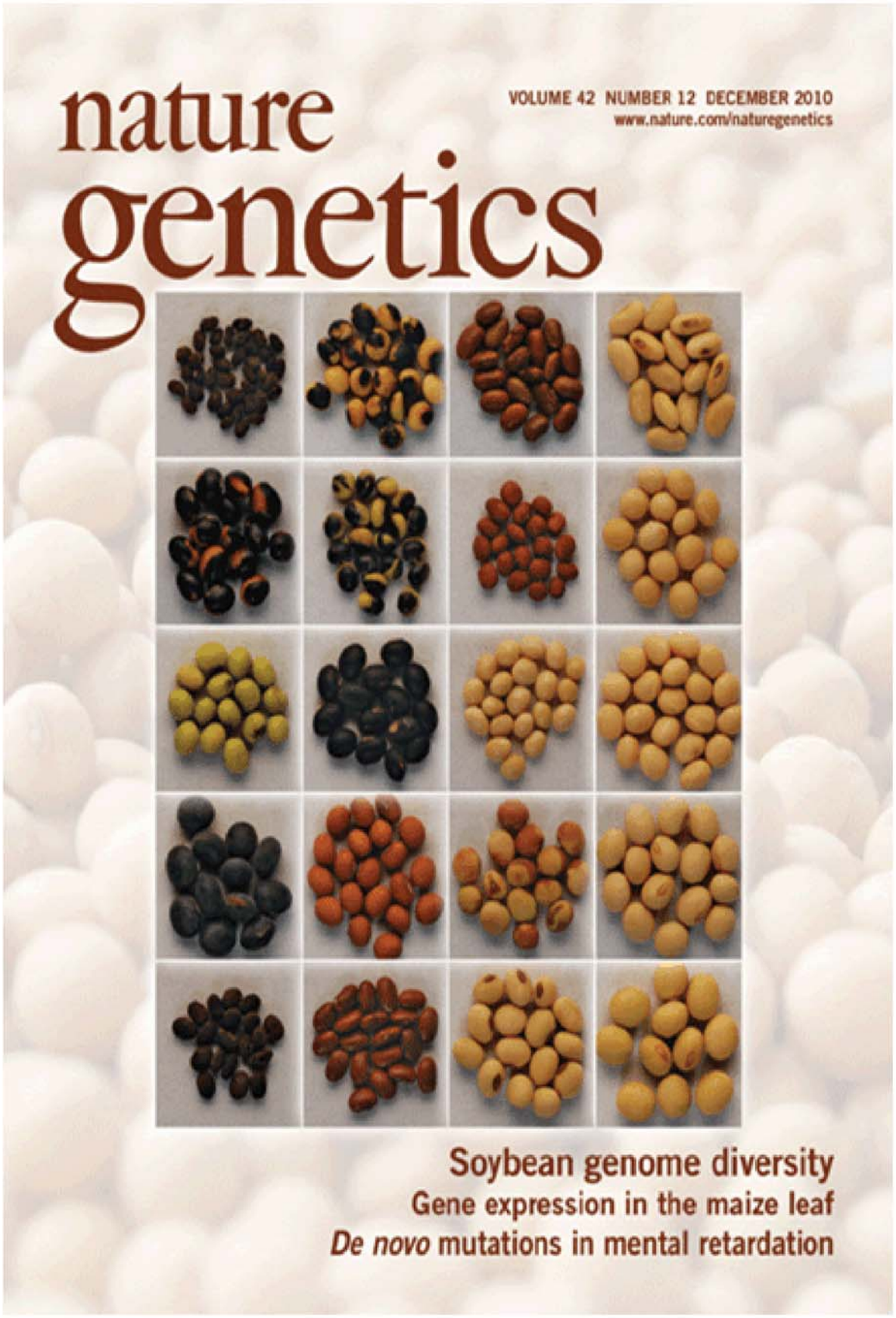
About the cover

December 2010 Volume 42 No 12



Cover photos by Yihan Jiang and Hon-Ming Lam

nature genetics

Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection

Hon-Ming Lam[1,6], Xun Xu[2,3,6], Xin Liu[1,2,6], Wenbin Chen[2,6], Guohua Yang[2,6], Fuk-Ling Wong[1], Man-Wah Li[1], Weiming He[2], Nan Qin[2], Bo Wang[2], Jun Li[2], Min Jian[2], Jian Wang[2], Guihua Shao[1,4], Jun Wang[2,5], Samuel Sai-Ming Sun[1] & Gengyun Zhang[2,3]

We report a large-scale analysis of the patterns of genome-wide genetic variation in soybeans. We re-sequenced a total of 17 wild and 14 cultivated soybean genomes to an average of approximately ×5 depth and >90% coverage using the Illumina Genome Analyzer II platform. We compared the patterns of genetic variation between wild and cultivated soybeans and identified higher allelic diversity in wild soybeans. We identified a high level of linkage disequilibrium in the soybean genome, suggesting that marker-assisted breeding of soybean will be less challenging than map-based cloning. We report linkage disequilibrium block location and distribution, and we identified a set of 205,614 tag SNPs that may be useful for QTL mapping and association studies. The data here provide a valuable resource for the analysis of wild soybeans and to facilitate future breeding and quantitative trait analysis.

Cultivated soybean (*Glycine max*) was domesticated in China ~3,000–5,000 years ago[1] and introduced to the United States in 1765 (ref. 2). Since then it has become an important cash crop, providing 69% and 30% of dietary protein and oil, respectively (see URLs). Given its economic importance, soybean productivity has garnered a great deal of attention in the scientific arena[3,4], and this has resulted in the recent sequencing of a cultivated soybean genome[5].

As a member of the Fabaceae family, soybean exhibits stringent cleistogamy (closed flower pollination). This characteristic may have a strong impact on maintaining genome homogeneity and reducing genomic variation, which may have been further exacerbated by the domestication process. Wild soybean (*Glycine soja*) may have retained genetic information before domestication and artificial selection, making it a unique resource for studying the impact of human selection on genetic variation in the soybean genome.

To obtain a comprehensive overview of the sequence variation of soybean at the population level, we resequenced the genomes of a diverse group of 17 wild and 14 cultivated soybean accessions. Using these data, we identified two unique features of the soybean genome that are distinct from other crop plants: they have exceptionally high linkage disequilibrium (LD) and a high ratio of average nonsynonymous versus synonymous nucleotide differences (Nonsyn/Syn). We also found that wild soybeans have retained allelic diversity that seems to have been lost in cultivated soybeans. These data and analyses should provide a valuable resource for recovering useful alleles and genes from wild soybeans.

RESULTS

Sequencing and variation calling

Samples for resequencing were taken from soybean accessions that originated or were popularized in different Asian and international regions (**Supplementary Fig. 1** and **Supplementary Table 1**). The advanced lines have been bred independently and have no known history of common ancestral lines. Additionally, some of these accessions have been used extensively as parental lines in breeding programs.

Resequencing of the 17 wild and 14 cultivated soybean accessions generated a total of 901.75 million (M) paired-end reads of 45-bp or 76-bp read length (180 Gb of sequence), with most to an approximately ×5 depth and >90% coverage (**Supplementary Table 1**). All sequence reads were aligned against the reference genome Williams 82 (ref. 5) using SOAP2 (ref. 6) with parameters that included sequence similarity, pair-end relationships and sequence quality. We called SNPs using SOAPsnp, filtered them[7] and identified present and absent variations (PAVs). From this analysis, we identified a total of 6,318,109 SNPs and 186,177 PAVs.

Previous reports have shown that the SNP calling accuracy from resequencing data is ~95–99% (refs. 8,9). Using the *de novo* sequencing data of the accession W05 (approximately ×80, data not shown), we estimated the SNP false-positive and false-negative rates to be ~1.79% and ~3.46%, respectively. This high accuracy provided a solid foundation for our data analyses and makes available high-quality data for future data mining.

[1]State Key Laboratory of Agrobiotechnology and School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. [2]BGI-Shenzhen, Shenzhen, China. [3]Key Laboratory of Genomics, Ministry of Agriculture, BGI-Shenzhen, China. [4]Institute of Crop Sciences, The Chinese Academy of Agricultural Sciences, Beijing, China. [5]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [6]These authors contributed equally to this work. Correspondence should be addressed to G.Z. (zhanggengyun@genomics.cn), S.S.-M.S. (ssun@cuhk.edu.hk), Jun Wang (wangj@genomics.cn) or H.-M.L. (honming@cuhk.edu.hk).

Received 21 June; accepted 22 October; published online 14 November 2010; doi:10.1038/ng.715

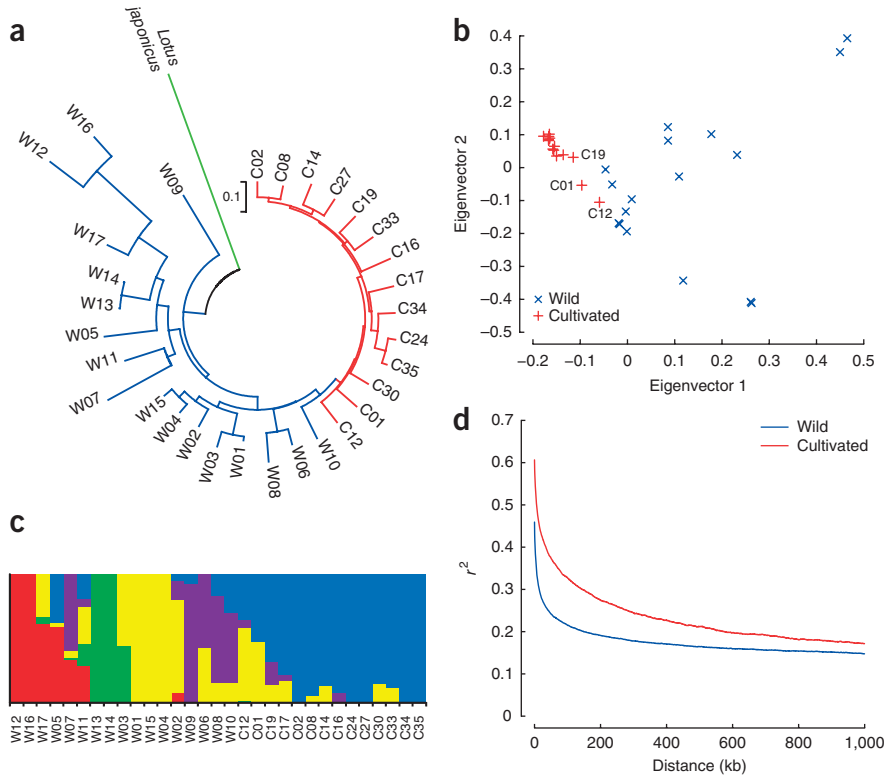


Figure 1 Analysis of the phylogenetic relationship, population structure and LD decay of wild and cultivated soybeans. (**a**) A neighbor-joining phylogenetic tree constructed using SNP data. (**b**) Principal component analysis of cultivated (red) and wild (blue) soybeans. (**c**) Bayesian clustering (STRUCTURE, $K = 5$) of soybean accessions. (**d**) LD decay determined by squared correlations of allele frequencies (r^2) against distance between polymorphic sites in cultivated (red) and wild (blue) soybeans.

Divergence between wild and cultivated soybeans

Cultivated soybeans have been under artificial selection to retain phenotypic variation that favored their mode of cultivation, harvesting and consumption (**Supplementary Table 2**). However, most of these phenotypes are quantitative traits that are influenced by environmental factors. To observe the divergence between wild and cultivated soybeans at the genomic level, we constructed a rooted phylogenetic tree using *Lotus japonicus* as the outgroup. The phylogenetic tree showed that the cultivated accessions formed a subclade within a larger mixed clade (**Fig. 1a**), and that wild and cultivated soybeans probably originated from a common ancestor. A principle component analysis (PCA) provided similar results (**Fig. 1b**), with cultivated soybeans forming a tight cluster that is clearly separate from wild soybeans. Using the Bayesian clustering program STRUCTURE[10], with K changing progressively from 2–7 (**Supplementary Fig. 2a**), subpopulations appeared in the wild population, whereas the cultivated population remained relatively uniform. The average value of ln likelihood was highest for the model $K = 5$; hence, we presented the clusters of $K = 5$ in **Figure 1c**. We did, however, find that multiple cultivated accessions showed evidence of admixture, with the most extreme cases evident in accessions C01, C12, C19 and C17 (**Fig. 1c**). This indicated that there was a recent history of introgression from wild soybean. This finding was also consistent with the PCA data (**Fig. 1b**), as shown by the separation of three cultivated accessions (C01, C12 and C19) from the main cultivated cluster.

Because the wild soybeans had a predominant effect in the STRUCTURE analysis, we performed the same analysis using only the cultivated soybeans. Here, the cultivated soybeans segregated into different groups that reflected their geographical distribution (**Supplementary Fig. 2b**). Phylogenetic, PCA and population structure analyses all indicated the heterogeneous nature of the genetic background of wild soybeans. In comparison to the wild soybeans, the cultivated soybeans showed a relatively homogenous genetic background, with some of the cultivars having genomic regions that were introgressed from wild soybeans. These findings indicated that human selection probably had a strong impact on the genetic diversity in the cultivated soybeans.

Whole-genome SNP analysis, using the parameter θ_π (ref. 11) (**Table 1**), also identified a lower level of genetic diversity in cultivated soybeans compared to wild soybeans (cultivated soybean: 1.89×10^{-3}; wild soybean: 2.97×10^{-3}). Additionally, the distribution of genome-wide diversity was significantly lower for cultivated soybeans compared to wild soybeans (**Supplementary Fig. 3**; $P < 0.01$ by paired t-test), which indicated the occurrence of a bottleneck in the genetic pool during domestication and under human selection. The total number of SNPs was much higher in wild soybeans, and wild-specific alleles (35%) were more abundant than cultivated-specific alleles (5%) (**Supplementary Fig. 4a**,**b**). The heterozygous rates of all accessions were low, reflecting the lack of cross-pollination resulting from cleistogamy (**Supplementary Fig. 4c**). The number of fixed loci in the wild and cultivated soybeans was 463,409 and 2,148,585, respectively (**Supplementary Table 3**).

Table 1 Statistics of SNPs in whole genome and genic regions of wild and cultivated soybean accessions

Whole genome

	Number of SNPs	θ_π (10^{-3})	θ_w (10^{-3})	Non-synonymous SNPs	Synonymous SNPs	Nonsyn/Syn
Wild soybean	5,924,662	2.966	2.307	106,716	78,701	1.36
Cultivated soybean	4,127,942	1.894	1.689	77,291	55,883	1.38

Genic regions

	CDS			UTR			Intron		
	Number of SNPs	θ_π (10^{-3})	θ_w (10^{-3})	Number of SNPs	θ_π (10^{-3})	θ_w (10^{-3})	Number of SNPs	θ_π (10^{-3})	θ_w (10^{-3})
Wild soybean	185,145	1.063	0.829	74,476	1.768	1.415	621,432	2.002	1.582
Cultivated soybean	132,976	0.723	0.626	53,730	1.118	1.073	426,897	1.318	1.180

We expected that the domestication bottleneck would yield a reduction in low-frequency alleles in the cultivated compared to wild accessions, and this has been seen previously for a few genomic regions[4]. However, our genome-wide analyses showed the opposite: we found that the low-frequency alleles were less abundant among the wild as compared to the cultivated accessions (**Supplementary Fig. 5**). To explain this unexpected observation, we inferred soybean history using a maximum-likelihood analysis based on the joint-allele frequency[12,13]. This analysis indicated that the most probable history was that the cultivated soybean population had expanded after domestication, whereas the wild soybean habitat area had been reduced (**Supplementary Fig. 6**). The allele frequencies simulated here were similar to those in our experimental data, with the singleton SNPs underestimated because of the high stringency of filtering we used in SNP calling.

To control for biases that might have been introduced by the use of a cultivated soybean genome as a reference for resequencing, we performed a similar analysis using a wild soybean (W05) *de novo* reference genome and saw the same pattern (data not shown). Of further interest, in comparison with other crops, SNP analysis showed that the cultivated soybean exhibited a lower diversity (cultivated soybean: 1.89×10^{-3}; rice: 2.29×10^{-3}; corn: 6.6×10^{-3})[14,15].

High linkage disequilibrium in the soybean genome

The stringent cleistogamy and relatively long generation time of soybeans suggested that there would be high LD in the soybean genome. To understand the specific LD block patterns in wild and cultivated soybeans, we used Haploview[16] to carry out an LD analysis. In general, both wild and cultivated soybeans exhibited high LD (**Fig. 1d**), and the average distance over which LD decays to half of its maximum value in soybean was substantially longer than that of all plants analyzed to date (cultivated soybean: ~150 kb; wild soybean: ~75 kb; maize: <1 kb; wild and cultivated rice: <1 kb; and *Arabidopsis thaliana*: ~3-4 kb)[15,17,18]. Unlike animals, plants rarely have such long LD patterns[19–22]; therefore, soybean may make a good plant model for studying the effect of extreme LD in genomic and population structures.

Our study showed that the pattern of LD block distribution differed between wild and cultivated soybeans. We found that the frequency of occurrence of LD blocks of lengths <20 kb was higher in wild soybeans than in cultivated soybeans, and the number of small LD blocks in wild soybeans was double that in cultivated soybeans (LD blocks of <1 kb: wild = 26,827, cultivated = 12,652; LD blocks of 1–2 kb: wild = 10,973, cultivated = 5,425). There was a general reversal of this trend as block size increased: the number of LD blocks of >150 kb in wild soybeans was about half that of cultivated soybeans, and the longest LD block we found in wild soybeans was ~500 kb, whereas the longest LD block in cultivated soybeans was ~1 Mb. Additionally, both the percentage and combined length of these long blocks were higher in cultivated soybeans (cultivated: 1.5%, total length 57.7 Mb; wild: 0.6%, total length 35.7 Mb) (**Supplementary Fig. 7**).

SNP analyses in the LD blocks showed that there was a lower SNP ratio in long LD blocks as compared to the whole genome in both wild (θ_w (ref. 23) = 1.82×10^{-3} versus 2.29×10^{-3} for the whole genome, $P < 0.01$ by Wilcoxon rank-sum test) and cultivated ($\theta_w = 1.56 \times 10^{-3}$ versus whole genome: 1.69×10^{-3}, $P < 0.01$ by Wilcoxon rank-sum test of all the LD blocks in two populations) soybeans. To determine the underlying cause of SNP loss in the long LD blocks, we calculated Tajima's *D* (ref. 24) values in cultivated and wild soybeans. The *D*-value distribution of cultivated soybeans was significantly higher than the average (0.2 in the whole genome compared to 0.8 in the LD blocks) (**Supplementary Fig. 8**; $P < 0.01$ by the Wilcoxon rank-sum test), indicating a significant loss of rare SNPs, which may be due to reduced recombination within the LD blocks. In contrast, the *D*-value calculations for wild soybean did not show an increase (1.1 for whole genome and 0.82 for long LD blocks), indicating that the reduced number of SNPs in wild soybeans was not related to the loss of rare SNPs but instead due to random loss of SNPs. These findings are again consistent with a history of population expansion of cultivated soybeans after domestication and a loss of habitat of wild soybeans (**Supplementary Fig. 6**).

Given the high LD in soybeans, only a small subset of SNPs would be required for marker-assisted breeding. We therefore defined a set of 205,614 tag SNPs that can be used to facilitate such future studies. It is important to note, however, that the high LD of soybeans also creates resolution limitations for association studies using genetic populations.

Selection and introgression

We used our soybean whole-genome sequence data to assess genome-wide patterns of nucleotide diversity. This analysis revealed that the allelic diversity in wild soybeans was higher than in cultivated soybeans across the entire genome (**Fig. 2**). We also identified conserved genomic segments shared by both, indicating regions that are potentially essential for the survival of both wild and cultivated soybeans. Calculation of the divergence index (F_{ST}) value between wild and cultivated soybeans allowed us to identify genomic regions of large F_{ST} value, which signified areas having a high degree of diversification between wild and cultivated soybeans. These regions may contain or be associated with loci related to domestication (**Fig. 2**). In total, we identified 369 subregions (100-kb non-overlapping regions) with high F_{ST} (higher than 0.45) and 101 subregions with low F_{ST} (lower than 0.02), and we found that these regions are distributed on all linkage groups and encompassed ~5% of the total genome.

Using F_{ST} values, we carried out a more detailed analysis on the genomic regions that span previously identified domestication QTLs[25] to identify segments within these QTLs that had high F_{ST} values in wild versus cultivated soybeans (**Supplementary Fig. 9**). Domestication soybean QTLs usually span large regions of the genome, with several having lengths >10 cM[25]. Based on our analysis, we identified subregions of high F_{ST} values within the genomic regions containing these domestication QTLs. These findings may aid in narrowing the functional subregions within the QTLs (**Supplementary Fig. 9**). For example, we identified several segments of exceptionally high F_{ST} values in the QTL for the twinning trait on chromosome 2 (**Supplementary Fig. 9a**) and in the QTL for stem elongation–related traits (plant height, twinning trait, maximum internode length and number of nodes) on chromosome 18 (**Supplementary Fig. 9g**). Subregions that have very high F_{ST} values may provide an indication of the functional genes or alleles involved.

We analyzed in greater detail two genomic regions with extreme patterns of diversity and differentiation. In the first, we found LD blocks in an overlapping region on chromosome 5 in both the wild and cultivated soybean genomes (**Fig. 3a–c**; position ~6.2 Mb to 6.4 Mb) that showed low diversity (θ_π of wild: 0.69×10^{-3}, cultivated: 0.097×10^{-3}; **Supplementary Fig. 10a**,**b**) as well as a low divergence index (F_{ST} ~ 0.00083; **Supplementary Fig. 9c**). This suggested that an inherited functional constraint was present in this region; thus, they were retained in both wild and cultivated soybeans through selective sweep in their common ancestor.

In a second example, and in contrast, we identified a region on chromosome 10 of cultivated soybeans that had two consecutive long LD blocks that were absent in wild soybeans (**Fig. 3d–f**; position ~42.6 Mb to 42.8 Mb). The diversity in this region was substantially lower in the cultivated soybeans. The mean loss of diversity (LoD) value, given by (θ_π of wild – θ_π of cultivated)/θ_π of this region

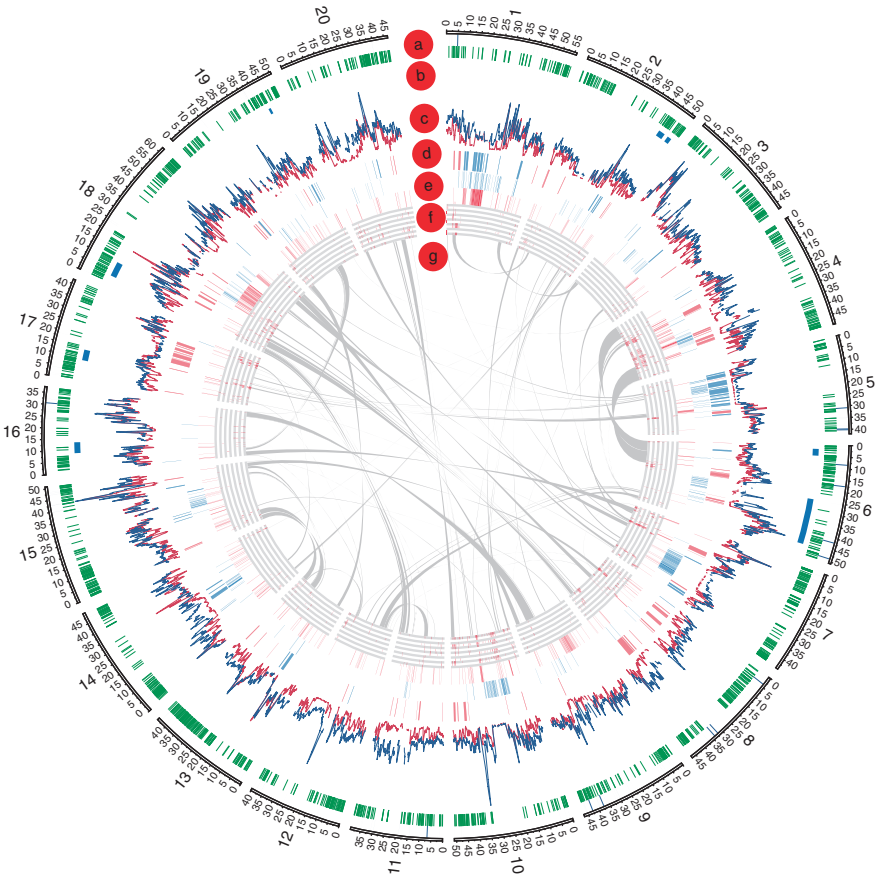


Figure 2 Summary of resequencing data of 17 wild and 14 cultivated soybean accessions. The average genome coverage is ~90%. Concentric circles show the different features that were drawn using the Circos program[39]. The 20 chromosomes are portrayed along the perimeter of each circle. (**a**) Insertion or deletion in the reference cultivated soybean genome[5] (unique genome in blue) and the wild accession W05 (unique genome in green). (**b**) QTLs of domestication-related traits[25] (blue blocks). (**c**) Genomic diversity (θ_π) of wild soybeans (blue) and cultivated soybeans (red). (**d**) F_{ST} value of wild versus cultivated soybeans (red, >0.4; blue, <0.03). (**e**) LD blocks (>50 kb) of wild soybeans (blue) and cultivated soybeans (red). (**f**) Introgression of wild genomic regions (red) into cultivated soybean accessions. (**g**) A graphical view of duplicated annotated genes is indicated by connections between segments.

in wild soybeans was 0.94 (**Supplementary Fig. 10c**). We also found that the F_{ST} value between wild and cultivated soybeans in this region was higher than average (0.511 versus 0.199 for the whole genome; **Supplementary Fig. 10d**). Notably, this LD region is close to the simple sequence repeat (SSR) marker Satt592, which is associated with important agronomic traits, such as biomass accumulation, apparent harvest index, yield and vitamin E content[26,27]. This indicated that the selection processes acting on cultivated soybeans could be different from those acting on wild soybeans.

The elite modern soybean germplasms used for current soybean crops are the result of extensive breeding and artificial selection. A genome-wide sequencing comparison to reveal haplotype sharing could provide a unique tool to identify introgression events in the history of these cultivars (**Fig. 2**). We used a sliding window of 100 kb (Online Methods) on the cultivated soybeans and identified a total of 431 potential regions of introgression (total 43.1 Mb). The cultivated soybean accessions C01, C12 and C19 possessed the most extensive introgression of the high F_{ST} regions (wild versus cultivated), occupying 29% (12.5 Mb), 36% (15.3 Mb) and 14% (6 Mb), respectively. There were also introgression regions shared between these accessions: C01 versus C12 (62%; 7.7 Mb), C01 versus C19 (43%; 2.6 Mb), and C12 versus C19 (43%; 2.6 Mb). Previous studies have indicated that conserved regions of introgression may indicate selection events[28]. To explore this in the future, it will be useful to sequence a more extensive collection of elite and phenotypically characterized cultivated soybean germplasms, which could provide information for developing better breeding programs that use wild germplasms.

Deleterious mutations accumulated in soybeans

The coding regions occupy ~6% of the soybean genome[5], but we found that only ~3% of the total SNPs identified were present in these regions. The remaining ~97% SNPs were in noncoding regions (**Table 1**).

The average Nonsyn/Syn ratios in the genome of both wild and cultivated soybeans (wild total: 1.36; wild specific SNPs: 1.36; cultivated total: 1.38; cultivated specific SNPs: 1.61) are the highest that have been reported among all plants so far (rice: 1.2; *A. thaliana*: 0.83)[28,29]. When compared to relatively conserved genes in rice (ratio of average Nonsyn/Syn < 1), ~84% of the soybean orthologs exhibited a higher Nonsyn/Syn value ($P < 0.01$ by paired *t*-test; **Supplementary Table 4**).

We also found that SNPs that are likely to have a major impact on gene function (large-effect SNPs) were present in 4,648 soybean genes (10%), which is higher than in *A. thaliana* (1,614 genes, 6.1%; ref. 29). These soybean genes included 3,018 that have premature stop codons (**Supplementary Table 3**). A total of 1,467 (wild: 1,421; cultivated: 834) gene categories contained large-effect SNPs, but these gene categories had different proportions of large-effect SNPs (**Supplementary Fig. 11**). The presence of a higher Nonsyn/Syn value at the whole-genome level and more large-effect mutations suggested that the soybean genome had accumulated a higher ratio of deleterious mutations.

High LD would result in the lack of effective recombination; consequently, deleterious mutations could not be eliminated and would accumulate. We looked at all the long LD blocks (>50 kb) of wild soybeans, some of which also existed in cultivated soybeans, and found that the average ratio of Nonsyn/Syn was higher than that of the whole-genome average (**Supplementary Table 5**). For long LD blocks that were specific to cultivated soybeans, this ratio was similar to the whole-genome average (**Supplementary Table 5**). These LD blocks might have been formed recently during the domestication process and under artificial selection, and would, therefore, not have accumulated a significant number of new mutations.

At the whole-genome level, we looked at cultivated-specific SNPs compared to wild-specific SNPs and found that the accumulation of deleterious (radical change) mutations (**Supplementary Table 3**) was slightly higher in cultivated soybeans.

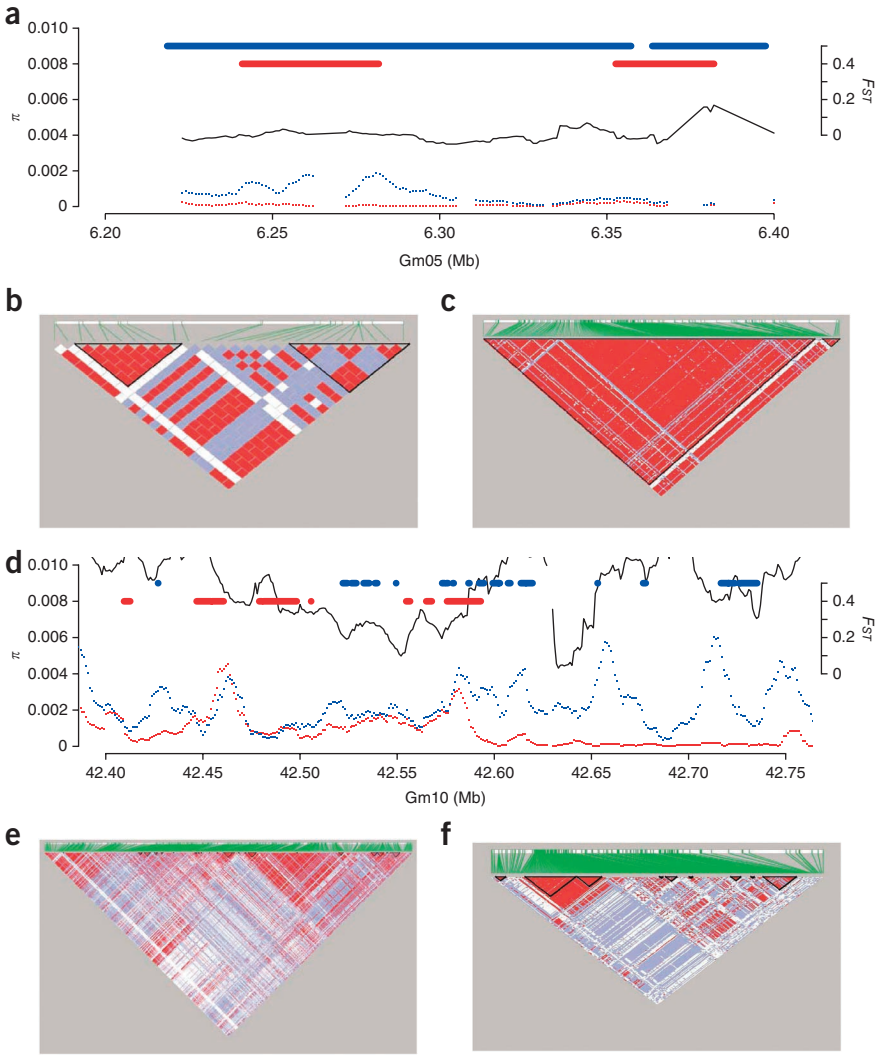


Figure 3 Patterns of LD blocks in two genomic regions. (**a**–**c**) LD blocks in chromosome 5 (~6.2–6.4 Mb). (**d**–**f**) LD blocks in chromosome 10 (~42.6–42.8 Mb). Location of LD blocks for wild (blue segments) and cultivated (red segments) soybeans, F_{ST} value (black line), and genomic diversity of wild (blue dotted line) and cultivated (red dotted line) soybeans are shown in **a** and **d**. Red and white spots indicate strong ($r^2 = 1$) and weak ($r^2 = 0$) LD, respectively, for wild (**b** and **e**) and cultivated (**c** and **f**) soybeans.

We assessed gene functional categories (selected groups are shown in **Supplementary Fig. 12**) of genes that had an average Nonsyn/Syn ratio that deviated significantly from the whole-genome average. Overall, we found that genes that had essential functions (for example, genes encoding enzymes for essential metabolism, transcription, translation, histones and ubiquitin-pathway components) tended to have a low ratio (χ^2 test with Bonferroni correction; $P < 0.01$), which is similar to previous findings in bacteria[30]. In contrast, genes that were required for regulatory processes or recognition of external signals (for example, proteins with leucine-rich repeats (LRRs) and the nucleotide binding adaptor (NB-ARC) domains that mediate protein-protein interactions and functions that recognize different external stimuli, such as strain-specific pathogens[31]) exhibited a high ratio (χ^2 test with Bonferroni correction; $P < 0.01$), which is consistent with previous findings in *A. thaliana*[29]. Many of these large-effect SNPs were associated with proteins containing LRRs and NB-ARC, which serve to recognize different external stimulants (for example, strain-specific pathogens)[31]; this is consistent with our findings from our Nonsyn/Syn ratio analysis.

Previous studies have indicated that whole-genome duplication (WGD) events can cause gene loss and rapid functional diversification[32,33]. WGD is considered an important source for promoting evolution because the extra genes can be mutated without risking loss of original gene function; this can potentially produce new genes and functions. Although most of these genes will be silenced within a few million years, a few survivors may be subjected to strong purifying selection[34]. Given that the last soybean WGD occurred relatively recently (~13 million years ago) in comparison to that of all other sequenced plants[5], we had an opportunity to study the impact of duplicated genes on genome evolution.

We determined the average Nonsyn/Syn ratio for duplicated regions and found that it was marginally lower than the whole-genome average (1.16 versus 1.37, respectively), which indicated that the high average Nonsyn/Syn ratio of the soybean genome cannot solely be attributed to gene duplication. We then calculated the ratio in each member of 1,237 annotated gene pairs and categorized them into three groups: (i) LL, in which both members were lower than average, including 460 pairs (37%); (ii) HL, in which one member was higher and the other lower than average, including 592 pairs (48%); and (iii) HH, in which both members were higher than average, including 185 pairs (15%).

To understand how duplicated gene pairs evolved, we determined the ratio of fixed nonsynonymous (N_F) versus synonymous (S_F) nucleotide differences of all gene pairs and the ratio of polymorphic nonsynonymous (N_P) versus synonymous (S_P) nucleotide differences of each gene member in the population. We deduced fixation by comparison between two members of each duplicate gene pair. A total of 362 gene pairs had a significantly lower N_P/S_P ratio than N_F/S_F ratio (Fisher's exact test; $P < 0.01$), and, of the 362 pairs, 38 pairs were within the LL group described above. Both members of the LL group were relatively conserved (low Nonsyn/Syn) and, hence, may have evolved new functions after duplication. Some of the 38 pairs (**Supplementary Table 6**) might have undergone neofunctionalization and been subjected to purifying selection.

Gene content variation

A pan-genome refers to the identification of individual- or population-specific sequences that may contain important information relevant to the subject's uniqueness[35]. To better understand the genetic changes associated with domestication, we set out to identify unique genomic differences between wild and cultivated soybeans. We compared *de novo* sequencing data of W05 (wild) with the reference cultivated soybean genome and identified 186,177 insertions or

deletions (>50% smaller than 5 bp) that passed our filtration criteria (Online Methods). A total of 4,444 and 1,148 large PAVs (>500 bp) were absent in the reference and W05 genomes, respectively (**Fig. 2**). We annotated the large PAVs using the AUGUSTUS and Genewise programs[36,37] and identified 856 genes. These fell into different gene categories (**Supplementary Fig. 13**), with a higher proportion (>40%) of genes relating to metabolic and catalytic processes, binding and other cellular processes. Additionally, we found that 28 gene fragments (**Supplementary Table 7**) that were absent in all cultivated accessions were primarily related to disease resistance and metabolism. The presence or absence of these and other genes may be indicative of different selective forces acting on or promoting the survival of wild and cultivated soybeans given their different habitats and the breeding practices during domestication.

DISCUSSION

This study provides the first comprehensive resequencing data of wild and cultivated soybean genomes and of Fabaceae family members. The availability of this data, generated from 31 wild and cultivated soybean genomes, along with a tag SNP set for QTL mapping and association studies, will aid in carrying out future in-depth studies of population genetics, marker-assisted breeding and gene identification in soybeans. For breeding applications, our identification of the high LD nature in the soybean genome indicates that marker-assisted breeding is a better choice for soybean improvement, whereas map-based cloning using genetic populations will be challenging.

Our finding of higher genomic diversity in wild soybeans as compared to cultivated soybeans is consistent with there being a negative effect caused by a genetic bottleneck and/or influenced by human selection in cultivated soybeans. The unusual Nonsyn/Syn ratio of SNPs in soybeans may be due to the high LD nature of the soybean genome, which could lead to an indirect consequence of continuing strong selection on a linked locus that permits newly derived 'hitchhiking' alleles to accumulate. The elevated average Nonsyn/Syn ratio of SNPs specific to cultivated soybeans and their greater accumulation of deleterious mutations can probably be attributed to the domestication-associated Hill-Robertson effect[38].

The information we provide on LD block locations in wild and cultivated soybean genomes can also facilitate the identification of genes related to the domestication and human selection processes. The presence of high LD in general in the soybean genome indicates that soybeans would serve as a good model for studying the genomes of crops with extreme LD.

Our data also indicate that the formation of cultivated-specific long LD blocks may have resulted from a combination of the lower genetic diversity of cultivated soybeans and a low frequency of genetic recombination.

Additionally, the nature of soybean fertilization, which results in high inbreeding and thus a reduction in recombination, may have promoted low genome diversity in the soybean and high LD. This could be further aggravated by the domestication process. The prevalent use of specific purebred cultivated soybeans, resulting in increased acreage of the same variety, has probably created further constraints on genetic recombination. The impact of soybean breeding along with selection forces during domestication may also have increased hitchhiking of deleterious mutations and, as a consequence, resulted in loss of fitness in the soybean[38].

As there is no sexual barrier between wild and cultivated soybeans, on the basis of our analyses, the availability of wild germplasms could be an important tool to expand the allelic pool of cultivated soybeans through introgression. The potential importance of wild soybeans for maintaining and improving cultivated soybean production and evidence of the shrinkage of its natural habitat makes it essential that steps be taken to protect wild soybeans.

URLs. Statistics of soybean, http://www.soystats.com/; *Glycine max* genome, http://www.phytozome.net/soybean.php; *Lotus japonicus* genome, http://www.kazusa.or.jp/lotus/; SOAP and SOAPsnp, http://soap.genomics.org.cn/; LASTZ, http://www.bx.psu.edu/miller_lab/; JGI, http://genome.jgi-psf.org/soybean/soybean.download.html.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The sequence data has been deposited in NCBI Short Read Archive with accession number SRA020131. The whole-genome SNP data set has been deposited in NCBI dbSNP with accession number records from ss244318098 to ss250607844.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

T. Han, X. Yan, H. Liao, B. Zhuang and Y.-K. Lau provided valuable advice, information and other aid. This work was partially supported by the Hong Kong RGC General Research Fund 468610 (to H.-M.L.), the Hong Kong UGC AoE Center for Plant and Agricultural Biotechnology Project AoE-B-07/09 and a special fund from the Resource Allocation Committee, The Chinese University of Hong Kong (to H.-M.L. and S.S.-M.S.). We also acknowledge the funding support from the National Natural Science Foundation of China (30725008), the Chinese 973 program (2007CB815703; 2007CB815705), Chinese Ministry of Agriculture (948 program), the Shenzhen Municipal Government of China and grants from Shenzhen Bureau of Science Technology & Information, China (ZYC200903240077A; CXB200903110066A). We thank L. Goodman for assistance in editing the manuscript.

AUTHOR CONTRIBUTIONS

H.-M.L., G.Z., S.S.-M.S. and Jun Wang managed the project. H.-M.L., X.X., X.L, N.Q. and G.Y. designed the experiments and led the data analysis. W.H., B.W., J.L., W.C., M.J. and Jian Wang contributed to DNA sequencing and bioinformatics. F.-L.W., M.-W.L. and G.S. prepared samples and contributed to data analysis. H.-M.L., X.X. and X.L. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Hymowitz, T. On the domestication of soybean. *Econ. Bot.* **24**, 408–421 (1970).
2. Hymowitz, T. & Harlan, J.R. Introduction of soybean to North America by Samuel Bowen in 1765. *Econ. Bot.* **37**, 371–379 (1983).
3. Hyten, D.L. *et al.* Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**, 1937–1944 (2007).
4. Hyten, D.L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA* **103**, 16666–16671 (2006).
5. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
6. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
7. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
8. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
9. Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**, 433–436 (2009).
10. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
11. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
12. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).

13. Hernandez, R.D. *et al.* Demographic histories and patterns of linkage disequilibrium in Chinese and Indian Rhesus Macaques. *Science* **316**, 240–243 (2007).
14. Caicedo, A.L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
15. Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
16. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
17. Kim, S. *et al.* Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39**, 1151–1155 (2007).
18. Zhu, Q., Zheng, X., Luo, J., Gaut, B.S. & Ge, S. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* **24**, 875–888 (2007).
19. Flint-Garcia, S.A., Thornsberry, J.M. & Buckler, E.S. IV. Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* **54**, 357–374 (2003).
20. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
21. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
22. The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**, 528–532 (2009).
23. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
24. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
25. Liu, B. *et al.* QTL mapping of domestication-related traits in soybean (*Glycine max*). *Ann. Bot. (Lond.)* **100**, 1027–1038 (2007).
26. Li, H. *et al.* Identification of QTL underlying vitamin E contents in soybean seed among multiple environments. *Theor. Appl. Genet.* **120**, 1405–1413 (2010).
27. Huang, Z.-W., Zhao, T.-J., Yu, D.-Y., Chen, S.-Y. & Gai, J.-Y. Correlation and QTL mapping of biomass accumulation, apparent harvest index, and yield in soybean. *Acta. Agron. Sin.* **34**, 944–951 (2008).
28. McNally, K.L. *et al.* Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. USA* **106**, 12273–12278 (2009).
29. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
30. Jordan, I.K., Rogozin, I.B., Wolf, Y.I. & Koonin, E.V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).
31. Dangl, J.L. & Jones, J.D.G. Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826–833 (2001).
32. Blanc, G. & Wolfe, K.H. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**, 1679–1691 (2004).
33. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 5454–5459 (2005).
34. Lynch, M. & Conery, J.S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
35. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
36. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
37. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
38. Lu, J. *et al.* The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *TIG* **22**, 126–131 (2006).
39. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

ONLINE METHODS

Sample preparation and sequencing. Seeds of soybean accessions (**Supplementary Table 1**) were germinated at 25 °C for 5 d on vermiculite in a dark chamber. After the 5 d, etiolated seedlings were collected for genomic DNA extraction using a standard CTAB (cetyl trimethylammonium bromide) protocol[40]. Sequencing libraries were constructed according to the manufacturer's instructions (Illumina). Short reads were generated by applying the base-calling pipeline, SolexaPipeline-0.3 (Illumina).

Short Oligonucleotide Alignment Program 2 (SOAP2)[6] was used to map raw pair-ends reads onto the JGI Glycine max reference genome (Glycine_max_Williams_82 8x Release v1.01). On the basis of the mapping results, reads were classified into three categories: 'uniquely aligned', 'repeatedly aligned' and 'unaligned'. The trimming strategy for mismatches was as described in the supplementary methods of a previous report[8]. Duplicated reads caused by the PCR process were removed by a PERL script. For each accession, more than 85% of the reads were properly aligned to the reference genome.

SNP detection and validation. SNPs were detected in four consecutive steps. (i) SOAPsnp[7] was used to calculate the likelihood of genotypes of each individual. (ii) All the individual likelihood files were then integrated to produce a pseudo-genome for each site by maximum likelihood estimation followed by filtering using criteria that included copy number (≤1.5), sequencing depth (according to average depth of each accession) and quality. SNPs that passed the rank-sum test ($P \geq 0.005$) were included in the final SNP set. (iii) Using the final SNP set as prior information, SNP calling was performed for both the wild (*G. soja*) and the cultivated (*G. max*) soybeans to generate two subsets. (iv) Base types were allocated back to each individual depending on genotypes of the final SNPs and each individual likelihood file.

Three methods were applied to validate the identified SNPs. First, *de novo* assembly of the genome of W05 was performed with a total depth of ×80. The SNPs detected using data from *de novo* sequencing and resequencing of W05 were compared. Of the W05 SNPs (~2.0 M) detected by resequencing data, 63.15% were identical to the SNPs detected using *de novo* data. Of the remaining SNPs, 35.06% were removed either by our filtration criteria or because the sequencing depth was too low for detection. The false-positive SNP detection rate was estimated to be 1.79%. Conversely, 3.46% of the SNPs found in *de novo* W05 sequencing data were not detected by the group SNP calling, giving a false-negative SNP detection rate of 3.46%.

Second, we used the resequencing data of accession C08 (which is closely related to the reference genome) for SNP evaluation. In the final SNP set, there were 229,104 SNPs in C08, of which 50,620 SNPs are homozygous. A total of 14,873 homozygous SNPs were in genomic regions with greater than ×4 depth; thus, these SNPs may be the result of sequencing errors or false SNP detection. The maximum possible sequencing error was about 1.5 kb per whole genome and the estimated false detection rate of SNPs was 0.24%. As C08 is not identical to the reference genome, the actual false detection rate is likely to be overestimated.

Third, we selected 30 rare SNPs and 100 random SNPs in C08 for Sanger sequencing, and determined that SNP calling had an accuracy of ~97%.

Population analysis. To construct the phylogenetic tree, we used *Lotus japonicus* as the outgroup. The genome of *L. japonicus* was obtained online (see URLs), and we used BLASTZ[41] to identify homologous regions between *G. max* and *L. japonicus*. SNPs within these regions were extracted, and genotypes of *L. japonicus* were used to provide the outgroup information at corresponding positions. The neighbor-joining tree was constructed by MEGA4 (ref. 42) under the *p*-distances model using these SNPs. Excluding SNPs from individuals that had missing data or heterozygous genotypes, 966,612 SNPs were used to construct the population structure using the program STRUCTURE[10]. The length of the burn-in period was set to 30,000. The number of the MCMC reps after burn-in was set to 10,000. The number of populations considered was set from 2–7.

Simulation of possible population changes. Parameter inference was done with the software package ∂a∂i (version 1.2.3)[12] using the folded joint-allele frequency of the synonymous SNPs (total: 83,559) in wild and cultivated soybeans. We established a model with a bottleneck in the cultivated population after splitting from the wild population, followed by population recovery. We also permitted possible changes in population size of the wild population (**Supplementary Fig. 6a**). After fitting the model (**Supplementary Fig. 6b**), we used the software *ms*[43] to simulate the frequency of SNPs under these demographic parameters (**Supplementary Fig. 6c**,**d**).

LD decay detection. Correlation coefficient (r^2) of alleles was calculated to measure LD level in both wild and cultivated soybeans using Haploview[16]. The parameters were set as follows: –maxdistance 1000 –dprime –minMAF 0.1 –hwcutoff 0.001. The average r^2 value was calculated for each length of distance, and LD decay figures were drawn using R script for both cultivated and wild soybean populations.

To find LD blocks in both wild and cultivated soybean populations, the parameters '–blockoutput GAB –pairwiseTagging' were added to the program. The maxdistance was first set to 250 and the blocks were then gradually extended (by setting a higher maxdistance value and re-running the program) to determine the best maxdistance for each LD block.

Identification of introgression. Introgression of genomic segments from wild soybean to cultivated soybean was identified. SNPs with missing data and heterozygous genotypes in individual accessions were excluded. The genotypes of SNPs in a sliding 100-kb window were scored for each individual and the ratio of shared genotype in cultivated versus wild soybeans was calculated in each window. Regions with a ratio lower than 0.5 were defined as introgressions.

SNP diversity and F_{ST} calculation. The average pairwise divergence within a population (θ_π) and the Watterson's estimator (θ_w)[23] were estimated for the whole genome of both wild and cultivated soybean populations. Sliding windows of different sizes (10 kb, 100 kb and 500 kb) that had a 90% overlap between adjacent windows were used to estimate θ_π, θ_w and Tajima's *D* (ref. 24) for the whole genome. In each window, these parameters were calculated with an in-house PERL script. To display the pattern in the whole genome, a window of 500 kb was used. To measure the population differentiation, F_{ST} was calculated[44].

Analysis of duplicate genes. Annotated genes of *G. max* were from the JGI website (see URLs), from which we performed a self-to-self BLAST. For each best hit, a four-fold degenerate transversion (4DTv) ratio was calculated. According to the distribution of the 4DTv ratio of all the gene pairs, the 4DTv ratio of recently formed duplicate genes was identified. Gene pairs in which both genes had a 4DTv ratio lower than 0.12 were identified as recently duplicated.

The CDS sequence of each selected duplicated gene was aligned by BLASTZ[41] to identify nonsynonymous and synonymous mutations between the gene pair. Using the identified SNPs, a McDonald Kreitman test[45] was performed to compare the variations within the gene and between the two duplicate genes.

Identification of present and absent variations (PAVs). The same procedure described in building the human pan-genome[35] was used to identify the PAVs between wild and cultivated soybeans. We made use of the *de novo* assembled genomic sequence of one wild soybean accession (W05; data not shown) in this analysis. All the assembled contigs were aligned to the *G. max* reference genome using BLAT[46] with the –fastmap option enabled. Using the alignment results, the location of the scaffold for each contig was determined. The alignment with the longest length in linear orientation between a scaffold and the reference was chosen as the 'best-hit' of the scaffold. Subsequently, the scaffolds were aligned against the located regions on the *G. max* genome by LASTZ (see URLs). The unmapped sequences derived from the LASTZ alignment were identified and re-aligned with the *G. max* reference using BLASTn[47]. Scaffold fragments with identity lower than 90% to any regions of the reference genome were defined as new sequences to identify the PAVs between wild and cultivated soybeans.

The PAVs and the flanking sequences were extracted from the reference genome. Raw reads of each individual were mapped back to these sequences. By comparing the depth of sequences between PAVs and the respective flanking sequences on the reference genome, the PAVs were assigned to each individual.

doi:10.1038/ng.715

40. Doyle, J.J. & Doyle, J.L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
41. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
42. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
43. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
44. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
45. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
46. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
47. 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).

Supplemental Information

Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection

Hon-Ming Lam[1*]†, Xun Xu[2,3,*], Xin Liu[1,2,*], Wenbin Chen[2,*], Guohua Yang[2,*], Fuk-Ling Wong[1], Man-Wah Li[1], Weiming He[2], Nan Qin[2], Bo Wang[2], Jun Li[2], Min Jian[2], Jian Wang[2], Guihua Shao[1,4], Jun Wang[2,5]†, Samuel Sai-Ming Sun[1]†, Gengyun Zhang[2,3]†

[1] State Key Laboratory of Agrobiotechnology and School of Life Sciences, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong SAR.

[2] BGI-Shenzhen, Shenzhen, China.

[3] Key Laboratory of Genomics, Ministry of Agriculture, BGI-Shenzhen, China.

[4] Institute of Crop Sciences, The Chinese Academy of Agricultural Sciences, Beijing, China.

[5] Department of Biology, University of Copenhagen, Copenhagen, Denmark

* These authors contributed equally to this work.

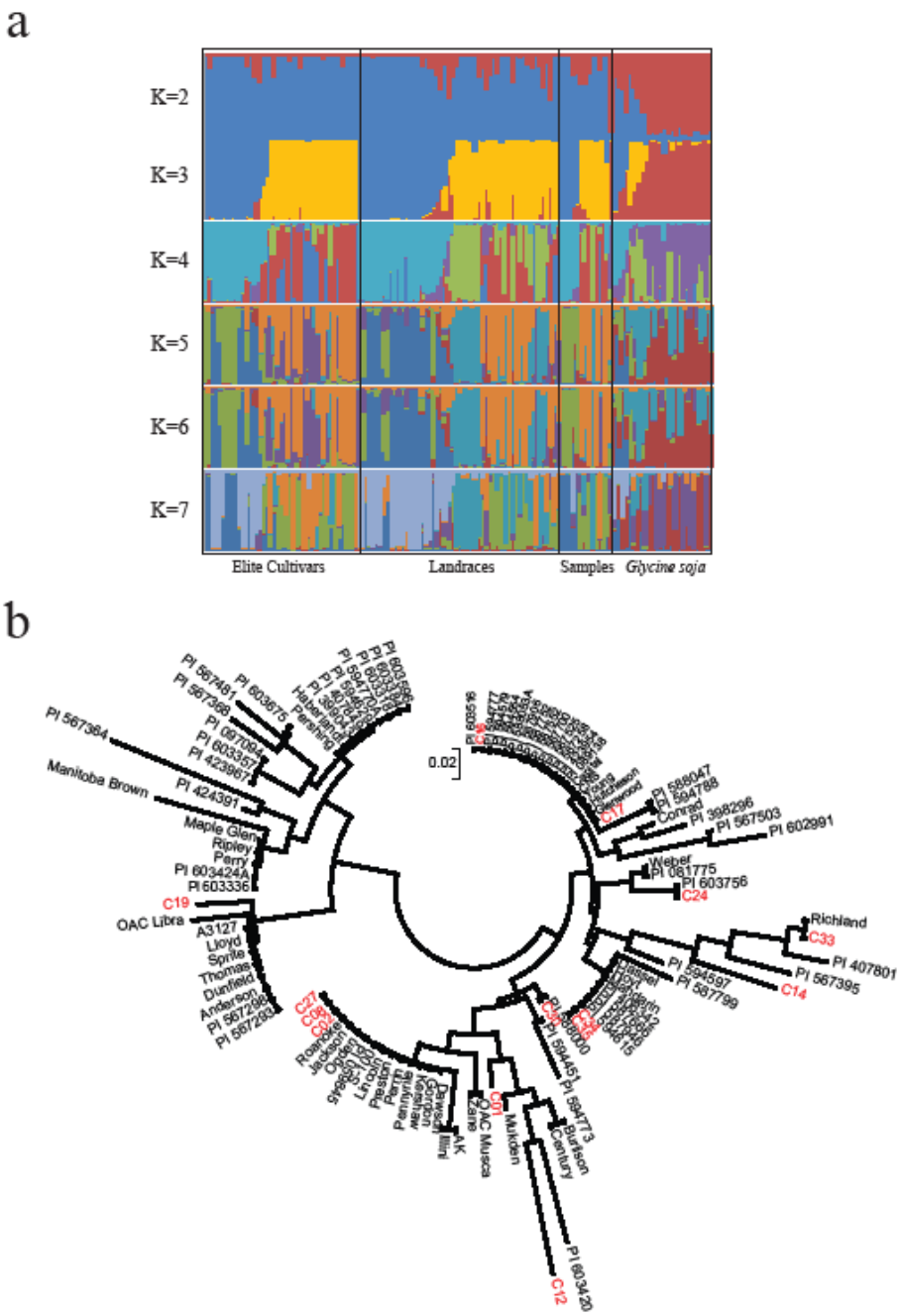
† To whom correspondence should be addressed.

This file contains supplementary figures and tables in the following orders:

- Supplemental Figures
  - **Supplementary Figure 1** Analysis of the population structure and phylogenetic relationship of cultivated accessions with soybeans of known genetic structure.
  - **Supplementary Figure 2** Bayesian clustering of samples using the STRUCTURE program.
  - **Supplementary Figure 3** Comparison of genome diversity between wild and cultivated soybeans.
  - **Supplementary Figure 4** Statistics of SNPs of wild and cultivated soybeans.
  - **Supplementary Figure 5** Occurrence of minor SNP alleles.
  - **Supplementary Figure 6** Comparison of distribution of minor alleles between simulated model and actual data.
  - **Supplementary Figure 7** Proportion of LD blocks with different block sizes.
  - **Supplementary Figure 8** Boxplot of Tajima's *D* values in cultivated soybean (whole genome and high LD region).
  - **Supplementary Figure 9** F_{ST} values in genomic regions of domestication QTLs.
  - **Supplementary Figure 10** Frequency distribution of θ_{π}, F_{ST}, and LoD values.
  - **Supplementary Figure 11** Distribution of large effect SNPs in different gene categories.
  - **Supplementary Figure 12** Ratio of average nonsynonmymous versus synonmymous nucleotide changes in annotated genes.
  - **Supplementary Figure 13** Gene categories and distribution of large Present/Absent Variations (PAVs; >500 bp) between the wild accession W05 and the reference genome Williams 82.

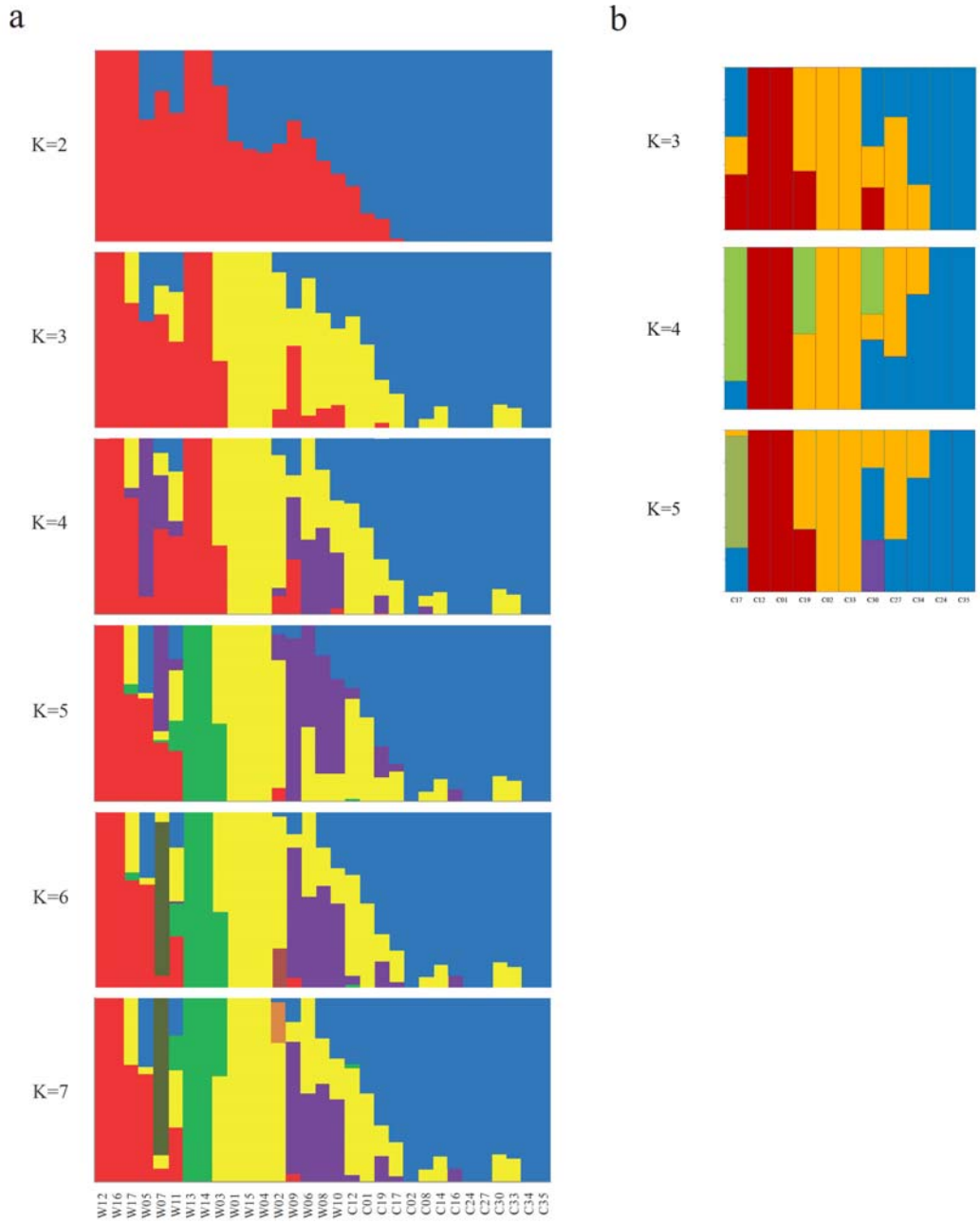
- Supplementary Tables
  - **Supplementary Table 1** Summary of soybean accessions: regions of collection/popularization and coverage and mean depth of re-sequencing.
  - **Supplementary Table 2** General phenotypic differences between wild and cultivated soybeans.
  - **Supplementary Table 3** Fixed, deleterious, and large-effect SNPs in wild and cultivated soybeans.
  - **Supplementary Table 4** Comparison of nonsynonymous/synonymous (Nonsyn/Syn) SNP ratio of soybean versus rice.
  - **Supplementary Table 5** Nonsynonymous/synonymous (Nonsyn/Syn) SNP ratio in LD blocks.
  - **Supplementary Table 6** Recently duplicated gene pairs that may have undergone neofunctionalization and subjected to purification selection.
  - **Supplementary Table 7** Present/Absence Variations (PAVs) absent in all cultivated soybeans.

SUPPLEMENTARY FIGURES

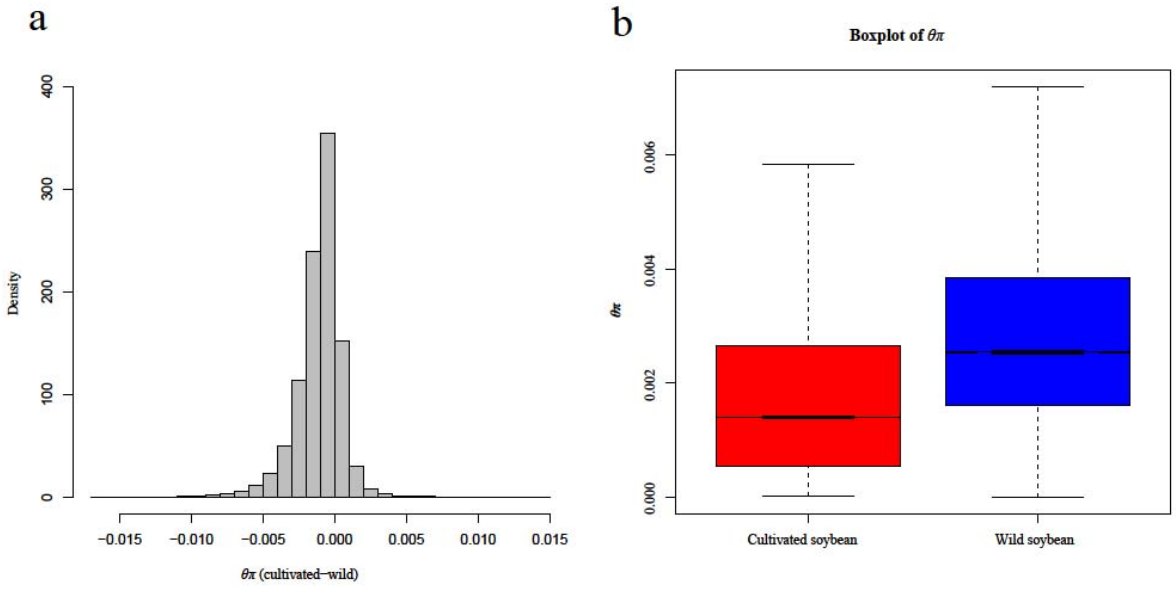


Supplementary Figure 1 Analysis of the population structure and phylogenetic relationship of cultivated accessions with soybeans of known genetic structure. (**a**) Bayesian clustering (STRUCTURE, *K*=2-7). (**b**) Neighbor-joining phylogenetic tree. SNP data that can be located on the reference genome Williams 82 was extracted from a previous soybean genetic structure analysis[1]. Analysis was performed after including the cultivation accessions of this study (samples in (**a**); colored red in (**b**)) to the original population.

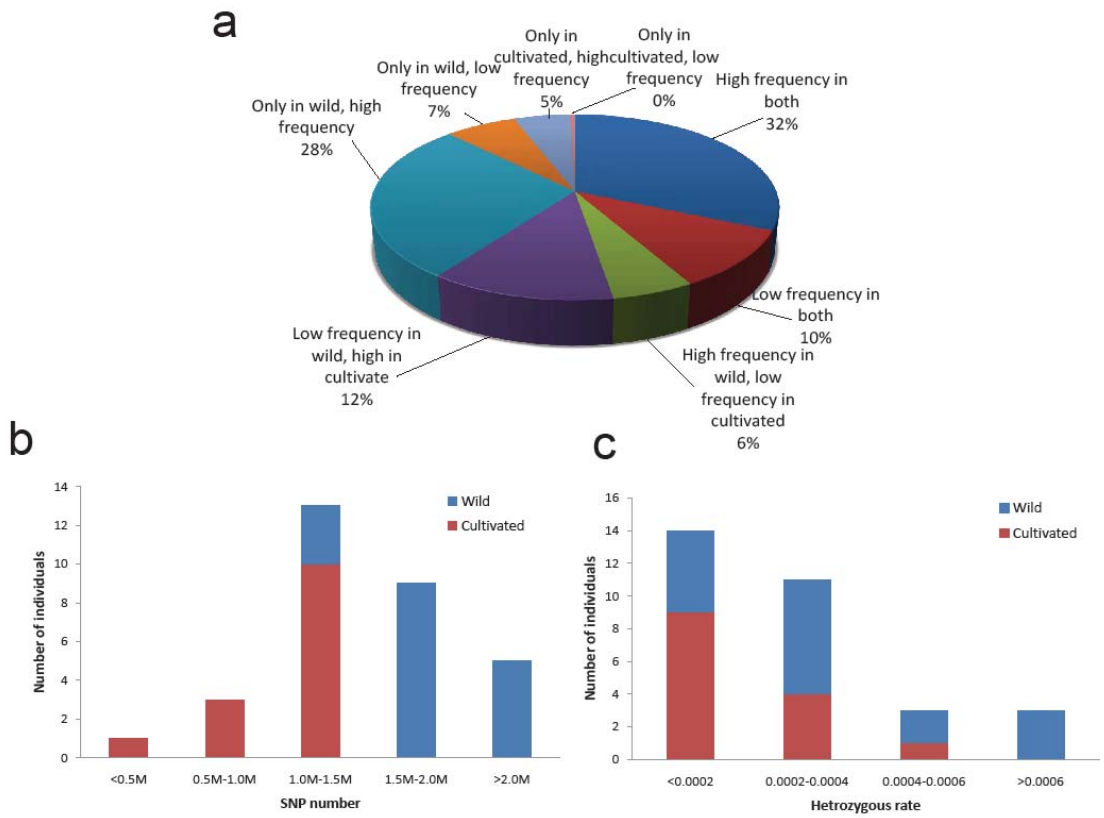
1. Hyten, D.L. et al. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**, 1937-1944 (2007).



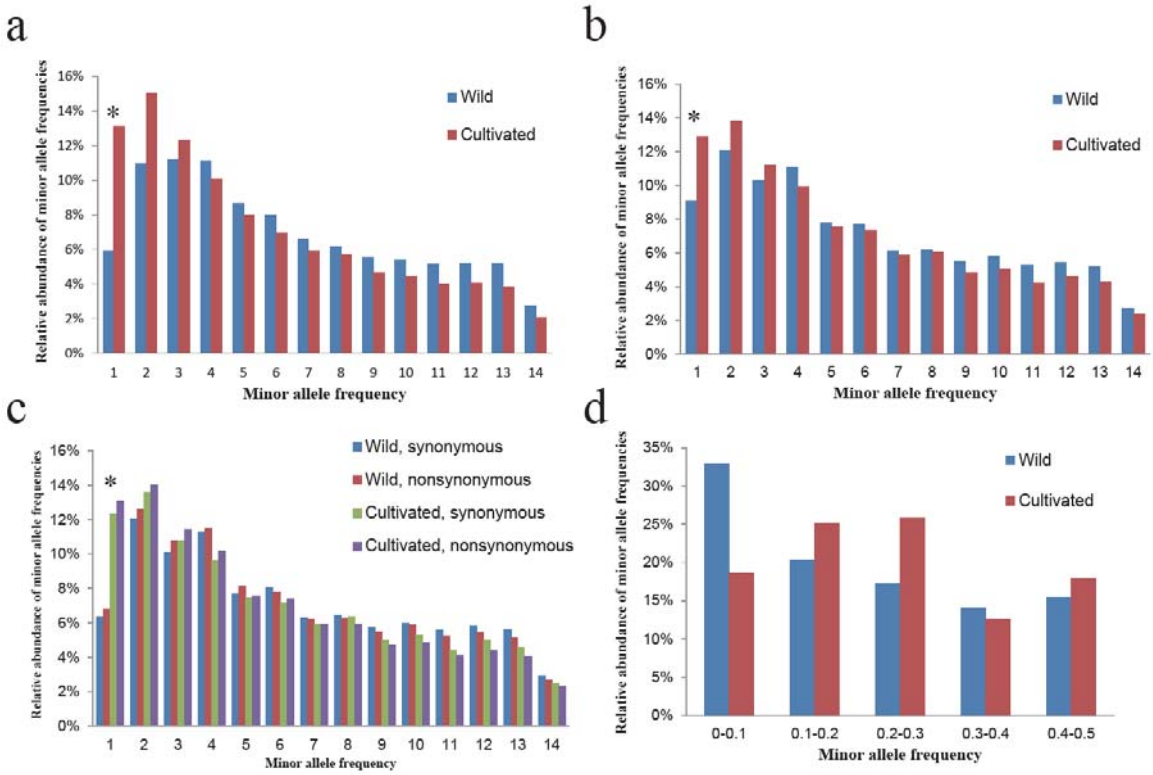
Supplementary Figure 2 Bayesian clustering of samples using the STRUCTURE program. (**a**) STRUCTURE analysis of wild and cultivated soybean (K=2 to 7). The average value of ln likelihood when *K* changed from 2 to 7 was -25011999, -22733525, -21552723, -20689707, -20763282, and -20859494, respectively. (**b**) STRUCTURE analysis using only cultivated soybean accessions from Mainland China (K=3 to 5). The average value of ln likelihood when *K* changed from 3 to 5 was -7187762, -7012090, and -7079800, respectively. The correlation to geographical distribution was shown. For examples, C24, C34, and C35 were popularized in three adjacent provinces in southern China; C01 and C12 were popularized in two adjacent provinces in central part of China; C19, C02, and C33 were popularized in three adjacent provinces in northeast China.



Supplementary Figure 3 Comparison of genome diversity between wild and cultivated soybeans. (**a**) Distribution of difference of θ_π between cultivated and wild soybeans. (**b**) Boxplot of θ_π of cultivated and wild soybeans.

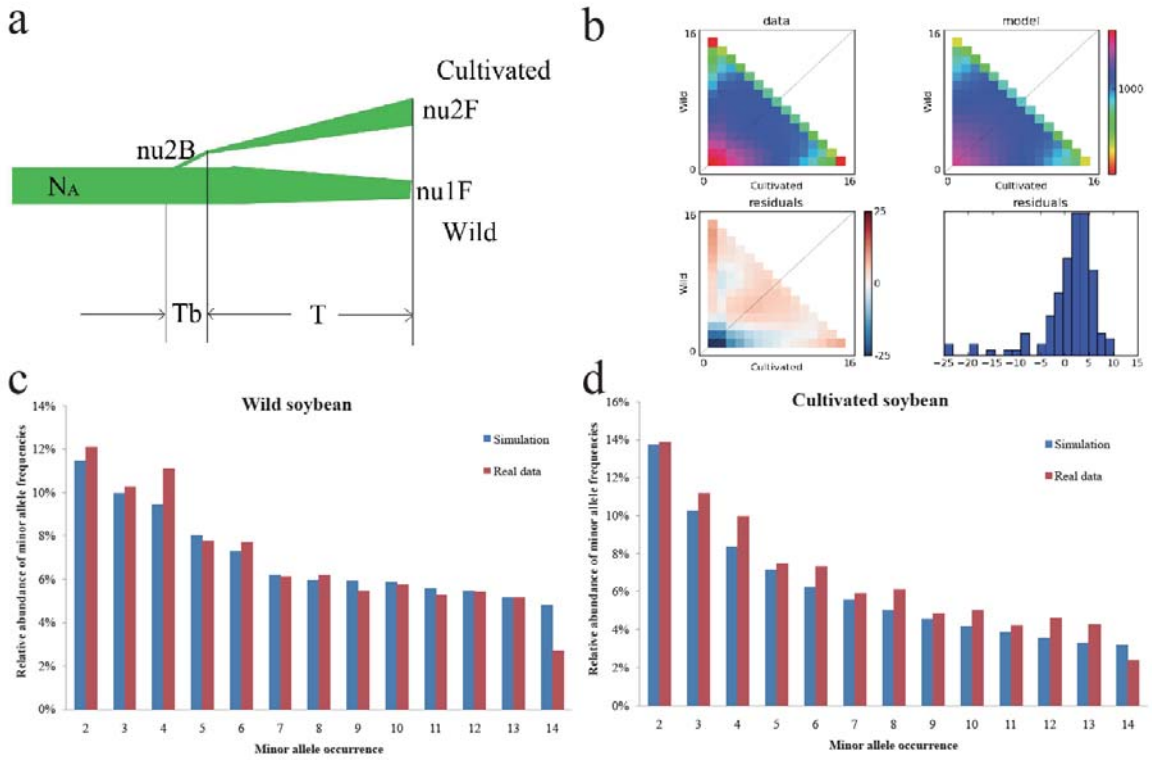


Supplementary Figure 4 Statistics of SNPs of wild and cultivated soybeans. (**a**) Frequency of common, wild soybean-specific, and cultivated soybean-specific SNPs. (**b**) Statistics showing higher number of total SNPs in wild soybeans than cultivated soybeans. (**c**) Heterozygous rate of SNPs in 31 soybean accessions.

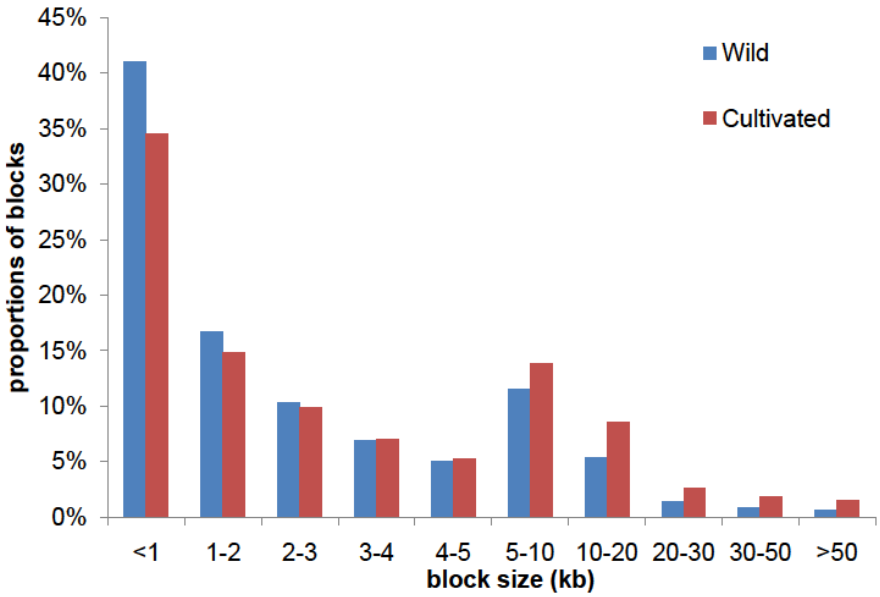


Supplementary Figure 5 Occurrence of minor alleles. Complete SNP data of 14 wild and 14 cultivated soybean accessions with best genomic coverage were analyzed. (**a**, **b**) Proportion of SNPs in the (**a**) whole genome or (**b**) genic regions was plotted against occurrence of minor alleles. (**c**) Distribution of synonymous and nonsynonymous SNPs in genic regions. (**d**) Proportion of SNPs plotted against minor alleles using a selected gene set as described in a previous report[2]. These results were comparable when the same set of genes was employed for the analysis, resulting in a higher proportion of minor alleles in wild soybeans. By contrast, when the analysis was extended to the (**a**) whole genome or (**b**) complete genic regions, the proportion of minor alleles was higher in cultivated soybeans. *There was an underestimation of singleton SNPs due to high stringency filtering during SNP calling.

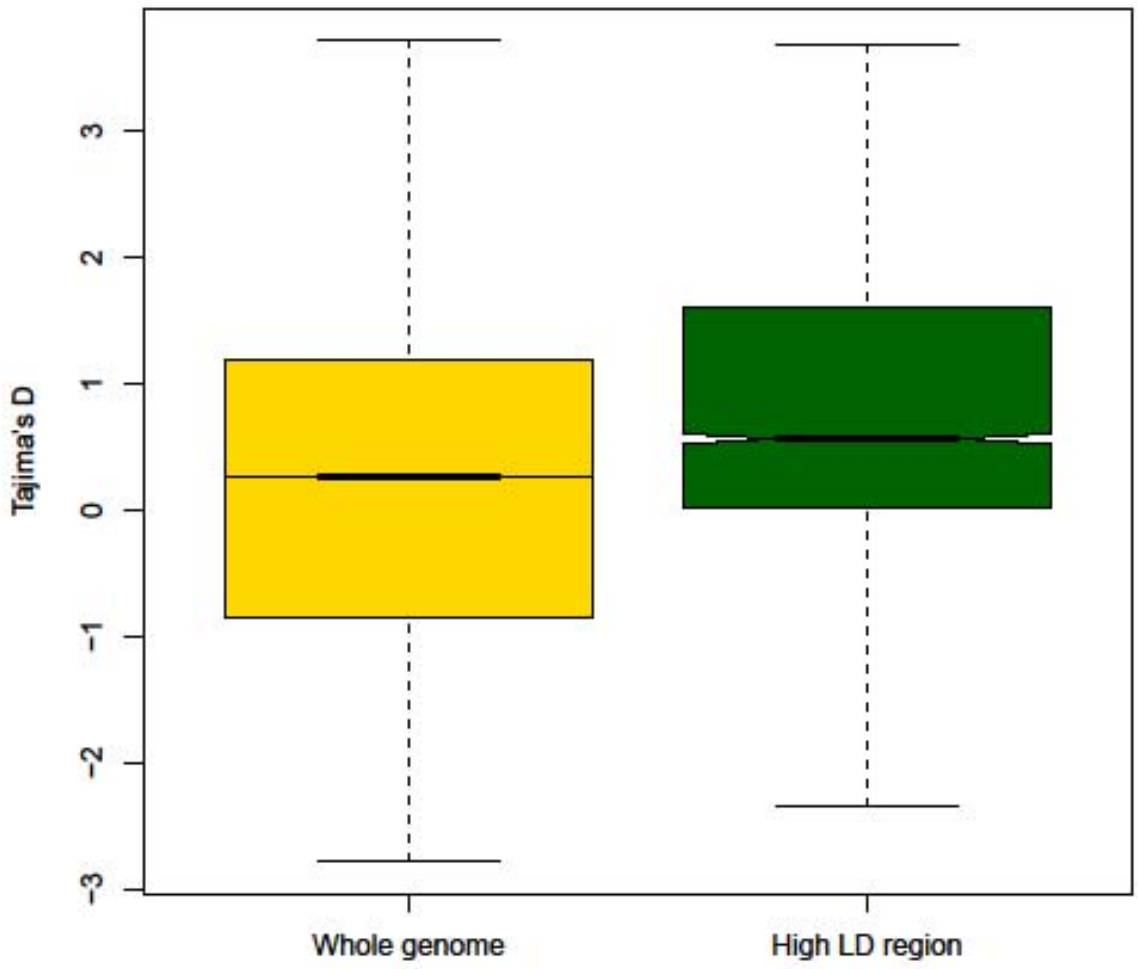
2. Hyten, D.L. et al. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA* **103**, 16666-16671 (2006).



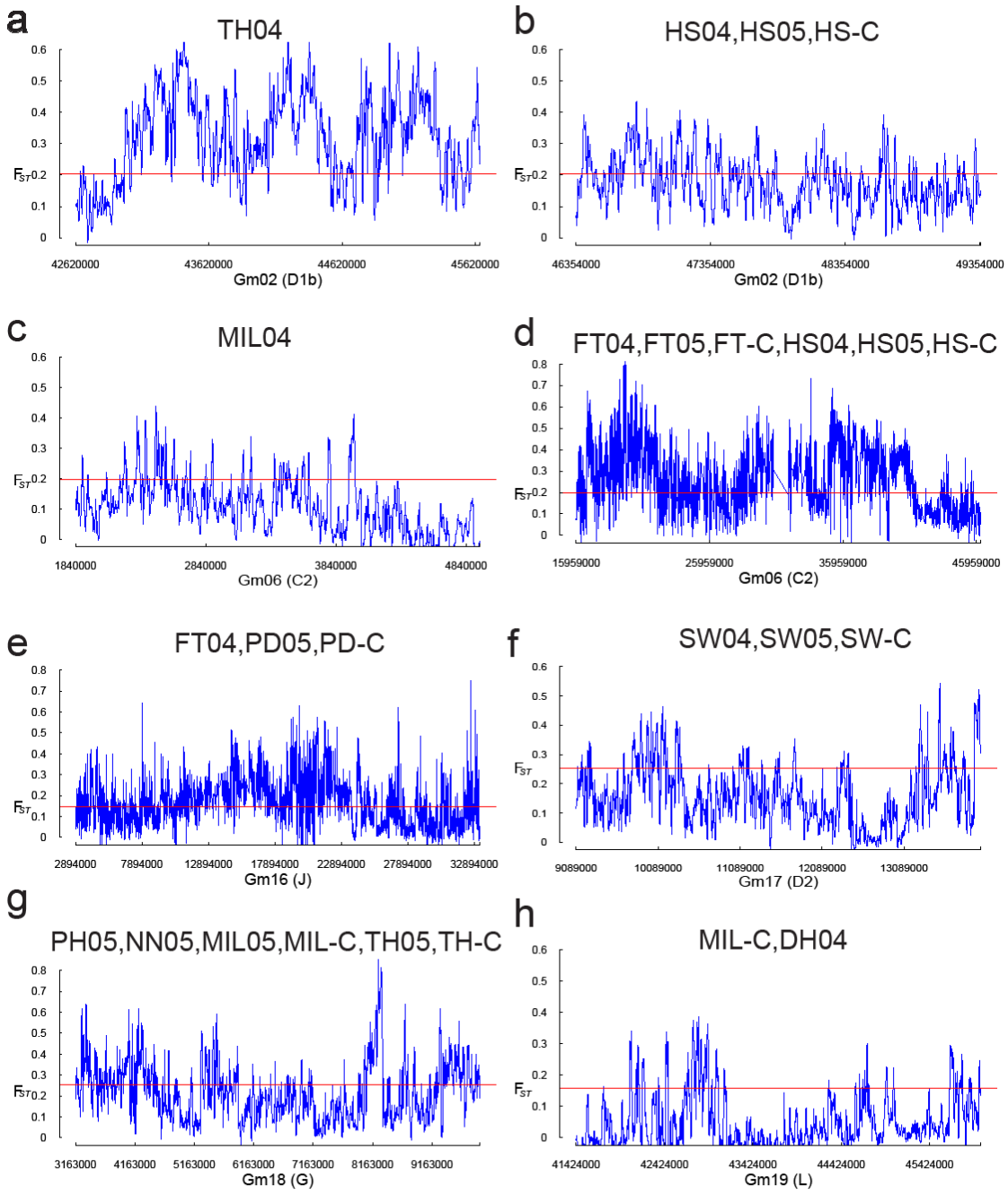
Supplementary Figure 6 Comparison of distribution of minor alleles between simulated model and actual data. The model depicted that after separation of wild and cultivated soybeans, the natural habitat of wild soybeans shrank and the growth areas of cultivated soybean increased. N_A: is the ancestor's effective population size. nu2B: the bottleneck size of cultivated soybeans, which is inferred to be $0.0165*N_A$; nu1F: the final effective population size of wild soybeans, which is inferred to be $0.732*N_A$. nu2F: the final effective population size of cultivated soybeans, which is inferred to be $0.0435*N_A$. Tb: the duration time of bottleneck in cultivated soybean, which is inferred to be $0.00078*4N_A$ generations. T: the time after the bottleneck till now, which is inferred to be $0.176*4N_A$ generations. (**a**) The model established to simulate the data. (**b**) The best fitting of the model to the parameter. The upper panel stands for the joint allele frequency spectrum of the data and the model with the inferred parameters. The lower panel shows the difference between the spectra of actual data and the model. (**c**) The minor allele frequency spectrum of wild soybean was compared to the simulated allele frequency spectrum. (**d**) The minor allele frequency spectrum of cultivated soybean was compared to the simulated allele frequency spectrum. The programs using for this simulation were described in the Online Methods.



Supplementary Figure 7 Proportion of LD blocks of different sizes.

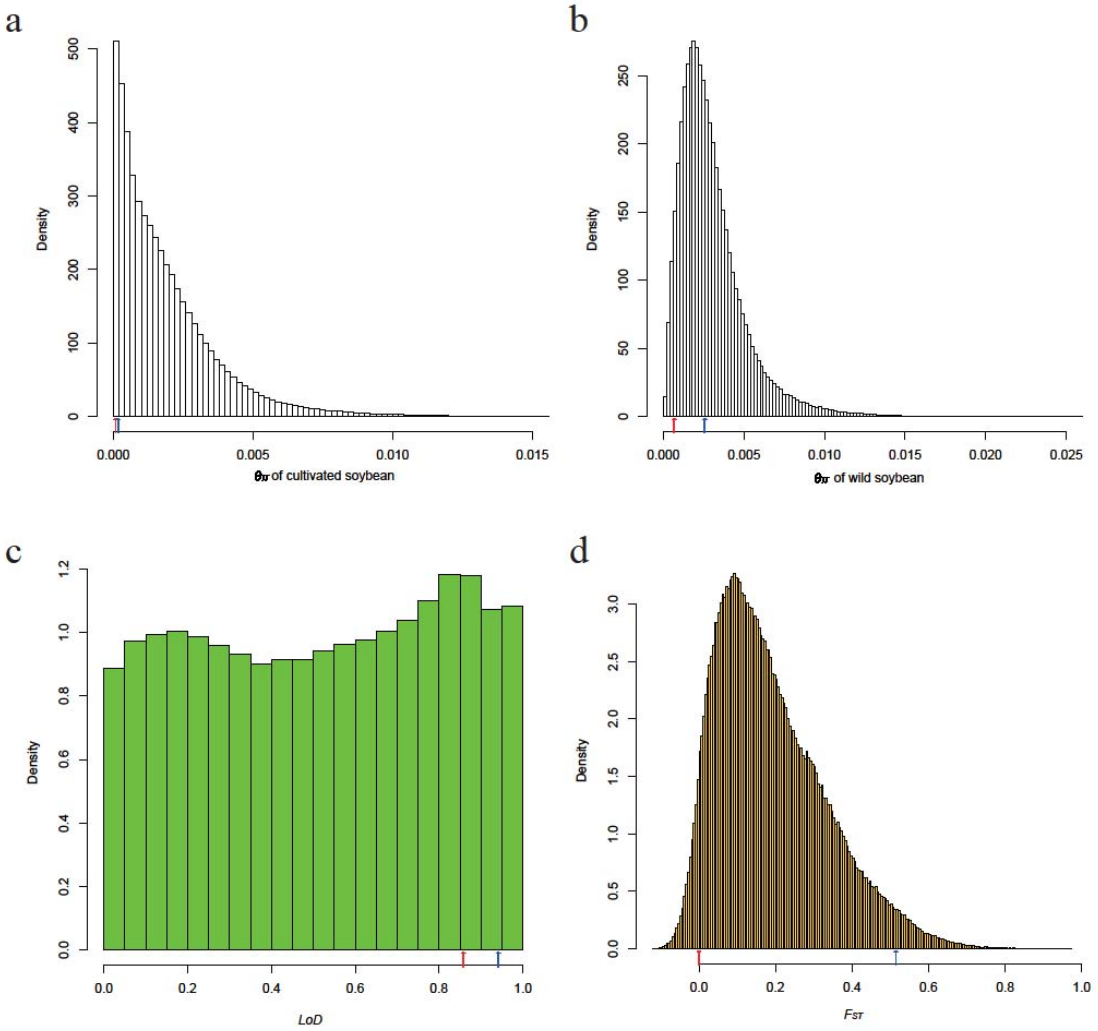


Supplementary Figure 8 Boxplot of Tajima's *D* values in cultivated soybean (whole genome and high LD region)



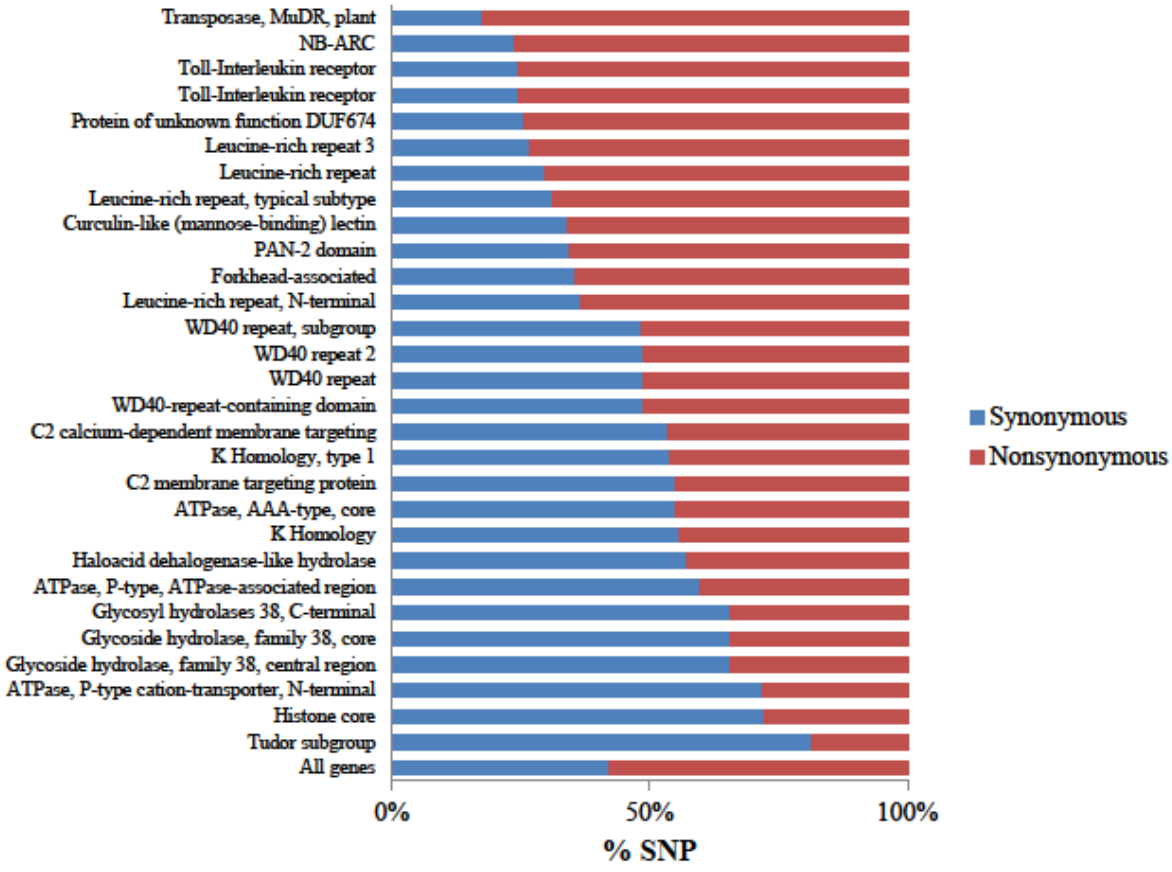
Supplementary Figure 9 F_{ST} values in genomic regions of domestication QTLs. The F_{ST} values (wild versus cultivated soybeans) were plotted for the genomic regions with known domestication QTLs. The approximate genomic locations of QTLs were estimated using the linkage map published by Liu et al.[3] DH: determinate habit; FT: flowering time; HS: hard seededness; MIL: maximum internode length; NN: number of nodes; PD: pod dehiscence; PH: plant height; SW: 100-seed weight; TH: twinning habit. The red line indicates the average F_{ST} value of each relevant chromosome.

3. Liu, B. et al. QTL mapping of domestication-related traits in soybean (*Glycine max*). *Ann. Bot.* **100**, 1027-1038 (2007).

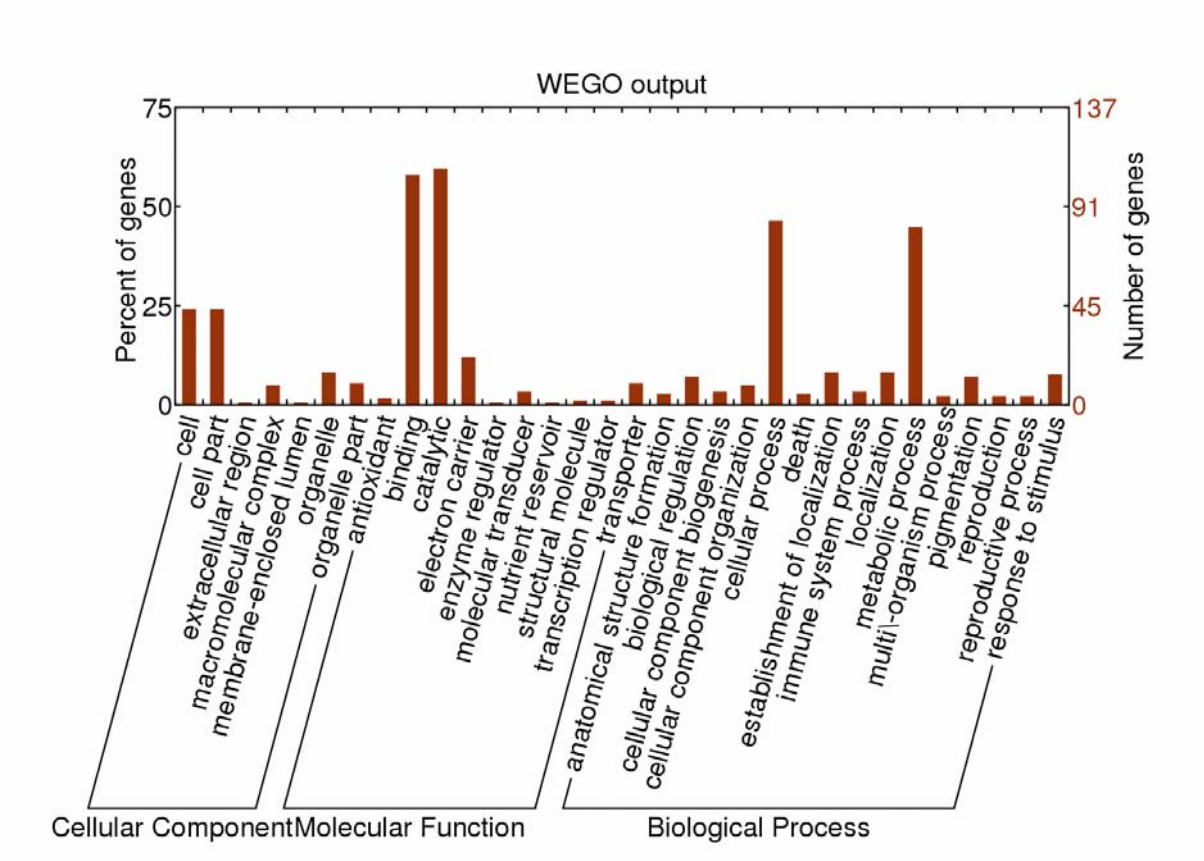


Supplementary Figure 10 Frequency distribution of θ_π, F_{ST}, and LoD values. Frequency distribution of $\theta\pi$ of cultivated soybeans (**a**) and wild soybeans (**b**), Loss of Diversity/LoD [$\theta\pi$ in cultivated soybeans-$\theta\pi$ in cultivated soybeans)/$\theta\pi$ in wild soybeans] in cultivated soybeans (**c**), and F_{ST} between wild and cultivated soybeans (**d**). The blue and red arrows indicated the positions on the frequency distribution curves, of the common high LD region on chromosome 5 or the cultivated specific high LD region on chromosome 10, respectively. (with reference to **Fig. 3**).

Supplementary Figure 11 Distribution of large effect SNPs in different gene categories.



Supplementary Figure 12 Ratio of average nonsynonmymous versus synonmymous nucleotide changes in annotated genes. Selected gene categories with ratios significantly different (χ-square test with *Bonferroni's* correction; p value <0.01) from genome average are shown.



Supplementary Figure 13 Gene categories and distribution of large Present/Absent Variations (PAVs; >500 bp) between the wild accession W05 and the reference genome Williams 82.

SUPPLEMENTARY TABLES

Supplementary Table 1 Summary of soybean accessions: regions of collection/popularization and coverage and mean depth of re-sequencing[a]

Accession	Description	Genome coverage	Mean depth
W01	Wild; Beijing area, PRC	0.928114	4.985707
W02	Wild; Liaoning, PRC	0.945107	4.722718
W03	Wild; Inner Mongolia, PRC	0.95598	5.784117
W04	Wild; Henan, PRC	0.927056	4.375092
W05	Wild; Henan, PRC	0.951029	7.894194
W06	Wild; Heilongjian, PRC	0.744567	1.634187
W07	Wild; Liaoning, PRC	0.907823	5.628903
W08	Wild; Heilongjian, PRC	0.94428	5.558055
W09	Wild; Liaoning, PRC	0.816146	2.17832
W10	Wild; Heilongjian, PRC	0.941307	4.742872
W11	Wild; Shanxi, PRC	0.941344	5.285625
W12	Wild; Anhui, PRC	0.954091	7.462448
W13	Wild; Inner Mongolia, PRC	0.95922	7.986427
W14	Wild; Inner Mongolia, PRC	0.825547	2.218275
W15	Wild; Henan, PRC	0.946831	5.250471
W16	Wild; Heilongjian, PRC	0.92242	4.423702
W17	Wild; Liaoning, PRC	0.799289	1.993777
C01	Advanced bred line; popularized in Shandong, PRC	0.919422	4.43853
C02	Advanced bred line; popularized in Liaoning, PRC	0.951359	5.28649
C08	Advanced bred line; popularized in USA	0.977795	7.851686
C12	Advanced bred line; popularized in Shanxi, PRC	0.935454	4.811598
C14	Advanced bred line; popularized in Brazil	0.945589	5.222104
C16	Japan neutron-mutated line adapted to Taiwan	0.953999	5.624765
C17	Landrace; Sichuan, PRC	0.926871	4.653919
C19	Landrace; Jilin, PRC	0.94893	5.524911
C24	Advanced bred line; popularized in Jiangxi, PRC	0.939643	4.783565
C27	Advanced bred line; popularized in Hebei, PRC	0.948785	5.621014
C30	Advanced bred line; popularized in Henan, PRC	0.949428	5.731478
C33	Advanced bred line; popularized in Heilongjiang, PRC	0.92603	4.242454
C34	Landrace; Guangxi, PRC	0.941575	4.76983
C35	Landrace; Guangdong, PRC	0.931602	4.780462

[a] The wild soybeans were collected from different geographical regions in Mainland China. Several cultivated soybeans were assessments that had been popularized in representative soybean cultivation regions in Mainland China (Northern eco-region, Huang-Huai eco-region, and Southern eco-region). Some advanced lines were used extensively as parental lines in breeding programs. We checked the pedigree to ensure that there was no known history of common parental lines. We also included cultivated germplasms popularized in Taiwan

(originated from Japan), USA and Brazil. To link our result to known soybean population structure, we performed STRUCTURE and phylogenetic analysis using SNP data from a previous study on the cultivated accessions (see **Supplementary Fig. 1**).

Supplementary Table 2 General phenotypic differences between wild and cultivated soybeans[a]

	Wild Soybeans	**Cultivated Soybeans**
Growth type	Mostly trailing	Mostly erect
Stem diameter	Thin	Thick
Branching number	Plenty	Fair
Leaves size	Generally smaller	Generally larger
Inflorescence type	Majority is infinite	Majority is finite
Flower color	Mostly purple	White and purple
Pod size	Small-Medium	Medium-Large
Seed coat color	Mostly black or brown	Mostly yellow
100-Seed Weight	Low	High
Seed Size	Small	Large
Seed protein content	High	Medium
Seed oil content	Low	High

[a] Most phenotypic differences (except flower and seed color) are quantitative traits which are strongly subjected to the influences of environmental conditions

Supplementary Table 3 Fixed, deleterious, and large-effect SNPs in wild and cultivated soybeans

Fixed SNPs

	Fixed location		
	Coding region	Inter-genic region	Total
Wild	15014	448395	463409
Cultivated	64224	2084361	2148585

Deleterious (radical change) SNPs

	Deleterious SNPs	**SNPs in genic region**	**Ratio of deleterious SNPs**	**Specific deleterious SNPs**	**Specific SNPs in genic region**	**Ratio of deleterious SNPs in specific SNPs**
Wild	40343	185417	21.8%	14048	64224	21.9%
Cultivated	29808	133174	22.4%	3678	15014	24.5%

Large effect SNPs

	Stop codon		**Start codon**	**Splice sites**	**Radical change**
	Premature stop	Stop codon to non-stop codon	Start codon to non-start codon		
Wild	2715	1088	380	1842	40343
Cultivated	2036	867	317	1299	29808
Total	3018	1156	420	1966	43876

Supplementary Table 4 Comparison of nonsynonymous/synonymous (Non/Syn) SNP ratio of soybean genes versus rice genes

Non/Syn in rice	Percentage in rice	Percentage of soybean Non/Syn greater than rice
<=1	12.9%	84%
1-1.37	49%	46%
>1.37	38%	17%

Supplementary Table 5 Nonsynonymous/synonymous (Nonsyn/Syn) SNP ratio in LD blocks

LD blocks	whole	>20 kb	>50 kb[a]	whole	>50 kb	>70 kb	>100 kb
		wild			cultivated		
Total SNPs	5831773	218694	80202 (44490)	4146597	172040	129656	89300
Syn	78701	1396	302 (194)	55883	1371	968	520
Nonsyn	106716	2194	617 (374)	77291	1894	1344	708
Nonsyn/Syn	1.36	1.57	2.04 (1.93)	1.38	1.38	1.39	1.36

[a] Some large (>50 kb) LD blocks in wild soybeans are shared with cultivated soybeans. The SNP information of cultivated soybeans in these blocks is given in the parenthesis.

Supplementary Table 6 Recently duplicated gene pairs that may have undergone neofunctionalization and been subjected to purification selection

Recently duplicated gene pairs		Functional annotation
Glyma01g06970	Glyma02g12870	IPR001732, IPR04026
Glyma01g34580	Glyma03g02580	IPR005175
Glyma02g01150	Glyma10g01200	IPR000719, IPR001245, IPR002290, IPR020635
Glyma02g35450	Glyma10g10040	IPR001356, IPR006563
Glyma03g02120	Glyma07g08740	IPR001736
Glyma03g27260	Glyma07g14770	IPR007196
Glyma04g01660	Glyma06g01750	IPR005828
Glyma04g07580	Glyma06g07700	IPR003618, IPR012921, IPR017890
Glyma04g09740	Glyma06g09830	IPR001461
Glyma04g09950	Glyma06g10000	IPR004443, IPR011576, IPR019576
Glyma04g34370	Glyma06g20200	IPR004014, IPR005834, IPR008250
Glyma04g36620	Glyma06g18290	IPR003347, IPR003349, IPR007087, IPR013129, IPR015880
Glyma04g38010	Glyma06g17050	IPR002999, IPR006021, IPR008191, IPR018351
Glyma04g39030	Glyma06g15950	IPR002041, IPR003577, IPR003578, IPR003579, IPR006688, IPR013753
Glyma04g39270	Glyma06g15650	IPR008889
Glyma04g39610	Glyma06g15270	IPR000719, IPR001611, IPR002290, IPR003591, IPR013610, IPR017442, IPR020635
Glyma04g41510	Glyma06g13320	IPR000719, IPR002290
Glyma04g42790	Glyma06g11980	IPR003593, IPR003959, IPR013748
Glyma06g46000	Glyma12g10710	IPR001356, IPR002913,
Glyma06g46300	Glyma12g10490	IPR001478, IPR008915
Glyma09g24410	Glyma16g29750	IPR003594, IPR020576
Glyma09g29600	Glyma16g34180	IPR005140, IPR005141, IPR005142
Glyma09g37070	Glyma18g49600	IPR004087, IPR004088, IPR018111
Glyma09g38740	Glyma18g47590	IPR000357, IPR000719
Glyma10g41290	Glyma20g25960	IPR004328
Glyma10g41450	Glyma20g25790	IPR000741
Glyma10g41880	Glyma20g25160	IPR015803
Glyma10g43230	Glyma20g23660	IPR000048, IPR001609
Glyma13g04760	Glyma19g01890	IPR000602, IPR011682, IPR015341
Glyma13g06470	Glyma19g04020	IPR006626
Glyma13g25480	Glyma15g35240	IPR012919
Glyma13g29140	Glyma15g09920	IPR001810
Glyma13g32760	Glyma15g06550	IPR007015
Glyma13g42760	Glyma15g02680	IPR000719, IPR002290, IPR017442, IPR020635
Glyma13g43180	Glyma15g02170	IPR000642, IPR003593, IPR003959
Glyma13g43350	Glyma15g01960	IPR001356, IPR002913
Glyma16g05750	Glyma19g26740	IPR005202
Glyma16g06050	Glyma19g25950	IPR006514

Supplementary Table 7 Present/Absence Variations (PAVs) absent in all cultivated soybeans

PAV_id	Chromosome	Start	Length (bp)	Present in the following individuals
SoyPAV0088	Gm03	41860519	1277	W02, W04, W05, W10, W11, W12, W13, W15
SoyPAV0136	Gm05	15394250	830	W13, W15
SoyPAV0145	Gm05	32040262	932	W03, W05, W07, W11, W12, W13,
SoyPAV0165	Gm06	13460926	1702	W05
SoyPAV0198	Gm06	8472891	1379	W05, W12, W13, W16
SoyPAV0247	Gm07	7504790	746	W01, W05
SoyPAV0297	Gm08	43695762	849	W05
SoyPAV0321	Gm08	8191148	1220	W05
SoyPAV0343	Gm09	40337456	1308	W05, W11, W12, W16
SoyPAV0359	Gm09	45061199	505	W05
SoyPAV0401	Gm10	50629412	2062	W05, W07, W12
SoyPAV0411	Gm11	24467239	1415	W03, W05, W07, W08, W10, W12, W13, W14, W16
SoyPAV0413	Gm11	31669392	3526	W03, W05, W12, W13, W16
SoyPAV0443	Gm12	13669774	777	W05, W10, W12, W13
SoyPAV0444	Gm12	13669774	777	W05, W10, W12, W13
SoyPAV0509	Gm13	36782052	2127	W05
SoyPAV0510	Gm13	36786091	2238	W05
SoyPAV0513	Gm13	38345033	4329	W02, W05, W13, W14, W16
SoyPAV0557	Gm14	44730779	2811	W03, W05, W08, W11, W12, W13, W16
SoyPAV0558	Gm14	45540077	2010	W05, W11, W12, W16
SoyPAV0560	Gm14	45924951	910	W05, W11, W12, W16
SoyPAV0562	Gm14	46690119	6289	W01, W04, W05, W11, W12
SoyPAV0624	Gm16	3559166	1014	W02, W05, W11, W12, W13, W16
SoyPAV0644	Gm17	11002027	1223	W05, W11, W13, W16
SoyPAV0705	Gm18	51705000	1500	W05, W16
SoyPAV0776	Gm20	35824028	4074	W05, W12
SoyPAV0777	Gm20	35824028	4074	W05, W12
SoyPAV0788	Gm20	44416735	3341	W01, W03, W05, W08, W10, W12, W16