New Phytologist

Research

Domestication footprints anchor genomic regions of agronomic importance in soybeans

Yingpeng Han[1]*, Xue Zhao[1]*, Dongyuan Liu[2]*, Yinghui Li[3]*, David A. Lightfoot[4]*, Zhijiang Yang[2], Lin Zhao[1], Gang Zhou[2], Zhikun Wang[1], Long Huang[2], Zhiwu Zhang[5], Lijuan Qiu[3], Hongkun Zheng[2] and Wenbin Li[1]

[1]Key Laboratory of Soybean Biology in Chinese Education Ministry, Northeast Agricultural University, Harbin 150030, China; [2]Bioinformatics Division, Biomarker Technologies Corporation, Beijing 101300, China; [3]Institute of Crop Science, National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI) Chinese Academy of Agricultural Sciences, Beijing 100081, China; [4]Illinois Soybean Center for Excellence in Soybean Research, Southern Illinois University, Carbondale, IL 62901, USA; [5]Department of Crop and Soil Science, Washington State University, Pullman, WA 99164-6420, USA

Authors for correspondence:
Wenbin Li
Tel: +86 0451 55190778
Email: wenbinli@neau.edu.cn

Hongkun Zheng
Tel: +86 010 57045000
Email: zhenghk@biomarker.com.cn

Lijuan Qiu
Tel: +86 010 62186650
Email: qiulijuan@caas.cn

Received: *14 February 2015*
Accepted: *12 July 2015*

New Phytologist (2016) **209:** 871–884
doi: 10.1111/nph.13626

Key words: divergence of soybean species, genome-wide association mapping, origin, selective sweeps, sequencing.

Summary

- Present-day soybeans consist of elite cultivars and landraces (*Glycine max*, fully domesticated (FD)), annual wild type (*Glycine soja*, nondomesticated (ND)), and semi-wild type (semi-domesticated (SD)). FD soybean originated in China, although the details of its domestication history remain obscure.
- More than 500 diverse soybean accessions were sequenced using specific-locus amplified fragment sequencing (SLAF-seq) to address fundamental questions regarding soybean domestication.
- In total, 64 141 single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAFs) > 0.05 were found among the 512 tested accessions. The results indicated that the SD group is not a hybrid between the FD and ND groups. The initial domestication region was pinpointed to central China (demarcated by the Great Wall to the north and the Qinling Mountains to the south). A total of 800 highly differentiated genetic regions and > 140 selective sweeps were identified, and these were three- and twofold more likely, respectively, to encompass a known quantitative trait locus (QTL) than the rest of the soybean genome. Forty-three potential quantitative trait nucleotides (QTNs; including 15 distinct traits) were identified by genome-wide association mapping.
- The results of the present study should be beneficial for soybean improvement and provide insight into the genetic architecture of traits of agronomic importance.

Introduction

Although the domestication of soybean (*Glycine max*) has been traced to *c.* 6000–9000 yr ago in China (Kim *et al.*, 2010), many of the details underlying its domestication remain unresolved. A set of semi-domesticated (SD) soybean accessions found throughout China bear seeds that are larger than those of nondomesticated (ND) accessions (*Glycine soja*) but smaller than those of fully domesticated (FD) accessions (*G. max*; Hymowitz, 1970). Furthermore, the genetic and evolutionary relationships between the FD, SD and ND germplasms are unclear. The morphologically intermediate SD group, which typically produces larger seeds (the 100-seed weight is > 3.0 g greater than that of common ND seeds) and has a semi-erect stem (Hymowitz, 1970), was taxonomically described by Skvortzow (1927) as a distinct species: *Glycine gracilis*. Hermann (1962) hypothesized that *G. gracilis* is a variant of the FD group, whereas Broich & Palmer (1980, 1981) considered *G. gracilis* to be a semi-cultivated soybean and thus proposed its designation as *Glycine max forma gracilis*. Multiple evolutionary processes for the FD, SD and ND groups have been postulated. For example, Fukuda (1933) considers SD accessions to be evolutionary intermediates between ND and FD accessions. However, Hymowitz (1970) hypothesized that SD soybeans originated as hybrids between ND and FD soybeans, and based on their distribution in China and their morphological characteristics, Wang *et al.* (1983) asserted that SD soybeans are variants of ND soybeans. Furthermore, although several hypotheses have been proposed, the exact geographical origin of the FD groups in China has not been established. The hypothesized origins include northeastern China, the Huang-Huai Valley (central China; Vavilov, 1982; Xu, 1986), the Yangtze River region (south China; Wang, 1947; Gai *et al.*, 2000), and various other sites within China (Lü, 1978).

Genome-wide association studies (GWASs) have been used to dissect the genetic basis of traits underlying domestication in a wide range of organisms (Huang *et al.*, 2010; Voight *et al.*, 2010; Burdon *et al.*, 2011; Li *et al.*, 2013; Zhang *et al.*, 2014).

*These authors contributed equally to this work.

However, only a few studies have investigated domestication-related traits in soybean (Zhang *et al.*, 2014); thus, information on soybean domestication remains limited. As the SD and ND groups may retain genetic information before the domestication of soybean, these accessions are valuable for exploring the impact of domestication on genomic variation. Indeed, GWASs and genomic variations among the FD, SD and ND groups offer a unique setting for the recovery of useful alleles or genes from SD and ND soybeans in efforts to improve the traits of FD soybeans.

Whole-genome sequencing (WGS) is the most straightforward method for the genome-wide identification of domesticated variants. However, genotyping hundreds of samples by WGS is not affordable for many investigators, even when considering the marked decrease in cost as a result of innovation and technological progress. In comparison with WGS, reduced representation sequencing has many advantages, such as reducing genome complexity and lower cost, and has thus been applied to the study of evolutionary genomics, GWASs, and marker-assisted molecular breeding (Mamanova *et al.*, 2010). This manuscript presents a comprehensive view of genome-wide sequence variation among the genomes of a diverse group of 512 soybean accessions that represents the spectrum of FD, SD and ND germplasm collections. The accessions were analyzed using specific-locus amplified fragment sequencing (SLAF-seq; Sun *et al.*, 2013), a reduced representation sequencing technology. Using this method, high-quality single nucleotide polymorphisms (SNPs) evenly distributed along the soybean chromosomes were obtained at a low cost. The evolutionary dynamics and geographical origins of the FD groups were then defined by analyzing the patterns of diversity and fixation detected. A GWAS of the three soybean groups (FD, SD, and ND) allowed the dissection of the genetic architecture of 15 important domestication traits to a resolution of nearly one SNP per gene.

Materials and Methods

Genotype selection and sampling

A total of 512 soybean (*Glycine max* (L.) Merr.) accessions were evaluated in the present study, 470 of which were selected from the 2000 core germplasms collected to represent the genetic and geographical diversity of > 30 000 FD, SD and ND germplasm collections in China. The collection sites ranged from 19°N to 50°N and from 73°E to 135°E (Supporting Information Table S1; Fig. 1a,b). The phenotypes of these accessions were determined based on existing data that described a variety of yield, quality and morphological traits. The samples included 404 FD, 36 SD and 72 ND soybeans. Additionally, 42 accessions collected from the USA, Canada, Japan, and various European countries were also selected to represent exogenous soybean germplasms.

Trait analyses

Soybeans seeds, which were derived from a single plant of each of the 512 lines, were grown in 2011 at Harbin, China. The seeds were planted with four replications in a randomized design. Two

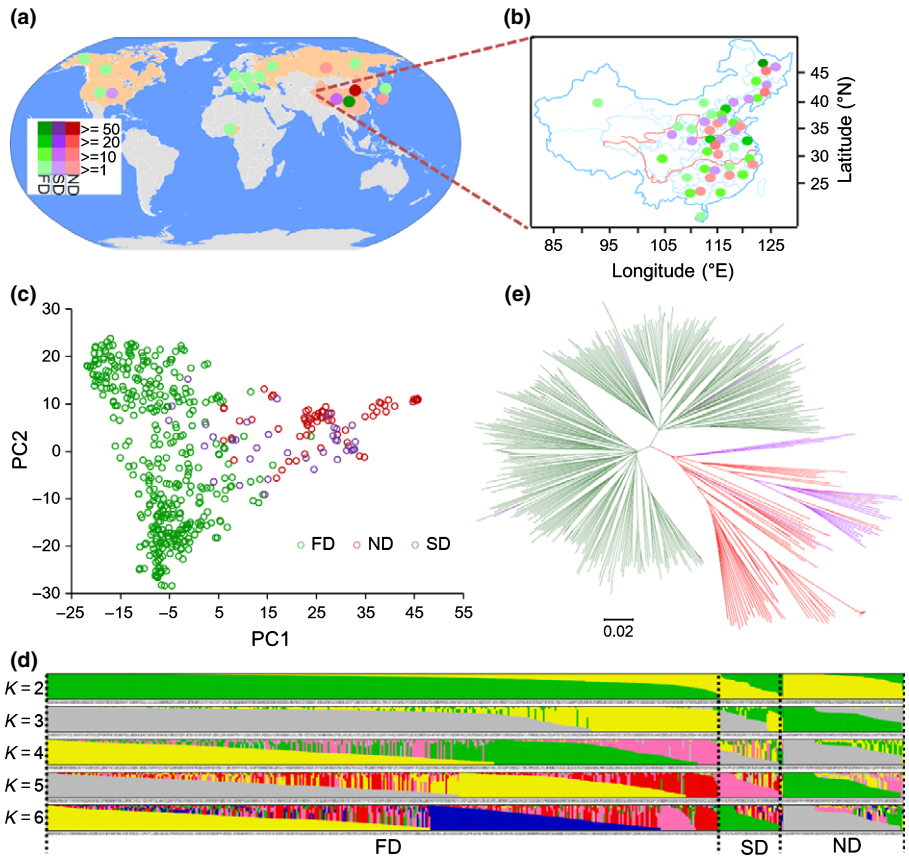

Fig. 1 Sample distribution and divergence. (a) Location of the soybean samples around the world, highlighting China. (b) The different colors indicate the number of individuals sampled. Regarding the map of China, the provinces are delineated by blue lines, and the upper and lower red lines indicate the Yellow River basin and Yangtze River basin, respectively. (c) Scatter plots of the first two principal components. Each dot represents an accession. (d) Population structure of soybean accessions. The accessions were divided into three groups: fully domesticated (FD) soybeans, semi-domesticated (SD) soybeans and nondomesticated (ND) soybeans. Within each group, the accessions were ordered according to the genetic component when $K = 2$ (Each accession shown as a vertical line partitioned into K colored components represents inferred membership in K genetic clusters). (e) Phylogenetic tree of the 512 accessions constructed with 64 141 single nucleotide polymorphisms (SNPs). The green, purple and red lines represent FD, SD and ND soybeans, respectively.

plants per pot and four pots per genotype replicate were analyzed. The pot size was 30 × 30 cm (diameter × depth) and contained PRO-MIX soil (a ready-made, peat-based growing medium containing 75–85% soil by volume, limestone for adjustment of the pH to 7.1, Perlite™, a wetting agent, and Vermiculite™; all Australian Perlite Pty Ltd, Sydney, NSW, Australia). The seeds were planted on 28 March 2011, in a glasshouse. Five quality traits, namely, hilum color, flower color, leaflet shape, pubescence color, and seed coat color, were assessed. Additionally, all of the accessions were planted in the field on 13 December 2011 and 20 December 2012, in Sanya in China. The plants were grown in a randomized complete block design with four replications (four-row plots). The rows were 2 m in length and located 70 cm apart, with plant spacing of 6 cm. Four plants from each replicate were selected to evaluate 10 domestication-related quantitative traits, including plant height, flowering time, seed protein and oil contents, 100-seed weight, seed weight per plant, node number on the main stem, pod number, and branch number per plant.

SLAF sample preparation for sequencing

Genomic DNA was isolated from the fresh leaves of a single plant per accession. Genomic DNA was analyzed according to SLAF-seq (Sun *et al.*, 2013). To obtain > 50 000 SLAF tags per genome which were evenly distributed in unique genomic regions of the genome, different restriction enzyme combinations were tested through *in silico* digestion prediction. Different length fragments of genomic DNA after digestion were simulated *in silico*. Enzymes were selected based on uniqueness and uniformity of simulated fragment alignments to the reference genome sequence of Williams 82 (NSRL, Champaign, IL, USA; Schmutz *et al.*, 2010). Two restriction enzymes (*Mse*I and *Hae*III) were selected. The 45-bp sequence read at both ends of each simulated 500–550-bp fragment was identified by a PERL script program. Predicted reads were aligned to the reference genome of Williams 82. The results showed that the expected percentage (unique mapped reads/total reads) was 74.74%. The restriction enzyme digestion protocol was proved to be effective in the validation of the 512 accessions. A total of 10 μg of genomic DNA (≥ 100 ng μl^{-1}) from each accession was incubated at 37°C with *Mse*I (New England Biolabs (NEB), Ipswich, MA, USA), T4 DNA ligase (NEB), ATP (NEB) and a 12-bp *Mse*I adapter. Restriction-ligation reactions were heat-inactivated at 65°C, and then digested with the second restriction enzyme *Hae*III at 37°C. PCR was performed with the restriction-ligation samples (diluted), dNTPs, *Taq* DNA polymerase (NEB) and a 12-bp *Mse*I-primer containing one of 96 unique barcodes. The PCR products were purified with the E.Z.N.A.® Cycle Pure Kit (Omega Bio-Tek, Norcross, GA, USA). Samples were incubated at 37°C with *Mse*I, T4 DNA ligase, ATP and Solexa™ adapters (Illumina Co., San Diego, CA, USA), purified with a Quick Spin column (Qiagen, Hilden, Germany), then electrophoresed on a 2% (w/v) agarose gel at 100 V for 1 h. Fragments from 500 to 550 bp were isolated using a Gel Extraction Kit (Qiagen). The fragments were used in PCR amplification with Phusion™ Master Mix (NEB) and Solexa amplification primer mix. Phusion™ PCR settings followed guidelines provided by Illumina. Samples were re-purified after PCR by electrophoresis and excising of DNA of 500–550 bp. DNA samples were diluted for sequencing.

Data processing and SNP calling

All reads were processed for quality control and filtered using Seqtk (https://github.com/lh3/seqtk). Then the high-quality paired-end reads were mapped onto the soybean genome (Wm82.a2.v1) using the Burrows-Wheeler Aligner (BWA; Li & Durbin, 2009). The reference genome sequence was downloaded from the Phytozome database (http://phytozome.jgi.doe.gov/). The mapping results were processed by sorting and duplicate marking using functions in SAMTOOLS (Li *et al.*, 2009) and PICARD (http://broadinstitute.github.io/picard/). Realigner Target Creator and InDel-Realigner in GATK (McKenna *et al.*, 2010) were used to realign InDels and Unified Genotyper was used to call genotypes across the 512 accessions using the default parameters. All data filter processing was performed following the 'best practices' workflow developed by the GATK team (McKenna *et al.*, 2010). Sequencing depths of each sample were as calculated using the 'Depth of Coverage' module of GATK. SNPs with minor allele frequency (MAF) > 5% were excluded from the genotype data sets of all the accessions.

Construction of phylogenetic trees and rooted phylogenetic trees

A total of 64 141 SNPs were used to calculate genetic distances among the 512 accessions using the *p*-distance method (Jin & Nei, 1990). A bootstrap consensus tree was obtained from 500 replicates. Phylogenetic trees were constructed using the neighbor-joining method in MEGA5 (Tamura *et al.*, 2011).

Lucerne was used as the outgroup when constructing rooted phylogenetic trees, and BLASTZ was used to identify homologous regions between soybean and Lucerne. SNPs within these regions were extracted, and the genomic sequences of Lucerne were used to provide outgroup information at corresponding positions. The neighbor-joining tree was constructed using MEGA5 under the *p*-distances model with the orthologous SNPs. The bootstrap consensus tree was constructed from 500 replicates.

Three-population test

The three-population test is capable of analyzing genomic mixture among different evolutionary groups (Myles *et al.*, 2011). SD soybean (X) was compared with two putatively parental populations, namely, ND (Y) and FD (W) soybeans, to determine whether X, Y and W are related in a simple tree or whether X is a mixture of Y and W. The *f*3 statistic, *f*3 (X, Y, W), was defined as the normalized product of the frequency difference between populations X and Y and the frequency difference between populations X and W averaged over all 64 141 SNPs. A detailed analysis of three-population tests is described by Reich *et al.* (2009).

Estimation of genomic characteristics and linkage disequilibrium (LD)

The genetic parameters π and θ (for details of the calculation, see http://cran.r-project.org/web/packages/popgen/index.html) were used to describe the diversity of a particular population based on the same sample size at a genome-wide scale (Watterson, 1975). Fu and Li's *D** and *F** statistics were used to estimate and compare the selection experienced by the evolutionary groups during the process of evolution and domestication investigated in this study (Fu, 1997). Population structure was calculated using ADMIXTURE software (Alexander *et al.*, 2009). The number of genetic clusters *K* was predefined as 2–10 to explore the population structure of tested accessions. This analysis provided maximum likelihood estimates of the proportion of each sample that was derived from each of the *K* populations. The divergence index, *F*-statistics (F_{ST}), is a measure of population differentiation or genetic distance based on genetic polymorphism data (Hudson *et al.*, 1992). F_{ST} was calculated via the PopGen package in BIOPERL (Wright, 1951) based on 100-kb sliding windows in 10-kb steps . The genomic windows where the average F_{ST} fell in the top 5% of the empirical F_{ST} distribution were defined as the F_{ST} outliers. Adjacent windows probably represent the effect of a single divergence region. An LD analysis was conducted using the PLINK software (specific parameters: MAF > 0.05, r^2, ld-window 999999, ld-window-r^2 0, Purcell *et al.*, 2007).

Gene flow pattern inference

We inferred the gene flow between soybean species and between subpopulations of *G. max*. The gene flow parameter *Nm* (the number of migrants coming into population) was evaluated among the FD, SD and ND groups using the method described by Wright *et al.* (2005). The directions of the gene flow between soybean species and between subpopulations within *G. max* were estimated using MIGRATE (Beerli & Palczewski, 2010). We assumed the following three models: a full model with two population sizes and two migration rates (from popA to popB and from popB to popA); a model with two population sizes and one migration rate to popB; and a model with two population sizes and one migration rate to popA. For each model, marginal likelihood was evaluated using the thermodynamic integration method. Among each model at least four heated chains were used to predict the optimal result using the suggested temperature scheme. The number of recorded steps in the chain was set to 500 000 and the replicate parameter was set to 3. Other parameters were set by default. Finally, the marginal likelihoods of all models were compared to infer the direction of gene flow.

Estimation of parameters for three assumed origins of FD soybeans

According to the results of other studies (Li, 1994), we postulated three candidate areas of origin for FD soybeans within China: northeastern China (> 45°N), the Huang-Huai Valley (26–44°N), and southern China (19–25°N). The π, θ, and F_{ST} values for the three geographical groups of FD soybeans were compared. Genetic components were analyzed using ADMIXTURE (Alexander *et al.*, 2009) and principal component analysis (PCA) was performed using GAPIT software (Lipka *et al.*, 2012).

Analyses of differentiation and putative selective sweeps

To detect genomic regions that are potentially differentiated from ND to FD soybeans, a total of 506 accessions (401 FD, 69 ND and 36 SD soybeans) were used for F_{ST} estimation according to the method described by Lam *et al.* (2010). These accessions encompassed the regions where soybeans are grown in China, and the accessions selected presented no direct kinship relationships. To evaluate F_{ST} and the π ratio for each combination of tested groups (401 FD, 69 ND and 36 SD soybeans), the F_{ST} and π values were plotted using 100-kb sliding windows with 10-kb steps. Genomic regions at which the F_{ST} value reached the critical value, corresponding to a 5% significant level, were taken as significantly differentiated genomic regions between two groups among the three soybean species. The top 5% values for the π ratio between ND and SD, between SD and FD and between ND and FD in each 100-kb sliding window with 10-kb steps were used to determine potential selective sweeps from highly differentiated regions (Li *et al.*, 2013).

Evaluation of highly differentiated and selected genomic regions

QTLs published in SoyBase (http://www.soybase.org) or genes located in highly differentiated or positive selective sweeps were analyzed using the method described by Wright *et al.* (2005).

Genome-wide association analyses of domestication traits

A total of 46 336 high-integrity SNPs from the domesticated soybean group were used to implement association analyses with a compressed mixed linear (MLM) model in GAPIT (Lipka *et al.*, 2012). The Bonferroni method at $\alpha \leq 0.05$ (corresponding to $P \leq 1.1 \times 10^{-6}$) was used as the threshold to determine whether a significant association existed (Holm, 1979). Candidate genes located within the 200-kb region upstream or downstream of peak SNPs were identified.

Results

Soybean accession sampling

A total of 512 soybean accessions, including 404 FD (*G. max*), 36 SD (*G. gracilis*), and 72 ND (*G. soja*) accessions, were selected based on their genotype, phenotype and geographical distribution. Of the 512 accessions analyzed, 470 were selected from a collection of *c.* 30 000 soybean accessions originating from China, and an additional 42 accessions were selected to represent soybeans originating outside of China (Table S1; Fig. 1a,b).

Sequencing and SNP calling

Using a bar-coded multiplex sequencing approach with an Illumina Genome Analyzer II, 392 million reads (each of which was 45 bp in length) were generated, encompassing 17.64 Gb of soybean genomic DNA sequence. For each accession, 59 494 high-quality tags (or SLAFs) were identified from 299 million paired-end reads after sequence alignment with the 'Williams 82' reference genome (Table 1). The high-quality SLAFs were selected using a total depth threshold of 1000-fold (Fig. S1), and the average sequencing depth was 6.14-fold for each of the 512 accessions (Table 1). Two samples of the reference 'Williams 82' genome, each from separate plants, were also sequenced as an internal control to evaluate accuracy. A total of 64 141 SNPs with an MAF $\geq$ 0.05 were identified from 59 494 SLAFs among the 512 accessions (Table 1; Fig. S2).

Genome-wide variation uncovered divergence among soybean species

To ascertain the divergence of the FD, SD and ND groups during evolution, principal component, population structure, phylogenetic relationship, and species-specific allele frequency analyses were performed. Additionally, the sequence diversity of the FD, SD and ND germplasms was evaluated. All of the analyses indicated strong divergence between the different soybean groups. The principal component and population structure analyses (Fig. 1c,d) showed that the FD, SD and ND collections were clearly distinguished, although different degrees of introgression were detected in these groups.

The FD, SD and ND groups were divided into different branches according to the phylogenetic relationships of the 512 accessions and based on 64 141 SNPs (Fig. 1e). Lucerne (*Medicago sativa* (L.) Millsp.) was used as an outgroup to construct a rooted phylogenetic tree that could be used to infer evolutionary relationships among the three soybean groups. Tree construction with Lucerne as the root was based on 2273 orthologous SNPs from soybean and Lucerne (Fig. S3). The evolutionary dynamics among the FD, SD, and ND accessions were clearly resolved in the rooted trees. Analysis of the observed genetic distances for the three groups revealed that the ND group is more closely related to the two outgroups than the SD and FD groups. Watterson's estimator (θ), the average pairwise divergence within a population (π), and the total and specific SNP numbers within the ND group were used to estimate genetic diversity. All three measures presented higher values for the ND group compared with the SD and FD groups (Fig. 2a–c). The LD patterns of these three groups revealed that the distance of LD decay in the ND group is shorter than those in the SD and FD groups (Fig. 2d). Indeed, the SD and FD germplasms may be under higher selective pressure than the ND germplasm, as consistently reflected by the results obtained from neutrality tests, including Fu and Li's D^* and F^* statistics (Fig. 2e,f). Therefore, ND soybeans were confirmed to be the progenitor of SD and FD soybeans.

The ND accessions possessed the largest number of group-specific SNPs. However, appreciable numbers of group-specific alleles were also identified in the FD and SD accessions, which is consistent with the maintenance of features that distinguish the SD groups from FD and ND relatives. The number of SNPs shared between the ND and SD groups was greater than the number of SNPs unique to the SD group but less than the number of SNPs unique to the ND group, indicating that the gene flow direction is from the ND to the SD group. The number of SNPs shared between the SD and FD accessions was greater than that of their respective group-specific SNPs and was also greater than the number of SNPs shared between the FD and ND accessions. In addition, the number of SNPs shared between the FD and ND accessions was lower than the number shared between the SD and ND accessions. At the whole-genome level, *Nm* values (F_{ST}-based gene flow rate) > 1 were obtained between the ND and SD groups (1.94), between the SD and FD groups (1.40), and between the ND and FD groups (1.02) (Fig. 2h). This finding indicates a relatively strong gene flow between ND and SD (1.94) and between SD and FD (1.40) and a relatively weak gene flow between ND and FD (1.02). Thus, FD soybeans may have been synchronously domesticated from SD soybeans and from ND soybeans.

Based on the analyses presented above, the SD group should be an evolutionarily intermediate species and not be a product of hybridization between the ND and FD groups. To further confirm the evolutionary relationships among the FD, SD and ND accessions during soybean domestication, a three-population test was performed (Myles *et al.*, 2011). A negative $f3$ value suggests that the tested group is a mixture of the other two groups; a positive value indicates that the groups are not mixtures. The $f3$

Table 1 Sequencing and single nucleotide polymorphism (SNP) calling

Accession group	FD	ND	SD	Average	Total
Sample size	404	72	36		512
Total reads (bp)	319 753 568	44 718 875	27 877 374		392 349 817
Pair-end reads (bp)	280 140 473	39 178 818	24 423 748		343 743 039
Properly pair-end reads (bp)	243 556 605	34 062 411	21 234 223		298 853 239
SLAF number	57 418	53 603	55 368		59 494
SLAF depth	6.29	5.77	6.36	6.14	
Heterozygosity	0.02	0.01	0.01	0.01	

FD, fully domesticated; ND, non-domesticated; SD, semi-domesticated; SLAF, specific-locus amplified fragment.

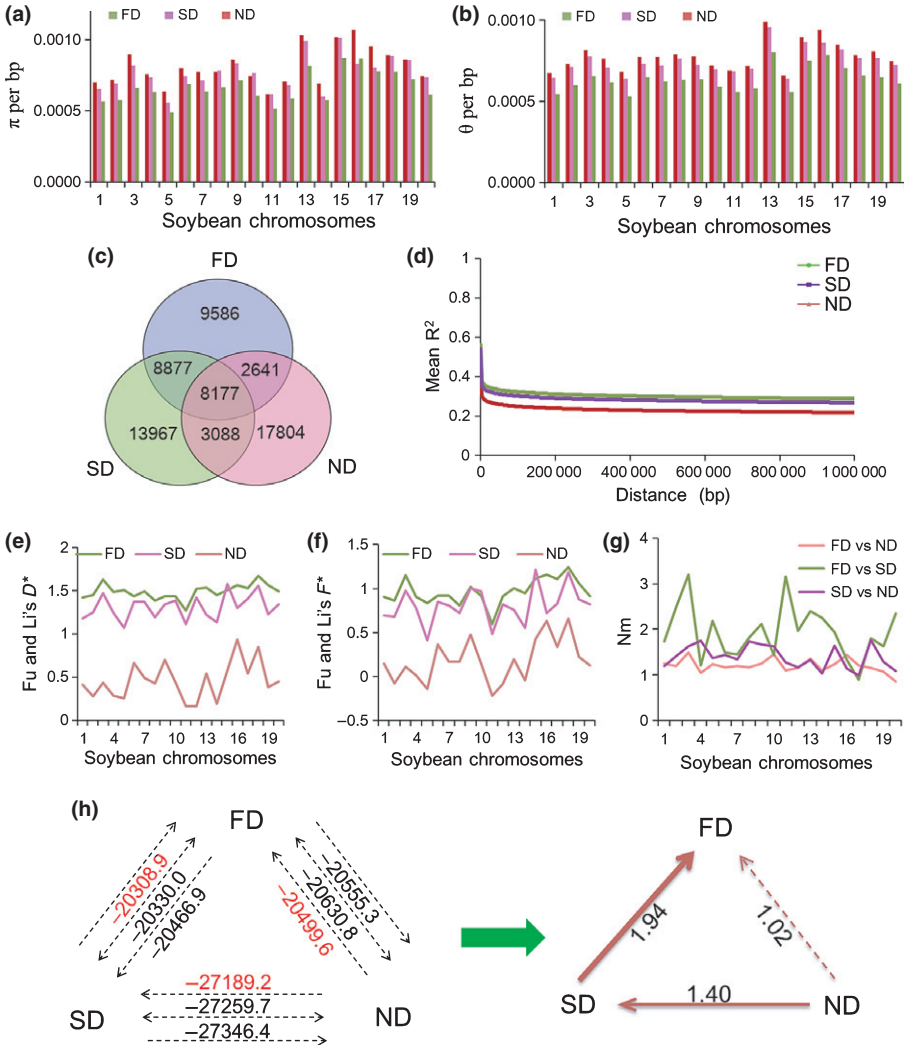

Fig. 2 Evolutionary relationship among fully domesticated (FD), semi-domesticated (SD) and nondomesticated (ND) soybeans. (a, b) Population parameters π and θ of FD, SD and ND soybeans. (c) Number of single nucleotide polymorphisms (SNPs) shared by FD, SD and ND soybeans. (d) Linkage disequilibrium (LD) patterns of FD, SD and ND soybeans. (e–g) Neutrality tests, including Fu and Li's *D** and *F** statistics, and the number of migrants coming into the population (*Nm*) values, of FD, SD and ND soybeans. (h) Hypothetical evolutionary relationships among FD, SD and ND soybeans were indicated by gene flow. All of the probable evolutionary patterns (dashed arrows) among the FD, SD and ND groups are summarized on the left of the figure, and the proven evolutionary patterns are shown on the right. The numbers on the left of the figure represent the marginal likelihoods of each probable evolutionary pattern among the FD, SD and ND groups. High values of marginal likelihood suggest a high possibility of the corresponding pattern, and arrows point in the direction of the inferred gene flow. The numbers on the right of the figure are the *Nm* values reflecting the power of gene flow among the groups. The dashed arrows indicate low gene flow (*Nm* near 1.0). Solid arrows with different widths indicate high gene flow (*Nm* > 1.0).

values for the three soybean groups were evaluated using 100 replicates, each of which randomly sampled one-half, one-third, two-thirds, and all of the SD accessions, along with equal numbers of the FD and ND accessions. All of the *f*3 values calculated for the FD, SD and ND groups estimated from the 100 replicates were positive (Fig. S4), indicating that none of the accessions are the product of hybridization between the other two soybean groups. This finding further suggested that the SD germplasm emerged as a transitional group during the evolutionary process of soybean domestication rather than from the direct hybridization of FD and ND soybeans.

The evolutionary relationships and directions of domestication among the FD, SD and FD groups were summarized by six probable patterns (Fig. 2h), and the gene flow directions were analyzed using the Bayesian approach in MIGRATE to investigate the migration rates (Beerli & Palczewski, 2010). The marginal likelihood of two out of the total of three models supported a unidirectional migration from ND to SD and to FD and from SD to FD. Therefore, this information indicates that SD soybeans are more ancient than FD soybeans and that a substantial portion of the genetic constitution of FD soybeans flowed from SD soybeans during domestication (Fig. 2h).

Geographical origin of FD soybean

The FD germplasms sampled in China were classified into three major subgroups according to their geographical distribution: northeastern China (FD3), the Huang-Huai Valley (FD2; central China) and southern China (FD1) (Fig. 3a; Wang, 1947; Vavilov, 1982; Xu, 1986; Gai *et al.*, 2000). The genotypic data set of 60 000 SNPs from 417 accessions (including all of the SD and ND soybeans as well as three FD subgroups) was analyzed to obtain the geographical origin of FD soybeans. The first principal component was correlated with latitude; the second principal component was also found to be associated with the latitude of the samples from northeastern China (Fig. 3a,b).

The ND-specific genomic component in the genome of the SD accessions and the three geographically distinct groups of FD accessions was evaluated based on genetic admixture analyses. The results showed that genetic introgression occurred from the ND accessions to the SD and FD accessions (originating from the three geographical regions). Of these three geographically based FD soybean groups, the accessions from the Huang-Huai Valley present greater genetic introgression from the ND accessions than those from the other two FD subgroups (Fig. 3c),

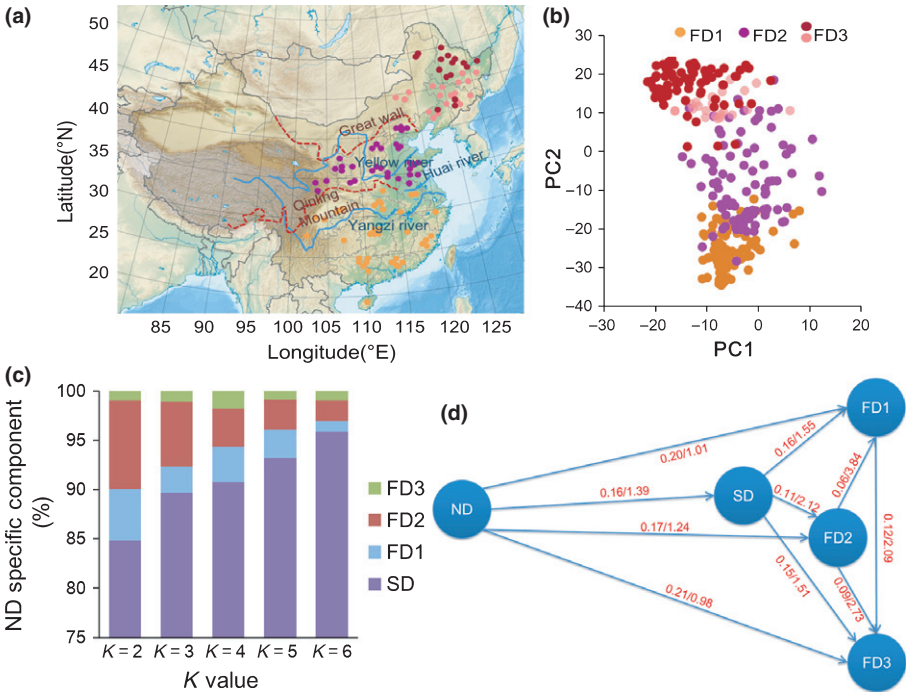

Fig. 3 Origin of soybean domestication in China. (a) Latitude and longitude distribution of fully domesticated (FD) soybean accessions in China. These accessions were classified to three hypothetical regions in China. FD1 (orange points) represents the soybean groups from southern China (south of the Qinling Mountains and Huai River, including both sides of the Yangtze River), FD2 (purple points) represents the soybean groups from the Huang-Huai Valley (between the Great Wall and Qinling Mountains, including both sides of the Yellow River), and FD3 (dark and light red points) represents the soybean groups from northeastern China (north of the Great Wall). (b) Differentiation of the three geographical groups of FD based on the first two principal components derived from genetic markers, illustrating a clear distinction between the groups from northeastern and southern China. The Huang-Huai Valley group presents an overlap with the other two groups. (c) Proportions of nondomesticated (ND)-specific components in the semi-domesticated (SD) group and the three groups of Chinese FD soybeans obtained through a population structure analysis. The Huang-Huai Valley (FD2) has more ND components than the other FD groups. (d) *F*-statistics (F_{ST}) based differentiated distance and the number of migrants coming into the population (*Nm*)-based gene flow among ND, SD and Chinese FD soybeans. The F_{ST} and *Nm* values are shown before and after the forward slash ('/'), respectively.

indicating that the FD subgroups from the Huang-Huai Valley may be the oldest cultivars.

Additional evidence of differential introgressions involves the degree of relatedness between the ND and FD groups within the three geographical regions. The FD group from the Huang-Huai Valley was genetically closer to the SD and ND accessions than the FD groups from northeastern and southern China. The population differentiation levels (as measured by F_{ST}) between the FD germplasm from the Huang-Huai Valley and the SD and ND germplasms were estimated to be 0.11 and 0.17, respectively. Markedly higher F_{ST} values were obtained from similar comparisons with the FD germplasms from northeastern and southern China (Fig. 3d). When calculated using the *p*-distance method, similar evolutionary distances were observed between the three FD groups and ND and SD soybeans. Specifically, closer evolutionary distances were obtained between the FD accessions from the Huang-Huai Valley and the SD and ND accessions than between the FD accessions from northeastern or southern China and the SD and ND accessions (Table 2).

Based on π and θ calculations, the genetic diversity values for the FD soybean subpopulations from northeastern ($\pi_{FD3} = 7.50 \times 10^{-4}$; $\theta_{FD3} = 7.57 \times 10^{-4}$) and southern China ($\pi_{FD1} = 7.75 \times 10^{-4}$; $\theta_{FD1} = 7.62 \times 10^{-4}$) were found to be slightly reduced compared with the genetic diversity of the FD soybeans from the Huang-Huai Valley ($\pi_{FD2} = 8.66 \times 10^{-4}$; $\theta_{FD2} = 7.63 \times 10^{-4}$). This result suggests that the FD germplasm may have experienced a modest reduction in genetic diversity as it expanded northward and southward.

Table 2 Evolutionary distances among subpopulations

	FD1[a]	FD2[b]	FD3[c]	SD[d]
FD2	0.167			
FD3	0.190	0.183		
SD	0.210	0.194	0.213	
ND[e]	0.212	0.198	0.227	0.203

[a,b,c]Fully domesticated (FD) soybeans from southern China, central China and northeastern China, respectively. [d,e]Semi-domesticated (SD) and nondomesticated (ND) soybeans, respectively.

According to these results, we deduced that FD soybeans may have expanded northward and southward from central China (the Huang-Huai Valley). To test this hypothesis, gene flow patterns among the ND and SD accessions and the three groups of FD accessions were estimated. The marginal likelihoods strongly support a unidirectional migration from both ND and SD to FD. Analysis of the relationships among the three FD soybean groups through a Bayesian comparison of migration models showed that the gene flow probably occurred from FD2 (the

Huang-Huai Valley) to FD1 and FD3, suggesting that the Huang-Huai Valley may be the origin of domesticated soybean in China.

The total number of SNPs (38 110) identified in the FD group from the Huang-Huai Valley was significantly greater (MAF ≥ 0.05) than the number of SNPs identified in the other two subgroups (37 494 SNPs in the FD subgroup from southern China and 36 114 SNPs in the FD subgroup from northeastern China). However, the number of SNPs unique to the Huang-Huai Valley FD accessions (12 964) was more than the number of unique SNPs identified in the FD accessions from southern China (12 652) but was less than that for northeastern China (13 212). This finding suggests that many new mutations were generated either as a result of the domestication process or during the introduction of soybeans from their region of origin to new environments.

Genome-wide association analyses of loci underlying domestication traits

Notable changes in soybean morphology, including flower color, seed coat color, seed weight, and seed composition, have emerged since its domestication. In this study, a high-density haplotype map was constructed for genome-wide association mapping. The monogenic, binary, oligogenic and quantitative domestication traits of soybean under different degrees of selection pressure were analyzed to identify the underlying genetic loci and putative genomic changes.

However, ND and FD soybeans were unsuitable for GWAS at the same time because of the greater morphological differences between these two groups. In addition, all of the ND soybean alleles would be inferior to all of the FD soybean alleles and would often have opposite beneficial alleles within the separate gene pools. Therefore, to evaluate the GWAS performance, we analyzed 15 traits from one FD soybean group that contained cultivars showing diversity among traits that are highly penetrant between the FD and ND groups (Figs S5, S6). The 15 traits included five binary traits (hilum, seed coat, flower color, pubescence color, and shape of ternately compound leaves) that were domesticated during the evolutionary transition from the ND to the FD group and 10 quantitative traits associated with soybean yield, seed quality and environmental adaptability to growth conditions. Forty-three association peaks obtained based on these 15 traits reached the corrected P value according to the Bonferroni method (P value $\leq 1.1 \times 10^{-6}$ at $\alpha = 0.05$; Fig. 4; Table S2).

Of the 18 peak SNPs associated with the five binary traits, five SNPs co-localized to known loci. Furthermore, 200-kb genomic regions around each peak SNP were analyzed, and six candidates genes were identified (Table 3).

Both of the peak SNPs responsible for pubescence and seed coat color were localized near the *T* locus of chromosome Gm06. This *T* locus is known to contain a gene encoding flavonoid 3′ hydroxylase (*F3′H*), which determines pubescence color (Yang *et al.*, 2010). *Glyma.06g202300*, located *c.* 14.01 kb from the peak SNP for pubescence color, is the closest candidate gene, consistent with the molecular function of *F3′H*, and is annotated as having flavonoid 3′-monooxygenase activity. Significant SNPs underlying flower color are found on chromosome Gm13 and are tightly linked to the *W1* locus, which was previously reported to control flower color in soybean (Yang *et al.*, 2010). The genes nearest to the peak SNP are also distributed into three classes of genes, including *v-myb avian myeloblastosis viral oncogene homolog* (*MYB*) transcription factors, which are known to regulate the early and late steps of anthocyanidin biosynthesis, genes involved in anthocyanin modification and transport, and genes encoding tissue-specific *Chalcone Synthase* (*CHS*, Feldbrugge *et al.*, 1997; Mathews *et al.*, 2003; Gillman *et al.*, 2011). A significant peak for leaflet shape was localized to chromosome Gm20. Both the locus and the candidate gene tightly linked to this peak SNP were previously identified in a fine mapping study of soybean leaflet shape (Jeong *et al.*, 2011).

Of all 25 QTNs associated with 10 quantitative traits, three are located at previously mapped QTLs or genes. For example, the signals associated with first flowering time link to *GIGANTEA* (*Glyma.01G100100*, *GI*) gene which is controlled by the circadian clock (Fig. 4; Table 3).

The GWAS was more effective in clarifying the genetic basis of the binary selected traits than that of quantitative traits. Only some major loci were detected for quantitative traits, whereas minor loci with small effects were difficult to identify. This result can be attributed to the difficulty that current GWAS approaches have in estimating the variable influences of environmental factors on the phenotypes of quantitative traits. To discover the genetic basis of more domesticated traits, particularly quantitative traits, the development of modified or new GWAS approaches is needed to estimate the effects of genotype and interactions with the environment.

Genomic changes and target regions associated with differentiation and selection

The differentiation of ND into FD soybeans via domestication may have partly occurred as a consequence of selective processes that induced genomic changes, which could be measured using genotypic data. In this study, the FD, SD and ND germplasm pools were sufficiently large to identify genomic regions that appear to contain selective sweeps caused by artificial selection. Subsequently, we determined the regional distribution of these sweeps in the genome. The average F_{ST} values between the FD and SD groups, between the FD and ND groups, and between the SD and ND groups were 0.12, 0.17 and 0.16, respectively.

A total of 425 highly differentiated genomic regions between the FD and ND accessions were defined using an F_{ST} threshold of 0.51 (determined by the 5% right tails of the empirical F_{ST} distribution calculated from 100-kb sliding windows with 10-kb steps). Between the SD and ND groups, 266 genomic regions had F_{ST} values > 0.35, revealing significant differentiation between these groups. Similarly, there were 204 genomic regions with F_{ST} values > 0.33 between the FD and SD groups, revealing significant differentiation between these groups (Fig. S7). These analyses indicate that selective sweeps caused by artificial selection during soybean domestication caused significant losses in diversity in highly differentiated genomic regions.

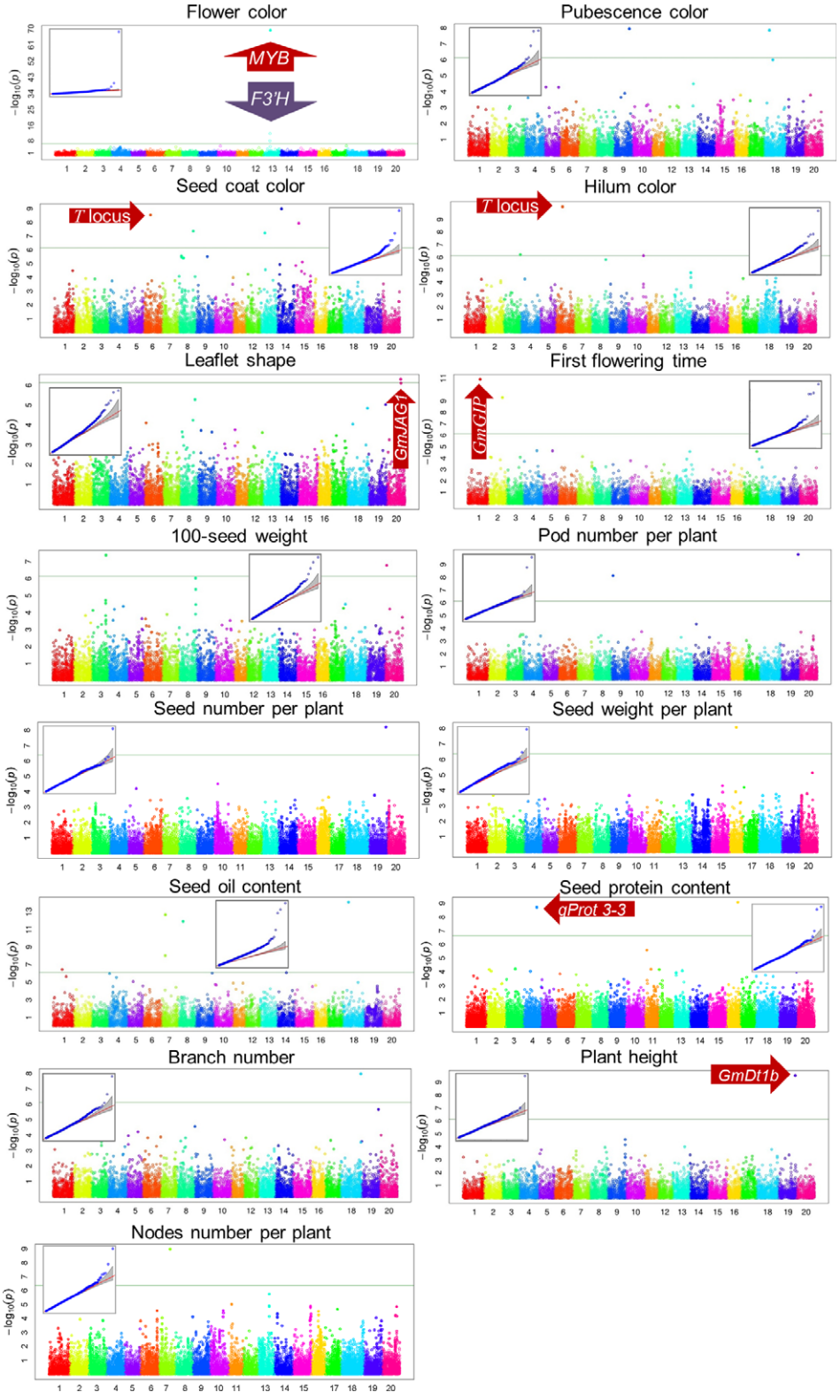

Fig. 4 Genome-wide association study (GWAS) of traits associated with domestication in soybean. Manhattan plots with the matching small QQ plots are shown in the same figure for each of the 15 traits. The associated single nucleotide polymorphisms (SNPs) that overlap with reported genes or quantitative trait loci (QTLs) are marked by red arrows. The Bonferroni multiple test threshold is shown as a solid green line (at $P \leq 0.01$).

To identify selective sweeps related to evolutionary bottlenecks and domestication in the three evolutionary groups, sequence diversity (π) ratios (π_{FD}/π_{ND} and π_{SD}/π_{ND}) were analyzed (Fig. S7). A total of 111 FD and 39 SD selective footprints in highly differentiated genomic regions presented significant reductions in diversity compared with the ND group. An additional 44 (FD) regions that showed significant reductions in diversity compared with the SD group were also found. These regions are the products of strong selective sweeps (Fig. 5a–c). The three sets of regions partially overlap (52 regions), such that a total of

New Phytologist

Table 3 Associated single nucleotide polymorphisms (SNPs), known loci/quantitative trait loci (QTLs) and candidate genes for qualitative traits of soybean

Trait	SNP position (bp)	Known loci/QTL	Candidate gene	Distance to SNP (kb)	Functional annotation
Flower color	Gm13_17625097	*W1*	*Glyma.13g073400*	59.6	Myb superfamily
Leaflet shape	Gm20_36514656	*Lflt shape 6-9*	*Glyma.20g116200*	684.5	Homology with *JAGGED* gene, which regulates lateral organ development
Seed coat color	Gm06_18724008	*T*	*Glyma.06g202300*	14.01	Flavonoid 3′-monooxygenase activity
Pubescence color	Gm06_18724008	T	*Glyma.06g202300*	14.01	Flavonoid 3′-monooxygenase activity
First flowering time	Gm01_33700831	*FT*	*Glyma.01G100100*	368.33	AtGCP3 interacting protein 1

AtGCP3, γ-tubulin complex proteins 3; Myb, v-myb avian myeloblastosis viral oncogene homolog.

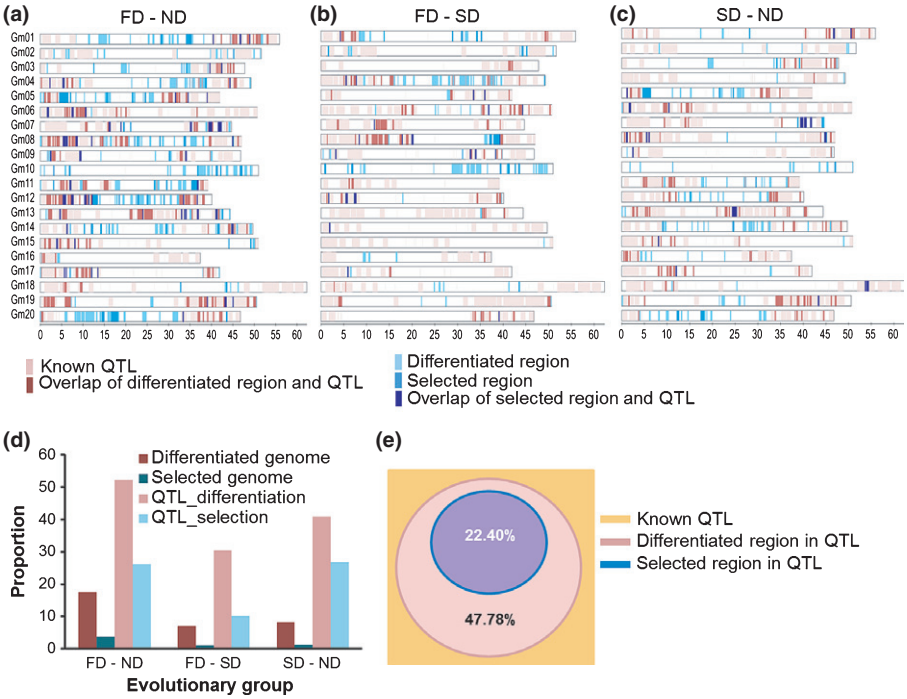

Fig. 5 Genomic characteristics of soybeans. (a–c) The distributions of differentiated regions, selected regions, and known quantitative trait loci (QTLs) as well as their overlaps obtained from two-group comparisons between fully domesticated (FD), semi-domesticated (SD), and nondomesticated (ND) soybeans. (d) Proportions of known QTLs in differentiated and nondifferentiated regions. (e) Proportions of known QTLs in selected and nonselected regions. The values on the *x*-axis are based on three groups of known QTLs, which include 258 QTLs in each group sorted by different intervals in QTL length (0–1, 1–2 and 2–5 Mb, respectively). The values on the *y*-axis are based on proportions of genomic regions, either QTL regions or differentiated/selected regions.

142 regions were inferred to represent strong selective sweeps. These regions are defined by 1023 SNPs (Fig. 5a). The genomic regions inferred to present stronger selective sweep signals in the FD group relative to the ND group comprised 36.94 Mb, corresponding to 3.19% of the genome. This area was nearly threefold larger than the regions representing selective sweeps in the SD group (12.90 Mb, corresponding to 1.29% of the genome) compared with the ND group. This result quantitatively reflects the impact that artificial selection may have had in modern soybeans (FD). Consistent with this observation, apparent selective sweeps were also found in the FD group (11.03 Mb, corresponding to 1.10% of the genome) relative to the SD group. Therefore, selective sweeps may have altered the domesticated soybean genome more than either the SD or ND soybean genome. Phenotypic changes in FD and SD soybeans, possibly associated with local adaptation, may result from these genomic loci.

Evaluation of selective sweeps in the context of modern breeding

The highly differentiated and selected genomic regions associated with the FD group were evaluated using existing soybean breeding resources to further define the genetic basis of domestication. Among the 2597 QTLs previously reported in soybean (SoyBase; http://www.soybase.org), 915 were selected for an analysis based on limit of detection (LOD) values and confidence intervals < 5 Mb.

The total length of the highly differentiated genomic regions is *c.* 278 Mb, which corresponds to 27.8% of the soybean genome. However, these regions contain the genetic information of 70% of the 915 known QTLs associated with domestication-related traits, including maturity date, flower number, plant height, seed weight, and seed protein and oil contents. The total length of the genomic regions encompassed by selective sweeps (constituting 10-kb steps that reached critical values for the 5% right tails of the empirical F_{ST} and π ratios from 100-kb sliding windows) was *c.* 46 Mb. This region represents 4.6% of the entire soybean genome and contains the genetic information of 22% of the known QTLs (Fig. 5a–c). Using a set of finely mapped QTLs (≤ 1 Mb in length), the results also showed that a highly differentiated or selected genomic region is, respectively, three- or twofold more likely to harbor a known domestication-related QTL than the rest of the soybean genome (Fig. 5d,e).

Two known loci, namely, *I* and *T*, control color traits and are encompassed by selection footprints that are inferred to have occurred during domestication (Fig. S8). Seventeen nearly fixed nonsynonymous SNPs among the FD, SD and ND groups (with an allele frequency > 0.8 in one group and < 0.2 in the others) were found in these regions (Table S3). These loci or genes may reflect, to a certain extent, the domestication history of soybean through the control of various domestication traits, particularly plant defense, morphology, and growth regulation.

Discussion

Genomic data can provide novel insight into species domestication processes (He *et al.*, 2011; Molina *et al.*, 2011; Myles *et al.*, 2011). Certain selection signatures were identified using genomic data obtained from dog and chicken (Boyko *et al.*, 2010; Rubin *et al.*, 2010). However, fundamental questions underlying the relatedness of FD, SD and ND soybeans have not been addressed to date through large-scale analyses of genomic data. Two previous studies conducted by Lam *et al.* (2010) and Li *et al.* (2013) provide preliminary analyses of the diversity among only *c.* 30 FD and ND accessions. In the present study, we used nearly 500 soybean accessions collected from all growing regions of China, ensuring the accuracy of the domestication analysis. Furthermore, all of the soybean accessions were analyzed using reduced representation sequencing technology (the SLAF-seq method), with an averaged sequencing depth of 6.14-fold. The results of other studies have verified that 5-fold sequencing depths can ensure the accuracy of domestication analyses (He *et al.*, 2011; Li *et al.*, 2013). The mean distance between markers was *c.* 28 kb, which is markedly less than the LD decay distances, over which LD decays to half of its maximum value in FD and ND soybeans (*c.* 150 and 75 kb, respectively; Lam *et al.*, 2010). Therefore, the marker density in this study was of sufficient resolution for general evolution analyses and GWAS (Morris *et al.*, 2013). Moreover, the sequenced markers avoided repetitive sequences and were randomly distributed throughout the genome (Fig. S2) and thus represented most regions of the soybean genome.

The evolutionary processes of soybeans have long been puzzling to biologists (Fukuda, 1933; Hymowitz, 1970). Fukuda (1933) hypothesized that the SD group is an evolutionary intermediate between the ND and FD groups, whereas Hymowitz *et al.* proposed that SD soybeans originated from hybridization between ND and FD soybeans, which was supported by limited lines of evidence (Hymowitz, 1970; Sisson *et al.*, 1978; Wang *et al.*, 1983; Xu *et al.*, 2002; Adams & Wendel, 2005). However, the lines of evidence cited above are insufficient to fully resolve the evolutionary process from SD soybean to FD soybean (Fukuda, 1933; Kim *et al.*, 2010). As genomic data have been successfully applied to analyze the evolutionary process of rice (Xu *et al.*, 2012), this strategy may also be used to analyze the evolutionary process of soybean domestication. Three previous studies have preliminarily evaluated relatedness between the FD and ND groups (Kim *et al.*, 2010; Lam *et al.*, 2010; Li *et al.*, 2014). In our study, the results from phylogenic, population structure and species-specific genotype frequency analyses based on > 60 000 SNPs showed that SD accessions possess unique genomic regions that differ from those found in ND and FD accessions. This result indicates that SD soybeans played an important role in the domestication and evolution of soybean. A three-population test and a gene flow pattern analysis further demonstrated that SD soybeans represent a transitional group that emerged during the domestication of FD from ND soybeans rather than a product of hybridization between ND and FD soybeans. This result provides the first explanation of the evolutionary position of SD soybeans based on large-scale sequencing. Thus, according to the present findings, SD soybeans are probably a beneficial source for introducing genetic diversity into new varieties in soybean breeding programs.

The geographical origin of FD soybeans has been the subject of intense debate, with numerous and contradictory domestication hypotheses, including predictions of single and multiple origins (Hymowitz, 1970; Vavilov, 1982; Xu, 1986; Li, 1994; Gai *et al.*, 2000; Xu *et al.*, 2002; Xu & Gai, 2003). The postulated geographical origins of FD soybeans include northeastern China, the Huang-Huai Valley (central China) and southern China (south of the Yangtze River). Based on large-scale sequencing data, our study revealed that the FD group from the Huang-Huai Valley is more closely related to the ND group and possesses higher sequence diversity than the FD groups from northeastern China and southern China. Additionally, the results from an admixture analysis support a model that indicates that the FD group of the Huang-Huai Valley presents greater genetic introgression from the ND group of the Huang-Huai Valley compared with the FD groups from northeastern and southern China. Thus, the results strongly suggest that FD soybeans were initially derived from ND soybeans in the Huang-Huai Valley. Through genetic distance, sequence diversity and gene flow analyses, we defined the boundaries of the region of origin in the Huang-Huai Valley for FD soybean as the Great Wall to the north and the Qinling Mountains to the south (central China), revealing a region that presents diverse climates and has experienced thousands of years of soybean cultivation history (Zhang & Hu, 1913; Yu & Nan, 2009).

The differentiation of soybean species during their evolution was mainly associated with genome-wide duplications, mutations, selection and drift (Kim *et al.*, 2010; Li *et al.*, 2013). In this study, selective sweeps among the three soybean groups were measured among genomes using F_{ST} values. A region consisting of a total of 46 Mb (*c.* 4.6% of the total genome) was found to be subjected to selection pressure. The existence of selective sweeps indicated the occurrence of primary selection pressure in the FD groups, suggesting that the effect of modern breeding is responsible for the loss of genetic diversity observed in FD soybeans. Moreover, > 800 highly differentiated regions in the soybean genome were identified based on analyses of different soybean species that may have been subjected to selection during the domestication process. The SNPs in these differentiated regions will be valuable for the marker-assisted selection of important traits during soybean breeding.

In this study, a high-density haplotype map of the loci underlying 15 domestication traits was constructed by GWAS. Based

on this haplotype map, a total of 36 association peaks were identified. All of the QTNs for five quality traits were found to be co-located at known loci or genes, such as *I*, *Chalcone Synthase* (*CHS*), *F3′H* and *MYB* transcription factors. Among the association peaks for 10 quantitative traits, seven association peaks were found to be located in directly related QTLs. These analyses provide proof of concept that haplotype mapping is effective. In the future, more agronomic traits will be analyzed using this haplotype map to rapidly identify candidate genes for validation.

In addition, > 900 reported QTLs were used to analyze the selective sweeps potentially responsible for domestication and selection traits. The results suggest that selective sweeps are approximately twofold more likely to harbor a known QTL associated with a domesticated trait compared with a group of QTLs finely mapped within a relatively short interval in the rest of the soybean genome (Fig. 5e). Thus, selective sweeps may be valuable for the domestication of and improvements in FD soybean traits using modern breeding techniques. However, our results also indicate that the effective evaluation of overlapping regions between selective sweeps and known QTLs is dependent on the mapping accuracy of QTLs. For example, when QTL intervals are greater than the size of the selective sweeps, the examined overlap of selective or nonselective regions with QTLs may be unreliable. Additionally, selective sweeps can be detected in any location in the soybean genome, but the reported QTLs are unevenly distributed in the genome. Additionally, QTLs are particularly difficult to identify near the centromere-associated regions of chromosomes (Schmutz *et al.*, 2010) because their detection depends on recombination events derived from crossing in certain populations (Lynch & Walsh, 1997). As most pericentromeric regions present a low degree of recombination (Ott *et al.*, 2011), selective sweep regions may harbor more undetected QTLs than reported QTLs.

Nonetheless, the availability of inferred selective sweep regions, which present higher probabilities that the harbored QTL is associated with domesticated and selected traits, may facilitate genomic selection. Moreover, the clarification of the soybean evolutionary process and the geographical origin of soybean presented in this study suggest that the ND and SD accessions in the Huang-Huai Valley represent a crucial source of new alleles for the future improvement of FD cultivars.

Acknowledgements

We thank X. Zhou, S. Wei, S. Wang, X. Wei, S. Song, R. Xu, D. Zhu, L Zhang, X. Liu, W. Lu and H. Nian for providing the soybean samples. This study was conducted in the Key Laboratory of Soybean Biology of the Chinese Education Ministry, the Soybean Research & Development Center, CARS and the Key Laboratory of Northeastern Soybean Biology and Breeding/Genetics of the Chinese Agriculture Ministry and was financially supported by the Chinese National Natural Science Foundation (31301339, 60932008 and 31201227), the National Core Soybean Genetic Engineering Project (2014ZX08002), the National Supporting Project (2014BAD22B00 and 2011BAD35B06), the National 863 Project (2013AA102602 and 2012AA101106-1-9), the Agricultural Science and Technology Innovation Program (ASTIP) of the Chinese Academy of Agricultural Sciences and the Provincial/National Education Ministry for the Team of Soybean Molecular Design. All data have been uploaded to ftp://ftp.biomarker.com.cn/soybean/.

References

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**: 135–141.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**: 1655–1664.

Beerli P, Palczewski M. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* **185**: 313–326.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM *et al.* 2010. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biology* **8**: e1000451.

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

Broich S, Palmer R. 1981. Evolutionary studies of the soybean: the frequency and distribution of alleles among collections of *Glycine max* and *G. soja* of various origin. *Euphytica* **30**: 55–64.

Burdon KP, Macgregor S, Hewitt AW, Sharma S, Chidlow G, Mills RA, Danoy P, Casson R, Viswanathan AC, Liu JZ *et al.* 2011. Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nature Genetics* **43**: 574–578.

Feldbrugge M, Sprenger M, Hahlbrock K, Weisshaar B. 1997. PcMYB1, a novel plant protein containing a DNA-binding domain with one MYB repeat, interacts *in vivo* with a light-regulatory promoter unit. *Plant Journal* **11**: 1079–1093.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

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

Gai J, Xu D, Gao Z, Shimamoto Y, Abe J, Fukushi H, Kitajima S. 2000. Studies on the evolutionary relationship among eco-types of *G. max* and *G. soja* in China. *Acta Agronomica Sinica* **26**: 513–520 (in Chinese).

Gillman J, Tetlow A, Lee J-D, Shannon J, Bilyeu K. 2011. Loss-of-function mutations affecting a specific *Glycine max* R2R3 MYB transcription factor result in brown hilum and brown seed coats. *BMC Plant Biology* **11**: 1–12.

He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu CI, Shi S. 2011. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genetics* **7**: e1002100.

Hermann F. 1962. *A revision of the genus* Glycine *and its immediate allies.* Washington, DC, USA: USDA Tech Bull.

Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**: 65–70.

Huang XH, Wei XH, Sang T, Zhao Q, Feng Q, Zhao Y, Li CY, Zhu CR, Lu TT, Zhang ZW *et al.* 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics* **42**: 261–267.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.

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

Jeong N, Moon JK, Kim HS, Kim CG, Jeong SC. 2011. Fine genetic mapping of the genomic region controlling leaflet shape and number of seeds per pod in the soybean. *TAG. Theoretical and Applied Genetics.* **122**: 865–874.

Jin L, Nei M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* **7**: 82–102.

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 *et al.* 2010. 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.

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.

Li F. 1994. Study on origin and evolution of soybean. *Soybean Science* **13**: 61–66 (in Chinese).

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, Wang T, Yeung CK, Chen L, Ma J *et al.* 2013. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature Genetics* **45**: 1431–1438.

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

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**: 2397–2399.

Lü SL. 1978. Discussion on the original region of cultivated soybean in China. *Scientia Agricultura Sinica* **4**: 90–94 (in Chinese).

Lynch M, Walsh B. 1997. *Genetics and analysis of quantitative traits*. Sunderland, MA, USA: Sinauer Associates Incorporated.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**: 111–118.

Mathews H, Clendennen SK, Caldwell CG, Liu XL, Connors K, Matheis N, Schuster DK, Menasco DJ, Wagoner W, Lightner J *et al.* 2003. Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. *Plant Cell* **15**: 1689–1703.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, Huang P, Jackson S, Schaal BA, Bustamante CD *et al.* 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences, USA* **108**: 8351–8356.

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE *et al.* 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences, USA* **110**: 453–458.

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D *et al.* 2011. Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences, USA* **108**: 3530–3535.

Ott A, Trautschold B, Sandhu D. 2011. Using microsatellites to understand the physical distribution of recombination on soybean chromosomes. *PLoS ONE* **6**: e22306.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**: 559–575.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature* **461**: 489–494.

Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S *et al.* 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587–591.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J *et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183.

Sisson VA, Brim CA, Levings CS. 1978. Characterization of cytoplasmic diversity in soybeans by restriction endonuclease analysis. *Crop Science* **18**: 991–996.

Skvortzow B. 1927. The soybean-wild and cultivated in Eastern Asia. *Proceedings of the Manchurian Research Society, Natural History Section Publication Series A* **22**: 1–8.

Sun X, Liu D, Zhang X, Li W, Liu H, Hong W, Jiang C, Guan N, Ma C, Zheng H. 2013. SLAF-seq: an efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. *PLoS ONE* **8**: e58700.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731–2739.

Vavilov H. 1982. *The world origin centers of main cultivated crops*. Beijing, China: Agricultural Press (in Chinese).

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G *et al.* 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42**: 579–589.

Wang J. 1947. Evolution of soybean traits. *Agriculture Journal* **12**: 6–11 (in Chinese).

Wang L, Wu H, Yao Z, Lin H. 1983. Investigation and research of the wild soybean in Heilongjiang Province (China). *Bulletin of Botanical Research* **3**: 116–130.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256–276.

Wright S. 1951. The genetical structure of populations. *Annals of Eugenics* **15**: 323–354.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. *Science* **308**: 1310–1314.

Xu B. 1986. Three new evidences for origin of soybean. *Soybean Science* **5**: 123–130 (in Chinese).

Xu DH, Gai JY. 2003. Genetic diversity of wild and cultivated soybeans growing in China revealed by RAPD analysis. *Plant Breeding* **122**: 503–506.

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

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L *et al.* 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology* **30**: 105–111.

Yang K, Jeong N, Moon JK, Lee YH, Lee SH, Kim HM, Hwang CH, Back K, Palmer RG, Jeong SC. 2010. Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. *Journal of Heredity* **101**: 757–768.

Yu W, Nan W. 2009. Isohyet 400 mm, the Great Wall and the dividing line between farming tribes and nomadic peoples. *Journal of Shanghai Jiaotong University (Philosophy and Social Sciences)* **17**: 46–52 (in Chinese).

Zhang D, Song H, Cheng H, Hao D, Wang H, Kan G, Jin H, Yu D. 2014. The acid phosphatase-encoding gene *GmACP1* contributes to soybean tolerance to low-phosphorus stress. *PLoS Genetics* **10**: e1004061.

Zhang X, Hu E. 1913. *New literature geography*. Changsha Hunan, China: Chinese Geographical Society (in Chinese).

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Sequencing depth distribution of SLAF tags used for SNP calling on 20 soybean chromosomes.

Fig. S2 Distribution of polymorphic SNPs on 20 chromosomes of soybean based on 512 accessions.

Fig. S3 Rooted phylogenetic tree of the 512 accessions.

Fig. S4 Distribution of three-population test statistics ($f3$ value).

Fig. S5 Phenotypes of qualitative traits for the genome-wide association study in soybean.

Fig. S6 Phenotypes of quantitative traits for the genome-wide association study in soybean.

Fig. S7 Empirical distribution of the π ratio, F_{ST} value and selection of soybean genome.

Fig. S8 Examples of genes with strong selective sweep signals in FD relative to ND soybeans.

Table S1 The list of 512 soybean accessions sampled in this study

Table S2 Summary of peak SNPs for the genome-wide association study on 15 soybean traits

Table S3 Fixed nonsynonymous SNPs between each two of fully domesticated soybean, semi-domesticated soybean and nondomesticated soybean

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.