G3, 2022, 12(4), jkac017
https://doi.org/10.1093/g3journal/jkac017
Advance Access Publication Date: 21 February 2022
Genome Report

Genome assembly of the JD17 soybean provides a new reference genome for comparative genomics

Xinxin Yi [1], Jing Liu,[1,2] Shengcai Chen,[1,2] Hao Wu,[1] Min Liu,[1] Qing Xu,[1] Lingshan Lei,[1] Seunghee Lee [3], Bao Zhang,[2] Dave Kudrna,[3] Wei Fan,[1,2] Rod A. Wing,[3] Xuelu Wang,[2] Mengchen Zhang,[4] Jianwei Zhang [1,3,*], Chunyan Yang,[4,*] Nansheng Chen[1,5,*]

[1]National Key Laboratory of Crop Genetic Improvement, Center of Integrative Biology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China,
[2]State Key Laboratory of Crop Stress Adaptation and Improvement, Henan University, Kaifeng 475004, Henan, China,
[3]Arizona Genomics Institute and BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA,
[4]Institute of Food and Oil Crops, Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang 050031, Hebei, China,
[5]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

*Corresponding author: National Key Laboratory of Crop Genetic Improvement, Center of Integrative Biology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China. Email: jzhang@mail.hzau.edu.cn; *Corresponding author: Institute of Food and Oil Crops, Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang 050031, Hebei, China. Email: chyyang66@163.com; *Corresponding author: National Key Laboratory of Crop Genetic Improvement, Center of Integrative Biology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China. Email: chenn@qdio.ac.cn

Abstract

Cultivated soybean (*Glycine max*) is an important source for protein and oil. Many elite cultivars with different traits have been developed for different conditions. Each soybean strain has its own genetic diversity, and the availability of more high-quality soybean genomes can enhance comparative genomic analysis for identifying genetic underpinnings for its unique traits. In this study, we constructed a high-quality de novo assembly of an elite soybean cultivar Jidou 17 (JD17) with chromosome contiguity and high accuracy. We annotated 52,840 gene models and reconstructed 74,054 high-quality full-length transcripts. We performed a genome-wide comparative analysis based on the reference genome of JD17 with 3 published soybeans (WM82, ZH13, and W05), which identified 5 large inversions and 2 large translocations specific to JD17, 20,984–46,912 presence–absence variations spanning 13.1–46.9 Mb in size. A total of 1,695,741–3,664,629 SNPs and 446,689–800,489 Indels were identified and annotated between JD17 and them. Symbiotic nitrogen fixation genes were identified and the effects from these variants were further evaluated. It was found that the coding sequences of 9 nitrogen fixation-related genes were greatly affected. The high-quality genome assembly of JD17 can serve as a valuable reference for soybean functional genomics research.

Keywords: soybean cultivar Jidou 17; comparative genomics; symbiotic nitrogen fixation; assembly; genome

Introduction

Soybean (*Glycine max*) is an important crop for protein and dietary oil and is ranked the fourth largest crop in production in the world. The current soybean reference genome, which was based on the Williams 82 (WM82) line (Schmutz *et al.* 2010) has greatly enhanced the identification of genes underlying important traits and facilitated research on the function and expression of soybean genes.

Recent studies using high-throughput sequencing has revealed extensive genetic diversities in soybean (Zhou *et al.* 2015). Pan-genome study on wild and cultivated soybeans has uncovered numerous genetic differences among soybean strains (Liu *et al.* 2020), suggesting that a single reference genome is inadequate for representing the genetic richness of soybean lines.

The latest progress in sequencing technologies has greatly advanced the ability to construct high-quality genome assemblies with chromosome-level continuity with dramatically reduced cost (Risse *et al.* 2015; Deschamps *et al.* 2018). The application of Iso-Seq protocol has enhanced genome annotation (Jiang *et al.* 2017; Jiao *et al.* 2017; Li *et al.* 2018; Magrini *et al.* 2018). Benefiting from improved technologies, more and more reference genomes of various soybean materials have been sequenced and reported, including the Chinese cultivar ZH13 (Shen *et al.* 2019), 3 reference genomes (*G. max* WM82v4, *G. max* Lee, and *Glycine soja* PI 483463) (Valliyodan *et al.* 2019), the wild strain W05 genome (Xie *et al.* 2019), the pan-genome constructed from 26 different soybean species (Liu *et al.* 2020), and the recently published Korean Hwangkeum genome (Kim *et al.* 2021) and 8 soybean genomes (Chu *et al.* 2021).

JD17 is a major soybean cultivar in the Huang-Huai-Hai region of China, and it is also the main soybean variety recognized by the Ministry of Agriculture since 2010. It is the offspring of Hobbit

Received: November 11, 2021. **Accepted:** January 11, 2022

(maternal parent) and Zao 5241 [7476 × 7527-1-1 (Yanli × Williams)] (paternal parent) (Qin *et al.* 2014), and is famous for its lodging resistance, high yield, and strong adaptability (Zhao *et al.* 2013, 2015). The goal of this project was to construct a high-quality genome assembly and provide annotations for JD17. Based on comparative analysis, we aim to obtain JD17-specific genes, with the goal to explain differences of important traits. Based on the analysis of genes for nodulation and symbiotic nitrogen fixation in legume (Roy *et al.* 2020), are there any JD17-specific SNF genes that may dictate JD17-specific phenotypes (such as rhizobia number, morphology, and nitrogen fixation capacity).

In this study, we extracted genome DNA from developing underground tissues of JD17, and used PacBio single-molecule real-time (SMRT) sequencing and Hi-C mapping technologies to construct a high-quality soybean reference genome. After being annotated more completely, the JD17 pseudomolecules were used to identify structural differences in a comparative analysis with 3 published soybean reference genomes. We identified JD17-specific presence–absence variations (PAVs) and a large number of SNPs and Indels, and also resolved the influence of these variants on the coding structure of nitrogen fixation-related genes.

Materials and methods

Plant and sample preparation

Soybean seeds of *G. max* cv. JD17 used in this study were from Hebei Academy of Agricultural and Forestry Sciences. The seeds are planted and extracted in Xuelu Wang's Laboratory in Huazhong Agricultural University. The seeds were sterilized with chlorine gas [5 ml of 32% (w/w) HCl to 100 ml 4–5% (wt/vol) sodium hypochlorite in a beaker] for 15 h (Kereszt *et al.* 2007) and then left in a sterile hood for 2 h. The sterilized seeds were sown in growth bottles filled with sand after being soaked in sterile Milli-Q water for 30 min and watered with sterile Fahraeus solution (Fåhraeus 1957) containing 2 mM KNO_3. Seeds were grown in a growth chamber (light) at 28°C and 8 h (dark) at 23°C with 60% humidity for 16 h.

RNA preparation and sequencing

Underground tissues of inoculation and uninoculation from the 9 timepoints (1, 4, 6, 8, 10, 15, 20, 25, and 30 day post inoculation, dpi) were collected and used for RNA extraction, respectively. After that, 9 RNA samples of inoculation and uninoculation were mixed equally as a sample, respectively. In addition, we also selected different tissues including root, nodule, stem, leaf, pod, seed, and flower for mixed RNA-Seq. All the RNA was extracted by TRIzol reagent (Invitrogen 15596026). we performed RNA-Seq on illumina platform and produced approximately 10 Gb raw data with 150 bp pair-end reads.

Whole-genome sequencing using SMRT technology

A total of 10-dpi root tissue of plants was used for SMRT whole-genome sequencing. Underground tissues were collected for genomic DNA preparation with modified CTAB method (Bergman and Quesneville 2007). Using 97 μg DNA, PacBio sequencing libraries were produced following manufacturers protocols as described for the greater than 30 kb-SMRTbell Libraries Needle Shearing (SMRTbell Template Prep Kit 1.0) with Blue Pippin size selections (Sage Science, http://www.sagescience.com/; Last accessed: 22 January 2022), and the SMRTbell libraries were constructed through Pacific Biosciences SMRTbell Template Prep Kit 1.0 (http://www.pacb.com/; Last accessed: 22 January 2022). SMRT sequencing was performed on a PacBio RSII instrument using P6/C4 sequencing chemistry (DNA/Polymerase Binding Kit P6) and 6 h movies. We used a total of 118 SMRT cells and produced 127.3 Gb of raw data with an average subread length of 15 kb (Supplementary Table 1). At the same time, we also used a part of DNA samples for resequencing on the Illumina HiSeq 2500 platform for 150 bp paired end reads, with a sequencing depth of approximately 45×, for a total of 44.5 Gb. These data are mainly used for the evaluation of genome size and its heterozygosity, post-assembly error correction, and genome quality assessment.

Hi-C library construction and sequencing

For samples used for Hi-C-assisted assembly, leaves fixed in 1% (volume/volume) formaldehyde were used for library construction. Cell lysis, chromatin digestion, proximity ligation treatment, DNA recovery, and subsequent DNA manipulation were performed as previously described (Lieberman-Aiden *et al.* 2009). The restriction enzyme used in chromatin digestion is MboI. Finally, the Hi-C library was sequenced on the Illumina HiSeq X 10 platform for 150 bp paired end reads, with a sequencing depth of approximately 150×, for a total of 152.0 Gb.

De novo genome assembly of JD17

To perform de novo assembly of the JD17, we combined 3 different assemblers, including CANU (Koren *et al.* 2017) (v1.4), FALCON (v0.3.0) (https://github.com/PacificBiosciences/FALCON-integrate; Last accessed: 22 January 2022) and HGAP4 (SMRT Link v 5.0.1.9585) (Supplementary Table 2). The main assembly was performed on whole SMRT sequenced long reads. All assembly softwares were performed with a presumed 1-Gb genome size. If not specified, all programs in our study were run with default parameters. CANU was run with default parameters, and FALCON was run with "length_cutoff = -1" for initial mapping of seed reads for the error-correction phase (Supplementary Fig. 1). For a better FALCON assembly, we additionally optimized parameters as "DBsplit = -x500, –s400, pa_HPCdaligner = -v -B128 -t16 -e.70 -l1000 -s1000 -T8 -M24, ovlp_HPCdaligner = -v -B128 -t32 -h60 -e.96 -l500 -s1000 -T8 -M24 and overlap_filtering = –max_diff 60 –max_cov 60 –min_cov 2." The stats of 3 initial assemblies were show in Supplementary Table 2. We subsequently used the CANU assembly as the working set because it was able to generate more accurate and more contiguous genomes compared to FALCON and CANU. Subsequently, the draft assembly was polished twice using Quiver (SMRT Link v 5.0.1.9585) and finally corrected using about 45× Illumina short reads with Pilon (v1.22) (Supplementary Fig. 2). Then, these contigs were aligned to the NT library using blastn, and those identified as not belonging to plants were filtered out.

Continuation and connection of contigs

Due to the complementarity among 3 different assembly results from CANU (Koren *et al.* 2017), FALCON, and HGAP4, we optimized our assembly results using the GPM (Zhang *et al.* 2016) pipeline to extend and connect contigs for better contiguity. First, GPM loaded WM82 as a reference genome, and CANU assembly as the back-bone contigs. These contigs were ordered and located on chromosome based on WM82 (Glycine_max_v2.0) using blastn. This step produced a draft chromosome assembly, as JD17 v0.1. Second, we loaded FALCON assembly result and aligned with CANU assembly using blastn (Camacho *et al.* 2009), and got the potential overlapping relationships between CANU and FALCON contigs. Then CANU contigs were extended or

connected by FALCON contigs as needed to produce the JD17 v0.2 assembly. Third, we repeated the above steps with HGAP4 assembly and the JD17 v0.3 was generated. Fourth, we loaded contigs from an assembly performed with the longest 70 Gb PacBio reads (extracted from the total 120 Gb sorted raw reads) using CANU. Alignment with the v0.3 version contigs to get the correspondence, then extend and connect contigs from v0.3 version. The optimization contigs is defined as JD17 v0.4. Finally, the contigs of chloroplast and mitochondrial sequences were identified and removed from JD17 v0.4 (Supplementary Table 3). All connection and extension events were validated by aligning subreads against the assembled contigs. Then, the JD17 v0.4 was re-polished using Quiver over twice iterations and corrected using Illumina short reads with Pilon (Walker *et al.* 2014). The finalized JD17 genome assembly (named as Glycine_max_JD17v1.0) was 995.0 Mb in size, with a contig N50 of 18.0 Mb (Table 1). The assembly processing details are shown in Supplementary Fig. 3.

To anchor hybrid contigs into chromosome, the Hi-C sequencing data were aligned into contigs using bwa. According to the orders and orientations provided by the alignment, those contigs were clustered into chromosomes by ALLHiC v0.9.8 (Zhang *et al.* 2019) with recommended params in https://github.com/tangerzhang/ALLHiC/wiki/ALLHiC:-scaffolding-of-a-simple-diploid-genome (Last accessed: 22 January 2022). According to the ALLHiC groups and assembly results create hic files, manual correction and validation were also performed by drawing contact maps with juicerbox (Durand *et al.* 2016). The genome assembly was finalized after this correction step (Supplementary Table 3).

The assessment of genomic heterozygosity and size is using the Genomic Character Estimator program (gce v 1.0.0, ftp://ftp.genomics.org.cn/pub/gce; Last accessed: 22 January 2022), and the heterozygous ratio based on kmer individuals is 0.029, and the corrected estimation of genome size is about 1.11 Gb.

Quality assessment of JD17 genome assembly

To assess the quality of the Glycine_max_JD17v1.0 assembly, we used our 65 Gb resequencing data. First, by aligning all reads to the assembly with BWA-MEM in BWA (v 0.7.17) (Li 2014), the mapping rate is over 98.8% and the coverage was over 99.65%, which shows the consistency between the assembly and reads. By using the GATK tools (v4.1.7.0) (McKenna *et al.* 2010; DePristo *et al.* 2011) for SNPs calling with JD17 resequencing data, we found 78,033 SNP, of which only 8,000 were homozygous, indicating that the JD17 genome has an accuracy of over 99.999%. Merqury is also used to assess the quality of our genome assembly (Rhie *et al.* 2020), results show that the genome assembly error rate is 3.99303 e-05, and the integrity as high as 94.9455%. The completeness of the assembly was estimated by BUSCO with default parameters.

Annotation of TE and ncRNA sequences

To investigate the JD17 genome sequence features, we identified transposable elements (TEs) and other repetitive elements by RepeatMasker (v4.1.0) (Bergman and Quesneville 2007). Miniature inverted transposable elements (MITEs) were collected by MITE-Hunter (Han and Wessler 2010) with all default parameters. In order to get as much reliable long terminal repeat (LTR) retrotransposons information as possible, we used the LTR_retriever (v2.7) (Ou and Jiang 2018) analysis process, which integrates the output of LTR_FINDER (v1.1) (Xu and Wang 2007) and LTRharverst tools in GenomeTools (v1.5.10) (Gremme *et al.* 2013). Masking sequence with RepeatMasker (version 4.0.8) (http://www.repeatmasker.org/; Last accessed: 22 January 2022) based on MITEs and LTR library that has been identified. The other tandem repeats were identified by constructing a de novo repeat library using Repeatmodeler (version 1.0.11) (http://www.repeatmasker.org/; Last accessed: 22 January 2022). RepeatMasker was run against the genome assembly again, with all above library as the query library.

Noncoding RNAs were predicted by the Infernal program (v1.1.4) using default parameters (Nawrocki and Eddy 2013) and comparing the similarity of secondary structure between the JD17 genome sequence and Rfam (Nawrocki *et al.* 2015) (v12.0) database.

Annotation of protein-coding genes

We performed gene calling analysis with Exonerate (v2.2.0) (Slater and Birney 2005), Trinity (v2.10.0) (Grabherr *et al.* 2011), and PASA (v2.4.1) (Haas *et al.* 2003), by using multisourced EST and protein sequences as evidences [including nonredundant soybean EST/Iso-seq (SRX7016448) (Chu *et al.* 2021)/protein sequences from NCBI, assembled JD17 RNA-Seq from mixed samples from underground samples, WM82 transcripts and proteins, *Lotus japonicus* and *Medicago truncatula* protein sequences], AUGUSTUS (v3.4.0) (Hoff and Stanke 2013), and GENEMARK (v4.59) (Brůna *et al.* 2020) as ab initial gene predictors, and a customized repeat library for RepeatMasker (Bergman and Quesneville 2007; Saha *et al.* 2008).

The RNA-Seq data was de novo assembled using Trinity to obtain the assembled cDNA sequence, and annotated with the

Table 1. Assembly statistics of Glycine_max_JD17 (JD17), Glycine_max_v4.0 (WM82), Gmax_ZH13 (ZH13), and W05.

	JD17	WM82	ZH13	W05
Assembly feature				
Estimated genome size (by K-mer analysis) (Mb)	1,109	1,115	—	—
Number of contigs	446	9,200	1,528	1,870
Total size of contigs (Mb)	995.0	952.5	1,007	998.6
Longest contig (Mb)	31.8	—	—	—
Number of contigs > 1 Mb	97	—	—	—
Number of contigs > 10 Mb	39	—	—	—
N50 contig length (Mb)	18.0 (PacBio)	0.4	CANU : 2.9 +Bionno : 18.0 +Hi-C : 22.6	3.3
L50 contig count	21	649	66	58
Anchored contigs				
Number of chromosomes	20	20	20	20
Number of contigs	411	—	—	772
Total size (Mb)	965.8	978.4	1,011.2	1,013.2
Number of gaps	391	7,221	448	750

PASA tool along with the nonredundant isoforms sequence from Iso-Seq. The annotation results will be used for AUGUSTUS model training and prediction. The WM82 protein sequence will be used for GENEMARK model training and prediction. Protein sequences of *Arabidopsis thaliana*, *Medicago sativa*, *Lotus corniculatus*, and 3 different strains of soybean (WM82, ZH13, W05) were used in the Exonerate deprediction protein-coding gene models. In order to obtain more accurate and complete annotation results, EVM (v1.1.1) was called to integrate the gene model prediction results from AUGUSTUS, GENEMARK, Exonerate, and PASA. After all, the PASA software is used to update these annotation results.

Gene functions were inferred according to the best match of the alignments to the National Center for Biotechnology Information (NCBI) Swiss-Prot (Boeckmann *et al.* 2003) protein databases using BLASTP (ncbi blast v2.6.0+) (Altschul *et al.* 1997; Camacho *et al.* 2009) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa *et al.* 2012) with an *E*-value threshold of 1E-5. Gene Ontology (GO) (Ashburner *et al.* 2000) IDs for each gene were obtained from Blast2GO (Conesa and Gotz 2008).

Genome-wide rearrangement and SV detection

To identify large-scale synteny among the 4 soybean lines, we created a genome-wide alignment using the Mauve aligner (February 13, 2015) with the progressiveMauve algorithm (Darling *et al.* 2010) with default parameters: default seed weights, determination of LCBs (minimum weights = default), and full alignment with iterative refinement (Chakraborty *et al.* 2021).

The PAV sequences, PAV clusters, and PAV genes between JD17 and the 3 genomes (WM82, ZH13, and W05) were identified using the sliding window method described in the 2018 publication by Sun et al. (2018), with the same slide window size, alignment software, and parameters as theirs.

We aligned WM82, ZH13, and W05 to JD17 using MUMmer 4.0 (NUCmer -maxmatch) (Marçais *et al.* 2018). MUMmer alignments were processed using SyRI (v1.4) (Synteny and Rearrangement Identifier) (Goel *et al.* 2019) (https://github.com/schneebergerlab/syri commit 3f16e01; Last accessed: 22 January 2022), which identifies syntenic regions, chromosome rearrangement events such as inversion and translocation, and also identified SNPs, Indel, Copygains, and Copylosses between chromosomes. The annotations of JD17 were indexed using the SnpEff tool (v4_3t) for construction, and then SNPs and Indels identified genome-wide were annotated to finally find genes subject to large genetic variation.

Identification and structural variation analysis of SNF genes

Identification of SNF gene sequences based on the literature published by Roy *et al.* in 2020 (Roy *et al.* 2020). These sequences were first aligned to the protein sequences of JD17, WM82, ZH13, and W05 by blastp tool and filtered by coverage and identity greater than 40% to obtain possible nitrogen fixation-related genes. These filtered genes were then subjected to clustering analysis by the OrthoFinder tool (v2.5.4) (Emms and Kelly 2019). The SNF genes in the 4 soybeans were finally identified by determining the gene cluster where the published SNF genes were located. To obtain the expansion and contraction of SNF genes, these single-copy SNF genes were further used to construct an evolutionary tree by RAxML tool (v8.2.11) (Stamatakis 2014) and finally combined with café (De Bie *et al.* 2006) for the analysis of expansion and contraction.

Then the SNPs and Indels loci between JD17 and them were combined to assess the protein encoding-affected SNF genes. These genes (and gene sequences from the article) were then aligned using mafft (Nakamura *et al.* 2018) software, constructing an evolutionary tree based on fasttree (Price *et al.* 2010) tools, and manually checking the evolutionary tree manually to determine that these genes whose structure was affected were homologous (protein-coding genes on the same chromosome; on the same branch of the evolutionary tree, and then expanded if genes of the species were not present on that branch).

Results and discussion

High-quality genome assembly and annotation

Genome assembly using PacBio SMRT data and Hi-C data (detail in Materials and Methods) resulted in a JD17 genome assembly with 411 contigs anchored to 20 pseudomolecules, with a total size of 965.8 Mb and contigs N50 up to 18.0 Mb (Table 1), accounting for 97.1% of the total contigs (Fig. 1). The genome accuracy rate was evaluated as 99.999% by using JD17 resequencing data and GATK tools (McKenna *et al.* 2010; DePristo *et al.* 2011) (see *Materials and Methods*). BUSCO score of the assembly was 93.2% (Simão *et al.* 2015) (Supplementary Table 4), indicating that the completeness of JD17 was higher than that of the WM82, ZH13, and W05 genomes (Shen *et al.* 2018; Xie *et al.* 2019).

In addition to the nuclear genome, we also constructed the full-length genomes of the soybean chloroplast (152.2 kb) and the complete genome of *Bradyrhizobium japonicum* USDA 110 (9.1 Mb), which are consistent with previously reported results (Supplementary Fig. 4) (Kaneko *et al.* 2002; Saski *et al.* 2005).

We identified 52,840 protein-coding genes and 74,054 full-length transcripts in the JD17 genome (Table 2). The average length of mRNA, 5′ UTR, CDS, and 3′ UTR were 4,465, 302, 1,183, and 487 bp, respectively. Our gene annotation is relatively in agreement with the other 3 genomes. After searching with existing databases and conserved structural domains for functional annotation, a total of 72,529 (97.94%) transcripts have known domains or functions, which suggested that high-confidence annotation of JD17 genome was performed.

Transposable elements

We identified 580.2 Mb (58.30% of the JD17 genome) repeat elements in JD17 genome (Fig. 2), which contain 1,138,787 intact TEs, including 734,061 class I RNA retrotransposons and 293,279 class II DNA transposons (Supplementary Table 5).

Among the genomic sequences of JD17, the highest proportion of repetitive sequences is the LTR retrotransposon, and a total of 631,056 LTR elements were identified, totaling 397.58 Mb, which is about 39.95% of the whole genome. Further by assessing the distribution and insertion time of LTRs, we found that the LTR enrichment was all in close proximity to the centromeric region, fewer in the gene enrichment region, and that the amplification of LTR retrotransposons in soybean occurred mainly within the past 1.5 million years and were all relatively young (Supplementary Fig. 5). When comparing the activity of LTR elements (Supplementary Fig. 6), Copia elements appear to be increasingly active and persistent over the past 3 million years. Similarly, Gypsy elements were relatively active during this period. In fact, between 1.4 and 3 million years, their activity levels do not differ much, but since 1.4 million years, the Gypsy element has become less active compared to the Copia element, while the unknown element has also become active over the last 3 million years, but at a much lower level compared to both the Gypsy and

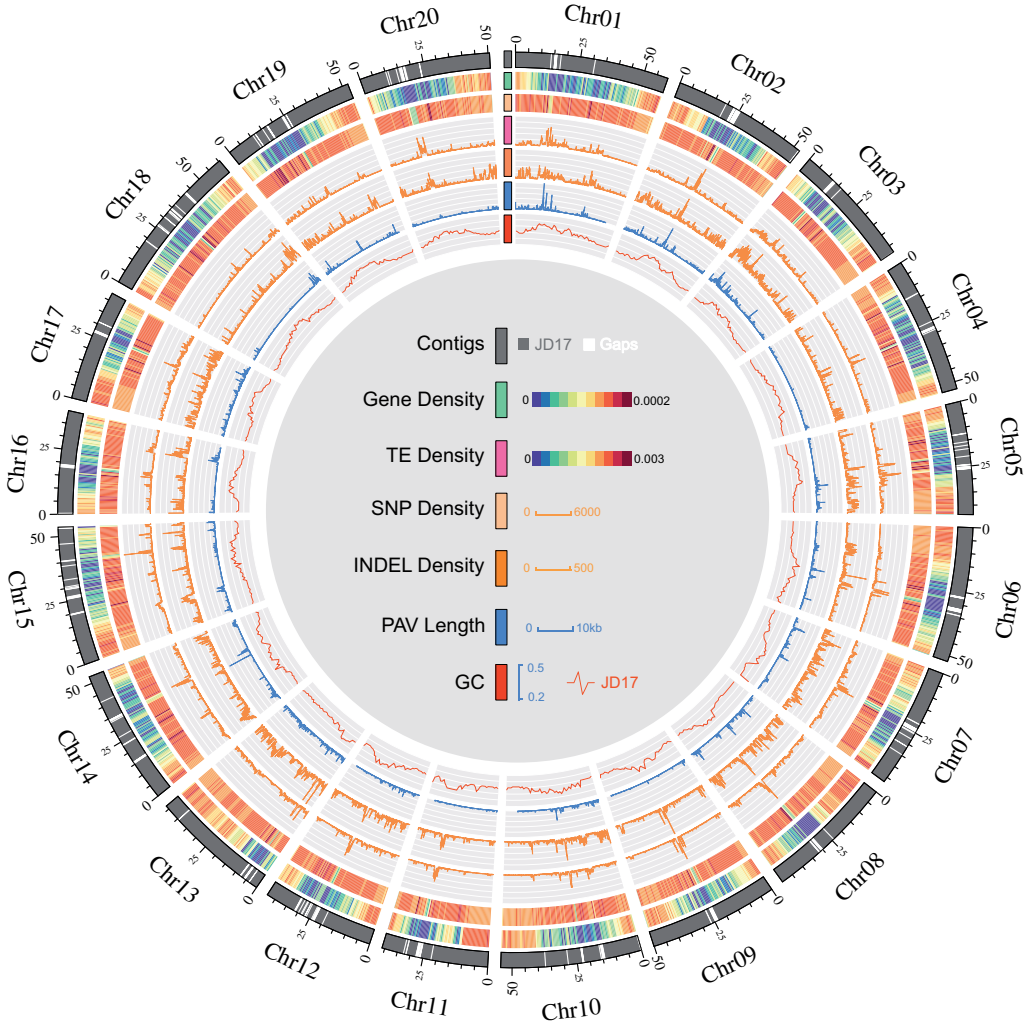

Fig. 1. Overview of the JD17 reference genome. Tracks from outer to inner circles indicate: the chromosome of the genome; the gene density map; the repeat sequence density map; density distribution of SNPs between JD17 and WM82; density distribution of InDel between JD17 and WM82. PAV length distribution between JD17 and WM82; GC content of JD17.

Table 2. Comparison of genome annotation of JD17, WM82, ZH13, and W05.

	JD17	WM82	ZH13	W05
Number of genes	52,840	52,872	55,573	47,201
Number of transcripts	74,054	86,256	96,496	69,277
Average number of transcripts per gene	1.4	1.6	1.7	1.5
Average length of transcript (bp)	4,465	4,889	5,230	5,198
Average exons number per transcript	5.9	6.5	6.5	6.6
Average length of 5′ UTR (bp)	302	294	395	252
Average length of 3′ UTR (bp)	487	448	562	336
Number of single exon mRNA	10,057	12,065	8,466	7,803

Copia elements. Based on the timing of the recent occurrence of WGD in soybean (~13 MYA) (Schmutz *et al.* 2010), it appears that these LTR insertions and expansions are different from gene duplications, but occur mainly after WGD.

To compare the differences in the composition and distribution of repetitive sequences between them, we annotated the repetitive sequences of other 3 genomes using the same approach. The results indicates that the composition of their TEs is essentially the same, except that WM82 has slightly less LTR content (Fig. 2 and Supplementary Table 5 and Figs. 7–9). Meanwhile, we found some blank regions in the genomes of W82 and W05 in the distribution of TEs, both of which are N sequences in the genome sequence. To determine whether these blank regions are the centromeric regions, we identified the centromeric regions of the JD17 genome by using the soybean-specific centromeric satellite repeats sequences (CentGm-1 and CentGm-2) (Gill *et al.* 2009). The evidence suggests that the location of JD17 centromeric regions is consistent with that of the blank regions in WM82 and W05 (Fig. 2 and Supplementary Fig. 10). In general, the assemblies of JD17 and ZH13 are more complete in the centromeric regions, and the TEs of W05 has slightly richer components than the other strains.

Genomic rearrangements

To explore how many and how large-scale discordant regions of genetic mapping exist between the JD17 reference genome and the 3 published genomes, we performed a genome-wide comparative analysis. When the pseudochromosomes of JD17 were aligned to the other pseudochromosomes of WM82, ZH13, and W05, a total of 3,727 syntenic blocks were identified

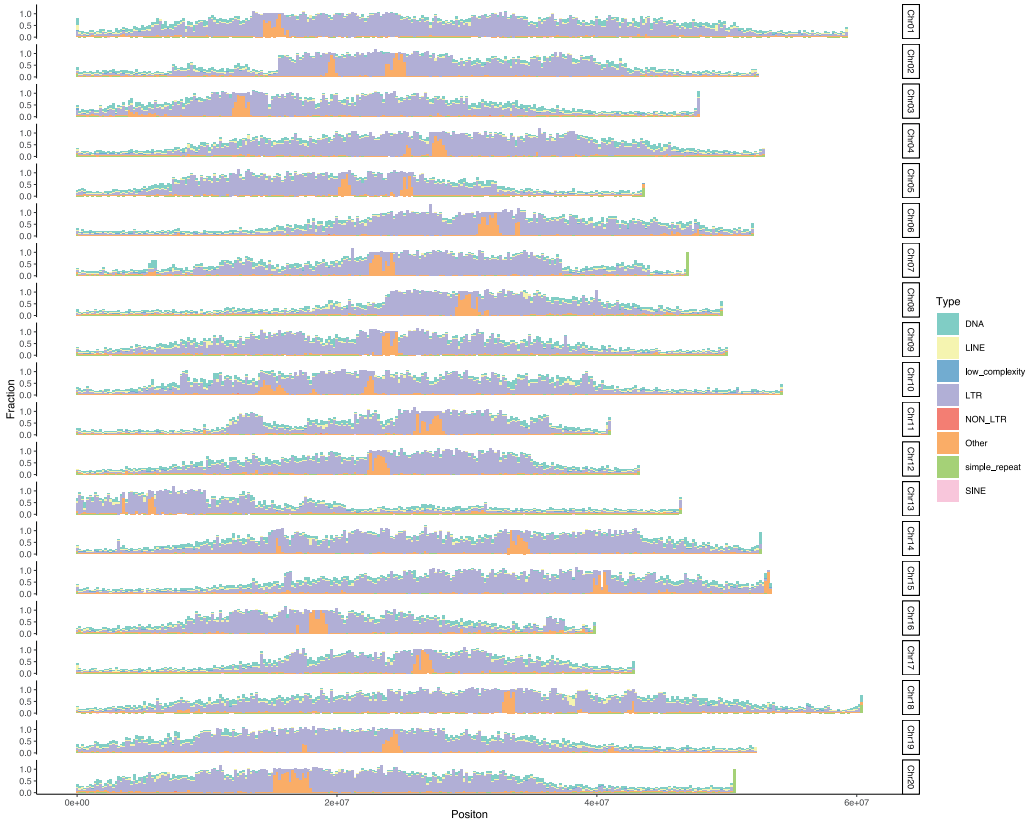

Fig. 2. Chromosome distribution map and percentage of different types of TEs. Chromosomes were split into 200 kb bins without overlap, and the percentage of major types of TE elements (include DNA, DNA transposons; LINEs, long interspersed nuclear elements, low complexity; LTR, long terminal repeat retrotransposons; NON LTR, nonlong terminal repeat retrotransposons; other, simple repeat; SINE, short interspersed nuclear elements) in each bin was counted.

(Supplementary Fig. 11), of which 1,368 were common syntenic blocks among the 4 strains (Supplementary Table 6). Approximately 88.51% of the JD17 genome sequence shares the syntenic blocks with 88.12% of WM82, 86.10% of ZH13, and 81.88% of W05. We found a total of 251 syntenic blocks that exist between Wm82, ZH13, and W05, but not in JD17, with a size between 3.87–5.27 Mb. Similarly, there are 113 such syntenic blocks between JD17, WM82, and ZH13, but not in W05, and the size is between 2.10 and 3.33 Mb. The size of the specific sequence in each of the 4 strains (i.e. the sequence that does not match any other strains) is between 82.58 and 119.37 Mb, of which the wild-type W05 has the most specific sequences and the W82 has the least. This may be due to the loss of sequences in cultivars during the long domestication process of artificial selection.

The chromosomal differences between JD17 and the other 3 strains were further observed by comparative genome analysis. We observe that compared to WM82 and ZH13, JD17 has slightly more Copygains, 4.0 Mb and 3.8 Mb respectively, but W05 has more Copygains compared to JD17, with a combined total of about 7.1 Mb (Supplementary Table 7). We also found 56 inversions and 1,051 translocations events for the rearrangement events that occurred between JD17 and WM82 (Supplementary Table 7). Similarly, 86 and 1,273 inversions and translocations occurred between JD17 and ZH13, and 93 and 2,996 inversions and translocations occurred between JD17 and W05, respectively. We observed 5 large inversions (>1 Mb) on Chr04, Chr05, Chr06, Chr07, and Chr19 and 2 large translocation events (>200 kb) on Chr02 and Chr07 (Supplementary Figs. 12–14 and Table 8), respectively, that were present in JD17 compared to all other 3 species. Examination of the breakpoint loci of these variants by comparing the sequenced subreads to the genome showed that the assembly of JD17 was correct (Supplementary Figs. 15–21). Although the breakpoint at JD17 (Chr06:32,732,757) is an N sequence (i.e. gaps) and the subreads at Chr07:27,723,877 are poorly supported, the corresponding breakpoint positions in the WM82 genome are both gaps. This suggests that these structural differences such as inversions and translocations are more likely to be true genetic variation in JD17 relative to the other 3 strains.

Identified PAVs by comparison with WM82, ZH13, and W05

Comparison of the genomes of JD17 and WM82 revealed 20,984 JD17-specific fragments (total length: 13.13 Mb) and 22,635 WM82-specific genomic fragments (total length: 30.81 Mb) (Supplementary Table 9). Similarly, comparing the genomes of JD17 and ZH13 revealed 22,818 JD17-specific genomic fragments (13.68 Mb) and 23,456 ZH13-specific genomic fragments (33.27 Mb). Comparison of the JD17 and W05 genomes also revealed 36,658 JD17-specific genomic fragments (21.86 Mb) and 37,443 W05-specific genomic fragments (46.92 Mb). We further merged PAV sequences within 100 kb from the physical coordinates to identify PAV clusters (Sun *et al.* 2018) (Supplementary Table 10). The majority of these PAV sequences (99.8%) were shorter than 5 kb. Compared with Wm82, 23 PAV sequences longer than 5 kb in JD17 were identified with an average length of 7,195 bp, standard deviation is 1,789.74. These PAV sequences were unevenly distributed in the genome (Fig. 1), with some located in clusters. The largest of these PAV cluster fragments is a

1.6-Mb JD17-specific fragment containing 5 predicted genes located between 14.8 and 16.4 Mb on chromosome 1. However, there is a PAV cluster of less than 0.7 Mb that is very rich in genes, up to 73 coding genes, located between 45.3 and 46.0 Mb on chromosome 8. After GO enrichment analysis, they were found to be associated with NADH dehydrogenase, aspartate-type endopeptidase activity, histidine, protein, and sugar de metabolism (Supplementary Fig. 22).

Based on the criterion that a gene can be designated as a PAV gene if its coding sequence is covered by ≥75% of the PAV sequence (Sun *et al.* 2018), we compared JD17 with WM82, respectively, and we identified 100 JD17-specific and 75 WM82-specific PAV genes. Similarly, by comparing JD17 with ZH13, we identified a total of 80 JD17-specific and 156 ZH13-specific PAV genes, and between JD17 and W05, 101 JD17-specific and 67 W05-specific PAV genes were identified. It can be seen that there are large differences in genomic sequences between JD17 and WM82, ZH13, and W05.

By comparing these JD17-specific genes, it was found that only 20 JD17-specific genes were identified simultaneously (Supplementary Fig. 23), and most of the JD17-specific genes could still be found in different strains. The functional enrichment analysis of these 20 genes showed that these genes were related to signal transduction, response to stimulus, protein synthesis, and nitrogen metabolism. The results of functional annotation also showed that these genes were associated with powdery mildew and TMV resistance protein synthesis (Supplementary Fig. 24).

Identified SNPs and Indels by comparison with WM82, ZH13, and W05

To find SNPs and Indel between JD17 and the other 3 strains, we combined MUMmer and SyRI tools. The results showed that we identified 1,695,741 (2,675,463 in ZH13 and 3,664,629 in W05) SNP, 213,509 (324,109 in ZH13 and 391,977 in W05) insertions, and 233,180 (391,977 in ZH13 and 408,512 in W05) deletions variations in JD17 genome by comparative genome-wide analysis. Combined with the results of structural variation and PAVs, both showed the smallest difference between JD17 and WM82 and the largest difference with wild-type soybean W05. This may be due to the fact that JD17 has the lineage of WM82, but it also supported that wild soybean can provide richer genetic diversity.

Annotation of these variants showed that they mainly affect intergenic regions, but there are still 4,947 variants between JD17 and WM82 that have a very large impact on 1,785 protein-coding genes. Enrichment analysis of these affected genes revealed that the primary function of these genes is associated with sulfate transmembrane transport, regulation of organ growth, regulation of developmental growth, carbohydrate binding (Supplementary Fig. 25).

Similarly, a large effect of 7,363 variants on 2,567 protein-coding genes was observed between JD17 and ZH13, and a large effect of 6,674 variants on 3,611 protein-coding genes was observed between JD17 and W05.

Identification of SNF genes and differences among JD17, WM82, ZH13, and W05

To identify genes associated with SNF in the 4 genomes, we blast identified genes from published articles to these genomes. We identified 331, 360, 285, and 346 genes potentially related to nitrogen fixation in JD17, WM82, ZH13, and W05 genomes, respectively (Supplementary Table 11). Of these, 88.8% of the SNF genes were identified in all 4 soybeans, and in JD17, 96% of the genes were identified, while fewer genes were identified in ZH13 (Supplementary Fig. 26). Subsequently, by comparative analysis of SNF genes from these 4 species, there are 2 expansion genes (*LjCLE-RS2* and *MtNAC969*), 2 contraction genes (*LjCLC1* and *MtCP6*), and 5 lost genes (*GmENOD93*, *GmRj2_GmRFG1*, *LjENOD40-1/ENOD40-2*, and *LjHIP*, *MtCAS31*) (see Supplementary Table 12 for details) in JD17 relative to their common ancestor (Supplementary Table 12). By examining the structural variation of 331 genes potentially associated with nitrogen fixation in JD17, we finally found 18 genes with structural variation between WM82, ZH13, or W05 and affecting protein coding (Supplementary Table 13). Further by constructing an evolutionary tree to manually confirm the effect of these genes undergoing structural variation on them, we finally observed that 9 of them were not in the same branch of the evolutionary tree as the published sequence of WM82. For example, the *LjIGN1* genes (Glyma.10G291900 and Glyma.20G241200), which encodes the synthetic Ankyrin-Repeat Membrane Protein, which is important for symbiosis and nodulation (Kumagai *et al.* 2007). By evolutionary analysis of their homologous genes, we found that Glyma.10G291900 and Glyma.20G241200 are located in 2 branches, and the branch where Glyma.10G291900 is located has corresponding genes in all 4 species with essentially the same protein code (Supplementary Fig. 27). However, the branch of Glyma.20G241200 showed differences. By multiple sequence comparison, JD020G0232500 in fact encodes 192 amino acids that are very different from several other proteins (Supplementary Fig. 28).

These sequence differences may affect its expression specificity, recognition, and interaction, which may in turn affect the recognition, infection, and symbiotic nitrogen of rhizobia. This may be one of the reasons for the difference in nodulation phenotype between JD17 and WM82, which requires further research to verify.

Conclusions

Here, we constructed a high-quality de novo assembly of the soybean cultivar Jidou 17 (JD17) with contigs N50 approaching 18 Mb. Combined use of homologous, de novo, and transcriptome evidence, 52,840 gene models, and 74,054 full-length transcriptomes were predicted. The total number of repeats was about 580.22 Mb, accounting for 58.30% of the whole genome, of which LTRs were the most abundant, accounting for 39.95%. The centromeric regions in the JD17 reference genome were identified. Analysis of LTR insertion time showed that LTR insertion was more and more active, and the 2 main LTR elements, Copia and Gypsy, showed differences. LTR insertions and expansions are different from gene duplications, but occur mainly after WGD. Through whole genome comparison between 4 genomes, identification of a large number of JD17 specific sequences, variants, and genes, including 5 large insertions, 2 large translocations, and 20 PAV genes. These genes were related to signal transduction, response to stimulus, protein synthesis, and nitrogen metabolism. At the same time, the SNPS and INDELs between the JD17 relative to other 3 genome are identified, respectively. The protein-coding structures of 1,785 genes were found to be affected by the variations. Finally, we identified the SNF gene in JD17 and assessed their genetic differences, the results found that protein-coding structures of 9 SNF genes were significantly affected. We hope that our dataset will provide a valuable resource for comparative genomics and functional genomics of soybean.

Data availability

The sequencing data (PacBio Whole Genome Sequencing data for assembly, resequencing data, Hi-C data, and Mixed-Sample RNA-seq data for Annotation) used in this study have been deposited into National Center for Biotechnology Information under BioProject Number PRJNA412346 with accession number SRR12416523—SRR12416632, SRR12416444, SRR12416750 and SRR9643849—SRR9643851. The final assembly had been deposited at GenBank JACKXA000000000. Supplemental material is available at figshare: https://doi.org/10.25387/g3.17065499.

Acknowledgments

NC, JZ, and XW coordinated the research. XY, SC, HW, LL, BA, and WF prepared the DNA and RNA samples for sequencing. SL, DK, and RW performed PacBio sequencing. XY, JL, HW, ML, QU, and LL performed assembly analysis and annotation of the genome and transcriptome. CY and MZ provided the plant seed from Hebei Academy of Agricultural and Forestry Sciences. XY wrote the manuscript with input from all authors. All authors have read and approved the final manuscript.

Funding

The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. All costs of this work was supported by grant 2016YFD0100700 of the National Key Research and Development Program of China (to BZ) and 2015CB910200 of the National Key Basic Research Foundation of China (to XW), grants 31870257, 91535104, and 31430046 of the National Natural Science Foundation of China (to XW), and grant 31471522 of the National Natural Science Foundation of China (to MZ). The Start-up Fund of Huazhong Agricultural University (HZAU) to J.Z., and Bioinformatics Computing Platform of National Key Laboratory of Crop Genetic Improvement, HZAU.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–29.

Bergman CM, Quesneville H. Discovering and detecting transposable elements in genome sequences. Brief Bioinform. 2007;8(6): 382–392.

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31(1):365–370.

Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2020;2(2):lqaa026.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform. 2009;10:421.

Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. Evolution of genome structure in the *Drosophila simulans* species complex. Genome Res. 2021;31(3):380–396.

Chu JS-C, Peng B, Tang K, Yi X, Zhou H, Wang H, Li G, Leng J, Chen N, Feng X. Eight soybean reference genome resources from varying latitudes and agronomic traits. Sci Data. 2021;8(1):164.

Conesa A, Gotz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:1.

Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.

De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006; 22(10):1269–1271.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–498.

Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin H. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun. 2018; 9(1):4844.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3(1):99–101.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.

Fåhraeus GJM. The infection of clover root hairs by nodule bacteria studied by a simple glass slide technique. 1957;16(2):374–381.

Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, Stacey G, Jackson SA. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 2009;151(3):1167–1174.

Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20(1):277.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–652.

Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(3): 645–656.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Jr, *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–5666.

Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38(22):e199.

Hoff KJ, Stanke M. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 2013;41(Web Server issue):W123–W128.

Jiang X, Hall AB, Biedler JK, Tu Z. Single molecule RNA sequencing uncovers trans-splicing and improves annotations in *Anopheles stephensi*. Insect Mol Biol. 2017;26(3):298–307.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, *et al.* Improved maize reference genome with single-molecule technologies. Nature. 2017;546(7659): 524–527.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40(Database issue):D109–D114.

Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S, Watanabe A, Idesawa K, Iriguchi M, Kawashima K, *et al.* Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. DNA Res. 2002;9(6): 189–197.

Kereszt A, Li D, Indrasumunar A, Nguyen CDT, Nontachaiyapoom S, Kinkema M, Gresshoff PM. Agrobacterium rhizogenes-mediated transformation of soybean to study root biology. Nat Protoc. 2007;2(4):948–952.

Kim MS, Lee T, Baek J, Kim JH, Kim C, Jeong SC. Genome assembly of the popular Korean soybean cultivar Hwangkeum. G3 (Bethesda). 2021;11(10):jkab272.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5): 722–736.

Kumagai H, Hakoyama T, Umehara Y, Sato S, Kaneko T, Tabata S, Kouchi H. A novel ankyrin-repeat membrane protein, IGN1, is required for persistence of nitrogen-fixing symbiosis in root nodules of *Lotus japonicus*. Plant Physiol. 2007;143(3):1293–1305.

Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–2851.

Li Y, Wei W, Feng J, Luo H, Pi M, Liu Z, Kang C. Genome re-annotation of the wild strawberry *Fragaria vesca* using extensive Illumina- and SMRT-based RNA-seq datasets. DNA Res. 2018;25:61–70.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950): 289–293.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, *et al.* Pan-genome of wild and cultivated soybeans. Cell. 2020; 182(1):162–176.e13.

Magrini V, Gao X, Rosa BA, McGrath S, Zhang X, Hallsworth-Pepin K, Martin J, Hawdon J, Wilson RK, Mitreva M, *et al.* Improving eukaryotic genome annotation using single molecule mRNA sequencing. BMC Genomics. 2018;19(1):172.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14(1):e1005944.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9): 1297–1303.

Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34(14):2490–2492.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, *et al.* Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015; 43(Database issue):D130–D137.

Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–2935.

Ou S, Jiang NJPp. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176(2):1410–1422.

Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.

Qin J, Wang F, Gu F, Wang J, Chen Q, Zhang M. A genetic composition analysis of soybean sibling varieties Jidou17 and Ji nf58. Aust J Crop Sci. 2014;8(5):8.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):245.

Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, Watson M. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience. 2015;4(1):60.

Roy S, Liu W, Nandety RS, Crook A, Mysore KS, Pislariu CI, Frugoli J, Dickstein R, Udvardi MK. Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation. Plant Cell. 2020;32(1):15–41.

Saha S, Bridges S, Magbanua ZV, Peterson DG. Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res. 2008; 36(7):2284–2294.

Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, Jansen RK. Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol Biol. 2005;59(2):309–322.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, *et al.* Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178–183.

Shen Y, Du H, Liu Y, Ni L, Wang Z, Liang C, Tian Z. Update soybean Zhonghuang 13 genome to a golden reference. Sci China Life Sci. 2019;62(9):1257–1260.

Shen Y, Liu J, Geng H, Zhang J, Liu Y, Zhang H, Xing S, Du J, Ma S, Tian Z, *et al.* De novo assembly of a Chinese soybean genome. Sci China Life Sci. 2018;61(8):871–884.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19): 3210–3212.

Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9): 1312–1313.

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, *et al.* Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 2018;50(9):1289–1295.

Valliyodan B, Cannon SB, Bayer PE, Shu S, Brown AV, Ren L, Jenkins J, Chung CY-L, Chan T-F, Daum CG, *et al.* Construction and comparison of three reference-quality genome assemblies for soybean. Plant J. 2019;100(5):1066–1082.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.

Xie M, Chung CY-L, Li M-W, Wong F-L, Wang X, Liu A, Wang Z, Leung AK-Y, Wong T-H, Tong S-W, *et al.* A reference-grade wild soybean genome. Nat Commun. 2019;10(1):1216.

Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007; 35(Web Server issue):W265–W268.

Zhang J, Kudrna D, Mu T, Li W, Copetti D, Yu Y, Goicoechea JL, Lei Y, Wing RA. Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. Bioinformatics. 2016;32(20):3058–3064.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019;5(8):833–845.

Zhao Q, Yan L, Liu B, Di R, Shi X, Zhao S, Zhang M, Yang C. Breeding of high-yield widespread and high-quality soybean cultivar Jidou 17. Soybean Sci. 2015;34(4):000736–000739.

Zhao S, Zhao X, Tang X, Zhang J, Xu Y, Feng Y, Zhang M. High yield characteristics of summer sowing soybean varieties. Soybean Sci. 2013;168–175.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33(4):408–414.

Communicating editor: R. Dawe