A reference-grade wild soybean genome

Min Xie1, Claire Yik-Lok Chung1, Man-Wah Li1, Fuk-Ling Wong1, Xin Wang1, Ailin Liu1, Zhili Wang1, Alden King-Yung Leung1, Tin-Hang Wong1, Suk-Wah Tong1, Zhixia Xiao1, Kejing Fan1, Ming-Sin Ng1, Xinpeng Qi1, Linfeng Yang2, Tianquan Deng2, Lijuan He2, Lu Chen2, Aisi Fu3, Qiong Ding3, Junxian He1, Gyuhwa Chung4, Sachiko Isobe5, Takanari Tanabata5, Babu Valliyodan6, Henry T. Nguyen6, Steven B. Cannon7, Christine H. Foyer8, Ting-Fung Chan1 & Hon-Ming Lam1

Efficient crop improvement depends on the application of accurate genetic information contained in diverse germplasm resources. Here we report a reference-grade genome of wild soybean accession W05, with a final assembled genome size of 1013.2 Mb and a contig N50 of 3.3 Mb. The analytical power of the W05 genome is demonstrated by several examples. First, we identify an inversion at the locus determining seed coat color during domestication. Second, a translocation event between chromosomes 11 and 13 of some genotypes is shown to interfere with the assignment of QTLs. Third, we find a region containing copy number variations of the Kunitz trypsin inhibitor (KTI) genes. Such findings illustrate the power of this assembly in the analysis of large structural variations in soybean germplasm collections. The wild soybean genome assembly has wide applications in comparative genomic and evolutionary studies, as well as in crop breeding and improvement programs.
Genomic information is the essential foundation of current crop improvement programs. Accurate information is required for effective tracking of genomic variations, mapping important quantitative trait loci (QTLs), and discovering novel alleles. These tasks are intrinsically dependent on the availability of a range of genetic resources and, most crucially, high-quality reference genomes. Wild germplasm contributes a significant proportion of the genetic resources of major crop species. Although reference genomes are currently available for two soybean cultivars, the Williams 82 (Wm82) genome has been most commonly used for a range of applications. However, these approaches alone have limitations and cannot fully address questions regarding large structural variations or complex genomic rearrangements. A high-quality reference genome from wild soybean is also a crucial tool for use in such studies, because it increases the precision of population genetic analysis of complicated genomes. For example, genomic information is the essential foundation for understanding domestication-related events that involved wild germplasms.

We have previously reported whole-genome sequencing data for wild and cultivated soybeans, and demonstrated the high genome diversity in wild soybean populations compared with cultivated soybean. The genomic diversity of wild soybean has been confirmed and elaborated in reports by ourselves and others. Despite several previous attempts at whole-genome assembly in wild soybeans, a high-quality reference genome has remained elusive. In this study, we not only demonstrate the power of W05 reference genome but also highlight its applicability in a wide range of comparative genomic and evolutionary studies, using a range of examples, such as the identification of large structural variations, QTLs, genes, and alleles. The advantages of combining high-quality reference genomes and optical mapping (OM) in studying structural variations among multiple accessions is also described.

Results

De novo sequencing and assembly. State-of-the-art whole-genome sequencing technologies were used to assemble a high-quality reference genome for wild soybean accession W05. W05 has previously been employed to identify several agronomically important QTLs, together with the identification of the causal gene conferring salt tolerance in wild soybean. In this study, we not only demonstrate the power of W05 reference genome but also highlight its applicability in a wide range of comparative genomic and evolutionary studies, using a range of examples, such as the identification of large structural variations, QTLs, genes, and alleles. The advantages of combining high-quality reference genomes and optical mapping (OM) in studying structural variations among multiple accessions is also described.

The resulting Hi-C scaffolds comprised 1161 sequences, with a total length of 989.8 Mb and a scaffold N50 of 48.5 Mb (Supplementary Figure 1e). Superscaffolds were generated by merging hybrid scaffolds and Hi-C scaffolds (Supplementary Figure 1f).

After gap filling and polishing, the final assembly for W05 is 1013.2 Mb in length, with 988.6 Mb unambiguous bases (Supplementary Figure 1g and Table 1). In total, 95.7% of sequences are anchored to 20 superscaffolds, corresponding to 20 chromosomes, whereas 43.6 Mb in 1098 contigs remain unplaced. The contig N50 of the final assembly is 3.3 Mb (Table 1). The longest contig of the W05 final assembly is 23.2 Mb in length, spanning 47.7% of chromosome 6. The contiguity of the W05 assembly is approximately a 17-fold improvement over the current reference genome Wm82.v2 and of similar quality as the recently published Chinese cultivated soybean reference genome of ZH135 (Supplementary Table 3). Contig N50, scaffold N50, and total assembled genome size of other soybean genome assemblies were compared (Supplementary Table 3), but these genomes were not included in subsequent analysis, because they are highly fragmented.

Assembly evaluation. The completeness of this genome assembly was examined using the Benchmarking Universal Single-Copy

Table 1. Summary of W05 genome assembly and annotation

| Categories       | Type      | Length (Mb) | No. (Mb) | Percentage (%) |
|------------------|-----------|-------------|----------|----------------|
| Assembly         | Contigs   | 988.6       | 1870     |                |
|                  | Contig N50| 3.3         | 58       |                |
|                  | Contig N90| 0.4         | 432      |                |
|                  | Scaffolds | 1013.2      | 118      |                |
|                  | Scaffold N50| 50.7     | 10       |                |
|                  | Scaffold N90| 38.4     | 19       |                |
| Protein-coding   | Total transcripts | – | 89,477 | 100.0 |
| genes            |           |             |          |                |
| Non-coding RNAs  | miRNA     | 0.036       | 288      | 0.004          |
|                  | snRNA     | 0.216       | 1988     | 0.021          |
|                  | rRNA      | 0.032       | 147      | 0.003          |
|                  | tRNA      | 0.067       | 892      | 0.007          |
| Transposable     | Class I:  | 359.6       | –        | 35.5           |
| elements         | Retroelements |        |          |                |
|                  | SINEs     | 1.1         | –        | 0.1            |
|                  | LINEs     | 13.3        | –        | 1.3            |
|                  | LTR elements | 345.2 | –        | 34.1           |
|                  | Ty1/Copia | 93.5        | –        | 9.2            |
|                  | Ty3/gypsy | 248.0       | –        | 24.5           |
|                  | Others    | 3.8         | –        | 0.4            |
|                  | Class II: DNA transposons | 74.8 | – | 7.4 |
|                  | CMC-EnSpm | 29.7        | –        | 2.9            |
|                  | MULE      | 27.9        | –        | 2.8            |
|                  | TcMar     | 0.8         | –        | 0.1            |
|                  | hAT       | 8.7         | –        | 0.9            |
|                  | Helitron  | 4.2         | –        | 0.4            |
|                  | Others    | 3.5         | –        | 0.3            |
|                  | Satellites| 4.9         | –        | 0.5            |
|                  | Simple repeats | 44.1  | –        | 4.4            |
|                  | Low complexity | 3.1   | –        | 0.3            |
|                  | Unknown   | 59.8        | –        | 5.9            |
|                  | Total transposable elements | 546.4 | – | 53.9 |
Orthologs (BUSCO) evaluation score\(^{16}\). The completeness of W05 is comparable to the Wm82_v2 and ZH13 reference genomes (Supplementary Table 4). The mapping rate of the PacBio Isoform Sequencing (IsoSeq) full-length transcripts was 97.7% (Supplementary Table 5). Cent91/92 soybean-specific centromeric repeats\(^{17}\) were found in 19 chromosomes, except chromosome 1 (Fig. 1e). In addition, the Arabidopsis-type telomeric tandem repeat array\(^{18}\) (CCCTAAA/TTTAGGG repeats) was found at both distal ends for 13 chromosomes and at a single distal end for the remaining 7 chromosomes (Fig. 1e and Supplementary Table 6). In contrast, only 9 and 7 chromosomes of Wm82_v2 and ZH13, respectively, contain telomeric tandem repeats at both distal ends. The completeness of telomeres in the W05 genome is therefore improved compared with that of Wm82_v2 and ZH13 (Supplementary Table 6). GC content along each chromosome was calculated (Fig. 1c). Furthermore, using in silico PCR, 874 conserved, unique-site soybean simple sequence repeat markers were mapped to the W05 genome, which are evenly distributed on the 20 chromosomes (Fig. 1d and Supplementary Data 1).

**Genome annotation.** Protein-coding genes and alternative spliced isoforms were annotated by combining evidence generated from RNA-seq/PacBio IsoSeq transcript mapping, homology-based protein mapping, and ab initio prediction. In total, 234.7 Gb of Illumina RNA-seq reads were collected from 31 samples at various development and physiological stages (Supplementary Data 2). PacBio IsoSeq libraries were constructed in order to generate 414,750 full-length and non-chimeric transcripts (Supplementary Table 7). In total, 89,477 protein-coding transcripts were annotated for 55,539 gene loci, with 69,455 transcripts (77.6%) having 5′-untranslated region (UTR) and 71,271 transcripts (79.7%) having 3′-UTR (Supplementary Table 8). In addition, 82,567 transcripts (92.3%) encode proteins that contain predicted functional domains (Table 1). Features of the annotated transcripts in W05 are similar to those of Wm82_v2 and ZH13\(^{4,5}\) (Fig. 1b and Supplementary Table 8). BUSCO evaluation shows that completeness of the W05 annotated gene set was comparable to that of Wm82_v2 and ZH13 (Supplementary Table 9). A total of 288 microRNAs (miRNA), 1988 small nuclear RNAs (snRNA), and 147 ribosomal RNAs (rRNA) were identified in the wild soybean genome (Table 1). In addition, 892 transfer RNAs (tRNA) were identified, representing anti-codons for all 20 types of amino acids.

**Identification and refinement of QTLs.** One major application of the wild soybean reference genome is the identification of QTLs, genes, and alleles. We previously constructed a recombinant inbred (RI) population by crossing W05 to a cultivated soybean Union (C08)\(^{8}\), which shared the same recurrent parent with Wm82. To demonstrate that W05 can be used effectively as a reference genome for QTL mapping, we make use of our previous published phenotypic data together with new data of seed size and sequencing reads of 96 core RI lines\(^{8}\) to construct bin-maps, and map QTLs using W05 genome or Wm82_v2 genome as reference. The relative genomic location and span of QTLs are comparable when either genome was used as the reference, with very few discrepancies (Supplementary Data 3). For example, the growth period QTL on chromosome 11 spans a 3.55 Mb region in the Wm82_v2 genome, but only 500 Kb in the W05 genome. This may be due to the low-quality assembly of Wm82_v2 in this region, as an ~3.8 Mb sequence within this region, which was originally present in Wm82_v1, was not anchored to the chromosome in Wm82_v2 (Supplementary Data 3).

The quality of the assembly down to the nucleotide level was assessed by examining known traits associated genes. For Nt\(^{18,19}\), R2/Rfg\(^{20}\), and G\(^{21,22}\) loci, the alleles in W05 match the observed phenotypes (Table 2, Supplementary Data 3 and 4). In addition, we have also identified known polymorphisms and additional alleles of the causal genes in the QTLs controlling growth period, flower color, seed coat color, and pubescence color in W05 (Table 2, Supplementary Data 3, 4).

To investigate a more complex case, we examined the I locus on chromosome 8, which determines the pigmentation of the seed coat\(^{33}\), a major trait that was selected during domestication\(^{31}\). It was reported that the dominant allele in cultivated soybeans contains an inverted repeat of the chalcone synthase (CHS) gene cluster, which triggers posttranscriptional gene silencing (PTGS) and inhibits the expression of CHS gene family members in the seed coat; hence, resulting in colorless seed coat and yellow seeds\(^{34}\). Deletion that disrupts the inverted repeat CHS gene cluster in a revertant soybean accession resulted in seed coat color transition from colorless to pigmented\(^{33}\). W05 has a pigmented seed coat, whereas C08 has a colorless seed coat. A seed coat color QTL that overlaps with the known I locus was mapped (Supplementary Data 3). The W05 reference genome possesses the same inverted repeat of the CHS gene cluster as Wm82 (Fig. 2a), indicating that the inverted repeat is not sufficient to explain the seed coat color change during domestication.

To identify the genetic variation that caused the seed coat color change during domestication, the W05 genome was compared with the two published bacterial artificial chromosomes of Wm82 that were previously used in I locus studies\(^{23,25}\) (the I locus region is poorly assembled in Wm82_v2), as well as the recently published ZH13 genome\(^{5}\). This comparative analysis reveals a complex structural rearrangement next to the CHS gene cluster in Wm82 and ZH13, which includes both inversion and gene duplication (Fig. 2a and Supplementary Figure 2). The inversion brought the promoter and the first four exons of a subtilisin gene (Glysoja.08G020214 in W05) to a position next to the CHS gene cluster in Wm82 and ZH13 (Fig. 2a and Supplementary Figure 2). This finding indicates that the subtilisin promoter may drive the expression of a chimeric transcript that reads through the subtilisin gene fragment and anti-CHS1 gene region to cause PTGS of the CHS genes. This is in agreement with the previous speculation based on Expressed Sequence Tag data\(^{26}\). To validate this hypothesis, the subtilisin-anti-CHS1 chimeric transcript was successfully amplified with PCR primers specific to both the subtilisin gene fragment and the anti-CHS1 genes using complementary DNA from developing seed coat of Wm82 in a strand-specific manner (Fig. 2b). The subtilisin-anti-CHS1 chimeric transcript was able to form double-stranded RNA with the sense CHS mRNAs, hence causing PTGS to inhibit the expression of CHS genes (Fig. 2c). Deletion of the CHS gene cluster B adjacent to the subtilisin gene fragment will disrupt PTGS and lead to the transition from a colorless seed coat to a pigmented seed coat, which is in agreement with the previous report\(^{23}\) (Fig. 2c).

**Major structural changes compared with cultivated genomes.** Transposable elements (TEs) are repeated DNA sequences that make up a significant proportion of plant genomes\(^{27}\). They are important source of variations for natural and artificial selection. Combining homology-based searches and de novo prediction, we identified 546.4 Mb of repeat elements (53.9% of the total assembled W05 genome) (Table 1, Fig. 1a). The most abundant type of TEs in the W05 genome is the long terminal repeat (LTR)
Fig. 1 Distribution of W05 genomic features. The outer layer illustrates the 20 chromosomes of W05 in megabases (Mb). 

- **a** Repeat coverage was calculated by the occupancy of repeat sequence in 1 Mb window (step size: 500 Kb).
- **b** Gene coverage was calculated by the occupancy of coding sequence in 1 Mb window (step size: 500 Kb).
- **c** GC content was calculated in a 200 Kb window.
- **d** Position of simple sequence repeat (SSR) markers were indicated in purple. Marker information could be found in Supplementary Data 1.
- **e** Presence of telomeric tandem arrays and cent91/92 soybean specific type centromeric repeats were marked in pink and blue, respectively.
Table 2 Predicted phenotypes based on genomic assemblies and observed phenotypes

| Trait               | Locus | Allele type | Predicted phenotype | Observed phenotype | Wm82 Allele type | Predicted phenotype | Observed phenotype |
|---------------------|-------|-------------|---------------------|--------------------|------------------|---------------------|--------------------|
| Salt tolerance      | Ncl   | intact GmCHX1 | Salt tolerant       | Salt tolerant      | TE-inserted GmCHX1 | Salt sensitive      | Salt sensitive     |
| Nodulation          | Rj2   | rj2(rfg1)   | Do not restrict     | Do not restrict    | Restrict some     | Restrict some       |                     |
|                     |       |             | neither B. japonicum | neither B. japonicum | some strains of  | some strains of     |                    |
|                     |       |             | nor S. fredii       | nor S. fredii      | S. fredii but    | S. fredii but       |                    |
| Flower color        | W1    | W1          | Purple flower       | Purple flower      | w1                | White flower        | White flower       |
| Seed coat color     | i     | i           | Pigmented           | Pigmented          | i                | Colorless           | Colorless          |
| Seed coat color     | G     | G           | Stay green after seed maturation | Stay green after seed maturation | g | Do not stay green after seed maturation | Do not stay green after seed maturation |

*Italicized text denoted gene loci, gene alleles, or species names

Fig. 2 Causal structural variation that controls soybean seed coat pigmentation. 

a) Sequence comparison between W05 genome and Wm82 bacterial artificial chromosome (BAC) sequences at the I locus region. CHS genes and subtilisin gene/gene fragments are indicated with blue and orange, respectively.

b) Top panel: cartoon shows the exon structure of the subtilisin gene fragment (orange), the CHS1 gene (blue), and the Expressed Sequence Tag (EST) sequence Gm-c1069-6017. Positions of primers designed for PCR amplification of subtilisin-anti-CHS1 chimeric transcript are indicated with black arrows. Bottom panel: PCR amplification of the subtilisin-anti-CHS1 chimeric transcript. Experiment was repeated at least twice with independent samples. Marker: 1 Kb Plus DNA ladder (NEB, cat. N3200S). NTC, no template control. GmACT11 is used as a housekeeping control. Unprocessed gel image is provided in Source Data file.

c) Proposed model for the generation of siRNAs originated from a large structural rearrangement in the I locus. CHS genes and the subtilisin gene/gene fragments are illustrated as blue and orange, respectively. Arrowheads indicated the direction of transcription that causes the formation of double-stranded RNA. Cluster A and B are named according to a previous report23. IR-CHS gene cluster: inverted repeat of CHS gene cluster.
Reviewing large structural variations using OM. Although comparisons of the W05 and other reference genomes allows identification of many structural variations, these variations could be accession-specific. Therefore, we expanded our scope to analyze more soybean accessions from diverse origins, including cultivated soybeans popularized in China, the United States, and Japan, as well as wild soybeans originating from China and Korea (Supplementary Data 6). OM technology that can be used to study long-range genomic DNA up to 1 Mb in length was employed together with an in silico map based on our high-quality W05 genome to detect structural variations in diverse soybean accessions. We can thus demonstrate that the W05 reference genome enhances genome comparisons at the kilobase scale. As proof-of-concept, several loci exhibiting different structural features among soybean accessions were investigated, including inversion in the I locus, the translocation between chromosomes 11 and 13, a previously reported region specific to cultivated soybean on chromosome 15\(^7\), and gene copy number variation in a Kunitz trypsin inhibitor (KTI) gene cluster. Only soybean accessions with sufficient coverage and depth of optical contigs in the target regions were employed for comparative analysis (Supplementary Data 6).

For the I locus, comparative analysis revealed that soybean accessions exhibiting a pigmented seed coat have no inversion (similar to W05), whereas those with colorless seed coats share the same inversion as Wm82 and ZH13 (Fig. 3a). To further confirm the OM results, the inversion junction that creates the subtilisin-anti-CHS1 chimeric transcript was amplified from genomic DNA of different soybean accessions by PCR. Consistent with the OM results, inversion junction could be amplified from all soybean accessions with colorless seed coats but not from those with pigmented seed coats (Supplementary Figure 6).

Using Wm82\(_{v1}\), ZH13, and W05 as references, the translocation event between chromosomes 11 and 13 were also detected. One (W01) out of three (W02, K02) wild soybean accessions share the same chromosome topology with W05. In contrast, all cultivated soybean accessions shared the same topology with the Wm82\(_{v1}\) and ZH13 in silico maps, regardless of their origin whether they were popularized in China, the United States, or Japan (Fig. 3b). Furthermore, the regions at the junction of the translocation are reasonably well conserved among different accessions, which suggests that the translocation was a single evolutionary event.

Previously, a region specific to cultivated soybean (Gm15 46.0–46.5 Mb) was identified in a pan-genome study by mapping short reads of seven wild soybeans onto the Wm82\(_{v1}\) genome\(^7\). We revisited this region and compared the optical contigs of different soybean accessions. Interestingly, the targeted region appears to have diversified among the different cultivated accessions that we have tested from China, the United States, and Japan, whereas only Wm82 possesses such a large insertion (Fig. 3c). This previously reported Glycine max-specific region was confirmed as Wm82 specific.

Comparisons of multiple reference genomes, together with the OM data, allow identification of gene copy number variations. KTI are anti-nutritional factors and are hence not desirable traits for selection in breeding purposes. Marker-assisted breeding has been used to introduce a null allele of KTI from an exotic soybean germplasm into two commercialized lines\(^29\). A KTI gene cluster was identified on chromosome 8 of W05 that contains 17 KTI genes. The optical signaling patterns in this region are similar for cultivated soybean accessions and diversify for wild soybean accessions. Wild soybean accessions W05, W02, and K01 possess long fragments, whereas W01 and all of the cultivated accessions tested (including Wm82) exhibited reduced size in this genomic
In the annotated Wm82_v2 and ZH13 reference genomes, there are only 13 and 11 KTI genes, respectively, in this region (Supplementary Figure 7), implying that copy number variations of KTI genes might have occurred through artificial selection.

**Discussion**

The availability of multiple high-quality reference genomes from diverse genetic backgrounds is a prerequisite for effective mining of crop genomes30, especially in studies involving wild germplasm. The current Wm82_v2 reference genome alone cannot
provide conclusive answers to questions regarding large structural variations and complex genomic rearrangement. A new version of Wm82 genome is under construction (www.soybase.org), which will improve the quality over the current version. However, comparative genomic analysis using only the reference genome of Wm82 may still fail to uncover wild-specific genetic variations, as these may have been lost during domestication or artificial selection. Here we report the assembly of a high-quality reference genome of a wild soybean accession W05, which has many distinguishing features compared with the two existing reference genomes of cultivated soybeans (see Results).

Wild soybeans exhibit agronomic traits that are different to cultivated soybean (e.g., smaller seed size, higher pod number, and seed number per plants, etc.). We have identified QTLs related to key yield components. A seed size QTL was reported and the proper chromosomal locations of the QTLs regulating trailing growth, pod number per plant, and seed number per plant were assigned. These findings demonstrate the value of the W05 reference genome. The combination of this genomic information and a genetic population that includes wild soybeans will result in the step change required for future soybean breeding.

Several examples demonstrate the unique strength of the W05 reference genome. In particular, an inversion associated with seed coat color was identified at the I locus by comparing W05 with Wm82 and ZH13 genomes. The PCR results and OM data presented here demonstrate that the inversion is the source of PTGS (Figs 2, 3a). Furthermore, major structural changes were identified between the W05 genome and those of two cultivated soybeans. One example is the reciprocal translocation between chromosomes 11 and 13 (Supplementary Figures 3 and 4, Supplementary Table 11). Our findings successfully identify the translocation breakpoints, refining previous results obtained using FISH technology. Together with the high base-to-base fidelity, the W05 genome provides an important tool for future investigations of the variations in genes and alleles in wild soybeans, which may not exist in modern cultivars due to domestication bottlenecks and poorly informed selection criteria.

TEs are major drivers of plant evolution. They occupy more than half of the genomic sequences of many crop species. Accession-specific TE insertion events can be effectively identified by comparing de novo genome assembles, as repeated regions could cause ambiguities in re-sequencing approaches. Comparisons of the W05 genome with those of two cultivated soybeans revealed that genes with TE insertions that are found in the two cultivated soybean genomes are concentrated in metabolic pathways (Supplementary Data 5). Conversely, similar patterns were not observed in W05-specific TE-affected genes. As metabolic changes lead to variations that are important for soybean production and quality, we may speculate that TE-affected metabolic genes were unintentionally selected during domestication. However, more wild and domesticated germplasm should be tested to further explore this hypothesis. The highlighted importance of variations in metabolic genes during domestication is consistent with our previous observations showing that fixed single-nucleotide polymorphisms (SNPs) are concentrated in metabolic genes in the cultivated soybeans, compared with wild soybeans.

The data presented here also demonstrate the value of the W05 reference genome and the optical contigs of other soybean accessions. Together, these results provide a powerful tool that can be used to uncover large structural variations in soybean germplasm collections (Fig. 3). We present evidence showing that the inversion found in the I locus (Fig. 3a) and the translocation between Chromosomes 11 and 13 (Fig. 3b) are events that are found in multiple soybean genomes. Moreover, such comparisons reveal accession-specific genomic regions, e.g., a region on chromosome 15 that was previously reported as a feature of the cultivated soybean is in fact unique to Wm82 (Fig. 3c).

The strategy of mapping optical contigs to high-quality reference genomes will have a significant impact in soybean breeding programs. For example, new cultivars generated from radiation-driven mutations often contain major structural genomic changes and these structural variations could be effectively identified using this approach. This strategy also provides an elegant means to detect copy number variations, which are the basis of a number of phenotypic traits in humans, animals, and plants. We specifically investigated a region containing multiple trypsin inhibitor genes as an example of how the genome can be interrogated to identify genes that influence seed nutritional quality. Trypsin inhibitors are considered to have evolved as a protective measure against herbivores and bacterial infections. Trypsin inhibitor proteins account for 2–6% total protein of soybean seeds, which is the highest value found in a range of legumes, reducing the nutritional value of the beans as food and/or animal feed. The genomic analysis reported here reveals that a genomic region on chromosome 8, which presumably contains a KTI gene cluster, has shrunk in cultivated soybeans from China and other countries compared with most of the wild soybeans from China and Korea (Fig. 3d). This finding suggests that domestication may have involved selection for a reduced copy number of KTI genes.

In summary, this study provides information regarding the wild soybean that cannot be easily inferred from the reference genomes of the cultivated soybeans. As wild accessions are important genetic resources for crop improvement, the wild soybean genome reported here will be a valuable, if not indispensable, tool for use in a wide range of applications by legume researchers for comparative genomic and evolutionary studies, and soybean breeders for crop improvement programs.

**Methods**

**Sample preparation and sequencing.** W05 is a Chinese wild soybean accession originally collected in Henan Province. Union (C08) is a cultivated soybean bred in the United States by crossing Williams and SL12, and then backcrossing five times with Williams. W05 and C08 was used to generate a RI population for the identification of important QTLs and the causal gene of a major QTL for salt tolerance. Information of other soybean accessions used in this study is presented in Supplementary Data 6.

For Illumina sequencing, seeds of W05 were germinated on 0.8% water agar in sterile magenta box at 28°C in dark. Hypocotyls and radicals were collected 3–4 days after germination and snap frozen in liquid nitrogen. DNA was extracted from the hypocotyls and radicals of young W05 seedlings using DNeasy Plant Mini
Total RNA was extracted from nodules and the remaining roots growing in a greenhouse. Flower, 7-day pods, 14-day pods, 14-kb strand breaks were introduced to DNA molecules by nicking enzyme Nb.BssSI or pooled. Four SMRTbell libraries were constructed with size selection of 1 kb and sequenced on Illumina HiSeq4000 platform (Cat. 15596018).

The Dovetail Hi-C library was prepared and sequenced by Dovetail Genomics (Santa Cruz, CA)35. Briefly, chromatin in the nucleus of soybean young seedlings was fixed with formaldehyde and extracted. Fixed chromatin was digested with DpnII and sticky ends were filled in with biotinylated nucleotides and ligated. Crosslinks were then reversed and DNA was purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp and sequencing libraries were constructed using NEBNext Ultra® DNA Library Prep Kit (Illumina, Cat. E7370S). Biotin-containing fragments were enriched through streptavidin bead pulldown before PCR amplification of the library. The library was sequenced on Illumina HiSeq X platform (Illumina).

For Bioniao OM, young leaves of all the soybean germplasm used in this study were collected from 7- to 10-day-old seedlings grown in greenhouse and high-molecular weight (HMW) DNA was extracted following the Bioniao Irys Prep High Molecular Weight Tissue DNA Foundation USp method (Illumina Document Number: 30128). Extracted HMW DNA molecules were fluorescently stained using Nick-Label-Repair-Stain (NLRS) enzymatic reactions following the Bioniao Prep™ Labeling - NLRS Protocol (Illumina Document Number: 30024). Briefly, single-strand breaks were introduced to DNA molecules by nicking enzyme Nb.BstSI or NB.BstSI to generate sequence motif-specific patterns. Post-cleanup molecules were labeled with fluorescent nucleotides and repaired. Molecule backbones were also fluorescently stained with YO-YO-1 to visualize the full lengths.

NLRS reaction products were then run on Bioniao Saphyr (Bioniao Genomics, San Diego, CA) for W05 or Irys system (Bioniao Genomics) for the rest of the soybean germplasm, where DNA molecules were automatically stretched and imaged within nanochannel arrays. Distances between fluorescently labeled nicking sites form patterns as basis for alignment and assembly by Bioniao AutoDetect software (v2.1.4).

Embryos of W05 were collected 24 h after germination in distilled water, in a dark incubator. Cotyledons and hypocotyl were collected 3 days, whereas root, apical buds, and stems were collected 10 days after sowing in vermiculite with 70% water content and growing in a greenhouse. Flower, 7-day pods, 14-day pods, 14-day seeds, 40-day pods, and 40-day seeds were collected at reproductive stages from soybean plant grown in the greenhouse. Total RNA was extracted using RNAiso Plus reagent (Takara, Kyoto, Japan, Cat. 9108). Nodules and remaining roots were collected 28 days post inoculation with Sinorhizobium fredii strain CCBAU45436. Total RNA was extracted from nodules and the remaining roots using TRizol® reagent (Thermo Fisher Scientific, Waltham, MA, Cat. 1559960). RNA samples were sent out to BGI-Shenzhen for both RNA sequencing (RNA-Seq) and PacBio IsoSeq. For RNA-Seq, stranded RNA-seq libraries were constructed for each RNA sample and sequenced on Illumina HiSeq4000 platform (Illumina). For PacBio IsoSeq, equal amount of RNA from different tissues were pooled. Four SMRTbell libraries were constructed with size selection of 1–2 Kb, 2–3 Kb, 3–6 Kb, and 5–10 Kb using the BluePippin (Illumina, Santa Cruz, CA)36. Briefly, young, chromatin in the nucleus of soybean young seedlings was fixed with formaldehyde and extracted. Fixed chromatin was digested with DpnII and sticky ends were filled in with biotinylated nucleotides and ligated. Crosslinks were then reversed and DNA was purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp and sequencing libraries were constructed using NEBNext Ultra® DNA Library Prep Kit (Illumina, Cat. E7370S). Biotin-containing fragments were enriched through streptavidin bead pulldown before PCR amplification of the library. The library was sequenced on Illumina HiSeq X platform (Illumina).

For Bioniao OM, young leaves of all the soybean germplasm used in this study were collected from 7- to 10-day-old seedlings grown in greenhouse and high-molecular weight (HMW) DNA was extracted following the Bioniao Irys Prep High Molecular Weight Tissue DNA Foundation USp method (Illumina Document Number: 30128). Extracted HMW DNA molecules were fluorescently stained using Nick-Label-Repair-Stain (NLRS) enzymatic reactions following the Bioniao Prep™ Labeling - NLRS Protocol (Illumina Document Number: 30024). Briefly, single-strand breaks were introduced to DNA molecules by nicking enzyme Nb.BstSI or NB.BstSI to generate sequence motif-specific patterns. Post-cleanup molecules were labeled with fluorescent nucleotides and repaired. Molecule backbones were also fluorescently stained with YO-YO-1 to visualize the full lengths.

De novo genome assembly. Before de novo assembly, low-quality PacBio subreads with a read length shorter than 500 bp or a quality score lower than 0.8 were filtered.

De novo genome assembly was performed using PBSuite (v3.0.1 release v06082017, parameter settings: Iteration: 5; minlen: 150 kb; minsites: 9; initialAssembly: 1.00E−11; extendRef: 1.00E−11; merge: 1.00E−11). Two-enzyme hybrid scaffolding were performed using Bioniao Solve (v3.0.1 release v06082017, default parameters) with polished contigs and two set of de novo assembled optical contigs as input to generate hybrid scaffolds.

Superscaffolds were generated by merging Hi-C scaffolds and hybrid scaffolds. Linkages between adjacent contigs within superscaffolds were classified into three categories based on supporting evidence: (i) both: linkage supported by both Hi-C scaffolds and hybrid scaffolds; (ii) map: linkage supported only by hybrid scaffolds; and (iii) Hi-C: linkage supported only by Hi-C scaffolds.

To identify TEs that were absent in the RepBase library, a de novo repeat library was constructed using Repeatmodeler (version 1.0.10) (http://www.repeatmasker.org/). Tandem repeats in the genome assembly was identified using RepeatMasker (version 4.0.4). Well-characterized TEs were identified by searching against W05 genome assembly at DNA level and protein level using RepeatMasker (version 4.0.7) (https://www.repeatmasker.org/) and ProteinRepeatMask (version 4.0.7) (https://github.com/Bjornst/RPMSuite) as the query library. To identify TEs that were absent in the RepBase library, a de novo repeat library was constructed using Repeatmodeler (version 1.0.10) (http://www.repeatmasker.org/).
using BWA45 (version 0.7.15-r1140) with default parameters. Properly and a 0.5 cM step size. For each agronomical trait, logarithm of the odds (LOD) setting:

**Gene annotation.** For homology-based evidence generation, the sequences of proteins from Arabidopsis thaliana (Arabidopsis) and Lotus japonicus (Lotus japonicus) were mapped to G. soja W05 genome using BLASTx (-a)

RepeatMasker was run against the genome assembly again, with de novo ARTAlignments for tRNAs genes with default parameters. Annotated tRNA genes that were well supported by both homology-based evidence and expression-based annotation, transcripts using PASApipeline54 (version 2.2.0) with the following parameters:

Gene annotation or/). RepeatMasker was run against the genome assembly again, with de novo ARTAlignments for tRNAs genes with default parameters. Annotated tRNA genes that were well supported by both homology-based evidence and expression-based annotation, transcripts using PASApipeline54 (version 2.2.0) with the following parameters:

For expression-based evidence generation, low-quality RNA-seq reads were filtered and trimmed using trim_galore (version 0.4.1) (http://www.bioinformatics. babraham.ac.uk/projects/trim_galore/). Trimmed reads with read length shorter than 80 bp were discarded. Stranded and non-stranded RNA-seq reads were assembled into stranded and non-stranded unigenes using Trinity87 (version 2.4.0) with parameters. The assembled stranded and non-stranded unigenes were mapped to genome sequences and assembled into non-redundant transcripts using PASApipeline54 (version 2.2.0) with the following parameter setting: –MAX_INTRON_LENGTH 20000. Furthermore, consensus PacBio IsoSeq subreads were extracted, classified, and clustered to generate corrected consensus reads using SMRTAnalysis package (smrta, 2.0.14). Corrected consensus sequences with low-quality score were further corrected with stranded RNA-seq reads using proovread55 (version 2.13.12, parameter setting: –mode sr). After correction, the consensus sequences were mapped to the genome assembly and assembled into non-redundant transcripts using PASApipeline54 (version 2.2.0) with parameters: –MAX_INTRON_LENGTH 20000.

For evolution synthesis and alternative spliced isoform annotation, transcripts that were well supported by both homology-based evidence and expression-based evidence were selected for parameter training for ab initio prediction software Augustus86 (version 3.2.3). All the generated transcript evidences mentioned above and transcripts generated using Augustus were fed into the following parameters:

**ncRNA annotation.** ncRNA sequences of G. max were downloaded from NCBI Genbank database (GenBank ID: X15199.1, AJ009787.1, X02623.1, and AH001766.2). These ncRNA sequences were mapped to G. soja W05 genome assembly using BLASTX57 (version 2.2.31) with default parameters.

Potential functional was assigned to each annotated protein sequence using InterProScan58 (version 5.29–68.0) by searching all available databases with corresponding utilities. Completeness of the annotated gene set was evaluated by BUSCO16 (version 3.0.2, lineage dataset embryophyta_oDb9).

**Binnmap construction and QTL identification.** Binnmap construction and QTL identification was performed with population sequencing and phenotype datasets generated in our previous studies8 using a slightly modified analysis pipeline. Briefly, short reads from the 96 RI lines were mapped to the genome sequence using BWA85 (version 0.7.15–n1410) with default parameters. Properly and uniquely mapped reads were extracted to identify SNPs using Samtools86 (version 1.2) with parameters. The SNP quality value > 30 and < 3 SNPs were allowed in any 10 bp window. Heterozygous or non-parental line-derived SNPs were filtered. At least 20 RI lines were required to have SNPs at each SNP locus. Recombination breakpoints were identified using a modified sliding window approach (window size: 15 SNPs, step size: 1 SNP)56. Adjacent 50 Kb intervals were merged into bins if no recombination events were identified in a sequence region of genetic distance among bins was calculated using R/qtl package43 (version 1.41.6, default parameters). QTLs for each agronomical trait were identified using QTL Cartographer (http://statgen.ncsu.edu/qtlcart/) (version 1.17) with a 10 cM scanning window and a 0.5 cM step size. For each agronomical trait, logarithm of the odds (LOD) cutoff was determined by 1000 permutation of the phenotypes and genotypes with significance level of α < 0.05 using QTL Cartographer. Peaks having LOD value higher than the LOD cutoff were considered as significant. For green seed phenotype data, only RI lines of the green and yellow seeds were used. For seed size phenotype dataset, about 35 seeds from each RI line were measured using SmartProbe46 (version 1.2).

For the same QTL for each agronomical trait, sequence alignment was performed between two QTL regions (with referenced to different genomes) using the nuancer program from MUMmer package50 (version 4.0), with default parameter. Aligned block with length smaller than 1 Kb were filtered. The overall aligned region was defined as overlapped region.

**Whole-genome sequence comparison.** Whole-genome comparison between wild soybean W05 and two cultivated soybean genomes were performed using nuancer from MUMmer package50 (version 4.0) and visualized with mummerplot from MUMmer package (version 4.0). Large structural variations (>100 Kb) were identified based on synteny alignment blocks, with variation boundaries manually checked.

**Optical map analysis of multiple soybean accessions.** To enhance result quality, quality assessment using DataQualityCheck module was performed to exclude optical molecules from scans with possible anomalies. Throughput, signal-to-noise ratios and alignment rates in datasets were indentified using a reference in silico map. They were checked for deviations from the norm, which may indicate clogging of the nanochannel or other errors. 

OM de novo assembly was performed as follows. Filtered optical molecules were assembled into optical contigs using Bionano Solve (version 3.1) assembly pipeline (pipelineCL.py), with parameters optimized by Bionano (optArguments_nonhaplotype_ids.xml, minlen 150, minsize 9, merge: 1.00E-15, vgap 2.0). Whole genome optical molecules were aligned to the W05 reference genome using OM emBedExtEndeRe defined according to genome size to 1E 8 and 1E 9, respectively.

For better accuracy in downstream analysis, assembled optical contigs were preprocessed with the OMTools DataTools module before alignment. In consideration of the resolution of signal detection, signals within 1 Kb apart were merged for both optical molecules and optical contigs to improve alignment accuracy at regions of dense signals. Optical contigs containing only repetitive segments were removed by OMTools (version 1.4a) DataTools module with the lowden parameter. Processed optical molecules were aligned to retained optical contigs with OmBlast (version 1.4a)86. Optical contigs with optical molecule coverage lower than 30 at the target region were excluded from subsequent comparative analysis.

To locate the coordinates of the optical contigs, sequence assemblies of W05, Wm82, and ZH13 were in silico digested for alignment. Signals on the in silico maps within 1 Kb were merged. For each region of interest, optical contigs or in silico maps were aligned to the W05 reference in silico map with OMBlas96. Pairwise alignment was performed between optical contigs that aligned to the target region. Multiple alignment of the optical contigs was then performed to obtain linkage information of genome segments using Optical Map Multiple Alignment package (https://github.com/TF-Chan-Lab/OMTools) (version 1.4a). Signal patterns and linkage information were compared, to characterize genomic structural variations among germplasms.

**PCR verification of the inversion junction at the I locus.** To verify the inversion in the I locus of soybeans with colorless seed coat, junction of the inversion created the subtilisin-anti-CHSI chimeric transcript were amplified from genomic DNA of selected soybean accessions. For the detection of subtilisin-anti-CHSI chimeric transcript, RNA was extracted from the seed coat of developing seeds of Wm82 and W05 using Fast-TR™ for RNA purification (Takara, Cat. 9192) and RNAiso Plus reagent (Takara, Cat. 9108). First-strand cDNAs and subsequent PCR were done using One-Step TB Green™ PrimeScript™ RT-PCR Kit II (Takara, Cat. RR086). In the reverse-transcription step, only the reverse primer specific for either the subtilisin-anti-CHSI chimeric transcript or the GmACT11 gene68 (housekeeping gene) was added. Forward primers were then added for the subsequent PCR amplification. Primer information can be found in Supplementary Table 12.

**Hi-C contact frequency calculation.** Hi-C raw reads were mapped to W05 genome and Hi-C contact frequency between genomic loci was computed using Hic-circ69 (version 1.5.3) with window size of 100 Kb. The Hi-C reads contact frequency matrix was visualized using Juicebox (version 1.5.2).
Received: 10 January 2019 Accepted: 22 February 2019
Published online: 14 March 2019

References
1. Morrell, P. L., Buckler, E. S. & Ross-Ibarra, J. Crop genomics: advances and applications. Nat. Rev. Genet. 13, 85–96 (2012).
2. Munoz, N., Liu, A., Kan, L., Li, M. W. & Lam, H. M. Potential uses of wild germplasms of grain legumes for crop improvement. Int. J. Mol. Sci. 18, 328 (2017).
3. Zhang, H. Y., Mittal, N., Leamy, L. J., Baranani, O. & Song, B. H. Back into the wild-apply untagged genetic diversity of wild relatives for crop improvement. Evol. Appl. 10, 5–24 (2017).
4. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183 (2010).
5. Shen, Y. et al. De novo assembly of a Chinese soybean genome. Sci. China Life Sci. 61, 871–884 (2018).
6. Kim, M. Y. et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proc. Natl Acad. Sci. USA 107, 22032–22037 (2010).
7. Li, Y. H. et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32, 1045–1052 (2014).
8. Qi, X. P. et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat. Commun. 5, 4340 (2014).
9. Lam, H. M. et al. Reresequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42, 1053–1061 (2010).
10. Chung, W. H. et al. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. 21, 153–167 (2014).
11. Qiu, J. et al. Genome re-sequencing of semi-wild soybean reveals a complex Soja population structure and deep introgression. PLoS ONE 9, e108479 (2014).
12. Vallyvodan, B. et al. Landscape of genomic diversity and trait discovery in soybean. Sci. Rep. 6, 23598 (2016).
13. Zhou, Z. K. et al. Reresequencing 302 wild and cultivated accessions in the legume-genome database identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33, 408–412 (2015).
14. Li, M. W. et al. QTLs regulating the contents of antioxidants, phenolics, and flavonoids in soybean seeds share a common genomic region. Front. Plant Sci. 7, 854 (2016).
15. Shimomura, M. et al. The Glycine max cv. Entrei genome for improvement of Japanese soybean cultivars. Int. J. Genom. 2015, 358127 (2015).
16. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
17. Gill, N. et al. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 151, 1167–1174 (2009).
18. Kaul, S. et al. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000).
19. Guan, R. X. et al. Salinity tolerance in soybean is modulated by natural variation in GmsALT3. Plant J. 80, 937–950 (2014).
20. Yang, S. M., Tang, F., Gao, M. Q., Krishnan, H. B. & Zhu, H. Y. K. gene-controlled host specificity of the leghemoglobin symbiosis. Proc. Natl Acad. Sci. USA 107, 18735–18740 (2010).
21. Song, J. et al. Identification and validation of Loci Governing Seed Coat Color by Combining Association Mapping and Bulk Segregation Analysis in Soybean. PLoS ONE 11, e0159064 (2016).
22. Wang, M. et al. Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat. Genet. 50, 1435–1441 (2018).
23. Tuteja, J. H., Zabala, G., Varala, K., Hudson, M. & Vodkin, L. O. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in Glycine max seed coats. Plant Cell 21, 3063–3077 (2009).
24. Tuteja, J. H., Clough, S. J., Chan, W. C. & Vodkin, L. O. Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in Glycine max. Plant Cell 16, 819–835 (2004).
25. Tuteja, J. H. & Vodkin, L. O. Structural features of the endogenous CHS silencing and target loci in the soybean genome. Crop Sci. 48, 549–568 (2008).
26. Clough, S. J. et al. Features of a 103-kb gene-rich region in soybean. Science 301, 819–831 (2004).
27. Tian, Z. X. et al. Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. Plant Cell 24, 442–4436 (2012).
28. Findley, S. D. et al. A fluorescence in situ hybridization system for karyotyping soybean. Genetics 185, 727–744 (2010).
29. Maranville, J. A. et al. Introgression of null allele of Kunitz trypsin inhibitor from marker-assisted backcross breeding in soybean (Glycine max L. Merr.). BMC Genet. 17, 106 (2016).
30. Yao, W., Li, G. W., Zhao, H., Wang, G. W., Lian, X. M. & Xie, W. B. Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. 16, 187 (2015).
31. Zhao, S. C. et al. Impacts of nucleotide fixation during soybean domestication and improvement. BMC Plant Biol. 15, 81 (2015).
32. Dolatabadian, A., Patel, D. A., Edwards, D. & Batley, J. Copy number variation and disease resistance in plants. Theor. Appl. Genet. 130, 2479–2490 (2017).
33. Vagadia, B. H., Vanga, S. K. & Raghavan, V. Inactivation methods of soybean trypsin inhibitor - a review. Trends Food Sci. Tech. 64, 115–125 (2017).
34. Liu, Y. G. & Whittier, R. F. Rapid preparation of megabase plant DNA from nuclei in agarose plugs and microbeads. Nucleic Acids Res. 22, 2168–2169 (1994).
35. Doyle, J. J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. Bot. Soc. Am. 19, 11–15 (1987).
36. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).
37. Tian, C. F. et al. Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. Proc. Natl Acad. Sci. USA 109, 8629–8634 (2012).
38. Xiao, C. L. et al. MECAT : fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods 14, 1072–1074 (2017).
39. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/13033997 (2013).
40. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963 (2014).
41. Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26, 342–350 (2016).
42. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238 (2012).
43. English, A. C. et al. Mind the gap: upgrading genomes with PacBio long-read sequencing. Nature 500, 411–414 (2013).
44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
45. Leung, A. K. Y. et al. OMBlasT: alignment tool for optical mapping data. Bioinformatics 33, 2923–2935 (2017).
46. Kent, W. J. BLAT - the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002).
Acknowledgements
This work was supported by Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16), Lo Kwee-Seong Biomedical Research Fund; Lee Hysan Foundation, National Natural Science Foundation of China Major Projects (91125027), and CUHK VC Discretionary Fund (VCF2014004). We also thank the Worldwide University Network (WUN) for support. Professor Silun Zhong and Dr. Yun-Ru Chen of CUHK provided help in the preparation of 250 bp sequencing library. Shigemi Sasamoto of KDRI assisted in preparing DNA samples from Japanese soybean cultivars. We also thank Professor Scott Jackson of University of Georgia and Dr. Carolina A. Contador of CUHK for their constructive suggestions to this manuscript.

Author contributions
H.-M.L. and T.-F.C. coordinated and designed the experiments. F.-L.W., C.Y.-L.C., M.-W.L., S.-W.T., Z.X., K.F., A.F.D., M.-S.N., X.Q. and T.T. contributed to sample preparation and data generation. M.X., C.Y.-L.C., L.Y., T.D., L.H. and L.C. completed the de novo assembly of the W05 genomes. M.X. performed annotation for W05 genome. X.W. constructed the binmap and performed QTL mapping. M.-W.L. and M.X. analyzed the QTLs, genome, and alleles. A.L. and Z.W. performed PCR validation for the QTLs, gene, and alleles. A.L. and Z.W. performed PCR validation for the QTLs, gene, and alleles. A.L., C.Y.-L.C., and H.-M.L. analyzed and refined the scientific contents. M.X., C.Y.-L.C., M.-W.L., C.F., T.-F.C., and H.-M.L. wrote the first draft of this manuscript. M.X., M.-W.L., C.F., and H.-M.L. wrote the final draft.

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09142-9.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal Peer Review Information: Nature Communications thanks David Lightfoot and other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019