Research Article

Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections

Hiromi Kajiya-Kanegae,‡,†, Hideki Nagasaki,‡, Akito Kaga,‡, Ko Hirano,§, Eri Ogiso-Tanaka,§, Makoto Matsuoka,¶, Motoyuki Ishimori, Masao Ishimoto,§, Masatsugu Hashiguchi,§, Hidenori Tanaka,§, Ryo Akashi,§, Sachiko Isobe,§, and Hiroyoshi Iwata*1

1Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan, 2Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan, 3Institute of Crop Science, National Agriculture and Food Research Organization (NARO), Tsukuba, Ibaraki 305-8518, Japan, 4Bioscience and Biotechnology Center, Nagoya University, Nagoya, Aichi 464-8601, Japan, and 5Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan

*To whom correspondence should be addressed: Tel. +81-3-5841-5069. Fax. +81-3-5841-5068. Email: hiroiwata@g.ecc.u-tokyo.ac.jp (H.I.); Tel. +81-438-52-3928. Fax. +81-438-52-3934. Email: sisobe@kazusa.or.jp (S.I.)

‡Present address: Research Center for Agricultural Information Technology, National Agriculture and Food Research Organization, Chiyoda-ku, Tokyo 100-0013, Japan.

†The first three authors contributed equally to this work.

Received 7 August 2020; Editorial decision 22 December 2020; Accepted 4 January 2021

Abstract

We performed whole-genome Illumina resequencing of 198 accessions to examine the genetic diversity and facilitate the use of soybean genetic resources and identified 10 million single nucleotide polymorphisms and 2.8 million small indels. Furthermore, PacBio resequencing of 10 accessions was performed, and a total of 2,033 structure variants were identified. Genetic diversity and structure analysis congregated the 198 accessions into three subgroups (Primitive, World, and Japan) and showed the possibility of a long and relatively isolated history of cultivated soybean in Japan. Additionally, the skewed regional distribution of variants in the genome, such as higher structural variations on the R gene clusters in the Japan group, suggested the possibility of selective sweeps during domestication or breeding. A genome-wide association study identified both known and novel causal variants on the genes controlling the flowering period. Novel candidate causal variants were also found on genes related to the seed coat colour by aligning together with Illumina and PacBio reads. The genomic sequences and variants obtained in this study have immense potential to provide information for soybean breeding and genetic studies that may uncover novel alleles or genes involved in agronomically important traits.

Key words: Glycine max, soybean, genome diversity, next-generation sequencing
1. Introduction

There are more than 1,750 gene banks existing in the world that store genetic resources comprised of 7.4 million accessions. However, 1% of these accessions have been used in breeding. Underutilization of genetic resources is partly due to the accessions being characterized insufficiently. Recent advances in DNA genotyping and sequencing technologies have enabled molecular descriptors for these accessions with genome-wide markers and whole-genome sequences. In rice, whole-genome sequence data of more than 3,000 accession genetic resources were collected and analysed. In soybeans, which is the species being investigated in the present study, 42,509 single nucleotide polymorphisms (SNPs) were determined from 20,087 accessions of genetic resources. Genomic information on genetic resources will enhance the utilization of genetic resources in plant breeding. A genome-wide association study (GWAS) facilitates the identification of genes/alleles, which can be utilized in plant breeding, from genetic variations within a germplasm collection. Genomic selection is a method of predicting the genetic ability of individuals and selecting individuals based on the prediction, which allows useful accessions to be identified from a germplasm collection.

Whole-genome sequences collected from genetic resources provide useful information for plant breeding that can provide important clues and insights into the evolution and domestication process of crop species, subspecies, and cultivars. Whole-genome sequences also provide information that can determine genes and alleles that played important roles in local adaptation and artificial selection in the history of plant breeding. In addition, whole-genome sequences are expected to improve the power and precision of GWAS because they preserve rare variants and variants that are not in linkage disequilibrium (LD) with a reduced set of SNPs, such as SNPs genotyped with a genotyping array or a reduced-representation sequencing approach. Genome structure variants, including copy number variations (CNVs), are also identified by whole-genome sequences. CNVs are known to have contributed to domestication and natural or artificial selection in the history of plant breeding and has been linked to important agronomic traits. Moreover, whole-genome sequences of genetic resources can also be useful as a reference panel to compute whole-genome polymorphisms in large experimental/breeding populations genotyped with a reduced set of SNPs to suppress the time and cost of genotyping. From the above-mentioned various viewpoints, studies on whole-genome sequences of genetic resources have been performed in various crop species (e.g. rice, maize, sorghum, tomato, and soybean).

Globally, soybean [Glycine max (L.) Merr.] is the most important legume and is also the fourth in worldwide production after rice, wheat, and maize in terms of global crop production. The soybean is an important source of protein, an oil crop, and is used for both food and animal feed. Recently, numerous functional constituents to foods have been reported to occur because the compounds responsible for the colouration of the seed coat, hilum, pubescence, and flower in soybean are mostly related to the biosynthetic pathway of anthocyanins. The dominant I allele inhibits seed coat pigmentation and causes a uniformly yellow seed coat and hilum, whereas the i allele allows pigmentation. The other I and i restrict pigmentation to the hilum and to the saddle-shaped region around hilum, respectively. The mechanisms for I and i alleles have been characterized as RNA silencing of chalcone synthase (CHS) genes in the biosynthetic pathway of anthocyanins. With the combination of i allele, the R and r alleles at the R locus result in a black and brown hilum/seed coat, respectively. The R gene encodes the R2R3 MYB transcription factor, which might control the expression of UDP-glucose: flavonoid-3-O-glucosyltransferase (UGT73G1) in the final step of anthocyanin biosynthesis, and the r allele is caused by four types of loss-of-function mutations. The T and W1 loci are known to control pubescence.
and flower colour, respectively, in addition to epistatic interactions for the pigmentation of seed colour traits. The dominant T allele produces tawny pubescence, whereas the recessive t allele with a single-base deletion causes a loss of function in flavonoid 3′-hydroxylase and produces grey pubescence. The dominant W1 allele produces a purple flower and hypocotyl phenotype, while the recessive w1 allele with a single-base deletion causes a loss of function in flavonoid 3′,5′-hydroxylase and produces a white flower and green hypocotyl phenotype.

Understanding the genetic control of flowering time and maturity is indispensable to efficiently develop a new variety with a photoperiodic adaptation to different latitudes. For that reason, the genes responsible for E1, E2, E3, E4, E5, E6, E7, E8, E9, J, and J0 were isolated among the 10 major classical loci (E1–E9 and J). In addition, genes for other loci such as qDTF-1, E1-like-b, GmPRR3 were isolated. The next important task would be to accumulate allelic information from the breeding materials since different allelic combinations of these loci determine adaptations to different latitudes. However, functional redundancy between duplicated gene copies in the soybean genome makes it difficult to understand the relationship between genetic variation and agronomically important traits. Therefore, a resequencing effort for many accessions will provide a chance to comprehensively identify new alleles and genes that potentially affect agronomically important traits apart from flowering time and maturity.

In this study, we collected and analysed the whole-genome sequences of 198 soybean accessions. The accessions were mainly from two soybean mini-core collections from the National Agricultural and Food Research Organization (NARO) Genebank. The accessions in the mini-core collections were carefully selected from 1,603 accessions, based on the polymorphisms of 191 SNP markers and several agronomic traits, to ensure that the collections retained as many genetic variations as possible in all accessions. In this study, we characterized the polymorphisms found in the whole-genome sequences and investigated the subpopulation structure and levels of genetic differentiation in the accessions based on their polymorphisms. In addition, using the number of days to flowering (DTF) as an example trait, we evaluated the potential of GWAS using whole-genome sequences. Among the 198 accessions, we also employed 10 for long-read sequencing to analyse large structural variants. Copy number variants were also identified using Illumina reads to investigate the genome structure variance. The variations in colour related to the seed coat, hilum, pubescence, and flower were characterized as polymorphisms in the whole-genome sequences, and their relations with known genes were investigated. Through these analyses, we evaluated the potential significance of whole-genome sequences being prepared for soybean genetic resources. The whole-genome sequences collected for the genetic resources will facilitate the active use of genetic resources in soybean breeding programs.

2. Materials and methods

2.1. Plant materials

In this study, we utilized 198 soybean accessions for whole-genome sequencing (Supplementary Table S1): 192 accessions from Japanese and world soybean mini-core collections, an Indian cultivar ‘L.323’ (JP241838), and a Japanese cultivar ‘Misuzudaizu’ (JP28856) obtained from NARO Genebank (https://www.gene.affrc.go.jp/index_en.php, 15 January 2021, date last accessed); Japanese landrace ‘Houjaku Kuwazu’ (PJ416937) and a United States (US) cultivar ‘5002T’ (Pl634193) obtained from the USDA (United States Department of Agriculture) germplasm collection through GRIN (Germplasm Resources Information Network). A soybean cultivar ‘Norin2’ and a Glycine soja accession (B01167) were obtained from the National BioResource Project (https://www.legumebase.brc.miyazaki-u.ac.jp, 15 January 2021, date last accessed).

Two plants were grown with an inter-row spacing of 80 cm and a hill spacing of 20 cm in the field at NARO in Tsukuba, Ibaraki, Japan (36°01′25.6"N 140°06′59.1"E). Seeds were sown on June 1, 2010, and the DTF of 184 successfully germinated accessions were recorded for association analysis. Of the 198 accessions, 14 were excluded from the evaluation of DTF because six and eight accessions had not planted in the field and germinated late due to overseed, respectively. The colour of hypocotyl, flower, pubescence, leaf at maturity, and seeds of these plants were recorded while comparing that of the soybean reference cultivar Williams 82 (accession no. GmWMC115 in the present study) with yellow seed with black hilum, tawny pubescence, and white flowers (‘t’, ‘T’, ‘R’, ‘e1’).

2.2. Illumina whole-genome sequencing

Freeze-dried young leaves collected from a plant in each accession were ground using a mortar and pestle. Total DNA was extracted from the finely ground leaf tissue using the DNAeasy Plant Mini Kit (Qiagen, Hilden, Germany). The DNA was physically sheared into ~350 bp fragments using Covaris S2 (Covaris, Brighton, UK). The fragmented DNA was used for DNA library construction with the TrueSeq DNA PCR-Free Library Prep Kit (Illumina, San Diego, CA). The DNA libraries were sequenced using the Illumina HiSeq X Ten or HiSeq 4000 (Illumina).

2.3. PacBio whole-genome sequencing

PacBio whole-genome sequencing was performed for 10 accessions: ‘Misuzudaizu’ (JP28856), ‘Enrei’ (GmJMC025), ‘Houjaku Kuwazu’ (PJ416937), ‘Fakuyutaka’ (GmJMC112), ‘Moshidou Gong 503’ (GmWMC084), ‘Peking’ (GmWMC084), ‘PK 73-54’ (GmWMC071), ‘L323’ (JP241838), ‘5002T’ (Pl634193), and ‘Williams 82’ (GmWMC115; Supplementary Table S1). For PacBio sequencing, the total DNA was extracted from finely ground leaf tissue using an SDS-based DNA extraction method and was used for SMRTbell libraries (Pacific Biosciences, Menlo Park, CA). Sequences were generated using PacBio Sequel (Pacific Biosciences).

2.4. Reference genome sequences and annotation data used in this study

Gmax._v2.0 softmasked sequences and the genome annotation of Williams 82, which was a completely sequenced soybean accession, were used as the reference for the analyses of this study. The reference data were obtained from Phytome 12.1.56 and the gene annotations and IDs described in this article were also based on the descriptions in Phytome.

2.5. Variant call and diversity analysis using Illumina reads

The Illumina reads were trimmed with Trimomatic version 0.36 with the following parameters: ILLUMINACLIP: TruSeq3-PE-2.fa: 2:30:10 LEADING: 3 TRAILING: 3 SLIDINGWINDOW: 4:15 MINLEN: 36’. The trimmed reads were mapped on the reference sequence using the BWA-aln (release 0.7.17) algorithm with default options. The mapped reads were then sorted using SAMtools...
2.6. Genetic and genomic diversity analyses
The genetic structure of the population was estimated using phylogenetic analysis, principal component analysis (PCA), and ADMIXTURE analysis based on whole-genome sequences. For the phylogenetic analysis, we constructed a neighbour-joining (NJ) tree based on the whole-genome genetic distances among accessions, calculated with the Jukes and Cantor model in R. In constructing the NJ tree, the accession ‘B01167’, which is the only accession of *G. soja*, was treated as an outgroup. For PCA, we calculated whole-genome Euclidean distances among the accessions based on their genotypes and performed multi-dimensional scaling based on the distances using the ‘cmdscale’ function in R. We performed ADMIXTURE analysis with the models of one to eight subpopulations to estimate the ancestries of the accessions. A 5-fold cross-validation was performed for each number of the subpopulations to select the appropriate K value.

We calculated nucleotide diversity (\(\pi\)), pairwise and total \(F_{ST}\), and \(r^2\) to measure the LD for the entire genome with non-overlapping 500 kb windows, and to evaluate genome-wide pattern levels of polymorphisms, genetic differentiation, and LD. For this calculation, we used an in-house developed R program. To identify CNVs among the 198 soybean accessions, CNV-Seq with the last updated version in 2014 was performed based on the Illumina reads with the options to select the appropriate K value.

2.7. GWAS of flowering date and seed weight
We performed GWAS for the flowering date to demonstrate the viability of the identified variants. The numbers of days from sowing to first flowering of 184 accessions were used for the association analysis. SNPs with minor allele frequencies (2.5%) or whose missing rate was more than 5% were filtered out for the GWAS study. Imputation was conducted using Beagle 5.0 with default parameter settings. GWAS was performed using a linear mixed model implemented by the ‘association.test’ function in gaston package ver. 1.5.5 in R. In the linear mixed model, the first two principal components of marker scores were included as fixed effects. A genetic relationship matrix specifying a random additive effect was computed with the Jukes and Cantor model in R. In constructing the NJ tree, the accession ‘B01167’, which is the only accession of *G. soja*, was treated as an outgroup. For PCA, we calculated whole-genome Euclidean distances among the accessions based on their genotypes and performed multi-dimensional scaling based on the distances using the ‘cmdscale’ function in R. We performed ADMIXTURE analysis with the models of one to eight subpopulations to estimate the ancestries of the accessions. A 5-fold cross-validation was performed for each number of the subpopulations to select the appropriate K value.

We calculated nucleotide diversity (\(\pi\)), pairwise and total \(F_{ST}\), and \(r^2\) to measure the LD for the entire genome with non-overlapping 500 kb windows, and to evaluate genome-wide pattern levels of polymorphisms, genetic differentiation, and LD. For this calculation, we used an in-house developed R program. To identify CNVs among the 198 soybean accessions, CNV-Seq with the last updated version in 2014 was performed based on the Illumina reads with the options to select the appropriate K value.

2.8. Structural variation analysis
Structural variation (SV) detection was performed using PacBio reads from the 10 soybean accessions. The PacBio reads of each accession were mapped onto the reference genome sequence using NGMLR version 0.2.6. SV detection from mapped results was performed using SAMtools version 1.3.1 and Sniffles version 1.0.8. Genomewide distribution of SVs (insertions, deletions, and duplications) was filtered by length, where the value of SVLEN in VCF file was greater than or equal to 50 kb, and grouped into three categories: Japan, Primitive, and World. SVs were filtered by length, where the value of SVLEN in VCF file was greater than or equal to 50 kb, and grouped into three categories: Japan, Primitive, and World, which were performed using in-house Perl scripts. The integrated genome maps were then illustrated by CIRCOS 69.

2.9. Identifying the variation in PacBio and Illumina reads related to the I locus
The presence and absence of variation (PAV) related to the I locus were analysed based on the mapping results of PacBio and Illumina reads onto the reference genome. In addition, the genomic positions of the CHS genes of BAC77G7-a and BAC56G2 (GenBank accession numbers: EF623854 and EF623856) were compared to the I locus of Williams 82. PAV of *Gm-c1069-6017* and *GmICHS* of the I locus and *GmD21N* of the D1 locus were manually identified by comparing the alignments using IGVtools 2.4.1172 and CLC Genomics Workbench 12 (Qiagen, Hilden, Germany).

3. Results and discussion
3.1. Whole-genome sequencing of the soybean accessions
A total of 25 billion paired-end Illumina reads were obtained for the 197 *G. max* and one *G. soja* accessions (Supplementary Tables S1 and S2). The mean depth of the reads against the soybean genome ranged from ×6.8 to ×32.9 with an average of ×16.3. The mapping ratio onto the Williams 82 reference was 94.7% on average, and the mean-coverage ratio on the reference genome was 93.6%. A total of 12,952,387 variants, including 10,116,707 SNPs and 2,835,680 insertions/deletions (indels), were identified as the results of the variant call and filtering (Table 1). The information for variants among accessions can be compared using the multiple genome Browser TASUKE from https://daizutasuke275-core.daizu.dna.affrc.go.jp/.

3.2. Genetic diversity and population structure analyses
An NJ tree based on whole-genome SNPs was built to investigate the phylogenetic relationships among the 198 accessions (Fig. 1A). The NJ tree indicated that the 198 accessions were divided into three subgroups, two of which were not monophyletic. Based on the origins of accessions consistent with each subgroup, we named the subgroups as ‘Primitive’, ‘World’, and ‘Japan’. The ‘Primitive’ subgroup was comprised of the *G. soja* accession (B01167), ‘Peking’ (GmWMC084), ‘Moshidou Gong 503’ (GmWMC042), and 17 other accessions (Supplementary Table S1). The World subgroup consisted of ‘Williams 82’ (GmWMC115), ‘S002T’, ‘PK 73-54’ (GmWMC071), and 55 other accessions. The ‘Japan’ subgroup consisted of a monophyletic cluster of 120 accessions, which were mainly Japanese and Korean landraces/cultivars.
Whole-genome diversity in soybean mini-core collections

The 12,953,387 variants (10,117,707 SNPs + 2,835,680 indels) was estimated using SnpEff (Supplementary Table S3 and Fig. S5A) to investigate and exploit genotypic and phenotypic variations. Out of all variants, 29,655,355 (97.45%) were classified into ‘Modifier’ (intergenic variants), while 9,757 (0.24%) were classified into ‘High’ (frameshift variant, stop lost/gained), 358,102 (1.26%) into ‘Moderate’ (coding sequence variant), and 319,839 (1.05%) into ‘Low’ (synonymous variant). The number of private variants classified as ‘High’ was more than twice (20,283) in ‘Japan’ than those in ‘Primitive’ (9,540) and ‘World’ (8,993) (Supplemental Table S3). For the functional classification of SnpEff, 613,799 variants were classified as ‘Nonsense’ (9,737; 1.6%), ‘Missense’ (358,102; 58%), and ‘Silent’ (319,839; 50%) as commonly observed among the three groups. These variants were regarded as the ancient variants rather than ‘private’ variants, which identified single groups only. The private variants per subgroup were defined as variants that had been observed only in a specific group. The number of private variants per subgroup was higher in ‘Japan’ (2,299,690) than in ‘World’ (1,119,074) or ‘Primitive’ (2,145,920; Supplementary Fig. S4). These private variants were considered as candidate causal variants that ascribe group-specific phenotypes. The allele frequencies of variants in each group (subpopulations and shared categories) were calculated to elucidate genetic diversity. The average, median, and mode of allele frequencies were higher in ‘Primitive’ than in ‘World’ or ‘Japan’ (Supplementary Table S3 and Fig. S5B). The ‘Japan’ subgroup had the largest number of private variants classified into ‘Nonsense’ and ‘Missense’ among the three subgroups (Supplementary Fig. S5B). These variants may implicate phenotypic variations among populations.

We evaluated the genome-wide diversity in the whole-genome sequences by calculating the nucleotide diversity (π), genetic differentiation (pairwise and total FST), and linkage disequilibrium (r²) in 500 kb non-overlapping windows for each of the three subgroups and all subgroups together. As a result, the nucleotide diversity was

### Table 1. Number of variants identified on the 198 soybean accessions

| Chr   | Raw SNP | Raw INDEL | Total raw variant | Filtered SNP | Filtered INDEL | Total filtered variant |
|-------|---------|-----------|-------------------|--------------|---------------|------------------------|
| Chr01 | 809,046 | 139,387   | 948,433           | 519,110      | 133,213       | 652,323                |
| Chr02 | 685,281 | 138,448   | 823,729           | 464,550      | 133,803       | 598,355                |
| Chr03 | 918,813 | 165,525   | 1,084,371         | 581,996      | 160,415       | 742,411                |
| Chr04 | 858,671 | 147,249   | 1,005,920         | 574,337      | 141,912       | 706,249                |
| Chr05 | 560,582 | 110,072   | 670,654           | 361,072      | 105,493       | 466,565                |
| Chr06 | 925,223 | 171,679   | 1,096,902         | 579,010      | 166,158       | 745,168                |
| Chr07 | 712,969 | 143,049   | 856,018           | 463,300      | 138,337       | 601,637                |
| Chr08 | 693,324 | 147,969   | 849,018           | 462,054      | 129,213       | 600,302                |
| Chr09 | 734,949 | 141,684   | 885,633           | 481,385      | 136,143       | 617,528                |
| Chr10 | 714,099 | 134,919   | 849,018           | 462,054      | 129,213       | 600,302                |
| Chr11 | 438,702 | 95,978    | 534,680           | 303,411      | 92,734        | 396,145                |
| Chr12 | 595,004 | 115,856   | 710,860           | 391,525      | 111,748       | 503,273                |
| Chr13 | 722,750 | 160,407   | 883,157           | 481,921      | 156,154       | 638,075                |
| Chr14 | 902,138 | 151,259   | 1,053,397         | 590,557      | 145,723       | 736,280                |
| Chr15 | 1,048,862 | 175,748 | 1,224,610         | 648,774      | 170,120       | 818,894                |
| Chr16 | 861,304 | 165,506   | 1,026,810         | 526,504      | 160,445       | 686,949                |
| Chr17 | 642,407 | 125,991   | 768,398           | 425,846      | 122,022       | 547,868                |
| Chr18 | 1,272,177 | 224,567 | 1,496,744         | 794,961      | 217,759       | 1,012,720              |
| Chr19 | 804,396 | 147,789   | 952,184           | 516,977      | 137,773       | 652,750                |
| Chr20 | 767,005 | 136,478   | 903,483           | 506,619      | 131,009       | 637,628                |
| Total | 15,676,702 | 2,936,592 | 18,613,294         | 10,116,707   | 2,835,680     | 12,952,387              |
Figure 1. The phylogenetic relationships and Population structure of 198 soybean accessions. (A) A neighbour-joining dendrogram phylogenetic tree of the 198 soybean accessions. The accessions were classified into three subgroups: ‘Primitive’, ‘World’, and ‘Japan’. (B) Population genetic structure estimated by Admixture analysis. Results from $k = 2$ to 5 are shown. Green triangles indicate the non-Japanese accessions in the ‘Japan’ subgroup.
highest in ‘Primitive’ and lowest in ‘Japan’ in most genomic regions (Supplementary Fig. S6A). The private variants, defined as variants that had been observed only in a specific accession, were counted in each accession. The average number of private variants per accession was higher in ‘Primitive’ (62,329) than in ‘World’ (7,091) or ‘Japan’ (6,584). The ‘Japan’ subgroup had the largest number of accessions (120 of 198), so it is reasonable for it to have a larger number of private variants per subgroup. The genetic differentiation between sub-populations was generally largest in the comparison between ‘Primitive’ and ‘Japan’ and lowest in the comparison between ‘World’ and ‘Japan’ in most of the genomic regions (Supplementary Fig. S6B). Some genomic regions, however, showed different patterns from the general pattern; for example, the differentiation in the terminals of the long arms of chromosomes 3 and 5 was the lowest between ‘Primitive’ and ‘World’, while it was high between ‘Primitive’ and ‘Japan’ and ‘World’ and ‘Japan’. The LD was the highest in ‘Primitive’ and the lowest in ‘Japan’ (Supplementary Fig. S6C), and studies have reported that small populations have higher LD value than large populations. As in the case of genetic differentiation, some genomic regions showed different patterns from the general pattern in LD. For example, we only observed the peak of the LD in the middle (~2.5 Mb) of the chromosome 6 in ‘Japan’, while the peak of the LD in the middle (~10 Mb) of the chromosome 7 was in ‘World’.

3.3. GWAS for flowering date

A genome-wide association test was performed on DTF with 4,776,813 genome-wide SNPs (Fig. 2) to demonstrate the strength of whole-genome GWAS with mini-core collections. With a 1% threshold of FDR, three significant associations were detected at the 5,520,945 and 5,542,737 bp positions on chromosome 12, and at the 45,310,798 bp position on chromosome 10. The variants of e2, e3-tr, and the stop-loss variant (rs125308117) of two-component response regulator-like gene on Chr12 were significant in a gene-based association test for flowering time in the mini-core collection. Ogiso-Tanaka et al. estimated a large deletion on E3 by the coverage of four amplicons on the 4th exon. In this study, e3-tr could not be incorporated into the association analysis due to the difficulty of detecting a 15 kb deletion in E3 using whole-genome sequences.

The significant association detected in the SNP at 45,310,798 bp on chromosome 10 was in the Glyma.10G221500 coding region, which encodes the classical E2 gene, that is, a soybean GIGANTEA gene. The alleles of the SNP were ref (same as the reference sequence, ‘Williams 82’) ‘A’ and alt ‘T’ alleles, which corresponded to Lys (AAA) and a premature stop codon (TAA) at the 528 amino acid sequence position, respectively (Fig. 3A). The proportion of the alleles was different among the subgroups, especially when comparing ‘Primitive’ to ‘World’ and ‘Japan’ (Fig. 3B). Based on the relationships between DTF phenotypes and SNP alleles, the reference (ref: A) and alternative (alt: T) alleles were associated with late and early flowering phenotypes, respectively (Fig. 3C). Previously, three alleles, E2-in (Williams 82), E2-dl, and e2-ns, have been identified from 63 accessions covering several ecological types by sequencing of the genomic region of the E2 locus. Among them, three amino acid sequence haplotypes, H1 (e2-ns), H2 (E2-dl), and H3 (E2-in), have been reported to be in the cultivated soybean gene pool. Interestingly, five novel variants with amino acid changes were obtained in the present study (Supplementary Table S4). It is necessary to confirm whether these novel variants, especially the novel nonsense variant of Gln53stop in

Figure 2. Manhattan plots of GWAS for flowering time in 2010. Significantly positive SNPs (FDR <0.01) are highlighted in green. (A) Whole-genome, (B) chromosome 10, and (C) chromosome 12.
GmJMC041 and GmJMC044 that has no Lys528stop mutation, affect the flowering time or maturity and can be called a new allele. The SNP at 5,520,945 bp on chromosome 12 showed a significant association in the coding region of Glyma.12G073900, which encodes the clock-associated pseudo-response regulator 3 (GmPRR3b). This SNP, which is located at the first nucleotide of the termination codon 'TAA', changes the codon to 'CAA' of Gln on the alt allele. This caused a shift of the stop codon at the position of 627 aa in the ref allele (5,520,945 bp) to 795 aa in the alt allele (5,521,025 bp, Supplementary Fig. S7A). The proportion of the alleles was significantly different in 'Primitive' compared to 'Japan' and 'World' (Supplementary Fig. S7B). Accessions with the ref allele tended to flower early, while accessions with the alt allele tended to flower late (Supplementary Fig. S7C).

There are two landraces in the 'Japan' subgroup, 'Kurodaizu' and 'Hiku Anda' (GmJMC030 and GmJMC049), which originated from Okinawa prefecture located in the south most part of Japan. It was reported that these landraces belonged to a cluster that consisted of old cultivars known as the precocious summer-type soybean. We confirmed that the two landraces had early flowering alleles in the E2 gene and GmPRR3b, supporting that the two landraces belonged in the early flowering group in the previous study. The 'Japan' subgroup also includes three landraces and three breeding lines, 'Waseousode' (GmJMC005), 'Tokachi Nagaha' (GmJMC007), 'Shizunai Daizu' (GmJMC009), 'Ooyachi 2' (GmJMC021), 'Bansei Hikarikuro' (GmJMC033), and 'Yakumo Meaka' (GmJMC037), that originated from Hokkaido which is the northernmost prefecture of Japan. These six accessions had early flowering alleles in both genes, suggesting an adaptation to specific environmental conditions in the northern part of Japan.

Li et al. reported six variants of GmPRR3b that caused amino acid changes and eight haplotypes from 383 accessions, including wild soybean accessions. We were able to identify four novel variants causing amino acid changes in the current study (Supplementary Table S4). Among the eight haplotypes determined by Li et al., H1 and H6 encoded truncated peptides. The H6 haplotype is the most

---

**Figure 3.** Sequence variants on the GIGANTEA gene (Glyma.10G221500), the allele frequency of a targeted SNP, and distribution of flowering date. (A) The structures of Glyma.10G221500 encodes the soybean GIGANTEA gene and identified variants. Black arrow—gene direction, white box—coding exon, and UTR—grey box. Genomic positions represent variants identified by Illumina reads in the 198 soybean accessions. Those enclosed with solid and dotted lines were basic variants used in GWAS and SVs detected by PacBio, respectively. (B) Allele frequency in the ‘Primitive’, ‘World’, and ‘Japan’ subgroups of the 198 accessions of the SNP (Ch10-45310798) that showed the highest significance on the gene by the GWAS analysis. Ref (same as the reference sequence, Williams 82) and Alt alleles are ‘A’ and ‘T’, respectively. The numbers in the pie charts indicate that of accessions having the corresponding homozygous alleles. One accession in the ‘Japan’ subgroup showed hetero allele and exclude from the chart. (C) Boxplots of flowering date in the 184 accessions having Ref (left) and Alt (right) allele on the SNP (Ch10, 45310798). The vertical line indicates the days to flowering (DTF).
frequent haplotype in cultivated soybeans, including the reference ‘Williams 82,’ and was frequently observed as a ref allele in our study, while H1 was found only in wild soybeans. The next would be the H4 and H5 haplotypes, which encode longer amino acid sequences, that were reported to flower later than H1 and H6. Similarly, H4 and H5 haplotypes were second and third most frequent haplotypes (33 and 19 accessions, respectively) in our study. The H7 and H8 haplotypes, which encode the same length of amino acid sequence with a Ser100Leu mutation in the pseudo receiver domain, have been reported to flower earlier than H4 and H5.74 The novel mutation of Asp98Asn, which was only found in haplotype H5, was located in the same pseudo receiver domain. Together with the other novel frameshift variant of Gly556fs in GmWMC087, which causes a shorter truncated protein than the H6 haplotype, further studies are necessary to understand the allelic effect of the novel variants on flowering time and maturity.

3.4. SVs between nine soybean accessions

PacBio reads were obtained for 10 accessions originating from Japan, China, India, and the USA (Fig. 1, Supplementary Table S1). The total length of the subreads ranged from 10.3 to 17.7 Gb, representing 10.5 to 18.1 of the soybean genome. The average subread lengths ranged from 6,777 to 8,019 bp. SVs for the 10 soybean accessions with PacBio reads were identified and compared (Table 2). ‘Williams 82,’ the accession of the reference genome, was identified to have a total of 2,033 SVs. The number represents the degree of errors (false positive or miss-assembly of the reference genome). Overall, the number of SVs with complex structures such as duplications was less than insertions and deletions. A large number (a total of 16,363) of insertion and deletion polymorphisms existed in two accessions in the ‘Primitive’ subgroup. The polymorphisms of these two accessions were newly detected by long-read resequencing, which was at a moderate genetic distance from the reference genome (Supplementary Fig. S2), agreeing with the result of the PCA analysis. However, several exceptions, such as on 10–20 Mb on Chr03 and 10–30 Mb on Chr19 in the ‘Japan’ subgroup, were observed. The fewer π values and higher SVs in these regions tend to be negatively correlated, suggesting that it is difficult to detect an association between the phenotype and DNA polymorphisms using SNP-based analysis even if there is phenotypic diversification due to SVs in the regions.31 Furthermore, we focused on regions where genes related to domestication or breeding are located. For example, the regions at 37–41.5 Mb on Chr14 and 10–20 Mb on Chr20 showed lower polymorphisms in ‘World’ and ‘Japan’ than ‘Primitive’. According to Zhou et al.,12 these regions were considered as putative selective sweep regions of seed oil contents (Chr14: 37,550,001.41,300,000, Chr20: 10,270,001.18,460,000) caused by the breeding or domestication process.12 The lower polymorphic regions were also observed in ‘World’ and ‘Japan’ than ‘Primitive’ at 10–20 Mb on Chr05, 20–30 Mb on Chr10, and 10–20 Mb on Chr12; suggesting the possibility of selective sweeps during domestication or the breeding process. Furthermore, we focused on somewhat higher polymorphic regions, which are remarkable to the ‘Japan’ subgroup. The region at 4.5 Mb on Chr03, reported as the Rsfl region, includes the coiled-coil nucleotide-binding site leucine-rich repeat (CC-NBS–LRR)-type gene clusters for Phytophthora resistance.82 The 30 Mb region in Chr13, reported as the Rs1 region, includes the NBS-LRR resistance gene cluster for the soybean mosaic virus.83 Furthermore, the 47 Mb region in Chr14, the Rs3 gene region, includes the CC-NBS–LRR gene cluster for the soybean mosaic virus.84 It is consistent that these three regions include R gene clusters and higher amounts of SVs, and it is thought that genes in such SV rich regions may generate unique alleles in the soybean accessions of the ‘Japan’ subgroup.

3.5. Identification of known variations related to the I locus and loci related to colour traits

We further examined the relationship between SVs and seed coat colour phenotypes by comparing known variants related to the chalcone synthase (CHS) gene cluster of the l locus on Chr08. It has been reported that duplications or deletions of CHS genes influence seed coat pigmentation in Rosids.48,50,52,86 In soybean, a chimeric sequence consisting of subtilisin and CHS1 anti-sense of duplicated CHS clusters (Gm-c1069-6017), has been suggested to cause CHS gene silencing and change the yellow seed coat with pigmented hilum, the ‘so-called’ dominant i allele.41 Another chimeric sequence (GmiIRCHS) consisting of Dnaj and inverted CHS3 genes, causes PTGS of CHS genes and changes the yellow seed coat with pigmented hilum, the ‘so-called’ dominant I allele.40 The genomic region of the i allele is covered by two ‘Williams 82’ BAC clones, which are BAC77G7-a and BAC56G2,9 and their sequences are located in Chr08 of Gmax_275_v2.0 and are approximately 18 kb apart from each other; from the position of 8,410,306, where the 3’ end of BAC56G2 matched to the position of 8,428,210 where the 3’ end of BAC77G7-a complementary matched (Supplementary Fig. S9). Although the gene annotations on BAC56G2 were well conserved on Gmax_275_v2.0, the most important genomic region, including Gm-

Table 2. The numbers of the SVs detected from 10 soybean accessions with PacBio reads

| Accession   | Peking | Moshidou | Gong 503 | C1329 | PK 73-54 | S002T | Williams 82 | Houjaku | Enrei | Fukuyutaka | Misuzudaizu |
|-------------|--------|----------|----------|-------|---------|-------|-------------|---------|-------|-------------|--------------|
| Total       | 17,922 | 25,002   |          | 9,128 | 6,373   | 5,335 | 2,033       | 14,392  | 14,921 | 15,982      | 6,144        |
| Insertions  | 7,388  | 10,959   |          | 3,878 | 2,645   | 2,151 | 785         | 6,254   | 6,432 | 6,764       | 2,486        |
| Deletions   | 8,975  | 12,475   |          | 4,348 | 2,885   | 2,454 | 452         | 6,671   | 7,043 | 7,458       | 2,821        |
| Inversions  | 370    | 156      | 166      | 126   | 132     | 356   | 370         | 348     | 143   | 92          | 46           |

Translocations: 1,092 1,106 698 632 557 623 1,011 985 1,315 648 Inverted duplications: 0 0 1 150 1 307 356 370 348 143
c1069-6017 and CHS clusters A and B related to seed coat pigmentation on BAC77G7-a were not correctly assembled and thus were not identified on Gmax.275_v2.0. Similarly, the genomic region with the GmIRCHS sequence for the dominant I allele was not identified because of ‘Williams 82’ (the donor of Gmax.275_v2.0) having a different i allele. Interestingly, we identified chimeric sequences derived from Gm-c1069-6017 and GmIRCHS in the partially aligned PacBio reads on another copy of subsilin (Glyma.08G190000) and DnaJ (Glyma.08G191000), respectively (Supplementary Fig. S9). Furthermore, the short chimeric sequences in the partially aligned Illumina reads near the SV breakpoints were manually recorded as presence and absence variation in all accessions (Supplementary Table S5). The seed coat colour of 53 accessions without both chimeric sequences in the mini-core collections revealed brown, reddish-brown, and black, while the remaining yellow and green seed coat colour accessions had either of the chimeric sequences. Among accessions with yellow and green seed coat colours, no pigmentation was observed on the hilum of 22 accessions with chimeric GmIRCHS sequences, while 122 accessions with chimeric partial Gm-c1069-6017 sequence revealed pigmentation on the hilum. Thus, the PAVs related to the I locus in the partially aligned reads successfully explained the variation in seed coat pigment and pigmentation on the hilum of 197 accessions.

Among the classical loci that governed the seed coat colour of soybean, we identified new functional alleles at the R and K1 loci based on the read mapping data. Three known non-functional alleles, Gly63fs, Arg75fs, and splice site change (AGgt) on the read mapping data. Three known non-functional alleles, Gly63fs, Arg75fs, and splice site change (AGgt) at the R2R3 MYB gene Glyma.09G235100, which is classically called Gly63fs, Arg75fs, and splice site change (AGgt) on GmSGR1 encodes the flavonoid 3′-hydroxylase (F3′H) gene. The known stop-loss variant in the third exon caused by a 3′-bp insertion, which led to a frameshift in the psbM gene Glyma.15G208300. Stay-green controlled by cytG is known to be the same 3′-bp insertion on chloroplast psbM, which encodes small subunits of photosystem II. Since organelle genome sequences did not include read mapping in the present study, the reads including chloroplast psbM have been mapped to the nuclear psbM gene Glyma.15G208300 and were detected as heterozygous variant Ile25fs. The remaining stay-green accessions were characterized as having double recessive genes, d1 and d2. The D1 and d2 loci encode the GmSGR1 gene Glyma.11G027400 and GmSGR2 gene Glyma.01G214600, respectively. For the D2 locus, the non-functional known variant, Val60fs, and the new non-functional variant, Lys59fs, at GmSGR2 gene Glyma.01G214600 were observed at five and three accessions, respectively. In contrast, a non-functional known variant of the D1 locus by GmD2IN transposon insertion to GmSGR1 gene Glyma.11G027400 was identified from five accessions by manual inspection of the partially aligned Illumina reads to the corresponding genomic region (Chr11:1975880 or Chr11:1975350). Among them, only four accessions (GmWMC011, GmWMC018, GmWMC127, and GmWMC129) revealed a stay-green phenotype by the non-functional variant at both loci.

3.6. Survey of rearrangement between soybean genomes during progression based on CNV analysis

As investigations related to the diversity of soybean accessions advanced, we performed CNV analysis for the 198 accessions with Illumina reads, which indicated the trace of the genomic rearrangements (Supplementary Fig. S10). The distributions of CNVs were generally similar in the three subgroups, unlike sequence variances such as SNPs and indels. However, such rearrangements were also observed in some specific accessions such as ‘Bongchubalejama’ (GmWMC089) in 13–16 Mb of Chr01 in the ‘Japan’ subgroup. We confirmed read mapping of these regions, which included long gaps of more than 1 Mb with extremely low read coverage, and found that many genes in these regions have been lost. Accordingly, the phenotype of such accessions with a long gap would be influenced if the target gene of interest lacks in the region. A similar long gap was found in ‘DAIZU’ (GmJMC133) at 30.2–31.0 Mb on Chr08 of the ‘Japan’ subgroup; ‘HOUJAKU’ (GmJMC067) at 14.9–15.7 Mb on Chr10 of the ‘Japan’ subgroup; approximately 24–28 Mb on Chr12 of ‘MEGURO 1’ (GmJMC064) and ‘POCHAL’ (GmWMC020) of the ‘Japan’ subgroup and ‘GREEN PE POKE’ (GmWMC127) and ‘IPPON SANGOU’ (GmJMC076) of the ‘World’ subgroup. This CNV information will contribute to association analysis and gene functional analysis.
4. Conclusions
In the present study, we obtained the whole-genome sequences of 198 soybean accessions, which had been carefully selected to represent the genetic diversity of the worldwide 1,603 accessions and particularly represented the variations in Japanese accessions harbouring the distinct genetic and morphological characteristics from those in the Asian continent. The genetic diversity of the 198 accessions was analysed based on their whole-genome sequences, and then the existence of three subgroups that contain specific and shared polymorphisms was suggested. GWAS on DTF and the detailed comparison between sequence polymorphisms and phenotypes in colour-related traits were performed with the whole-genome sequences of the accessions. The variations in DTF and colour related to the seed coat, hilum, pubescence, and flower were mostly characterized by using the known and new variants observed in the whole-genome sequence reads. CNV analysis suggested that only a few large-scale genome rearrangements may have occurred during the domestication of soybeans. The results suggested that the genomic sequences and variants obtained for the 198 soybean accessions have great potential to provide information for soybean breeding and genetic studies to uncover novel alleles or genes involved in agronomically important traits.

Acknowledgements
We thank Kenta Shirasawa, Yoshie Kishita, and Shinobu Nakayama of KDRI for their technical assistance, and Yuichiro Katayose of the National Agriculture and Forestry and Fisheries of Japan (SFC-1008, BAC2004), and from the Genebank Project, National Agriculture and Food Research Organization (NARO).

Accession number
PRJDB7281.

Funding
This work was supported by Japan Science and Technology Agency (JST) CREST Grant Number JPMJCR1601 and JPMJCR1602, the Kazusa DNA Research Institute (KbRI) Foundation, in part by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (SFC-1008, BAC2004), and from the Genebank Project, National Agriculture and Food Research Organization (NARO).

Conflict of interest
None declared.

Data availability
The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files. The obtained genome sequence reads are available from the DDBJ Sequence Read Archive (DRA) under the BioProject accession number of PRJDB7281.

Supplementary data
Supplementary data are available at DNAES online.

References
1. Food and Agriculture Organization of the United Nations. 2010, The Second Report on the State of the World’s Plant Genetic Resources for Food and Agriculture. Rome, Italy: Food and Agriculture Organization.
2. Sharma, S., Upadhyaya, H.D., Varshney, R.K. and Gowda, C. 2013, Pre-breeding for diversification of primary gene pool and genetic enhancement of grain legumes, Front. Plant Sci., 4, 309.
3. Li, J.-Y., Wang, J. and Zeigler, R.S. 2014, The 3,000 rice genomes project: new opportunities and challenges for future rice research, Gigascience, 3, 7.
4. Wang, W., Mauleon, R., Hu, Z., et al. 2018, Genomic variation in 3,010 diverse accessions of Asian cultivated rice, Nature, 557, 43–9.
5. Fuentes, R.R., Chebotarov, D., Duitama, J., et al. 2019, Structural variants in 3000 rice genomes, Genome Res., 29, 870–80.
6. Song, Q., Hyten, D.L., Jia, G., et al. 2015, Fingerprinting soybean germplasm and its utility in genomic research, G3 (Bethesda), 5, 1999–2006.
7. Hamblin, M.T., Buckler, E.S. and Jannink, J.-L. 2011, Population genetics of germplasm-based crop improvement methods, Trends Genet., 27, 98–106.
8. Momotani, T.H.E., Hayes, B.J. and Goddard, M.E. 2001, Prediction of total genetic value using genome-wide dense marker maps, Genetics, 157, 1819–29.
9. Pace, J., Yu, X. and Liebhaberst, T. 2015, Genomic prediction of seedling root length in maize (Zea mays L.), Plant J., 83, 903–12.
10. Yu, X., Li, X., Guo, T., et al. 2016, Genomic prediction contributing to a promising global strategy to turbocharge gene banks, Nature Plants, 2, 16150.
11. Tanaka, R. and Iwata, H. 2018, Bayesian optimization for genomic selection: a method for discovering the best genotype among a large number of candidates, Theor. Appl. Genet., 131, 93–105.
12. Zhou, Z., Jiang, Y., Wang, Z., et al. 2015, Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean, Nat. Biotechnol., 33, 408–414.
13. Lye, Z.N. and Purugganan, M.D. 2019, Copy number variation in domestication, Trends Plant Sci., 24, 352–65.
14. Zmienko, A., Samelak, A., Kozłowski, P. and Figlerowicz, M. 2014, Copy number polymorphism in plant genomes, Theor. Appl. Genet., 127, 1–18.
15. Jannink, J.L., Iwata, H., Bhat, P.R., Chao, S., Wenzl, P. and Muelbauer, G.J. 2009, Marker imputation in barley association studies, Plant Genome, 2, 11–22.
16. Iwata, H. and Jannink, J.L. 2010, Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines, Crop Sci., 50, 1269–78.
17. Scheben, A., Batley, J. and Edwards, D. 2017, Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application, Plant Biotechnol. J., 15, 149–61.
18. Yao, Z., Zhao, H., Ren, L., et al. 2012, Genome-wide genetic changes during modern breeding of maize, Nat. Genet., 44, 812–5.
19. Mace, E.S., Tai, S., Gilding, E.K., et al. 2013, Whole-genome sequencing reveals untagged genetic potential in Africa’s indigenous cereal crop sorghum, Nat. Commun., 4, 2320.
20. 100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., et al. 2014, Exploring generic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing, Plant J., 80, 136–48.
21. Zhang, B., Chen, P., Florez-Palacios, S.L., Shi, A., Hou, A. and Ishibashi, T. 2010, Seed quality attributes of food-grade soybeans from the US and Asia, Euphytica, 173, 387–96.
22. Hill, J., Nelson, E., Tilman, D., Polasky, S. and Tiffany, D. 2006, Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels, Proc. Natl. Acad. Sci. U.S.A., 103, 11206–10.
23. Schmutz, J., Cannon, S.B., Schlueter, J., et al. 2010, Genome sequence of the palaeopolyploid soybean, Nature, 463, 178–83.
24. Lam, H.-M., Xu, X., Liu, X., et al. 2010, Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection, Nat. Genet., 42, 1053–9.
25. Li, Y.-H., Zhao, S.-C., Ma, J.-X., et al. 2013, Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing, BMC Genomics, 14, 579.
26. Chung, W.-H., Jeong, N., Kim, J., et al. 2014, Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes, DNA Res., 21, 153–67.
73. Kumagai, M., Nishikawa, D., Kawahara, Y., et al. 2019, TASUKE+: a web-based platform for exploring GWAS results and large-scale resequencing data, DNA Res., 26, 445–52.

74. Rafalski, A. and Morgante, M. 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size, Trends Genet., 20, 103–11.

75. Ogiso-Tanaka, E., Shimizu, T., Hajika, M., Kaga, A. and Ishimoto, M. 2019, Highly multiplexed AmpliSeq technology identifies novel variation of flowering time-related genes in soybean (Glycine max), DNA Res., 26, 243–60.

76. Tsubokura, Y., Watanabe, S., Xia, Z., et al. 2014, Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean, Ann. Bot., 113, 429–41.

77. Wang, Y., Gu, Y., Gao, H., et al. 2016, Molecular and geographic evolutionary support for the essential role of GIGANTEAa in soybean domestication of flowering time, BMC Evol. Biol., 16, 79.

78. Li, C., Li, Y-h., Li, Y., et al. 2020, A domestication-associated gene GmPRR3b regulates the circadian clock and flowering time in soybean, Mol Plant., 13, 745–59.

79. Zhou, Y., Minio, A., Massonnet, M., et al. 2019, The population genetics of structural variants in grapevine domestication, Nat. Plants, 5, 965–79.

80. Hind, D.A., Kloek, A.P., Jen, M., Chen, X. and Frazer, K.A. 2006, Common deletions and SNPs are in linkage disequilibrium in the human genome, Nat. Genet., 38, 82–5.

81. Gaut, B.S., Seymour, D.K., Liu, Q. and Zhou, Y. 2018, Demography and its effects on genomic variation in crop domestication, Nat. Plants, 4, 512–20.

82. Bhattacharyya, M., Narayanan, N., Gao, H., et al. 2005, Identification of a large cluster of coiled coil-nucleotide binding site–leucine rich repeat-type genes from the Rps1 region containing Phytophthora resistance genes in soybean, Theor. Appl. Genet., 111, 75–86.

83. Hayes, A., Jeong, S., Gore, M., et al. 2004, Recombination within a nucleotide-binding-site/leucine-rich-repeat gene cluster produces new variants conditioning resistance to soybean mosaic virus in soybeans, Genetics, 166, 493–503.

84. Su, S.J., Bowman, B.C., Jeong, N., et al. 2011, The Rsv3 locus conferring resistance to Soybean mosaic virus is associated with a cluster of coiled-coil nucleotide-binding leucine-rich repeat genes, Plant Genome, 4, 55–64.

85. Xu, M., Xu, Z., Liu, B., et al. 2013, Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean, BMC Plant Biol., 13, 91.

86. Matsushika, A., Makino, S., Kojima, M. and Mizuno, T. 2000, Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in Arabidopsis thaliana: insight into the plant circadian clock, Plant Cell Physiol., 41, 1002–12.

87. Cho, Y.B., Jones, S.I. and Vodkin, L.O. 2017, Mutations in Argonaute5 illuminate epistatic interactions of the K1 and I loci leading to saddle seed color patterns in Glycine max, Plant Cell, 29, 708–25.