32 resequenced domesticated and wild lines)	Genetic diversity (θ, π), phylogenetic tree, PCA, population structure, MAF, Tajima's *D*, F_{st}, LD decay, ROD, GWAS
Han *et al.* (2016)	512	Wild (72)/Semi-wild (36)/ Cultivated (404)		SLAF-seq (specific-locus amplified fragment sequencing)	Genetic diversity (θ_π), phylogenetic tree, population structure, PCA, Watterson's estimator (θ_w), Fu and Li's *D**, Fu and Li's *F**, F_{st}, LD analysis, three population test, gene flow (Nm), selective sweeps, GWAS
Song *et al.* (2015)	19 648	Wild (1168)/Domesticated (18 480)		The SoySNP50K Illumina Infinium II BeadChip	F_{st}, similarity analysis, cluster analysis, LD analysis, haplotype block analysis, GWAS

Lee *et al.* (2015)	89	Domesticated (9)	SoyNAM founder lines (41) (P. Cregan, unpublished);
			Wild (17)/Landrace (4)/Elite (9)/Neutron-mutated (1) Lam *et al.* (2010);
			Landraces (6) Cook *et al.* (2014);
			Wild (1)/Breeding line (1) Schmitz *et al.* (2013)
				Whole-genome resequencing	Genetic diversity (θ, π), phylogenetic tree, Tajima's *D*, F_{st}, LD decay, CNV, SNP, genetic structure, qPCR validation

Zhou *et al.* (2015a)	286	Wild (14)/Landrace (153)/ Elite (119)		RAD-seq genotyping	F_{st}, chi-squared test, *U*-test, population structures, GWAS
Zhou *et al.* (2015b)	302	Wild (62)/Landrace (130)/Elite (110)		Whole-genome resequencing	Genetic diversity (π), PCA, F_{st}, CNV, LD decay, RFD, selective sweeps, GWAS, XP-CLR
Fang *et al.* (2016)	302		Wild (62)/Landrace (130)/Elite (110) Zhou *et al.* (2015b)	Whole-genome resequencing	Genetic diversity (π), selective sweeps, phylogenetic tree
Qiu *et al.* (2014)	11	Wild (1)/Semi-wild (10)		Whole-genome resequencing	Genetic diversity (π), phylogenetic tree, population structure, selective sweeps (pooled heterozygosity, Hp), GO enrichment analysis
Li *et al.* (2014a)	7	Wild (7)		Whole-genome resequencing (*de novo*)	SNP, indel, PAV, CNV, dn/ds, GO enrichment analysis, gene clustering, phylogenetic tree, selective sweeps (maximum likelihood ratio test)
Qi *et al.* (2014)	97	Wild (1)/RIL (96)		Whole-genome resequencing	Genotyping-by-sequencing, GO analysis, QTL mapping

Table 1 (Continued)

Reference	No. of accessions studied	Accessions sequenced	Additional materials	Sequencing method	Analysis performed
Anderson *et al.* (2014)	41	Soybean NAM (41)		CGH and whole-genome resequencing	PAV, CNV, GO enrichment analysis, cross-validation, SFS
Cook *et al.* (2014)	43	Landraces (6)	Landraces (1) Cook *et al.* (2012); SoyNAM (35) Q. Song, B.W. Diers, & P. Cregan (unpublished data); Wild (1) Kim *et al.* (2010b)	Whole-genome resequencing	Alignment, indel identification, copy number estimation, network analysis
Chung *et al.* (2014)	16	Wild (6)/Landrace (4)/Elite (6)		Whole-genome resequencing	SNP and indel detection, identification of nonreference genes and gene loss event, phylogenetic tree, population structure, genetic diversity (θ),Watterson's estimator (θ_w), F_{st}, ROD, LD decay, GO term enrichment analysis, selective sweeps ($\pi_{cultivated}/\pi_{wild}$)
Li *et al.* (2013)	55	Wild (8)/Landrace (8)/Elite (9)	Wild (17)/Landrace (4)/Elite (9) Lam *et al.* (2010)	Whole-genome resequencing	SNP/InDel calling, population structure and phylogenetic analysis, genetic diversity (π), Tajima's D, F_{st}, selective sweeps ($\pi_{wild}/\pi_{cultivated}$), QTL mapping, PCA
Ha *et al.* (2012)	2	Wild (1)/Elite (1)		Whole-genome resequencing	MTP

SNP, single nucleotide polymorphism; GO, gene ontology; qRT-PCR, quantitative reverse transcription polymerase chain reaction; PCA, principal component analysis; PAV, presence/absence variation; CNV, copy number variation; LD, linkage disequilibrium; MAF, minor allele frequency ROD, reduction of diversity; GWAS, genome-wide association studies; RAD, restriction-site-associated DNA; RFD, relative frequency difference; XP-CLR, cross-population composite likelihood ratio test; RIL, recombinant inbred line; QTL, quantitative trait locus; NAM, nested association mapping; CGH, comparative genomic hybridization; SFS, site frequency spectrum; MTP, minimum tiling path.

identification of genomic regions that potentially underwent selection in geographic differentiation or modern breeding programs. Based on mean pairwise F_{st} values, wild soybeans, landraces and modern cultivars are distinct to some extent: the greatest differentiation is observed between the wild and modern cultivars ($F_{st} = 0.162$) and the least differentiation between the landraces and modern cultivars ($F_{st} = 0.0047$) (Li *et al.*, 2014b). A number of clustered selection hotspots were identified in the soybean genome, containing a large number of candidate genes that may have experienced artificial selection (Lam *et al.*, 2010; Li *et al.*, 2013; Zhao *et al.*, 2015). Approximately 4.4% of the total annotated genes are targeted by artificial selection based on F_{st} (Li *et al.*, 2013). At least 1188 and 489 genes contain nonsynonymous substitutions that are fixed in early domestication and modern improvement, respectively (Zhao *et al.*, 2015). Although in-depth functional characterization and evolutionary studies await those genes, a combination of statistical tests and GWAS or previously reported quantitative trait loci (QTL) have identified a handful of genes that play important roles in soybean domestication, diversification and improvement. Among them are *GmTFL1a*, *GmCRY1a* and *GmCOL7a* that are homologs of flowering-related genes as detailed later. In the following section, we list and review genes that potentially underlie domestication, diversification and improvement of soybean, identified by QTL mapping, candidate gene cloning, GWAS or population genetic study (Table 2).

Genes underlying domestication-related traits

The process of crop domestication encompasses a broad range of evolutionary changes that transition through multiple continuous stages (Meyer & Purugganan, 2013). With the aim of classifying genes according to their contribution to specific stages of the domestication process, in this review we define domestication genes as follows: the gene's function has been characterized and is known to underlie a trait, the gene has undergone positive selection, and the complete or near-complete fixation of the causative mutation should be observed in all lineages from a single domestication event (Meyer & Purugganan, 2013). Under this criterion, a gene that controls an important trait but has a causative mutation(s) that segregates in domesticated populations is considered a diversification or improvement gene that played a lineage-specific role in the crop's regional adaptation or subsequent improvement.

Pod shattering

Loss of pod shattering/seed dispersal is a key agronomic trait that was targeted by human selection and is regarded as a milestone of crop domestication. Indicating evolutionary parallelism, orthologous genes have been shown to control seed shattering in multiple cereals (Dong & Wang, 2015). In soybean, the genetic mechanisms underlying the evolution of the shattering-resistant phenotype of domesticated soybean appear different from that of cereals. Loss of pod shattering in soybean lies in the excessively lignified fiber cap cells (FCCs), and is promoted by a NAC (NAM, ATAF1/2 and CUC2) transcription factor, SHAT1-5 (Dong *et al.*, 2014). SHAT1-5 activates secondary wall biosynthesis and promotes the thickening of FCC secondary walls. The domesticated allele of this gene is expressed 15-fold higher than the wild allele, attributing to a 20 bp deletion that disrupts a repressive element in the regulatory region of *SHAT1-5*. Nucleotide diversity suggests that all domesticated soybeans carry *SHAT1-5* haplotypes derived from a single haplotype distinct from wild soybeans. In addition, the *SHAT1-5* locus shows a severe selective sweep across *c.* 116 kb. These results indicate that this locus has experienced artificial selection and was probably derived from a single domestication event, making *SHAT1-5* a prime domestication gene of soybean.

Pdh1, encoding a dirigent-like protein involved in lignification, is another gene that affects soybean's pod shattering phenotype (Funatsuki *et al.*, 2014). *Pdh1* promotes torsion of dried pods under low humidity, causing higher pod dehiscence. Shattering-resistant varieties carry a single nucleotide substitution at the beginning of the coding sequence that produces a stop codon. In clear contrast to the domestication gene *SHAT1-5*, the shattering-resistant allele of *Pdh1* is observed at low frequencies in Japanese and Korean landraces and cultivars and at moderate frequency in China, while *c.* 75% of South Asian landraces carry the resistant allele. Notably, most of modern North American cultivars possess the resistant allele, indicating that the *Pdh1* gene was utilized as an additional shattering-resistance locus in the modern breeding programs in North America.

Seed hardness

Seed hardness, which includes water permeability of dry seeds and hardness of cooked seeds, is another important trait for soybean domestication and improvement. The causal gene of a major QTL controlling water permeability has been identified as *GmHs1-1*, which encodes a calcineurin-like metallophosphoesterase transmembrane protein and is expressed in an epidermal layer of the seed coat (Sun *et al.*, 2015). Although its cellular functions are unknown, a large percentage of soybean landraces carry a SNP that causes an amino acid substitution and show low polymorphism in the *c.* 160 kb genomic region surrounding *GmHs1-1*, indicating a possible signature of artificial selection during soybean domestication. Water permeability is controlled by several additional QTLs (Liu *et al.*, 2007; Orazaly *et al.*, 2015); of those, the causal gene of the *qHS1* locus is shown to encode an endo-1,4-β-glucanase that controls the amount of β-1,4-glucans in the outer layer of palisade cells of the seed coat on the dorsal side of seeds, a point of water entrance (Jang *et al.*, 2015). Some of the seed permeability QTLs underlie seed coat cracking (Nakamura *et al.*, 2003). Although seed cracking is considered an unfavourable trait as it reduces soybean's commercial value, this trait appears among landraces probably because it causes more advantageous water permeability. It is therefore possible that different types of selection may have acted on seed hardness-related traits in early domestication and in modern soybean improvement. Concerning the hardness of cooked seeds, a potential causal gene of a major locus, *qHbs3-1*, was identified as a pectin methylesterase homolog (Toda *et al.*, 2015). Identification of additional genes underlying seed permeability and seed hardness and further evolutionary studies at

Table 2 Genes that potentially underlie domestication, diversification and improvement of soybean

Trait	Category	Gene	Orthologs in Arabidopsis	Gene loci	Gene category	Causative change	Prevalence	Gene identification method	Selection evidence	Reference(s)
Shattering	Domestication	***GmSHAT1-5***	*NST1/2*	Glyma.16G019400	Transcriptional regulator	*cis*-regulatory	All domesticates	Candidate gene	Hitchhiking effect	Dong *et al.* (2014)
Shattering	Improvement	***GmPdh1***		Glyma.16G141500	Dirigent protein	Premature stop	Subset of domesticates	Map-based cloning	NA	Funatsuki *et al.* (2014)
Hard-seededness	Domestication	***GmHs1-1***		Glyma.02G269500	Metallophosphoesterase	Point mutation leading to structural change of protein	Subset of domesticates	Mapping	Hitchhiking effect	Sun *et al.* (2015)
Determinate growth	Diversification	***GmDt1/ GmTFL1b***	*TFL1*	Glyma.19G194300	Transcriptional regulator	Amino acid change	Subset of domesticates	Candidate gene	GWAS	Liu *et al.* (2010); Tian *et al.* (2010); Zhou *et al.* (2015b)
Semi-determinate growth	Diversification	***GmDt2***	*AP1 /FUL /CAL*	Glyma.18G273600	Transcriptional regulator	*cis*-regulatory	Subset of domesticates	Mapping	NA	Ping *et al.* (2014)
Flowering time	Diversification	***GmTFL1a***	*TFL1*	Glyma.03G194700	Transcriptional regulator	NA	NA	Candidate gene	Reduced SNP	Li *et al.* (2013)
Flowering time	Diversification	***GmCRY1a***	*CRY*	Glyma.04G101500	Photoreceptor	NA	NA	Candidate gene	Reduced SNP	Zhang *et al.* (2008); Li *et al.* (2013)
Flowering time	Diversification	***GmCOL7a***	*COL*	Glyma.10G274300	Transcriptional regulator	NA	NA	Candidate gene	Reduced SNP	Li *et al.* (2013); Wu *et al.* (2014)
Flowering and maturity	Diversification	***E1***		Glyma.06G207800	Transcriptional regulator (with B3 domain)	Amino acid change	Subset of domesticates	Map-based cloning	Geographic differentiation (F_{st})	Xia *et al.* (2012); Langewisch *et al.* (2014); Zhou *et al.* (2015b)
Flowering and maturity	Diversification	***E2 (GmGIa)***	*GI*	Glyma.10G221500	Protein binding	Premature stop	Subset of domesticates	Map-based cloning	NA	Watanabe *et al.* (2011); Langewisch *et al.* (2014)
Flowering and maturity	Diversification	***E3 (GmPhyA3)***	*PHYA*	Glyma.19G224200	Photoreceptor	Single base deletion or single base change	Subset of domesticates	Map-based cloning	NA	Watanabe *et al.* (2009); Langewisch *et al.* (2014)
Flowering and maturity	Diversification	***E4 (GmPhyA2)***	*PHYA*	Glyma.20G090000	Photoreceptor	Retrotransposon insertion or single base deletion	Subset of domesticates	Mapping	NA	Liu *et al.* (2008); Langewisch *et al.* (2014)
Seed coat color, hilum color	Diversification	***I (GmCHS)***		Glyma.08G109200 (CHS4), Glyma. 08G110300 (CHS3), Glyma.08G109500 (CHS1)	Enzyme (chalcone synthase)	NA	NA	Mapping	Reduced SNP, GWAS	Tuteja *et al.* (2009); Li *et al.* (2013); Zhou *et al.* (2015b)
Flower color	Diversification	***W1 (F3′5′H)***		Glyma.13G072100	Enzyme (flavonoid 3′5′-hydroxylase)	Insertion in exon causing frameshift	Subset of domesticates	Mapping	GWAS	Zabala & Vodkin (2007); Zhou *et al.* (2015b)
Pod color	Improvement	***L1 (MYB)***		Glyma.19G101700	Transcriptional regulator	*cis*-regulatory	Subset of domesticates	Mapping	NA	He *et al.* (2015)
Pubescence color	Diversification	***T***	*F3′H*	Glyma.06G202300	Enzyme	Single base deletion	Subset of domesticates	Candidate gene	Geographic differentiation (F_{st})	Toda *et al.* (2002); Zhou *et al.* (2015b)

Table 2 (Continued)

Trait	Category	Gene	Orthologs in Arabidopsis	Gene loci	Gene category	Causative change	Prevalence	Gene identification method	Selection evidence	Reference(s)
Cyst nematode resistance	Improvement	***Rhg1* (amino acid transporter, α-SNAP & WI12), multiple copies**		Glyma.18G022400, Glyma.18G022500, Glyma.18G022600 & Glyma.18G022700	Amino acid transporter, disassembly of SNARE membrane trafficking complexes, a protein with a WI12 (wound-inducible protein 12) region	Copy number variation	Subset of domesticates	GWAS signal	GWAS	Cook *et al.* (2012); Zhou *et al.* (2015b)
Leaf shape and the four-seed pod	Improvement	***Ln***	*JAG*	Glyma.20G116200	Transcriptional regulator	Single base change	Subset of domesticates	Candidate gene	Geographic differentiation (F_{st})	Jeong *et al.* (2012); Zhou *et al.* (2015b)
Structural diversity of glycosylation	Improvement	***Sg-1***		Glyma.07G254600	Enzyme (glycosyltransferase)	Truncated protein or amino acid deletion	Subset of domesticates	Mapping	GWAS	Sayama *et al.* (2012); Zhou *et al.* (2015b)
Oil content	Improvement	***LPD1***		Glyma.15G143100	Enzyme	Nonsense mutation	Subset of domesticates	Mapping	Selective sweep	Diers *et al.* (1992); Scoles *et al.* (2006); Qi *et al.* (2011); Zhou *et al.* (2015b)

Gene IDs are based on *Wm82.a2.v1*. GWAS, genome-wide association studies; SNP, single nucleotide polymorphism; NA, information not available.

the population level will help to clarify the molecular mechanisms and precise modes of artificial selection for seed hardness-related traits.

Shoot architecture/stem growth habit

Determinacy is an agronomically important trait associated with the domestication process of soybean. Classic genetic analyses demonstrated that soybean stem growth habit was regulated by an epistatic interaction between two major loci, *Dt1* and *Dt2* (Bernard, 1972). The causal gene of the *Dt1* locus encodes the functional counterpart of Arabidopsis *TERMINAL FLOWER1* (*TFL1*), designated as *GmTfl1* or *GmTFL1b* (Tian *et al.*, 2010). Similar to Arabidopsis *TFL1*, the *GmTfl1* transcript accumulates in the shoot apical meristem (SAM) during early vegetative growth in both the determinate and the indeterminate lines, but thereafter is abruptly lost in the determinate line (Liu *et al.*, 2010). Four independent single nucleotide substitutions were identified in the *GmTfl1* gene, each of them leading to an amino acid substitution. These substitutions were found in a subset of *G. max* but not in *G. soja* (Tian *et al.*, 2010), indicating that selection for determinacy took place during soybean diversification.

The *Dt2* locus encodes a MADS box transcription factor in the APETALA1/SQUAMOSA (AP1/SQUA) subfamily (Ping *et al.*, 2014). In *Dt1*/*Dt1* homozygous genetic backgrounds, *Dt2*/*Dt2* genotypes produce semi-determinate phenotypes, whereas *dt2*/*dt2* genotypes produce indeterminate phenotypes. However, in *dt1*/*dt1* genetic backgrounds, the *dt2*/*dt2* phenotype is determinate, indicating an epistatic effect of the *dt1* allele on expression of the *Dt2* locus (Bernard, 1972). *Dt2* suppresses expression of the *GmTfl1* (*Dt1*) gene in the SAM to promote early conversion of the SAM into reproductive inflorescence (Ping *et al.*, 2014). Given that the semi-determinate and determinate stem growth habit phenotypes are rarely observed in *G. soja*, it is proposed that the dominant *Dt2* allele is a recent gain-of-function mutation that occurred after soybean domestication .

GmTFL1a is the closest paralog of *GmTfl1*/*GmTFL1b* (*Dt1*). Despite its high sequence similarity, *GmTFL1a* does not seem to function in the control of the stem growth habit. Contrary to *GmTfl1*/*GmTFL1b*, *GmTFL1a* is expressed mainly in immature seeds and slightly in the cotyledon and stem tip (Liu *et al.*, 2010). Although the function is unclear, *GmTFL1a* is reported to have undergone strong artificial selection during soybean domestication and improvement based on population genetic tests for selection (Li *et al.*, 2013).

Photoperiodic flowering

A major focus of the soybean domestication and diversification process was selection for adaptation to a particular latitudinal photoperiod (Cober & Morrison, 2010; Kim *et al.*, 2012a). As a short-day flowering plant, its latitudinal expansion requires loss of photoperiod sensitivities. Several genes underlying soybean's latitudinal adaptation have been identified. *GmCRY1a* and *GmCOL7a*, soybean homologs of the Arabidopsis blue light receptor *CRYPTOCROME 2* (*CRY2*) and the photoperiodic flowering regulator *CONSTANS* (*CO*), respectively, are reported to exhibit a strong signature of selection (Li *et al.*, 2013). *GmCRY1a* affects blue light-mediated inhibition of cell elongation and promotes floral initiation in soybean (Zhang *et al.*, 2008). The rhythmic expression of the GmCRY1a protein correlates with flowering time and latitudinal distribution of soybean cultivars under flowering-inhibitory long-day photoperiods. The function of *GmCOL7a* has not been characterized, but it is among the 26 soybean homologs of *CO* and some of these homologs are shown to function in photoperiodic flowering in soybean (Wu *et al.*, 2014; Cao *et al.*, 2015).

Analyses of QTLs that control photoperiod sensitivities of cultivated soybeans, known as maturity loci (*E* loci), identified nucleotide variation in flowering-associated genes. Among 180 cultivated soybeans surveyed, the percentages of recessive alleles at the major maturity loci *E1*, *E2*, *E3* and *E4* are 38.3, 84.5, 36.3 and 7.2%, respectively (Zhai *et al.*, 2014), suggesting that these maturity loci have contributed to diversification or local adaptation rather than soybean domestication. Among these *E* loci, *E1* shows a predominant effect on photoperiodic control of flowering and maturation. A transcription factor containing a plant-specific B3 domain was identified as the causal gene of *E1* (Xia *et al.*, 2012). The dominant *E1* allele delays flowering and maturation, and this effect is significantly enhanced by long-day photoperiods. The abundance of the *E1* transcript negatively correlates with that of *GmFT2a* and *GmFT5a*, functional orthologs of the flowering inducer *FLOWERING LOCUS T* (*FT*) (Xia *et al.*, 2012). The *E1* locus did not appear to be significant in a genomic scan for selective sweeps nor in GWAS, but the level of differentiation between subpopulations (F_{st}) indicates strong regional differentiation of *E1* alleles between Southern China and North America (Zhou *et al.*, 2015b). While the dominant *E1* allele is predominant among both the wild and the cultivated soybeans in China (Langewisch *et al.*, 2014), the *e1* allele carrying a nonsynonymous substitution is distributed mainly in the high latitudinal regions, such as the United States and Canada, northeastern China, Japan and Korea, where soybeans typically show shorter maturity periods compared with those in southern regions (Zhou *et al.*, 2015b).

Another major maturity locus *E2* encodes GmGIa, a homolog of Arabidopsis GIGANTEA (GI) that is a component of the circadian clock and a regulator of photoperiodic flowering (Watanabe *et al.*, 2011). The dominant *E2* allele delays flowering and maturity, while the homozygous *e2* alleles carrying an SNP causing a premature stop codon elevates expression of *GmFT2a*, leading to early flowering. Although these alleles segregate, the *E2* allele is prevalent in wild soybeans and the *e2* allele in cultivated soybeans (Langewisch *et al.*, 2014). The recessive *e2* haplotypes, H1, H2 and H3, display unique geographic patterns (Wang *et al.*, 2016b). H1 is widely distributed among cultivated soybeans, while H2 is present in Southern China. H3 is assumed to have been later introgressed from wild soybean independently and is restricted to the Northeast region of China. Among wild soybeans, H1 appears only in the Yellow River basin with a low frequency, supporting Central China as the origin of domesticated soybean. The photoreceptor gene *PHYTOCHROME A* (*PHYA*) was isolated as the causal gene of *E3* (Watanabe *et al.*, 2009) and *E4* (Liu *et al.*, 2008). Plants carrying the nonfunctional *GmPhyA3* allele carrying a 40 bp deletion in the first exon flower earlier than those with the functional allele (Watanabe *et al.*, 2009). *E3* and *E4* are considered to act redundantly in photoperiod sensitivity (Liu *et al.*, 2008; Wu *et al.*, 2013). In Japan, the geographic distribution of the *e4* allele appears restricted to the high-latitudinal regions (Kanazawa *et al.*, 2009). In addition, *GmFT2a*, a soybean homolog of the flowering inducer *FT*, is shown to be the causal gene of the minor maturity locus *E9* (Zhao *et al.*, 2016b).

Flower, seed coat and pod color

A set of traits that have been targeted by GWAS includes flower, seed coat and pod colors, identifying the *W1* and inhibitor (*I*) loci (Zhou *et al.*, 2015b). The *W1* locus is one of the six loci that control soybean flower pigmentation (Zabala & Vodkin, 2007). The causal gene of *W1* is the flavonoid 3′5′-hydroxylase (F3′5′H) gene. The recessive allele from the white flower isoline 'Williams' (w1) carries a structural rearrangement leading to a small insertion (65 bp) of tandem repeats in exon 3 that results in a premature stop codon. The F3′5′H gene is a rare single-copy gene in the soybean genome and is expressed at very low levels in all tissues examined, including flower and seed coats, but sufficient to account for the delphinidin-based anthocyanins and/or proanthocyanins in these tissues.

The dominant alleles of the *I* locus contain a cluster of duplicated and inverted chalcone synthase (*CHS*) genes encompassing a 27 kb region (Tuteja *et al.*, 2009). CHS is the first committed enzyme in the flavonoid synthesis pathway to create a diverse set of secondary metabolites, including the seed coat pigmentation of certain genotypes. Short interfering RNAs (siRNAs) created from the transcripts of this *CHS* inverted repeat target mRNAs of *CHS* genes on other genomic regions and silence these genes in a seed coat-specific manner. The CNVs containing the *CHS* cluster have been identified by GWAS, but for seed hilum color variation (Zhou *et al.*, 2015b).

Although it was missed by GWAS, the *T* locus that regulates pubescence color exhibits a strong regional differentiation signal in F_{st} (Zhou *et al.*, 2015b). The *T* locus encodes a flavonoid 3′-hydroxylase (F3′H) and the dominant *T* allele produces brown pubescence, while the recessive *t* allele makes it gray (Toda *et al.*, 2002). The frequency of the recessive allele increases from Southern to Northern China, indicating its potential role in chill adaptation (Zhou *et al.*, 2015b).

The pod colors of soybeans include black, brown and tan types, and are controlled by two classical genetic loci, *L1* and *L2* (Woodworth & Veatch, 1929; Bernard, 1967; Kiang, 1990). The potential causal gene of the *L1* locus, Glyma19g27460 (*Wm82.a1.v1*)/Glyma.19G101700 (*Wm82.a2.v1*), has recently been identified by fine mapping (He *et al.*, 2015). Glyma19g27460/Glyma.19G101700 encodes a MYB transcription factor and is expressed at high levels in black pods. Artificial selection might have preferred light-colored pods that could lead to pod-shattering resistance. The *L1* locus has not been identified by GWAS or other tests for selection, but several QTLs have been found near this locus that control diverse agronomic traits, such as seed weight and yield (Csanadi *et al.*, 2001; Guzman *et al.*, 2007) and resistance to bacterial leaf pustule disease (Kim *et al*., 2010a), providing a potential hotspot for artificial selection.

Resistance and other traits

The *Rhg1* locus exhibits a profound effect on soybean cyst nematode resistance (Cook *et al*., 2012). The *rhg1-b* allele has been widely used for resistance against soybean cyst nematodes. This allele contains a 31.2 kb tandem repeat of four genes that varies from one to 10 copies per haploid genome, with increased copy number conferring greater cyst nematode resistance (Cook *et al*., 2014; Lee *et al*., 2015). The *Rhg1* locus also contains variation in DNA methylation status that influences cyst nematode resistance (Cook *et al*., 2014).

Seed quality traits of soybeans are another target of artificial selection. A genomic region conferring an extended haplotype block overlaps with the *Sg-1* locus (Zhou *et al*., 2015b) that encodes a glycosyltransferase responsible for structural diversity of triterpenoid saponins (Sayama *et al*., 2012). Soybean saponins are the main cause of bitterness and astringent aftertastes, and thus undesirable components for human consumption. The strong selection signal identified on the *Sg-1* locus suggests recent artificial selection for loss-of-function alleles of this gene during the soybean improvement process. In addition, the oil content-related gene *LPD1* that encodes lipoamide dehydrogenase 1 has been identified by both GWAS and F_{st} analysis (Zhou *et al*., 2015b), indicating that this gene has experienced selection during soybean improvement.

Although no yield-determining genes have been isolated to date, the *Ln* locus may affect soybean's yield potential directly or indirectly (Lee *et al*., 2001). The *Ln* gene encodes a homolog of Arabidopsis JAGGED, an EAR motif-containing putative nuclear protein that regulates lateral organ development including flower and fruit patterning (Jeong *et al*., 2012). In soybean, *Ln* is responsible for leaf shape and the production of four-seed pods and shows a regional differentiation signal (Zhou *et al*., 2015b). The mutant *Ln* allele is mainly distributed in Northeastern and Northern China, consistent with the geographic distribution of leaflet shape (Chen & Nelson, 2004).

Convergent evolution of domestication-related traits

During the domestication process, rapid and directional changes in a similar set of traits occur in a variety of crop species in parallel due to similar human demands regarding cultivation, harvest and consumption. To try to understand the molecular bases of evolutionary parallelism, here we compare the genes and causative mutations that regulate domestication-related traits in a number of crop species with those in soybean, and discuss whether artificial selection has targeted specific genes in a convergent manner.

Genes and causative mutations for reduced seed shattering have been identified in several grain crops. Loss-of-function mutations in orthologs of the YABBY transcription factor *Shattering1* (*Sh1*) gene underlie this trait in multiple crops, including sorghum, maize, rice (Lin *et al*., 2012) and wheat (Katkout *et al*., 2015), although species-specific genes have been also reported. For example, in rice, the two major shattering QTLs *SH4* (Li *et al*., 2006) and *qSH1* (Konishi *et al*., 2006) are caused by transcription factors that are unique to rice. Similarly, in wheat, the major shattering locus encodes the AP2-like transcription factor Q, a shattering gene unique to wheat (Simons *et al*., 2006). The common shattering gene *Sh1* indicates the well-conserved gene network controlling seed shattering of cereal crops and that the evolution of this gene network tends to converge under artificial selection, although species-specific modifications of the gene network may exist. In soybean, loss of pod shattering is achieved by soybean-specific shattering genes: a regulatory mutation of the NAC transcription factor SHAT1-5 (Dong *et al*., 2014) and a nonsynonymous mutation of the dirigent-like protein Pdh1 (Funatsuki *et al*., 2014). This difference in the causes of nonshattering between cereals and soybean likely stems from the anatomical differences between monocot and eudicot fruit structure. Shattering in cereals is associated with rachis fragility caused by modification of the abscission layer, while shattering in soybean derives from the thickening of FCCs of seed pods and the torsion of lignified pod walls.

Loss of function of *TERMINAL FLOWER1* (*TFL1*) orthologs is observed to underlie determinant growth habit of inflorescence in a number of crop species including soybean, suggesting that artificial selection for determinacy is highly convergent at this gene. Examples include the *SELF-PRUNING* gene in tomato (*Solanum lycopersicum*) (Pnueli *et al*., 1998; Carmel-Goren *et al*., 2003), *PvTFL1y* in common bean (*Phaseolus vulgaris*) (Repinski *et al*., 2012) and *CcTFL1* in pigeon pea (*Cajanus cajan*) (Mir et al., 2014). Causal mutations of most of these examples are nonsynonymous. Although the function of TFL1 is deeply conserved among flowering plants including monocots (Wickland & Hanzawa, 2015), there currently is no clear sign of selection reported on this gene in grass crops. Contrasting selection on this gene suggests that a different mechanism is responsible for the evolution of shoot architecture under human selection in grass crops compared to eudicots. Indeed, the well-known players of grass domestication and improvement include Teosinte branched 1 (Tb1) in maize (Clark *et al*., 2004) and the GA_3 biosynthesis and signaling pathway in rice and wheat (Peng *et al*., 1999; Oikawa *et al*., 2004) that modulate shoot architecture, but the evolutionary roles of these factors in domestication of eudicot crops have not yet been reported.

In addition to seed shattering and shoot architecture, central genes in flowering time control appear to be recurrent targets of artificial selection. One such example is the known regulator of photoperiodic flowering *CONSTANS* (*CO*) that encodes a zinc finger transcription factor carrying two B-boxes (Putterill *et al*., 1995; Turck *et al*., 2008). In rice, multiple alleles of the *CO* ortholog *Heading date 1* that acquired mutations in the coding region display late flowering (Takahashi *et al*., 2009; Huang *et al*., 2012a). Similarly, an indel in the coding sequence and a splicing variant are found in the *CO* orthologs of sorghum and foxtail millet, respectively (Liu *et al*., 2015). In addition, a strong signature of selection appears on the soybean *CO* homolog *GmCOL7a* that localizes in a large cluster of selection hotspots (Li *et al*., 2013), although the function of this gene is currently unknown.

The depth of available information on the flowering inducer *FT* suggests the important roles of this gene in domestication and diversification of both monocot and eudicot crops. The sunflower *FT* ortholog *HaFT1* played an important role at the early stage of sunflower domestication (Blackman *et al.*, 2010). A frame shift mutation in *HaFT1* is found in most domesticated sunflowers and leads to later flowering than the functional wild allele. The domesticated *HaFT1* allele is widespread in domesticated sunflowers and exhibits a selective sweep. Supporting the role of *FT* in domestication, the soybean *FT* ortholog *GmFT2c* is shown to harbor a structural rearrangement in domesticated soybeans (Li *et al.*, 2014a). Additionally, one soybean *FT* homolog, *GmFT2a*, possesses a *Ty1*/*copia*-like retrotransposon in the first intron that underlies the minor maturity QTL *E9* (Zhao *et al.*, 2016b). In rice, promoter variation of the *FT* ortholog *Hd3a* is shown to contribute to flowering time diversity (Takahashi *et al.*, 2009; Tsuji *et al.*, 2011). Moreover, regulatory and nonsynonymous mutations of this gene underlie delayed flowering under flowering inhibitory long-day conditions in the late-flowering *indica* varieties (Ogiso-Tanaka *et al.*, 2013).

The photoperiodic flowering regulator *GI*, the causal gene of the soybean *E2* locus, is another general target of artificial selection among monocot and eudicot crops. In *Brassica rapa*, an amino acid substitution in the *B. rapa GI* ortholog underlies a major QTL for the allelic variation in circadian period (Xie *et al.*, 2015). Tilling alleles of this gene carrying missense mutations weaken the rhythmic movement of leaves and confer late flowering at high temperatures. Homologs of *GI* in bread wheat (Rousset *et al.*, 2011) and African sorghum (Bhosale *et al.*, 2012) also are shown to affect flowering time.

Although its soybean homologs have not been characterized, the circadian clock gene *EARLY FLOWERING 3* (*ELF3*) underlies flowering time variation in diverse crop plants. Orthologs of this gene affect flowering time variation in the temperate long-day legumes pea and lentil (Weller *et al.*, 2012); a frameshift indel mutation in *PsELF3* in pea and a splicing mutation in the lentil *ELF3* ortholog result in early flowering. Similarly, in barley, a deletion or rearrangement of the locus containing the *ELF3* homolog *EAM8* is responsible for early flowering of commercial barley varieties bred for short growing seasons (Faure *et al.*, 2012).

Conclusion

In this review, we have summarized the latest information on soybean domestication history and highlighted a number of candidate genes that may have played key roles in soybean domestication, diversification and improvement processes. While the commonly accepted single origin hypothesis of domesticated soybean has a strong genetic, genomic and geographic foundation, emerging evidence points to an extended transitional period of low-intensity cultivation of wild soybeans at multiple locations before the rapid domestication event took place. This complex model fits well with the previously proposed protracted model of crop domestication (Allaby *et al.*, 2008). Among the genes underlying soybean domestication, diversification and improvement processes, nearly half are involved in transcriptional regulation with the remaining half in a variety of structural roles. Only a few genes, including *SHAT1-5* involved in loss of pod shattering and *GmHs1-1* in seed hardness, can be considered domestication genes that were selected for at the early stage of soybean domestication, whereas other genes played a role in the subsequent diversification or improvement process. While several genes that control important agronomic traits appear to be recurrent targets of artificial selection in multiple crop species, unique genes have also been selected in soybean, likely reflecting the relative complexity of the gene network controlling a given trait, different selective pressures or fundamental morphological differences between species. A significant number of these discoveries were made possible by the recent progress in the *de novo* assembly and whole-genome resequencing of wild and cultivated soybean genomes. The ever-expanding reservoir of genomic information also allows the investigation into the evolution of specific classes of genetic components, including the chloroplast genome (Fang *et al.*, 2016) and miRNAs (Liu *et al.*, 2016).

Although the mode of soybean domestication is becoming more apparent, it is evident that much of its molecular basis remains unknown. For example, despite significant efforts to discover loci that govern soybean yield, including seed size and other yield components (Csanadi *et al.*, 2001; Guzman *et al.*, 2007; Liu *et al.*, 2007; Zhou *et al.*, 2015a; Wang *et al.*, 2016a), no causal genes controlling these QTLs other than the *Ln* gene (Jeong *et al.*, 2012) have been identified so far. This may be attributed in part to the quantitative nature of these traits controlled by many loci with small effects, some of which may act indirectly through regulation of other traits, such as nutrition uptake, nitrogen fixation, photosynthesis and sugar transportation. Redundant roles of homologous genes may contribute to this issue. The extended LD observed in soybean genomes would also hinder the identification of genes underlying agronomic traits; however, it may also provide unique opportunities for future investigation into the genomic landscape and improvement of soybean. Since the extended LD would probably elevate the levels of deleterious or weakly deleterious polymorphisms near advantageous alleles (Felsenstein, 1974), we may ask to what extent would these polymorphisms impact soybean's performance, and in what ways might they be removed to breed superior soybeans? Additionally, given the extended LD, it is important to assess the extent of epistatic interactions among loci controlling domestication-related traits to help design effective breeding strategies.

As the post-genome sequencing era quickly approaches, better understanding of the genetic and biochemical mechanisms underlying important agronomic traits at the molecular and systems levels remains a major obstacle. Translational approaches assist with this problem, taking advantage of the wealth of knowledge in the model species and other legumes. For example, soybean homologs of the CYP78A subfamily, which controls organ size and development in Arabidopsis, have been shown to affect seed size in soybean (Wang *et al.*, 2015; Zhao *et al.*, 2016a). It is also important to note that goals in modern soybean breeding programs have been drastically expanded beyond traditional domestication traits, addressing emerging new diseases and changing environments, optimizing seed protein and oil contents, improving nutritional values and developing specialty traits for specific consumption such as sprouting. A number of new tools and resources available for functional genomics, including over 20 000 mutant lines generated from fast neutron bombardment (Bolon *et al.*, 2011), transposon-based mutant lines generated by *Ac*/*Ds* (Mathieu *et al.*, 2009), *Tnt1* (Cui *et al.*, 2013) or *mPing* (Hancock *et al.*, 2011), the soybean Nested Association Mapping (NAM) panel (Stupar & Specht, 2013) and over 50 000 SNPs for 18 480 *G. max* and 1168 *G. soja* accessions (Song *et al.*, 2013, 2015), will enable the discovery of genes responsible for new and old agronomic traits including high yield. Moreover, systems-level modeling approaches to regulatory, metabolic and signaling networks that integrate accumulating omics data, polymorphisms and phenotypic data obtained from the field and controlled environments will further accelerate gene identification and our understanding of important agronomic traits in soybean, and highlight to specific genes and pathways for future breeding efforts. As genome-editing techniques are becoming more efficient and transgene-free (Woo *et al.*, 2015; Zhang *et al.*, 2016), future soybean breeding programs will take advantage of synthetic approaches beyond naturally occurring variation to introduce desired novel alleles and pathways that are designed for specific cultivation strategies and diverse consumer needs. Despite a number of technical challenges, soybean domestication research driven by the recent *de novo* assembly and whole-genome resequencing has taken a significant step closer to precision breeding.

Acknowledgements

This work was supported by the Plant Genome Research Program from the National Science Foundation (NSF-PGRP-IOS-1339388) to Y.H.

References

Abe J, Xu D, Suzuki Y, Kanazawa A, Shimamoto Y. 2003. Soybean germplasm pools in Asia revealed by nuclear SSRs. *Theoretical and Applied Genetics* **106**: 445–453.

Allaby RG, Fuller DQ, Brown TA. 2008. The genetic expectations of a protracted model for the origins of domesticated crops. *Proceedings of the National Academy of Sciences, USA* **105**: 13982–13986.

Anderson JE, Kantar MB, Kono TY, Fu F, Stec AO, Song Q, Cregan PB, Specht JE, Diers BW, Cannon SB. 2014. A roadmap for functional structural variants in the soybean genome. *G3: Genes, Genomes, Genetics* **4**: 1307–1318.

Bernard R. 1967. The inheritance of pod color in soybeans. *Journal of Heredity* **58**: 165–168.

Bernard R. 1972. Two genes affecting stem termination in soybeans. *Crop Science* **12**: 235–239.

Bhosale SU, Stich B, Rattunde HFW, Weltzien E, Haussmann BIG, Hash CT, Ramu P, Cuevas HE, Paterson AH, Melchinger AE *et al.* 2012. Association analysis of photoperiodic flowering time genes in west and central African sorghum [*Sorghum bicolor* (L.) Moench]. *BMC Plant Biology* **12**: 32.

Blackman BK, Strasburg JL, Raduski AR, Michaels SD, Rieseberg LH. 2010. The role of recently derived *FT* paralogs in sunflower domestication. *Current Biology* **20**: 629–635.

Bolon Y-T, Haun WJ, Xu WW, Grant D, Stacey MG, Nelson RT, Gerhardt DJ, Jeddeloh JA, Stacey G, Muehlbauer GJ. 2011. Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. *Plant Physiology* **156**: 240–253.

Cao D, Li Y, Lu S, Wang J, Nan H, Li X, Shi D, Fang C, Zhai H, Yuan X *et al.* 2015. *GmCOL1a* and *GmCOL1b* function as flowering repressors in soybean under long-day conditions. *Plant and Cell Physiology* **56**: 2409–2422.

Carmel-Goren L, Liu YS, Lifschitz E, Zamir D. 2003. The self-pruning gene family in tomato. *Plant Molecular Biology* **52**: 1215–1222.

Carter T Jr, Hymowitz T, Nelson R. 2004. Biogeography, local adaptation, Vavilov, and genetic diversity in soybean. In: Werner D, ed. *Biological resources and migration.* Berlin, Germany: Springer, 47–59.

Chen Y, Nelson RL. 2004. Evaluation and classification of leaflet shape and size in wild soybean. *Crop Science* **44**: 671–677.

Chung W-H, Jeong N, Kim J, Lee WK, Lee Y-G, Lee S-H, Yoon W, Kim J-H, Choi I-Y, Choi H-K. 2014. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Research* **21**: 153–167.

Clark RM, Linton E, Messing J, Doebley JF. 2004. Pattern of diversity in the genomic region near the maize domestication gene *tb1*. *Proceedings of the National Academy of Sciences, USA* **101**: 700–707.

Cober ER, Morrison MJ. 2010. Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. *Theoretical and Applied Genetics* **120**: 1005–1012.

Cook DE, Bayless AM, Wang K, Guo X, Song Q, Jiang J, Bent AF. 2014. Distinct copy number, coding sequence, and locus methylation patterns underlie *Rhg1*-mediated soybean resistance to soybean cyst nematode. *Plant Physiology* **165**: 630–647.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE. 2012. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* **338**: 1206–1209.

Csanadi G, Vollmann J, Stift G, Lelley T. 2001. Seed quality QTLs identified in a molecular map of early maturing soybean. *Theoretical and Applied Genetics* **103**: 912–919.

Cui Y, Barampuram S, Stacey MG, Hancock CN, Findley S, Mathieu M, Zhang Z, Parrott WA, Stacey G. 2013. *Tnt1* retrotransposon mutagenesis: a tool for soybean functional genomics. *Plant Physiology* **161**: 36–47.

Diers B, Keim P, Fehr W, Shoemaker R. 1992. RFLP analysis of soybean seed protein and oil content. *Theoretical and Applied Genetics* **83**: 608–612.

Dong Y, Wang YZ. 2015. Seed shattering: from models to crops. *Frontiers in Plant Science* **6**: 476.

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**: 3352.

Dong Y, Zhao L, Liu B, Wang Z, Jin Z, Sun H. 2004. The genetic diversity of cultivated soybean grown in China. *Theoretical and Applied Genetics* **108**: 931–936.

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**: 217–221.

Faure S, Turner AS, Gruszka D, Christodoulou V, Davis SJ, von Korff M, Laurie DA. 2012. Mutation at the circadian clock gene *EARLY MATURITY 8* adapts domesticated barley (*Hordeum vulgare*) to short growing seasons. *Proceedings of the National Academy of Sciences, USA* **109**: 8328–8333.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* **78**: 737–756.

Fukuda Y. 1933. Cytogenetical studies on the wild and cultivated Manchurian soybeans (*Glycine* L.). *Japanese Journal of Botany* **6**: 489–506.

Funatsuki H, Suzuki M, Hirose A, Inaba H, Yamada T, Hajika M, Komatsu K, Katayama T, Sayama T, Ishimoto M *et al.* 2014. Molecular basis of a shattering resistance boosting global dissemination of soybean. *Proceedings of the National Academy of Sciences, USA* **111**: 17797–17802.

Gizlice Z, Carter TE, Gerig TM, Burton JW. 1996. Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. *Crop Science* **36**: 753–765.

Guo J, Wang Y, Song C, Zhou J, Qiu L, Huang H, Wang Y. 2010. A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. *Annals of Botany* **106**: 505–514.

Guzman PS, Diers BW, Neece DJ, Martin SKS, Leroy AR, Grau CR, Hughes TJ, Nelson RL. 2007. QTL associated with yield in three backcross-derived populations of soybean. *Crop Science* **47**: 111–122.

Ha J, Abernathy B, Nelson W, Grant D, Wu X, Nguyen HT, Stacey G, Yu Y, Wing RA, Shoemaker RC. 2012. Integration of the draft sequence and physical map as a framework for genomic research in soybean (*Glycine max* (L.) Merr.) and wild soybean (*Glycine soja* Sieb. and Zucc.). *G3: Genes Genomes, Genetics* **2**: 321–329.

Han YP, Zhao X, Liu DY, Li YH, Lightfoot DA, Yang ZJ, Zhao L, Zhou G, Wang ZK, Huang L *et al.* 2016. Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytologist* **209**: 871–884.

Hancock CN, Zhang F, Floyd K, Richardson AO, LaFayette P, Tucker D, Wessler SR, Parrott WA. 2011. The rice miniature inverted repeat transposable element *mPing* is an effective insertional mutagen in soybean. *Plant Physiology* **157**: 552–562.

Hartman GL, West ED, Herman TK. 2011. Crops that feed the World 2. Soybean – worldwide production, use, and constraints caused by pathogens and pests. *Food Security* **3**: 5–17.

He Q, Yang H, Xiang S, Tian D, Wang W, Zhao T, Gai J. 2015. Fine mapping of the genetic locus *L1* conferring black pods using a chromosome segment substitution line population of soybean. *Plant Breeding* **134**: 437–445.

Ho P-t. 1975. *Cradle of the East: an enquiry into the indigenous origins of techniques and ideas of neolithic and early historic China, 5000–1000 B.C.* Hong Kong, China: Chinese University of Hong Kong.

Huang CL, Hung CY, Chiang YC, Hwang CC, Hsu TW, Huang CC, Hung KH, Tsai KC, Wang KH, Osada N *et al.* 2012a. Footprints of natural and artificial selection for photoperiod pathway genes in *Oryza*. *Plant Journal* **70**: 769–782.

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W. 2012b. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497–501.

Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**: 808–811.

Hymowitz T. 1970. On the domestication of the soybean. *Economic Botany* **24**: 408–421.

Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences, USA* **103**: 16666–16671.

Jang SJ, Sato M, Sato K, Jitsuyama Y, Fujino K, Mori H, Takahashi R, Benitez ER, Liu B, Yamada T *et al.* 2015. A single-nucleotide polymorphism in an endo-1,4-beta-glucanase gene controls seed coat permeability in soybean. *PLoS ONE* **10**: e0128527.

Jeong N, Suh SJ, Kim MH, Lee S, Moon JK, Kim HS, Jeong SC. 2012. *Ln* is a key regulator of leaflet shape and number of seeds per pod in soybean. *Plant Cell* **24**: 4807–4818.

Kanazawa A, Liu B, Kong F, Arase S, Abe J. 2009. Adaptive evolution involving gene duplication and insertion of a novel *Ty1/copia*-like retrotransposon in soybean. *Journal of Molecular Evolution* **69**: 164–175.

Katkout M, Sakuma S, Kawaura K, Ogihara Y. 2015. *TaqSH1-D*, wheat ortholog of rice seed shattering gene *qSH1*, maps to the interval of a rachis fragility QTL on chromosome 3DL of common wheat (*Triticum aestivum*). *Genetic Resources and Crop Evolution* **62**: 979–984.

Kiang Y. 1990. Linkage analysis of *Pgd1*, *Pgi1*, pod color (*L1*), and determinate stem (*dt1*) loci on soybean linkage group 5. *Journal of Heredity* **81**: 402–404.

Kim DH, Kim KH, Van K, Kim MY, Lee SH. 2010a. Fine mapping of a resistance gene to bacterial leaf pustule in soybean. *Theoretical and Applied Genetics* **120**: 1443–1450.

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

Kim MY, Shin JH, Kang YJ, Shim SR, Lee S-H. 2012a. Divergence of flowering genes in soybean. *Journal of Biosciences* **37**: 857–870.

Kim MY, Van K, Kang YJ, Kim KH, Lee S-H. 2012b. Tracing soybean domestication history: from nucleotide to genome. *Breeding Science* **61**: 445–452.

Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, Yano M. 2006. An SNP caused loss of seed shattering during rice domestication. *Science* **312**: 1392–1396.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B *et al.* 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* **42**: 1053–1059.

Langewisch T, Zhang H, Vincent R, Joshi T, Xu D, Bilyeu K. 2014. Major soybean maturity gene haplotypes revealed by SNPViz analysis of 72 sequenced soybean genomes. *PLoS ONE* **9**: e94150.

Lee G-A, Crawford GW, Liu L, Sasaki Y, Chen X. 2011. Archaeological soybean (*Glycine max*) in East Asia: does size matter? *PLoS ONE* **6**: e26720.

Lee KJ, Kim DS, Kim J-B, Jo S-H, Kang S-Y, Choi H-I, Ha B-K. 2016. Identification of candidate genes for an early-maturing soybean mutant by genome resequencing analysis. *Molecular Genetics and Genomics* **291**: 1561–1571.

Lee TG, Kumar I, Diers BW, Hudson ME. 2015. Evolution and selection of Rhg1, a copy-number variant nematode-resistance locus. *Molecular Ecology* **24**: 1774–1791.

Lee Y, Park T. 2006. Origin of legumes cultivation in Korean Peninsula by viewpoint of excavated grain remains and genetic diversity of legumes. *Korean Agricultural History Association* **5**: 1–31.

Lee SH, Park KY, Lee HS, Park EH, Boerma HR. 2001. Genetic mapping of QTLs conditioning soybean sprout yield and quality. *Theoretical and Applied Genetics* **103**: 702–709.

Li YH, Li W, Zhang C, Yang L, Chang RZ, Gaut BS, Qiu LJ. 2010. Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytologist* **188**: 242–253.

Li YH, Reif JC, Jackson SA, Ma YS, Chang RZ, Qiu LJ. 2014b. Detecting SNPs underlying domestication-related traits in soybean. *BMC Plant Biology* **14**: 251.

Li Y-h, S-c Zhao, J-x Ma, Li D, Yan L, Li J, Qi X-t, X-s Guo, Zhang L, W-m He. 2013. Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**: 579.

Li Y-h, Zhou G, Ma J, Jiang W, L-g Jin, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L. 2014a. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* **32**: 1045–1052.

Li CB, Zhou AL, Sang T. 2006. Rice domestication by reducing shattering. *Science* **311**: 1936–1939.

Lin Z, Li X, Shannon LM, Yeh C-T, Wang ML, Bai G, Peng Z, Li J, Trick HN, Clemente TE *et al.* 2012. Parallel domestication of the *Shattering1* genes in cereals. *Nature Genetics* **44**: 720–724.

Liu TF, Fang C, Ma YM, Shen YT, Li CC, Li Q, Wang M, Liu SL, Zhang JX, Zhou ZK *et al.* 2016. Global investigation of the co-evolution of *MIRNA* genes and microRNA targets during soybean domestication. *Plant Journal* **85**: 396–409.

Liu B, Fujita T, Yan ZH, Sakamoto S, Xu D, Abe J. 2007. QTL mapping of domestication-related traits in soybean (*Glycine max*). *Annals of Botany* **100**: 1027–1038.

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

Liu H, Liu H, Zhou L, Zhang Z, Zhang X, Wang M, Li H, Lin Z. 2015. Parallel domestication of the *Heading Date 1* gene in cereals. *Molecular Biology and Evolution* **32**: 2726–2737.

Liu B, Watanabe S, Uchiyama T, Kong F, Kanazawa A, Xia Z, Nagamatsu A, Arai M, Yamada T, Kitamura K. 2010. The soybean stem growth habit gene *Dt1* is an ortholog of Arabidopsis *TERMINAL FLOWER1*. *Plant Physiology* **153**: 198–210.

Mathieu M, Winters EK, Kong F, Wan J, Wang S, Eckert H, Luth D, Paz M, Donovan C, Zhang Z. 2009. Establishment of a soybean (*Glycine max* Merr. L) transposon-based mutagenesis repository. *Planta* **229**: 279–289.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM. 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology* **159**: 1295–1308.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews Genetics* **14**: 840–852.

Mir RR, Kudapa H, Srikanth S, Saxena RK, Sharma A, Azam S, Saxena K, Varma Penmetsa R, Varshney RK. 2014. Candidate gene analysis for determinacy in pigeonpea (*Cajanus* spp.). *Theoretical and Applied Genetics* **127**: 2663–2678.

Nakamura T, Yang D, Kalaiselvi S, Uematsu Y, Takahashi R. 2003. Genetic analysis of net-like cracking in soybean seed coats. *Euphytica* **133**: 179–184.

Obata H. 2011. *Jomon agriculture and paleoethnobotany in Northeast Asia.* Tokyo, Japan: Douseisya.

Obata H, Manabe A. 2011. Issues on the early agriculture in Korea and Japan, based on recent archaeobotanical studies. In: Kenkyukai KJ, eds. *Current research on the Neolithic period in Japan and Korea: Proceedings of the 9th Conference of the Kyushu Jomon Kenkyukai and the Korean Neolithic Research Society.* Iki, Japan: Kyushu Jomon Kenkyukai, 1–30.

Ogiso-Tanaka E, Matsubara K, Yamamoto S-i, Nonoue Y, Wu J, Fujisawa H, Ishikubo H, Tanaka T, Ando T, Matsumoto T. 2013. Natural variation of the *RICE FLOWERING LOCUS T 1* contributes to flowering time divergence in rice. *PLoS ONE* **8**: e75959.

Oikawa T, Koshioka M, Kojima K, Yoshida H, Kawata M. 2004. A role of *OsGA20ox1*, encoding an isoform of gibberellin 20-oxidase, for regulation of plant stature in rice. *Plant Molecular Biology* **55**: 687–700.

Orazaly M, Chen PY, Zeng AL, Zhang B. 2015. Identification and confirmation of quantitative trait loci associated with soybean seed hardness. *Crop Science* **55**: 688–694.

Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F *et al.* 1999. 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400**: 256–261.

Ping J, Liu Y, Sun L, Zhao M, Li Y, She M, Sui Y, Lin F, Liu X, Tang Z. 2014. *Dt2* is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. *Plant Cell* **26**: 2831–2842.

Pnueli L, Carmel-Goren L, Hareven D, Gutfinger T, Alvarez J, Ganal M, Zamir D, Lifschitz E. 1998. The *SELF-PRUNING* gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of *CEN* and *TFL1*. *Development* **125**: 1979–1989.

Putterill J, Robson F, Lee K, Simon R, Coupland G. 1995. The *CONSTANS* gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* **80**: 847–857.

Qi X, Li M-W, Xie M, Liu X, Ni M, Shao G, Song C, Yim AK-Y, Tao Y, Wong F-L. 2014. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nature Communications* **5**: 4340.

Qi Z-m, Wu Q, Han X, Sun Y-n, Du X-y, Liu C-y, Jiang H-w, Hu G-h, Chen Q-s. 2011. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. *Euphytica* **179**: 499–514.

Qiu J, Wang Y, Wu SL, Wang YY, Ye CY, Bai XF, Li ZF, Yan CH, Wang WD, Wang ZQ *et al.* 2014. Genome re-sequencing of semi-wild soybean reveals a complex *Soja* population structure and deep introgression. *PLoS ONE* **9**: e108479.

Repinski SL, Kwak M, Gepts P. 2012. The common bean growth habit gene *PvTFL1y* is a functional homolog of *Arabidopsis TFL1*. *Theoretical and Applied Genetics* **124**: 1539–1547.

Rousset M, Bonnin I, Remoue C, Falque M, Rhone B, Veyrieras JB, Madur D, Murigneux A, Balfourier F, Le Gouis J *et al.* 2011. Deciphering the genetics of flowering time by an association study on candidate genes in bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* **123**: 907–926.

Sayama T, Ono E, Takagi K, Takada Y, Horikawa M, Nakamoto Y, Hirose A, Sasama H, Ohashi M, Hasegawa H *et al.* 2012. The *Sg-1* glycosyltransferase locus regulates structural diversity of triterpenoid saponins of soybean. *Plant Cell* **24**: 2123–2138.

Schmitz RJ, He Y, Valdés-López O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G. 2013. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Research* **23**: 1663–1674.

Scoles G, Reinprecht Y, Poysa VW, Yu K, Rajcan I, Ablett GR, Pauls KP. 2006. Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. *Genome* **49**: 1510–1527.

Shimamoto Y, Abe J, Gao Z, Gai JY, Thseng FS. 2000. Characterizing the cytoplasmic diversity and phyletic relationship of Chinese landraces of soybean, *Glycine max*, based on RFLPs of chloroplast and mitochondrial DNA. *Genetic Resources and Crop Evolution* **47**: 611–617.

Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai YS, Gill BS, Faris JD. 2006. Molecular characterization of the major wheat domestication gene *Q*. *Genetics* **172**: 547–555.

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE* **8**: e54985.

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB. 2015. Fingerprinting soybean germplasm and its utility in genomic research. *G3: Genes, Genomes, Genetics* **5**: 1999–2006.

Stupar RM, Specht JE. 2013. Insights from the soybean (*Glycine max* and *Glycine soja*) genome: past, present, and future. *Advances in Agronomy* **118**: 177–204.

Sun L, Miao Z, Cai C, Zhang D, Zhao M, Wu Y, Zhang X, Swarm SA, Zhou L, Zhang ZJ *et al.* 2015. *GmHs1-1*, encoding a calcineurin-like protein, controls hard-seededness in soybean. *Nature Genetics* **47**: 939–943.

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research* **20**: 1689–1699.

Takahashi Y, Teshima KM, Yokoi S, Innan H, Shimamoto K. 2009. Variations in Hd1 proteins, *Hd3a* promoters, and *Ehd1* expression levels contribute to diversity of flowering time in cultivated rice. *Proceedings of the National Academy of Sciences, USA* **106**: 4555–4560.

Tian Z, Wang X, Lee R, Li Y, Specht JE, Nelson RL, McClean PE, Qiu L, Ma J. 2010. Artificial selection for determinate growth habit in soybean. *Proceedings of the National Academy of Sciences, USA* **107**: 8563–8568.

Toda K, Hirata K, Masuda R, Yasui T, Yamada T, Takahashi K, Nagaya T, Hajika M. 2015. Relationship between mutations of the pectin methylesterase gene in soybean and the hardness of cooked beans. *Journal of Agriculture and Food Chemistry* **63**: 8870–8878.

Toda K, Yang D, Yamanaka N, Watanabe S, Harada K, Takahashi R. 2002. A single-base deletion in soybean flavonoid 3′-hydroxylase gene is associated with gray pubescence color. *Plant Molecular Biology* **50**: 187–196.

Tsuji H, K-i Taoka, Shimamoto K. 2011. Regulation of flowering in rice: two florigen genes, a complex gene network, and natural variation. *Current Opinion in Plant Biology* **14**: 45–52.

Turck F, Fornara F, Coupland G. 2008. Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. *Annual Review of Plant Biology* **59**: 573–594.

Tuteja JH, Zabala G, Varala K, Hudson M, Vodkin LO. 2009. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. *Plant Cell* **21**: 3063–3077.

Valliyodan B, Dan Q, Patil G, Zeng P, Huang J, Dai L, Chen C, Li Y, Joshi T, Song L *et al.* 2016. Landscape of genomic diversity and trait discovery in soybean. *Scientific Reports* **6**: 23598.

Wang J, Chu S, Zhang H, Zhu Y, Cheng H, Yu D. 2016a. Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Scientific Reports* **6**: 20728.

Wang Y, Gu Y, Gao H, Qiu L, Chang R, Chen S, He C. 2016b. Molecular and geographic evolutionary support for the essential role of *GIGANTEAa* in soybean domestication of flowering time. *BMC Evolutionary Biology* **16**: 79.

Wang KJ, Li XH. 2011. Interspecific gene flow and the origin of semi-wild soybean revealed by capturing the natural occurrence of introgression between wild and cultivated soybean populations. *Plant Breeding* **130**: 117–127.

Wang X, Li Y, Zhang H, Sun G, Zhang W, Qiu L. 2015. Evolution and association analysis of *GmCYP78A10* gene with seed size/weight and pod number in soybean. *Molecular Biology Reports* **42**: 489–496.

Wang Y, Lu J, Chen S, Shu L, Palmer RG, Xing G, Li Y, Yang S, Yu D, Zhao T. 2014. Exploration of presence/absence variation and corresponding polymorphic markers in soybean genome. *Journal of Integrative Plant Biology* **56**: 1009–1019.

Watanabe S, Hideshima R, Xia Z, Tsubokura Y, Sato S, Nakamoto Y, Yamanaka N, Takahashi R, Ishimoto M, Anai T. 2009. Map-based cloning of the gene associated with the soybean maturity locus *E3*. *Genetics* **182**: 1251–1262.

Watanabe S, Xia Z, Hideshima R, Tsubokura Y, Sato S, Yamanaka N, Takahashi R, Anai T, Tabata S, Kitamura K. 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.

Weller JL, Liew LC, Hecht VF, Rajandran V, Laurie RE, Ridge S, Wenden B, Vander Schoor JK, Jaminon O, Blassiau C *et al.* 2012. A conserved molecular basis for photoperiod adaptation in two temperate legumes. *Proceedings of the National Academy of Sciences, USA* **109**: 21158–21163.

Wickland DP, Hanzawa Y. 2015. The *FLOWERING LOCUS T/ TERMINAL FLOWER 1* gene family: functional evolution and molecular mechanisms. *Molecular Plant* **8**: 983–997.

Woo JW, Kim J, Kwon SI, Corvalan C, Cho SW, Kim H, Kim SG, Kim ST, Choe S, Kim JS. 2015. DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nature Biotechnology* **33**: 1162–1164.

Woodworth C, Veatch C. 1929. Inheritance of pubescence in soy beans and its relation to pod color. *Genetics* **14**: 512.

Wu F-Q, Fan C-M, Zhang X-M, Fu Y-F. 2013. The phytochrome gene family in soybean and a dominant negative effect of a soybean *PHYA* transgene on endogenous Arabidopsis PHYA. *Plant Cell Reports* **32**: 1879–1890.

Wu F, Price BW, Haider W, Seufferheld G, Nelson R, Hanzawa Y. 2014. Functional and evolutionary characterization of the *CONSTANS* gene family in short-day photoperiodic flowering in soybean. *PLoS ONE* **9**: e85754.

Xia Z, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, Anai T, Sato S, Yamazaki T, Lü S. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. *Proceedings of the National Academy of Sciences, USA* **109**: E2155–E2164.

Xie QG, Lou P, Hermand V, Aman R, Park HJ, Yun DJ, Kim WY, Salmela MJ, Ewers BE, Weinig C *et al.* 2015. Allelic polymorphism of *GIGANTEA* is responsible for naturally occurring variation in circadian period in *Brassica rapa*. *Proceedings of the National Academy of Sciences, USA* **112**: 3829–3834.

Xu D, Abe J, Gai J, Shimamoto Y. 2002. Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. *Theoretical and Applied Genetics* **105**: 645–653.

Zabala G, Vodkin LO. 2007. A rearrangement resulting in small tandem repeats in the F3′5′H gene of white flower genotypes is associated with the soybean *W1* locus. *Crop Science* **47**(S2): S-113–S-124.

Zhai H, Lu S, Wang Y, Chen X, Ren H, Yang J, Cheng W, Zong C, Gu H, Qiu H *et al.* 2014. Allelic variations at four major maturity *E* genes and transcriptional abundance of the *E1* gene are associated with flowering time and maturity of soybean cultivars. *PLoS ONE* **9**: e97636.

Zhang Q, Li H, Li R, Hu R, Fan C, Chen F, Wang Z, Liu X, Fu Y, Lin C. 2008. Association of the circadian rhythmic expression of GmCRY1a with a latitudinal cline in photoperiodic flowering of soybean. *Proceedings of the National Academy of Sciences, USA* **105**: 21028–21033.

Zhang Y, Liang Z, Zong Y, Wang Y, Liu J, Chen K, Qiu JL, Gao C. 2016. Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nature Communications* **7**: 12617.

Zhao Z. 2004. Floatation: a paleobotanic method in field archaeology. *Archaeology* **3**: 80–87.

Zhao B, Dai A, Wei H, Yang S, Wang B, Jiang N, Feng X. 2016a. Arabidopsis *KLU* homologue *GmCYP78A72* regulates seed size in soybean. *Plant Molecular Biology* **90**: 33–47.

Zhao C, Takeshima R, Zhu J, Xu M, Sato M, Watanabe S, Kanazawa A, Liu B, Kong F, Yamada T *et al.* 2016b. A recessive allele for delayed flowering at the soybean maturity locus *E9* is a leaky allele of *FT2a*, a *FLOWERING LOCUS T* ortholog. *BMC Plant Biology* **16**: 20.

Zhao SC, Zheng FY, He WM, Wu HY, Pan SK, Lam HM. 2015. Impacts of nucleotide fixation during soybean domestication and improvement. *BMC Plant Biology* **15**: 81.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y. 2015b. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* **33**: 408–414.

Zhou L, Wang SB, Jian J, Geng QC, Wen J, Song Q, Wu Z, Li GJ, Liu YQ, Dunwell JM *et al.* 2015a. Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. *Scientific Reports* **5**: 9350.