

Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean

Zhengkui Zhou^{1,7}, Yu Jiang^{2,3,7}, Zheng Wang^{1,7}, Zhiheng Gou^{2,4,7}, Jun Lyu^{2,7}, Weiyu Li^{1,7}, Yanjun Yu⁵, Liping Shu², Yingjun Zhao^{2,4}, Yanming Ma¹, Chao Fang^{1,4}, Yanting Shen^{1,4}, Tengfei Liu^{1,4}, Congcong Li^{1,4}, Qing Li^{1,4}, Mian Wu¹, Min Wang^{1,4}, Yunshuai Wu¹, Yang Dong², Wenting Wan², Xiao Wang^{2,4}, Zhaoli Ding², Yuedong Gao², Hui Xiang², Baoge Zhu¹, Suk-Ha Lee⁶, Wen Wang² & Zhixi Tian¹

Understanding soybean (*Glycine max*) domestication and improvement at a genetic level is important to inform future efforts to further improve a crop that provides the world's main source of oilseed. We detect 230 selective sweeps and 162 selected copy number variants by analysis of 302 resequenced wild, landrace and improved soybean accessions at >11× depth. A genome-wide association study using these new sequences reveals associations between 10 selected regions and 9 domestication or improvement traits, and identifies 13 previously uncharacterized loci for agronomic traits including oil content, plant height and pubescence form. Combined with previous quantitative trait loci (QTL) information, we find that, of the 230 selected regions, 96 correlate with reported oil QTLs and 21 contain fatty acid biosynthesis genes. Moreover, we observe that some traits and loci are associated with geographical regions, which shows that soybean populations are structured geographically. This study provides resources for genomics-enabled improvements in soybean breeding.

Soybean (*Glycine max* [L.] Merr.) is a crop with substantial economic value, accounting for more than half of global oilseed production¹. It has been suggested that cultivated soybean was domesticated from wild soybean (*G. soja* Sieb. & Zucc.) in China 5,000 years ago² and may have been introduced to Korea, and then to Japan approximately 2,000 years ago, to North America in 1765, and to Central and South America during the first half of the last century¹.

Generally, a limited number of the best lines are used for breeding the next generation, which greatly reduces the genetic diversity³. An analysis of 111 fragments from 102 genes showed that several genetic diversity bottlenecks have occurred during soybean domestication and improvement⁴. The detection of genome-wide genetic diversity and the identification of genes contributing to domestication and improvement are essential for breeding superior varieties^{5–7}. So far, less than ten genes have been associated with domestication or improvement traits in soybeans (reviewed by Xia *et al.*⁸). Based on the soybean reference genome⁹, resequencing of dozens of soybean accessions led to an initial understanding of genetic variation patterns during soybean domestication and varietal improvement^{10–12}. However, the small sample sizes have hindered the correlation of selected loci with domestication and improvement traits. This is because a sufficient number of samples are required for a genome-wide association study (GWAS) to link genetic variations with a trait¹³, for instance,

115 and 278 lines were used in cucumber⁷ and maize⁶, respectively. In addition, analyses of more samples are needed to probe how soybean subpopulations have adapted to different geographic areas¹⁴ and to identify the large number of rare alleles that have been lost during soybean domestication and improvement⁴. These rare alleles could benefit future soybean improvement as they constitute excellent genetic variances, especially for resistant genes⁴. Therefore, sufficient wild and cultivated soybean genome sequences to cover all geographic areas and represent the population structure of each subpopulation are necessary for a comprehensive investigation of genetic structure and genes related to selection and breeding.

In this study, we resequenced 302 soybean accessions to more than 11× depth and analyzed genomic variation dynamics during soybean domestication, improvement and local breeding. This endeavor enabled the confirmation of and identification of multiple loci and genes for important agronomic traits. This genetic resource should facilitate future soybean cultivar breeding.

RESULTS

Genomic variation

A total of 302 soybean accessions, including 62 wild soybeans (*G. soja*), 130 landraces and 110 improved cultivars, were used in this study. The samples included 93 representative diverse accessions

¹State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China.

²State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ³College of Animal Science and Technology, Northwest A&F University, Yangling, China. ⁴University of Chinese Academy of Sciences, Beijing, China. ⁵Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. ⁶Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Korea.

⁷These authors contributed equally to this work. Correspondence should be addressed to Z.T. (zxtian@genetics.ac.cn) or W.W. (wwang@mail.kiz.ac.cn).

Received 10 April 2014; accepted 11 November 2014; published online 2 February 2015; corrected after print 4 February 2016; doi:10.1038/nbt.3096

Figure 1 Geographic distribution and population structure of 302 soybean accessions. **(a)** The geographic distribution of the 302 accessions, each of which is represented by a dot on the world map. **(b)** Phylogenetic tree of all accessions inferred from whole-genome SNPs, with *Medicago truncatula* as an outgroup. The layer rings indicate the group name of each clade. **(c)** PCA plots of the first two components of 302 accessions. **(d)** The geographic origin of each accession in the eight clades. J&K, Japan and Korea; SC, southern China; NC, northern China; NEC, northeastern China; NUS, northern United States; SUS, southern United States; CAN, Canada; Other, other country. Geographic region information is provided in **Supplementary Note 1**.

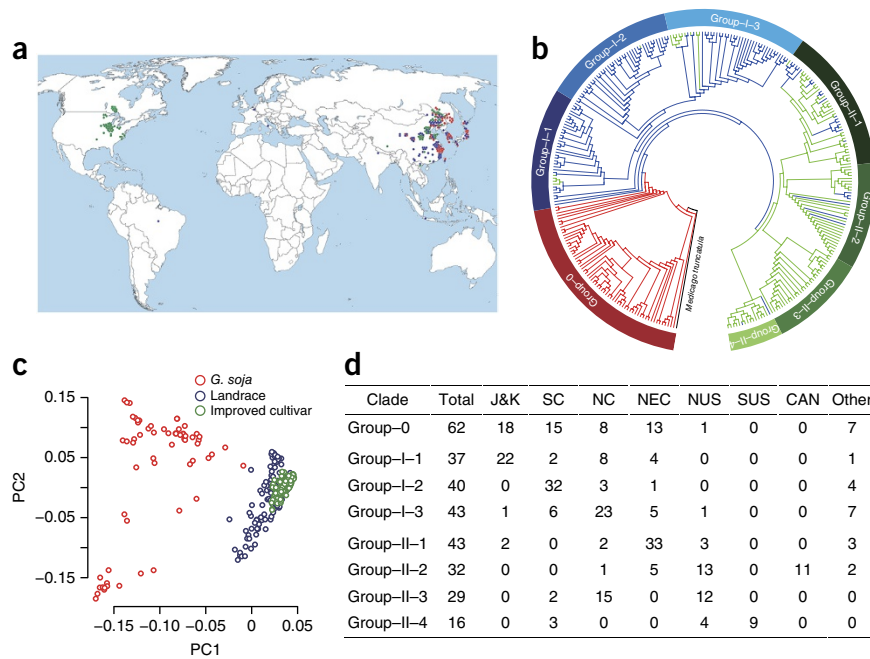
previously analyzed by Hyten *et al.*⁴ and 39 *G. soja*, 86 landraces and 84 improved cultivars from different countries or regions within countries (**Supplementary Table 1**). The geographic distributions of the accessions included in this report include China, Korea, Japan, Russia, the United States and Canada (**Fig. 1a** and **Supplementary Table 1**).

Resequencing of the 302 soybean accessions by an Illumina HiSeq 2000 sequencer generated a total of 33 billion paired-end reads of 100 bp in length (3.3 Tb of sequences), with an average coverage depth of more than 11× for each line. After mapping against the soybean Williams 82 reference genome⁹, we identified 9,790,744 single-nucleotide polymorphisms (SNPs) and 876,799 indels (shorter than or equal to 6 bp) (**Supplementary Table 2**). In addition, the depth of sequencing data allowed us to identify a total of 1,614 copy number variations (CNVs) and 6,388 segmental deletions comprising 15.14 Mb and 73.6 Mb sequences, respectively.

Considering the genomic diversity among wild and domesticated soybean, we performed additional mapping for our resequenced reads using the reported *G. soja* (*G. soja* var. IT182932) genome-specific sequences¹⁵ as references. The mapping ratio of the wild soybean accessions to the *G. soja* specific sequences was higher than that of the cultivated soybeans, and the mapping ratio of the landraces was slightly higher than that of the improved cultivars (**Supplementary Fig. 1a,b**). In particular, half of the annotated resistance-related sequences in *G. soja* (**Supplementary Table 3**) were lost in both the landrace and improved cultivar samples (**Supplementary Fig. 1c**).

Soybean population structure and linkage disequilibrium

Using *Medicago truncatula* as an outgroup, we explored the phylogenetic relationships among the 302 accessions through whole-genome SNP analysis (**Fig. 1b** and **Supplementary Fig. 2**). All wild lines clustered together although a few lines were also close to some cultivars as revealed by the phylogenetic tree. We therefore put all the wild soybean lines in one group, group 0. Together with principal component analysis (PCA) (**Fig. 1c**), these results support the hypothesis that all currently grown domesticated soybeans originated from a single domestication event¹⁶. Then the domesticated accessions were grouped into two main clades, designated as group I and group II. They included both landraces and improved cultivars, with group I biased to landraces and group II biased to improved cultivars. The genetic diversity (π) decreased from 2.94×10^{-3} in *G. soja* to 1.40×10^{-3} in landraces and to 1.05×10^{-3} in improved cultivars (**Supplementary Table 2**), suggesting that approximately half of the genetic diversity has been lost during soybean



domestication from *G. soja* to landraces. However, most genetic diversity was retained during improvement from landraces to improved cultivars. The genetic diversity in group 0, group I and group II were 2.94×10^{-3} , 1.44×10^{-3} and 1.10×10^{-3} , respectively.

Accessions in each domesticated group were further classified into several subclades, which exhibited strong geographic distribution patterns (**Fig. 1b,d**). Group I was classified into three subclades. Group I-1 mainly contained landraces from Japan and Korea as well as several landraces and improved cultivars from different areas of China; group I-2 and group I-3 mainly consisted of landraces from southern China and northern China, respectively. Group II was classified into four subclades. Group II-1 mainly consisted of the landraces and improved cultivars from northeastern China; group II-2 consisted of the landraces from northeastern China and improved cultivars from the northern United States and Canada. Group II-3 included improved cultivars from the northern United States and the improved cultivars from northern China, and group II-4 consisted of the landraces and improved cultivars from southern China and the improved cultivars from the southern United States (SUS). The different subgroups exhibited variations in genetic diversity (π) and population structure (**Supplementary Fig. 3** and **Supplementary Table 4**). Overall, group II had a lower within-subgroup genetic diversity than group I. The geographic clustering pattern from the population structure and phylogenetic tree shed light on the introduction and development processes during modern elite cultivar breeding. We also found that some cultivated soybeans have mixed ancestry, indicating that these lines might have experienced introgression or gene flow during breeding.

Linkage disequilibrium (LD; indicated by r^2) dropped to half of its maximum value at 420 kb for all the samples (**Fig. 2a**) but with variations among different populations. The LD extent in wild soybean was ~27 kb, similar to that of wild rice (*Oryza rufipogon*, 20 kb)¹⁷ and wild maize (*Z. mays ssp. parviglumis*, 22 kb)⁵. In the soybean landraces and improved cultivars, LD increased to 83 kb and 133 kb, respectively (**Fig. 2a**), similar to that of cultivated rice (123 kb and 167 kb in *indica* and *japonica*, respectively)¹⁸ but much higher than cultivated maize (30 kb)⁵. All three populations had low background LD values, but wild soybean had a relatively higher value (**Supplementary Fig. 4**).

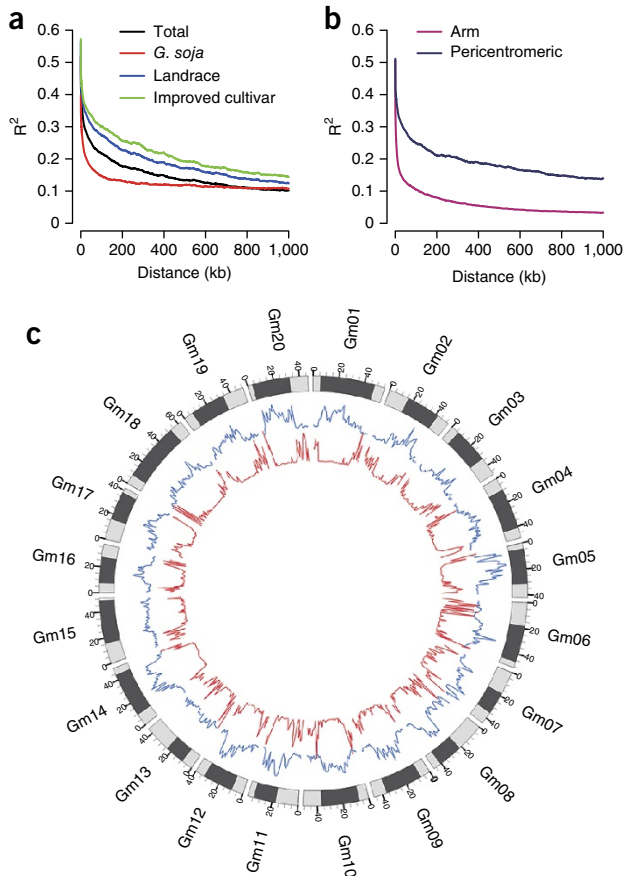


Figure 2 Decay of linkage disequilibrium (LD) in soybean genome. (a) Decay of LD in all samples, *G. soja*, landrace and improved cultivars. (b) Decay of LD in chromosome arms and pericentromeric regions for all samples. (c) Decay of LD (blue line) and recombination rate (red line) in contiguous 1 Mb subregions across soybean genome. The decay of LD was measured by r^2 . The outer rings indicate individual chromosomes, and the dark areas represent pericentromeric regions.

The LD decay in individual subgroups also exhibited variations, with the highest value found in group II–3 (Supplementary Fig. 5a).

LD decay is variable in different genomic regions¹⁹. We found that the LD of pericentromeric regions was much higher than that of arm regions (Fig. 2b and Supplementary Fig. 5b). Further investigation using 1 Mb contiguous subregions across each chromosome showed that LD was negatively correlated with genetic recombination rates ($r = -0.47$, $P = 1 \times 10^{-4}$) (Fig. 2c). We also observed that the LD extent for different chromosomes was variable (Supplementary Fig. 6 and Supplementary Table 5).

Selection signals during domestication and improvement

To identify potential selective signals during soybean domestication (wild soybeans versus landraces) and improvement (landraces versus improved cultivars), we scanned genomic regions with extreme allele frequency differentiation over extended linked regions using a likelihood method (the cross-population composite likelihood ratio, XP-CLR)²⁰. A total of 121 domestication-selective sweeps (Fig. 3a and Supplementary Table 6) and 109 improvement-selective sweeps (Fig. 3f and Supplementary Table 7) were detected.

In addition to SNPs, CNVs may be targets of artificial selection, which results in CNV frequency differences among populations. We adopted a statistical parameter, relative frequency difference

(RFD), to identify these CNVs. We used the highest 5% RFD value as the threshold for identifying potentially selected CNVs. In total, there were 162 potentially selected CNVs during domestication and improvement (Fig. 3i, Supplementary Fig. 7, Supplementary Tables 8 and 9).

Morphological features of soybeans were selected during domestication and improvement, resulting in significant variations among different populations. For instance, wild soybeans have small, coarse black seeds; landraces have large seeds with variable colors (black, yellow, green or striped); and improved cultivars have large shiny yellow seeds (Supplementary Fig. 8), suggesting seed size was mainly selected during domestication, whereas seed color was uniformly selected during improvement. These results indicated that dominant selection of different traits might have occurred at different evolutionary stages.

In our analyses, selection signals were detected in almost all reported domestication-related QTL regions²¹ (Fig. 3 and Supplementary Table 10). However, the sizes of the selected regions were smaller than the QTLs (Supplementary Fig. 9a and Supplementary Table 10). For instance, a 190-kb region that contains only 14 genes responsible for pod dehiscence was identified (Supplementary Fig. 9b), whereas the previous QTL region spanned 12 Mb. These defined regions will be helpful in identifying genes that govern domestication-related traits.

From 2011 to 2013, we assayed domestication-related morphological features of the accessions used in this study by surveying the plant height, flower color, stem determinacy, seed weight, seed coat color, hilum color and pubescence form (Supplementary Fig. 10). To further annotate the selected regions, we performed GWAS for these domestication traits (Supplementary Table 11). GWAS signals associated with stem determinacy and seed weight were detected at the previously reported *Dt1* locus on Chr. 19 (Fig. 3e and Supplementary Fig. 11a) and the *qSW* locus on Chr. 17 (Fig. 3d and Supplementary Fig. 11b), respectively²¹. We also detected a GWAS signal responsible for purple or white flower color at the *W1* locus²² and a GWAS signal corresponding to seed coat color variation at the *I* locus^{23–25}. We found that these two signals underwent selection during domestication and improvement, respectively (Supplementary Fig. 11c and Fig. 3g). In addition to these previously characterized loci, we identified two new GWAS signals responsible for seed coat color (Supplementary Fig. 11c and Supplementary Table 11), five new GWAS signals responsible for pubescence form (Supplementary Fig. 11d and Supplementary Table 11) and a new GWAS signal responsible for flower color (Supplementary Fig. 11e and Supplementary Table 11). The strongest GWAS signal responsible for pubescence form overlapped with a selective sweep originating on Chr. 18 during improvement (Fig. 3h).

Further, we performed GWAS using the CNV data, which allowed us to directly identify the genes controlling cyst nematode resistance and hilum color (Supplementary Table 12 and Supplementary Fig. 12). We found that a CNV on Chr. 18 was significantly associated with cyst nematode resistance and exactly overlapped with the reported *Rhg1* location²⁶ (Fig. 3l and Supplementary Fig. 12c). Interestingly, this signal experienced selection during domestication. We detected a strong GWAS CNV signal for hilum color on Chr. 08 and found that it was located exactly at a chalcone synthase cluster (Fig. 3j and Supplementary Fig. 12a). This GWAS signal also overlapped with a selected CNV during soybean domestication. For plant height, we detected four new GWAS CNV signals (Fig. 3k, Supplementary Table 12 and Supplementary Fig. 12b). One of the GWAS signals overlapped with a strong CNV selection signal during soybean domestication.



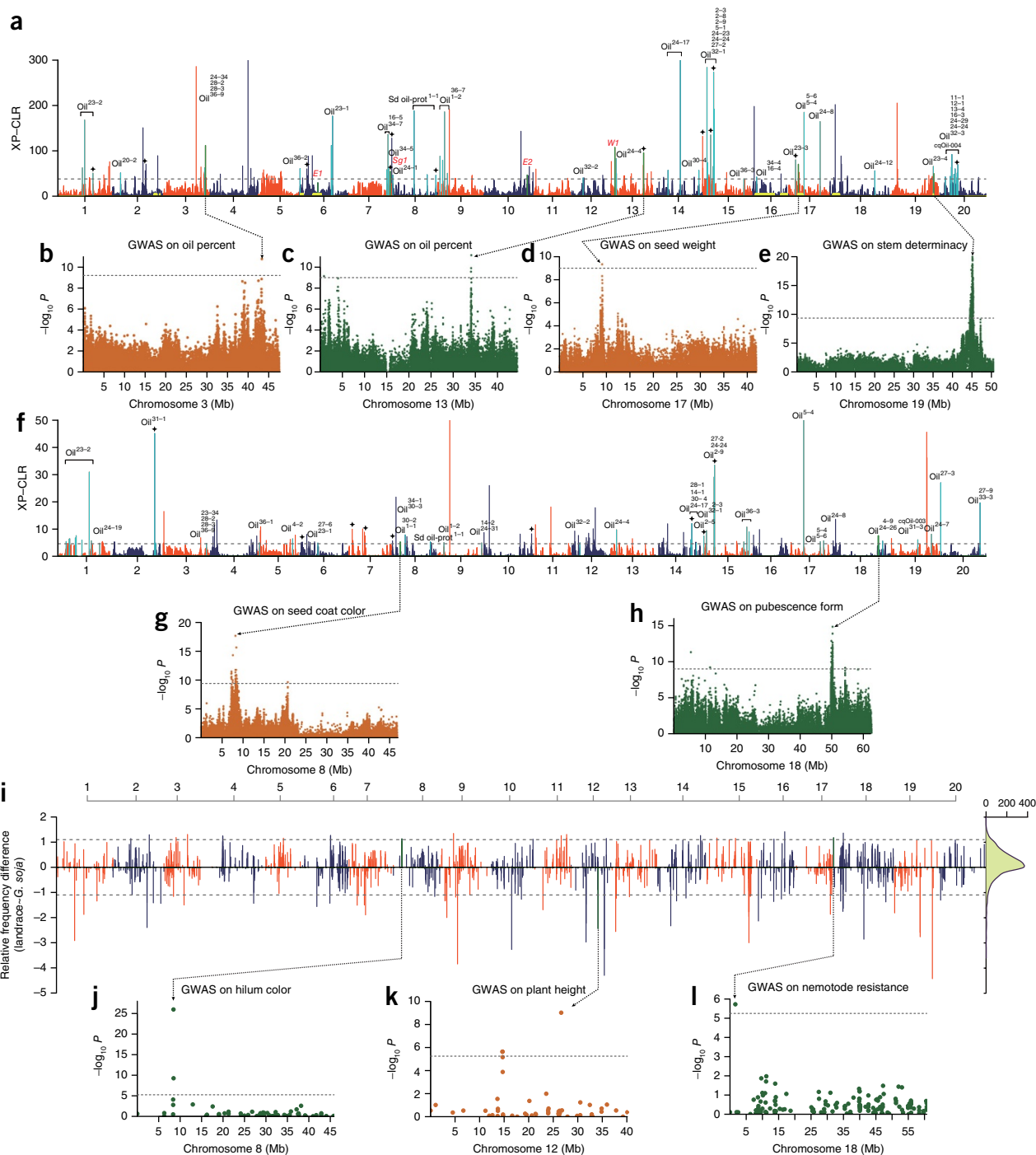
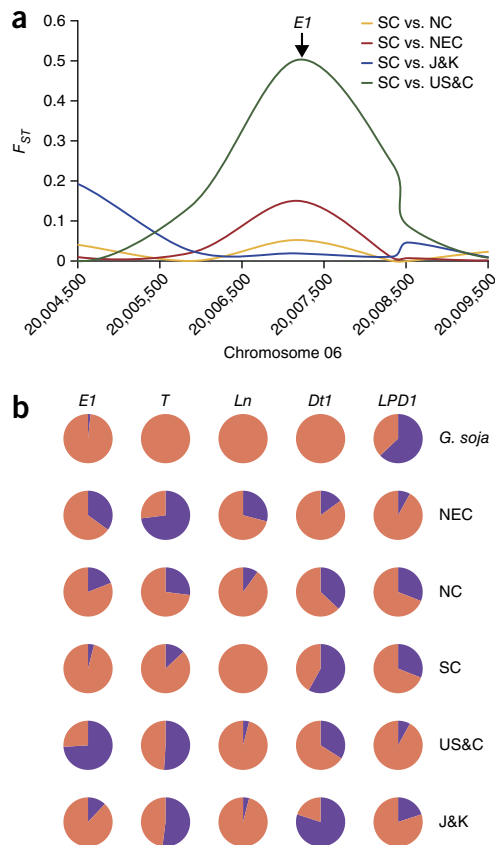


Figure 3 Genome-wide screening and functional annotations of selected regions during domestication and improvement. (a) and (f) show the whole-genome screening of selective signals in domestication and improvement, respectively. The XP-CLR values are plotted against the position on each of the 20 chromosomes. The horizontal dashed lines indicate the genome-wide threshold of selection signals. For domestication (a), the threshold is $XP-CLR \geq 38$, and for improvement (f), the threshold is $XP-CLR \geq 4.6$. The previously reported QTL regions²¹, which mapped 9 domestication-related traits in soybean (Supplementary Table 10), are marked with yellow bars above the x-axes. Four known causal genes that overlapped selective signals—*E1* (ref. 29), *Sg1* (ref. 43), *E2* (ref. 44) and *W1* (ref. 22) are labeled in red font, and the *E1* signal is moderately lower than the threshold. Selection signals that overlap soybean oil QTLs (Supplementary Table 11) are labeled with turquoise bars, and the QTL names assigned by SoyBase⁴⁵ (<http://www.soybase.org/>) are marked above. Selective sweeps overlapping characterized GWAS loci and known causal genes are shown in dark green. b–e, g and h show six GWAS results that overlapped strong selective signals. The gray horizontal dashed lines indicate the Bonferroni significance threshold of GWAS (1×10^{-9}). The 21 soybean fatty acid biosynthesis genes overlapped by selective sweeps are denoted by the stars above the line. (i) Genome-wide screening and functional annotations of selected copy number variations (CNVs) during domestication. The relative frequency difference (RFD) between landraces and *G. soja* is plotted against the position on each of the 20 chromosomes. The two horizontal dashed lines indicate the genome-wide threshold of selection signals, which showed the highest 5% absolute RFD value (≥ 1.10). Selected CNVs that overlapped with characterized GWAS loci are shown in dark green. (j–l) Three GWAS results that overlapped with the strong selective CNVs are shown: (j) hilum color, (k) plant height and (l) nematode resistance. The Bonferroni significance threshold (5.3×10^{-6}) is indicated by gray horizontal dashed lines.



Figure 4 Population genetics differentiation (F_{ST}) around the *E1* gene and five gene allele frequency differences across *G. soja* and five geographic groups. (a) The F_{ST} value from southern China (SC) relative to five geographic groups: northeastern China (NEC), northern China (NC), Japan & Korea (J&K) and the United States & Canada (US&C). (b) The spectrum of allele frequencies at the causal polymorphisms of *E1*, *T*, *Ln*, *Dt1* and *LPD1* in *G. soja* and five geographic groups.



Selected regions relevant to oil content

In addition to changes in morphology, cultivated and wild soybeans have different seed oil content. Typically, wild soybean seeds have lower oil content (Supplementary Fig. 13a) (calculated based on the publicly available resources from the US Department of Agriculture (USDA) GRIN database; <http://www.ars-grin.gov/>). The difference likely results from the fact that oil content is a priority in soybean breeding, imposing selective pressure on the genes/loci controlling oil biosynthesis. Several relevant oil QTLs have been reported (references are listed in Supplementary Table 13). When the physical loci of the oil QTLs were compared with the 230 selective sweeps, we found that 53 domestication-selective sweeps and 43 improvement-selective sweeps were located within known oil QTL regions (Fig. 3 and Supplementary Table 13). The fatty acid biosynthesis pathway is composed of complex steps, and the modification of fatty acid biosynthesis genes regulates oil content or the ratio of oil components^{27,28}. We identified 21 fatty acid biosynthesis genes in the selected regions (Supplementary Table 14). Moreover, 10 of the 21 fatty acid biosynthesis genes in the selected regions overlapped with oil content QTLs (Supplementary Table 13 and Fig. 3).

To further identify selected regions potentially related to oil content, we carried out a GWAS for oil content using 175 of the resequenced lines that had oil content records in the USDA GRIN database (<http://www.ars-grin.gov/>). We detected six strong GWAS signals (Supplementary Fig. 13b), five of which overlapped with previously identified oil content QTLs, and one that was newly identified by this study (Supplementary Table 11). We determined that two of the six GWAS signals, located on Chr. 13 and Chr. 03, overlapped with domestication-selective sweeps (Fig. 3c). The allelic distributions of the highest significant association sites also exhibited frequency difference between *G. soja* and cultivated soybeans (Supplementary Fig. 13c), supporting the hypothesis that these alleles have experienced selection. These results will be valuable for the functional characterization of genes responsible for oil content.

Local breeding and related traits

Our phylogenetic analysis showed that genetically close domesticated accessions tended to have close geographic origins (Fig. 1c). Similarly, many agronomic traits showed geographic distribution patterns. For example, most varieties in northeastern China showed indeterminate stems and a narrow leaf shape, whereas the varieties in southern China showed determinate stems and an ovate leaf shape. Thus, some traits and their underlying genes underwent selection during geographic differentiation or local breeding. We calculated the pairwise population differentiation level (F_{ST}) across different geographic groups (Supplementary Table 15 and Supplementary Fig. 14).

We identified some local differentiation signals that have not been detected previously in the screening of domestication and improvement selection, such as the *E1* gene, which is responsible for the major flowering time (*FT*) QTL^{29,30}. Typically, soybeans in northern regions have shorter maturity periods, whereas those in southern regions have

longer maturity periods³¹. In the domestication and improvement selection screening, there was only a weak selection signal at the *FT* locus (Fig. 3). However, through F_{ST} analysis of pairwise population differentiation, we detected a strong signal at the *E1* gene locus in accessions from southern China versus United States and Canada (Fig. 4a). Further investigation suggested that the mutant allele was mainly distributed in the high latitude regions, such as the United States and Canada, northeastern China, and Japan & Korea, and a few in southern China (Fig. 4b), which is consistent with the maturity period of different geographical areas.

Similarly, we found regional differentiation in the *Ln* gene, which is responsible for soybean leaf shape and the four-seed pod^{32,33}, and the *T* locus, which regulates pubescence color^{34,35}, that were not detected in the screening for domestication and improvement selection described above. The mutant *Ln* allele was mainly distributed in northeastern China and northern China (Fig. 4b), consistent with the geographical distribution of its characteristic leaflet shape³⁶. The *T* non-sense allele increased in frequency from southern to northern China, which may be related to the chill adaptation of the *T* gene³⁷.

In addition, some traits that experienced domestication and improvement selection were also detected by the geography differentiation analysis. *Dt1*, which regulates stem determinacy, showed obvious regional differentiation, as the mutant allele increased in frequency from the northern to the southern regions (Fig. 4b). Interestingly, we found that an oil content-related gene—*LPD1*, which encodes lipoamide dehydrogenase 1—appears to have experienced selection during improvement (Fig. 3) and exhibited a geographic distribution pattern as well (Fig. 4b).



DISCUSSION

It is predicted that current crop production must be doubled by 2050 to meet the food consumption demands of an increasing world population^{38,39}. To meet this requirement, crop yields need to increase at a rate of at least 2.4% per year⁴⁰. However, the low levels of genetic variation in most modern crop varieties greatly hamper the breeding of superior crop varieties⁴¹. Soybean is eaten by humans and animals and accounts for approximately 56% of global oil-seed production¹. The current rate of yield increase in soybean is approximately 1.3% per year, which is far from enough to meet the predicted requirement⁴⁰. A comprehensive evaluation and utilization of large-scale representative germplasm is important for crop improvement through identification and access to allelic variations affecting the crop phenotype. In this study, through genomic and GWAS analyses of large-scale populations, we successfully characterized several selective signals related to domestication and improvement traits. The allele distributions of these loci/genes in different populations constitute a valuable resource that can be used for design of future breeding strategies.

In addition, we have narrowed down the selective sweeps corresponding to the main domestication traits into small regions, which will be helpful for future characterization and determination of new soybean domestication genes. Research on other crops, for example, gramineous crops, have revealed interesting general patterns regarding crop domestication, such as the similar genetic basis underlying the same domestication traits and the tendency of transcription factors and a few key genes to participate in domestication⁴². Because legume crops and gramineous crops are both seed crops, it would be interesting to study whether convergent gene networks or different genetic mechanisms underlie the domestication of the two types of seed crops. With a more complete set of reliable soybean domestication genes, it should be possible, in future, to assess whether these general patterns still hold true for a wider breadth of species.

In conclusion, this study provides a resource to improve our understanding of the genetics of soybean domestication and to inform future studies on the allelic variation of relevant traits within genetic resource collections, thereby enabling soybean crop improvement.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Accession codes. SRA: [SRP045129](#). NCBI dbSNP: under the two batches: http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1061960 and http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1061961. The SNPs, indels and CNV fragment deletions have also been deposited into Figshare database (http://figshare.com/articles/Soybean_resequencing_project/1176133).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the USDA GRIN database, SoyBase, the Platform of National Crop Germplasm Resources of China, and L. Qiu for providing publicly available resources and R. Nelson at the University of Illinois for sharing some of the phenotype data. We thank X. Liu (BGI-Shenzhen) for helping to upload sequence data to NCBI. This work was supported by the National Key Basic Research Program (No. 2013CB835200), National Natural Science Foundation of China (grant nos. 31222042 and 91131005), “Strategic Priority Research Program” of the Chinese Academy of Sciences (grant no. XDA08020202), and “One-hundred talents” Startup Funds from Chinese Academy of Sciences for Z.T.

AUTHOR CONTRIBUTIONS

Z.T., W. Wang, and Z.Z. designed the experiments and managed the project. Z.Z., Y.J., Z.G., J.L., Y.S., T.L., S.-H.L., Z.T. and W. Wang performed the

data analysis. Z. Wang, W.L., Y.Y., Y.M., C.F., C.L., Q.L., M. Wu, M. Wang, Y. Wu and B.Z. performed the phenotyping and prepared DNA samples. Z.G., Y.D., W. Wan, X. Wang, Z.D., Y.G. and H.X. performed the genome sequencing. Z.T., W. Wang, Z.Z., Z.G., J.L. and Y.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

- Wilson, R.F. *Soybean: Market Driven Research Needs in Genetics and Genomics of Soybean* (Springer, 2008).
- Carter, T.E., Nelson, R., Sneller, C.H. & Cui, Z. *Soybeans: Improvement, Production and Uses* 3rd edn. (Madison, Wisconsin, USA, 2004).
- Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
- Hyten, D.L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA* **103**, 16666–16671 (2006).
- Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
- Jiao, Y. *et al.* Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).
- Qi, J. *et al.* A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**, 1510–1515 (2013).
- Xia, Z., Zhai, H., Lü, S., Wu, H. & Zhang, Y. Recent achievements in gene cloning and functional genomics in soybean. *ScientificWorldJournal* **2013**, 1–7 (2013).
- Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
- Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
- Chung, W.H. *et al.* Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. *DNA Res.* **21**, 153–167 (2014).
- Li, Y.H. *et al.* Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* **14**, 579 (2013).
- Ingvarsson, P.K. & Street, N.R. Association genetics of complex traits in plants. *New Phytol.* **189**, 909–922 (2011).
- Li, Y.H. *et al.* Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. *Theor. Appl. Genet.* **117**, 857–871 (2008).
- Kim, M.Y. *et al.* Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci. USA* **107**, 22032–22037 (2010).
- Guo, J. *et al.* A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. *Ann. Bot.* **106**, 505–514 (2010).
- Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
- Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
- Hyten, D.L. *et al.* Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**, 1937–1944 (2007).
- Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
- Liu, B. *et al.* QTL mapping of domestication-related traits in soybean (*Glycine max*). *Ann. Bot.* **100**, 1027–1038 (2007).
- Zabala, G. & Vodkin, L. 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 Sci.* **47**, 113–124 (2007).
- Todd, J.J. & Vodkin, L.O. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. *Plant Cell* **8**, 687–699 (1996).
- Clough, S.J. *et al.* Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of *CHS* genes comprising the *I* locus. *Genome* **47**, 819–831 (2004).
- Tuteja, J.H., Clough, S.J., Chan, W.C. & Vodkin, L.O. Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *Glycine max*. *Plant Cell* **16**, 819–835 (2004).
- Cook, D.E. *et al.* Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* **338**, 1206–1209 (2012).

27. Clemente, T.E. & Cahoon, E.B. Soybean oil: genetic approaches for modification of functionality and total content. *Plant Physiol.* **151**, 1030–1040 (2009).
28. Yamada, T., Takagi, K. & Ishimoto, M. Recent advances in soybean transformation and their application to molecular breeding and genomic analysis. *Breed. Sci.* **61**, 480–494 (2012).
29. Xia, Z. *et al.* Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci. USA* **109**, E2155–E2164 (2012).
30. Liu, B.H. *et al.* The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis TERMINAL FLOWER1*. *Plant Physiol.* **153**, 198–210 (2010).
31. Pathan, M.S. & Sleper, D.A. *Advances in Soybean Breeding in Genetics and Genomics of Soybean* (Springer, 2008).
32. Jeong, N. *et al.* *Ln* is a key regulator of leaflet shape and number of seeds per pod in soybean. *Plant Cell* **24**, 4807–4818 (2012).
33. Fang, C. *et al.* Cloning of *Ln* gene through combined approach of map-based cloning and association study in soybean. *J. Genet. Genomics* **40**, 93–96 (2013).
34. Toda, K. *et al.* A single-base deletion in soybean flavonoid 3'-hydroxylase gene is associated with gray pubescence color. *Plant Mol. Biol.* **50**, 187–196 (2002).
35. Toda, K., Akasaka, M., Dubouzet, E.G., Kawasaki, S. & Takahashi, R. Structure of flavonoid 3'-hydroxylase gene for pubescence color in soybean. *Crop Sci.* **45**, 2212–2217 (2005).
36. Chen, Y.W. & Nelson, R.L. Evaluation and classification of leaflet shape and size in wild soybean. *Crop Sci.* **44**, 671–677 (2004).
37. Funatsuki, H. & Ohnishi, S. Recent advances in physiological and genetic studies on chilling tolerance in soybean. *Jpn. Agric. Res. Q.* **43**, 95–101 (2009).
38. Foley, J.A. *et al.* Solutions for a cultivated planet. *Nature* **478**, 337–342 (2011).
39. Tilman, D., Balzer, C., Hill, J. & Befort, B.L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci. USA* **108**, 20260–20264 (2011).
40. Ray, D.K., Mueller, N.D., West, P.C. & Foley, J.A. Yield trends are insufficient to double global crop production by 2050. *PLoS ONE* **8**, e66428 (2013).
41. Tanksley, S.D. & McCouch, S.R. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**, 1063–1066 (1997).
42. Larson, G. *et al.* Current perspectives and the future of domestication studies. *Proc. Natl. Acad. Sci. USA* **111**, 6139–6146 (2014).
43. Sayama, T. *et al.* The *Sg-1* glycosyltransferase locus regulates structural diversity of triterpenoid saponins of soybean. *Plant Cell* **24**, 2123–2138 (2012).
44. Watanabe, S. *et al.* 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 (2011).
45. Grant, D., Nelson, R.T., Cannon, S.B. & Shoemaker, R.C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* **38**, D843–D846 (2010).

ONLINE METHODS

DNA sample preparation and sequencing. Soybean plants were grown at the Experimental Station of the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing. Young leaves were collected 3 weeks after planting and quickly frozen in liquid nitrogen. Total DNA was extracted with the cetyltrimethylammonium bromide (CTAB) method⁴⁶. At least 6 µg of genomic DNA from each accession was used to construct a sequencing library following the manufacturer's instructions (Illumina Inc.). Paired-end sequencing libraries with an insert size of approximately 300 bp were sequenced on an Illumina HiSeq 2000 sequencer at BerryGenomics company.

Variation calling and annotation. *Mapping.* Paired-end resequencing reads were mapped to the soybean Williams 82 reference genome⁹ with BWA⁴⁷ (Version: 0.6.1-r104) using the default parameters. SAMtools⁴⁸ (Version: 0.1.18) software was used to convert mapping results into the BAM format and to filter the unmapped and non-unique reads. Duplicated reads were filtered with the Picard package (picard.sourceforge.net, Version: 1.87). Using BWA software, all the unmapped reads in the cultivated soybean (Williams 82) genome were mapped to 1,537 *G. soja* (IT182932)-specific sequences¹⁵, which were provided by Suk-Ha Lee, Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University. The BEDtools⁴⁹ (Version: 2.17.0) coverageBed program was used to compute the coverage of sequence alignments. A sequence was defined as absent if coverage was lower than 90% and present if coverage was greater than 90%.

SNP calling. SNP detection was performed using the Genome Analysis Toolkit (GATK, version 2.4-7-g5e89f01)⁵⁰ and SAMtools. Only the SNPs detected by both methods were analyzed further. The detailed processes were as follows: (1) After BWA alignment, the reads around indels were realigned. Realignment was performed with GATK in two steps. The first step used the RealignerTargetCreator package to identify regions where realignment was needed, and the second step used IndelRealigner to realign the regions found in the first step, which produced a realigned BAM file for each accession. (2) SNPs were called at a population level with GATK and SAMtools. For GATK, the SNP confidence score was set as greater than 30, and the parameter -stand_call_conf was set as 30. The same realigned BAM files were used in SNP calling through the SAMtools mpileup package. (3) In the filter step, we chose the common sites identified by GATK and SAMtools with the SelectVariants package; SNPs with allele frequencies lower than 1% in the population were discarded.

Indel calling. Indel calling was similar to SNP calling but with the UnifiedGenotyper parameter -glm INDEL for the indel report only. Only insertions and deletions shorter than or equal to 6 bp were taken into account.

Annotation. SNP annotation was performed according to the soybean genome⁹ (Williams 82 assembly v1, gene model set 1.1, internal identifier v189 accessed from Phytozome on June 2013) using the package ANNOVAR⁵¹ (Version: 2013-08-23). Based on the genome annotation, SNPs were categorized in exonic regions (overlapping with a coding exon), splicing sites (within 2 bp of a splicing junction), 5'UTRs and 3'UTRs, intronic regions (overlapping with an intron), upstream and downstream regions (within a 1 kb region upstream or downstream from the transcription start site), and intergenic regions. SNPs in coding exons were further grouped into synonymous SNPs (did not cause amino acid changes) or nonsynonymous SNPs (caused amino acid changes; mutations causing stop gain and stop loss were also classified into this group). Indels in the exonic regions were classified by whether they had frame-shift (3 bp insertion or deletion) mutations. Correction of E1 annotation is described in **Supplementary Note 2**.

***G. soja* genome-specific sequence functional annotation.** A total of 1,537 *G. soja* genome-specific CDSs were obtained with the ORF finding module in the Trinity⁵² (Version: trinityrnaseq_r2013_08_14) program. For the *G. soja* genome-specific sequences, we used InterPro to annotate the motifs and domains against publicly available databases including Pfam, PRINTS, PROSITE, ProDom and SMART (all "-appl" parameters used in the iprscan program include blastprod, fprintscan, hmmpfam, hmmsmart, patternscan, proflescan, gene3d, seg, coils). CDS descriptions were presented by Gene Ontology, which was retrieved from InterPro. We also mapped the *G. soja*

genome-specific sequences to KEGG pathway maps by searching the KEGG databases and finding the best hit for each of them.

Identification of CNVs and segmental deletion using read depth of coverage. The CNV and deletion calling method from Bickhart⁵³ was improved as follows. The aligned read numbers were counted across the whole genome with 200 bp sliding windows and 100 bp slide steps using an in-house Perl script. The GC bias of the Illumina platform was corrected using LOESS smoothing toward a pattern of uniform coverage at all GC percentages, as previously described⁵⁴. CNV candidate windows were initially defined as having five out of seven or more sequential 200 bp overlapping windows with read depth values that differed significantly from the whole-genome average depth ($> \text{Mean} + 2 \times \text{Stdev}$). CNV regions were defined as ten or more continual CNV candidate windows that were independently identified in six or more individuals out of the 302 total accessions. Deletion candidate windows were initially defined as 200 bp windows with very low read depth (< 0.1 average whole-genome depth), within which at least six individuals showed normal read depth (> 0.5 average whole-genome depth). Deletion regions were defined as ten or more continual deletion candidate windows that were independently identified in six or more individuals out of the 302 accessions.

Population genetics analysis. To construct the phylogenetic tree with outgroup, we used BLASTN⁵⁵ (Version: 2.2.28) to identify orthologous regions between *G. max* and *M. truncatula*. We extracted each SNP and its 150 bp flanking sequences in soybean genome, then blasted to the *M. truncatula* genome sequence⁵⁶ with an *e* value $< 1e^{-6}$ and only kept the best hit with a conservative cut-off value of $\geq 70\%$ identity and 60 bp coverage (including the SNP). SNPs within orthologous regions were extracted, and genotypes of *M. truncatula* were used to provide outgroup information at corresponding positions. The neighbor-joining tree was constructed using PHYLIP 3.68 software⁵⁷ on the basis of a distance matrix. MEGA5 (ref. 58) was used to illuminate the phylogenetic tree. Principal component analysis (PCA) of whole-genome SNPs was performed with the EIGENSOFT 4.2 software⁵⁹ smartpca program, and the first two eigenvectors were plotted in two dimensions.

Linkage disequilibrium (LD) was calculated using PLINK⁶⁰ (Version 1.90) software. The pairwise r^2 values within and between different chromosomes were calculated following a previously reported method⁶¹. The LD for background was calculated using one million randomly selected interchromosomal SNP pairs. The LD for each chromosome was calculated using SNP pairs only from the corresponding chromosome. Regarding the LD for overall genome, the r^2 value was calculated for individual chromosomes using SNPs from the corresponding chromosome with parameter -ld-window-r2 0 -ld-window 99999 -ld-window-kb 1000, and then the pairwise r^2 values were averaged across the whole genome. The chromosome arm and pericentromeric region boundaries for each chromosome were previously described⁹.

Population structure was analyzed using the FRAPPE⁶² (Version: 1.1) program with a maximum likelihood method. We ran 10,000 iterations, and the number of clusters (K) was set from 2 to 7.

Genome scanning for selective signals. We performed a genome scan using an updated cross-population composite likelihood approach XP-CLR²⁰ (updated version, acquired from the author). Evidence for selection across the genome during domestication and improvement was evaluated in two comparisons: landraces versus *G. soja* for domestication and improved cultivars versus landraces for improvement. A 0.05 cM sliding window with 100 bp steps across the whole genome was used for scanning. Individual SNPs were assigned at positions along the genetic map from the SoyBase⁴⁵ (<http://www.soybase.org/>) by assuming uniform recombination between mapped markers. To ensure comparability of the composite likelihood score in each window, we fixed the number of SNPs assayed in each window to 250. The command line was `XPCLR -c freqInput outputFile -w1 gWin(Morgan) snpWin gridSize(bp) chrN`. Finally, we calculated the mean likelihood score in 100 kb sliding windows with a step size of 10 kb across the genome. Adjacent windows with high XP-CLR were grouped into a single region to represent the effect of a single selective sweep. The highest XP-CLR values, accounting for 5% of the genome, were considered as selected regions.

To confirm the selection signals identified by XP-CLR, we plotted the genetic diversity (π) values of *G. soja*, landrace and improved cultivar groups in 100 kb sliding windows with a step size of 10 kb. The π ratio ($\pi_{G. soja}/\pi_{landrace}$, $\pi_{landrace}/\pi_{improved cultivar}$) was calculated and compared with the domestication and improvement candidate regions identified by XP-CLR (Supplementary Fig. 15). We found that more than 80% of the selective sweeps identified by XP-CLR could also be identified by the π ratio approach, indicating that most of the selection regions can be identified by both methods, which are thus quite reliable.

To detect the selection signals caused by CNVs, the normalized population frequency was calculated. The RFD was used to measure CNV differentiation in populations based on variation frequency. The RFD is defined as follows:

$$\text{Relative Frequency Difference (RFD)} = \frac{F_{Landrace} - F_{G. soja}}{F_{population}}$$

where $F_{Landrace}$, $F_{G. soja}$ and $F_{population}$ represent the frequency of CNV in landrace, *G. soja* and the full population (all samples), respectively. CNVs with low frequency (both $F_{Landrace}$ and $F_{G. soja}$ were lower than 0.01) were removed. Improvement CNVs were identified with the same procedure. Gene expression analysis is described in Supplementary Note 3 and Supplementary Figure 16.

Phenotyping. For phenotyping, the 302 accessions were planted in two environmental conditions during the summer seasons of three years (2011 to 2013) at the Institute of Genetics and Developmental Biology, Changpin, Beijing, and the agricultural experiment station at Shanxi Agricultural University, Taigu, Shanxi. A QR-code scanning digital tool was developed by our group for rapid phenotyping. The phenotypic value of quantitative traits was the mean of five measurements. Our planting experiments found that 17 of the 302 lines were poorly adapted to these two locations. Thus, regarding the environmentally sensitive traits, such as the seed weight and plant height, we used the data from lines that are normally flowering and matured in these two locations for the subsequent GWAS analyses. Grain weight was obtained by weighing a total of 100 grains for each accession. Five batches of 100 randomly chosen grains were evaluated, and their mean was calculated. The angle between the pubescence and main veins in the reverse leaf side was measured to represent pubescence form. The oil percentage and cyst nematode resistance of some accessions were downloaded from the publicly available resources on the USDA GRIN database (<http://www.ars-grin.gov/>).

Genome-wide association study (GWAS). To minimize false positives and increase statistical power, population structure and cryptic relationships were considered. A compressed mixed linear model program, GAPIT⁶³ (Version: 2.12), was used for the association analysis. The first three PCA values (eigenvectors), which were derived from whole-genome SNPs, were used as fixed effects in the mixed model to corrects for stratification⁵⁹. The random effect was estimated from the groups clustered based on the kinship among all accessions. Kinship was derived from all SNPs or CNVs. We defined the whole-genome significance cutoff as the Bonferroni test threshold. For SNP GWAS, the threshold was set as 0.01/total SNPs ($\log_{10}(P) = -8.99$). For CNVs GWAS, the threshold was set as 0.01/total ($\log_{10}(P) = -5.27$). For GWAS using

CNVs data, we first converted the CNVs values into genotypes for individual CNV fragments (CNVRs) following a previously reported method⁶⁴; then, the genotypes for individual CNVRs were used for GWAS. In our study, the average SNP heterozygous rate in CNV regions was 3.17×10^{-5} , and the SNPs in CNV regions with high heterozygous rates greater than the 95th percentile of genome-wide heterozygous rate (1.3×10^{-4}) were treated as missing loci in the GWAS analysis.

Geographic differentiation analyses. Population nucleotide diversity (π) was calculated from the means of nonoverlapping 100 kb whole-genome fragments. The population divergence statistic F_{ST} was estimated for 100 kb sliding windows using a variance component approach with an in-house Perl code. Sliding windows with the highest 5% of F_{ST} values across different geographic groups were picked as candidate signals. Neighboring windows with contiguous signals were merged into one fragment. These regions were regarded as highly divergent across groups.

46. Murray, M. & Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4326 (1980).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
50. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
52. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
53. Bickhart, D.M. *et al.* Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* **22**, 778–790 (2012).
54. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
55. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
56. Young, N.D. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520–524 (2011).
57. Felsenstein, J. PHYLIP-phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
58. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
59. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
60. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
61. Mather, K.A. *et al.* The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* **177**, 2223–2232 (2007).
62. Tang, H., Peng, J., Wang, P. & Risch, N.J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
63. Lipka, A.E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
64. Kim, J.H. *et al.* CNVRuler: a copy number variation-based case-control association analysis tool. *Bioinformatics* **28**, 1790–1792 (2012).

Corrigendum: Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease

James E Dahlman, Omar O Abudayyeh, Julia Joung, Jonathan S Gootenberg, Feng Zhang & Silvana Konermann
Nat. Biotechnol. 33, 1159–1161 (2015); published online 5 October 2015; corrected after print 4 February 2016

In the version of this article initially published, when discussing the data in Figure 2b, on p. 1160, we wrote, “...targeting the same *HBG1/2* promoter and found they had 32 and 55 perturbed transcripts...” This should have been “31 and 55 perturbed transcripts” as in the sentence in the figure legend discussing the same data. The error has been corrected in the HTML and PDF versions of the article.

Erratum: Phosphoproteomics takes it easy

Paola Picotti

Nat. Biotechnol. 33, 929–930 (2015); published online 8 September 2015; corrected after print 4 February 2016

In the version of this article initially published, the page numbers in reference 1 were incorrect, and the reference thus linked online to the wrong article. The error has been corrected in the HTML and PDF versions of the article.

Erratum: Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean

Zhengkui Zhou, Yu Jiang, Zheng Wang, Zhiheng Gou, Jun Lyu, Weiyu Li, Yanjun Yu, Liping Shu, Yingjun Zhao, Yanming Ma, Chao Fang, Yanting Shen, Tengfei Liu, Congcong Li, Qing Li, Mian Wu, Min Wang, Yunshuai Wu, Yang Dong, Wenting Wan, Xiao Wang, Zhaoli Ding, Yuedong Gao, Hui Xiang, Baoge Zhu, Suk-Ha Lee, Wen Wang & Zhixi Tian

Nat. Biotechnol. 33, 408–414 (2015); published online 2 February 2015; corrected after print 4 February 2016

In the version of this article initially published, grant no. 91131005 from the National Natural Science Foundation of China was inadvertently omitted. The error has been corrected in the HTML and PDF versions of the article.