



OPEN

DATA DESCRIPTOR

Eight soybean reference genome resources from varying latitudes and agronomic traits

Jeffrey Shih-Chieh Chu^{1,2}, Bo Peng², Kuanqiang Tang¹, Xingxing Yi^{2,5}, Huangkai Zhou¹, Huan Wang², Guang Li¹, Jiantian Leng¹, Nansheng Chen^{3,4,6}✉ & Xianzhong Feng^{1,6}✉

Comparative analysis of multiple reference genomes representing diverse genetic backgrounds is critical for understanding the role of key alleles important in domestication and genetic breeding of important crops such as soybean. To enrich the genetic resources for soybean, we describe the generation, technical assessment, and preliminary genomic variation analysis of eight *de novo* reference-grade soybean genome assemblies from wild and cultivated accessions. These resources represent soybeans cultured at different latitudes and exhibiting different agronomical traits. Of these eight soybeans, five are from new accessions that have not been sequenced before. We demonstrate the usage of these genomes to identify small and large genomic variations affecting known genes as well as screening for genic PAV regions for identifying candidates for further functional studies.

Background & Summary

Soybean is an important crop that is responsible for about 50% of the world's oilseed production (www.fao.org), and a source of high-quality protein for animal feeds. Cultivation of soybean has experienced specific selections for the last 4,000 years, yielding over 45,000 *Glycine max* (*G. max*) accessions¹. The construction and subsequent targeted improvement of the reference genome assembly for *G. max* Williams 82 (W82), has greatly enhanced genome contiguity and has in turn promoted research on soybean^{2,3}. However, recent genomic advances in many plant and animal species have shown that a single reference genome is insufficient to capture and represent the variations that exist within the population of the corresponding species^{4–20}. Such is also true in soybeans. Resequencing and *de novo* assembly of wild and cultivated soybean accessions identified hundreds of genic presence-absence variations (PAVs) or SVs affecting agronomic genes compared to W82 genome^{7,21–26}. The recent construction of the soybean pan-genome also discovered over 120 K non-redundant SVs²⁷. Thus, we endeavor to enhance the genomic resources available for furthering soybean functional research and breeding by constructing reference-grade genome assemblies from cultivars found in different latitudes and exhibiting different agronomic traits. In this data descriptor, we report the sequencing, genome assembly, annotation, and genomic variation resources generated for the assembly of 8 reference-grade soybean genomes.

Methods

Sample selection, collection, and nucleic acid extraction. Eight soybean accessions were selected for this project (Table 1). Aside from IGA1003, which is a wild soybean, the remaining seven were cultivated soybeans. These eight soybeans showed different phenotypic traits including flowering color, pubescence color, maturity time, seed shape, and hundred-grain weight (Fig. 1a, Supplementary Table 1). The eight soybeans were typically grown in different latitudes with IGA1007 and IGA1005 typically grown in southern China (around 22°N) and Huanghuai region (around 31°N), China, respectively (Fig. 1b). In terms of traits, we have included a high salinity tolerant accession (IGA1001) and a high yielding accession (IGA1004). We do note that IGA1003, IGA1005, and IGA1008 were sequenced and assembled previously. IGA1003 was sequenced and assembled

¹Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. ²Wuhan Frasiergen Bioinformatics Inc., East Lake High-Tech Zone, Wuhan, China. ³Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China. ⁴Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, Canada. ⁵College of Life Science and Technology, Huazhong Agricultural University, Wuhan, China. ⁶These authors jointly supervised this work: Nansheng Chen, Xianzhong Feng. ✉e-mail: chenn@qdio.ac.cn; fengxianzhong@iga.ac.cn

Accession ID	Accession Name	Seed source
IGA1001	Wenfeng 7	National Soybean Preservation Center, ICS, CAAS
IGA1002	Hefeng 25	Jiamusi Branch of Heilongjiang Academy of Agricultural Sciences
IGA1003	GsojaF	National Wild Soybean Preservation Center, ICS, CAAS
IGA1004	Zhonghuang 35	National Soybean Improvement Sub-Center, ICS, CAAS
IGA1005	Zhonghuang 13	National Soybean Improvement Sub-Center, ICS, CAAS
IGA1006	Jingyuan	National Soybean Preservation Center, ICS, CAAS
IGA1007	Huaxia 3	South China Agricultural University
IGA1008	Williams 82	Northeast Institute of Geography and Agroecology, CAS

Table 1. Selected 8 soybean accessions. Abbreviations: ICS: Institute of Crop Sciences; CAAS: Chinese Academy of Agricultural Sciences; CAS: Chinese Academy of Sciences.

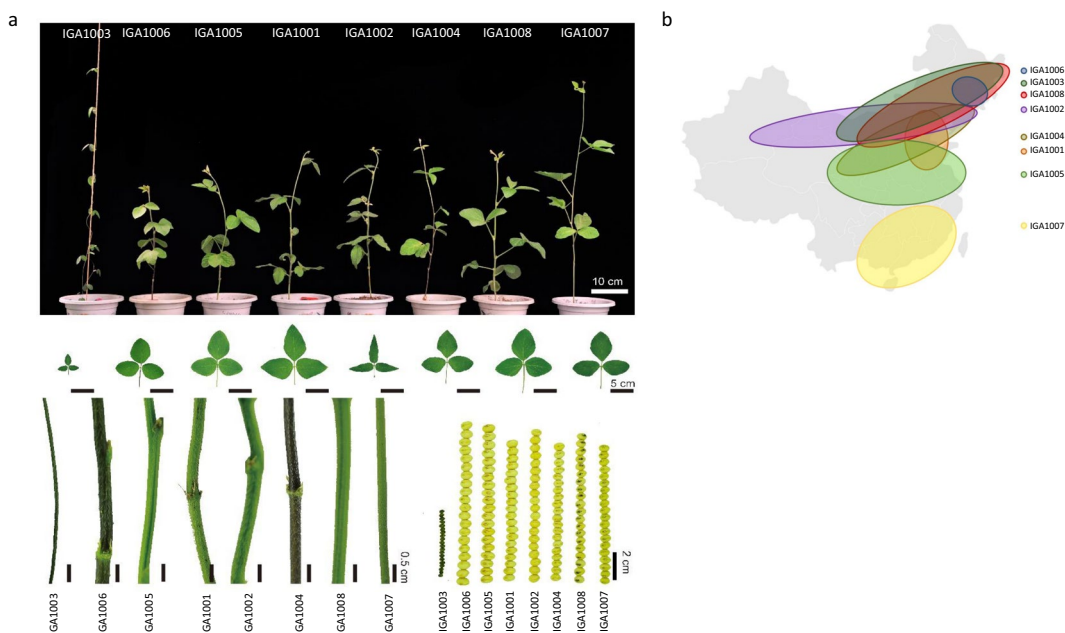


Fig. 1 Phenotypic variation and geographical distribution of the 8 soybeans. (a) 8 soybean plants after growing 35 days showing differences in growth, leaf shape, stem thickness, and seed color and size; (b) Geographical distribution of the 8 soybeans.

using only Illumina sequencing data⁷. IGA1005²⁸ and IGA1008³ had published genome assemblies but we have included these materials allowing for assessing genomic variations within a particular accession.

Seeds from each accession were grown in the greenhouse at Northeast Institute of Geography and Agroecology. Five seeds were planted with 14 hours of light and temperature controlled at 28 °C during the day time and 20 °C during the night time. Soil composition was 8:5:3 ratio of peat:vermiculite:perlite mixed with 5 g of phosphate. Plants were grown until the V2 growth phase and the leaves were harvested from the top.

Genomic DNA was extracted from 1 g of leaf tissue harvested from V2 growth period using CTAB method. Quality of the genomic DNA was assessed using the Qubit Fluorometric system (ThermoFisher) and gel electrophoresis system. Total RNA was isolated from root, stem and leaf separately using TRIzol. Quality of the RNA was assessed using Qubit Fluorometric system and Agilent 2100 Bioanalyzer (Agilent). A quality of RIN >7 was considered high quality RNA.

DNA sequencing library construction and sequencing. DNA fragments >20 kb was selected for PacBio library preparation using BluePippin (SAGE). PacBio library was prepared using SMRTbell Template Prep Kit-SPv3 following manufacturer's recommendations (Pacific Biosciences). The library QC was performed using Qubit and Agilent 2100. The final library was sequenced on the Pacific Biosciences Sequel system. For Illumina sequencing library construction, the genomic DNA was fragmented to 300–500 bp using Covaris M220 (Covaris). Illumina library was prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina following manufacturer's recommendations (NEB) and sequenced on the Illumina system with PE150 format. We generated 4–6.74 million PacBio reads^{29–36} with average subread length N50 of 14Kb and 296–368 million Illumina short reads (Table 2)^{37–44}. The estimated sequencing depth based on soybean genome size was 50X Illumina short reads and 50X of PacBio long reads.

Accession ID	PacBio raw data (G)	PacBio number of subreads (M)	Subreads length N50 (Kb)	Illumina DNA raw data (G)	Illumina RNA raw data (G)	Hi-C library raw data (G)
IGA1001	48.51	6.74	12.52	43.2	11.37	49.80
IGA1002	51.14	6.37	12.68	53.6	10.44	57.81
IGA1003	51.97	6.54	12.27	52.3	11.54	46.43
IGA1004	48.01	5.50	12.53	47.8	13.72	54.67
IGA1005	50.10	4.50	16.56	57.6	14.62	57.14
IGA1006	58.04	6.13	13.84	48.9	12.91	51.05
IGA1007	49.69	4.01	18.56	51.6	11.88	52.82
IGA1008	53.64	6.74	12.47	46.8	12.77	54.45

Table 2. Sequencing output of each library type for the 8 soybean accessions.

	Assembly Length (Mb)	Number of contigs	Contig N50 (Mb)	Number of Scaffolds	Scaffold N50 (Mb)	Number of Gaps	BUSCO %	Het SNPs (%)	Hom SNPs (%)	NCBI Accession
IGA1001	996	2623	1.74	334	50.7	2289	96.9	0.0055	0.0005	WIWY00000000
IGA1002	987	1968	2.92	249	49.51	1719	97	0.0051	0.0003	WIWX00000000
IGA1003	975	2406	1.55	320	48.77	2086	97.2	0.0088	0.0005	WIXD00000000
IGA1004	1001	3382	1.41	460	50.49	2922	97	0.0213	0.0005	WIWZ00000000
IGA1005	988	1169	4.65	181	50.28	988	97.1	0.0037	0.0002	WIXA00000000
IGA1006	995	1420	4.27	332	50.57	1088	96.7	0.004	0.0003	WIXB00000000
IGA1007	986	1038	6.16	208	50.07	830	96.7	0.0039	0.0002	WIXC00000000
IGA1008	993	2339	1.88	397	49.8	1942	97.3	0.0061	0.0005	WIWW00000000

Table 3. Genome assembly, BUSCO evaluation and remapping assessment of the 8 soybean accessions. Abbreviations: Het: Heterozygous; Hom: Homozygous.

RNA sequencing library construction and sequencing. RNA from each tissue was pooled in equal molar to produce a mixed RNA sample. mRNA was enriched using Poly(A) mRNA Magnetic Isolation kit (NEB) followed by RNA-seq library construction using NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB) following manufacturer's recommendations. Sequencing was performed on the Illumina system with PE150 format to generate 10.43–14.61 Gb of data (Table 2)^{45–52}.

Hi-C sample preparation, library construction and sequencing. About 1.5 g of young leaves were used for Hi-C library construction as described in previous reports with some modifications⁵³. Briefly, leaf samples were cross-linked with 3% formaldehyde for 45 minutes in vacuum at 4 °C and stopped using 0.4 M glycine. Leaf pellets were then pulverized in liquid nitrogen followed by resuspension in the nuclei isolation buffer (NIB). The cross-linked nuclei were treated with 0.3% SDS and neutralized with 3% Triton X-100. The resulting DNA was digested with MboI (NEB) overnight at 37 °C and the reaction was stopped with heat inactivation at 65 °C. Restriction fragment ends were fixed with Klenow and labeled with biotinylated cytosine nucleotides using biotin-14-dCTP (TriLINK). Blunt-end ligation was carried out using T4 DNA ligase incubated at 16 °C overnight. After ligation, the cross-linking was reversed by proteinase K (Thermo) overnight at 65 °C. DNA purification was performed using DNeasy Plant Mini Kit (Qiagen) according to manufacturer's instructions. Purified DNA was sheared to a length of ~400 bp using Covaris M220 (Covaris). Hi-C ligated junctions were captured by Dynabeads MyOne Streptavidin C1 (ThermoFisher) according to manufacturer's instructions. The Hi-C sequencing library was prepared using NEBNext Ultra II DNA library Prep Kit for Illumina (NEB) following manufacturer's instructions. Fragments between 400 and 600 bp were sequenced on the Illumina platform with PE150 format to generate 46–57 G of data (Table 2)^{54–61}.

De novo genome assembly. *De novo* assembly was performed with PacBio sequencing data using FALCON⁶² setting length_cutoff = 3000–13000 and length_cutoff_pr = 3000 and CANU⁶³ setting correctedErrorRate = 0.039–0.04. We further improved the assembly by merging complementing contigs between FALCON and CANU using CANU assembly as the basis⁶⁴. The final contigs were assembled into chromosomes with Hi-C data using LACHESIS v.c23474⁶⁵. The assembly was corrected with PacBio long reads using Arrow in SMRTLink 5.0⁶⁶ and Illumina short reads using Pilon v1.22⁶⁷. The final genome assemblies ranged between 986.1 Mb and 1001.3 Mb with more than 98.7% of the genome anchored to 20 chromosomes. The contig N50 were between 1.4 Mb–6.1 Mb and the scaffold N50 were between 48.8 Mb–50.7 Mb (Table 3)^{68–75}.

Genome annotation. Repetitive elements were first annotated before other features. Tandem Repeat Finder v.4.09⁷⁶ was used to annotate tandem repeats. LTR_FINDER⁷⁷ was used to build a LTR-retrotransposon library and RepeatModeler v.1.0.10 (<http://www.repeatmasker.org/RepeatModeler.html>) was used to build a de novo repetitive element library. The above libraries and Repbase⁷⁸ were used by RepeatMasker⁷⁹ to annotate repetitive elements. About 50% of each genome was annotated with repetitive elements (Table 4), which is similar to published soybean genomes^{3,22,28}. After repetitive sequences were masked, protein-coding gene annotation was

Accession		IGA1001	IGA1002	IGA1003	IGA1004	IGA1005	IGA1006	IGA1007	IGA1008
Protein coding genes		57505	58102	57961	58150	57474	58392	57396	57286
Complete BUSCO %		95.3	95.1	96.9	95.4	95.2	93.4	95.5	95.9
Non-coding genes	miRNA	279	282	283	286	277	281	279	280
	tRNA	1028	1163	1122	1028	1077	1100	1060	999
	rRNA	246	383	233	139	336	315	136	202
	snRNA	2645	2609	2643	2709	2612	2609	2605	2617
Repetitive elements	LTR-Retro-transposons (%)	39.74	39.2	38.78	39.39	39.84	40.05	39.55	39.6
	LINE (%)	1.93	1.95	1.87	1.84	1.87	1.8	1.81	1.91
	SINE (%)	0.07	0.05	0.08	0.1	0.06	0.04	0.04	0.07
	DNA	7.11	7.24	7.01	7.37	7.03	7.02	7.17	7.3
	Transposons (%)								
	Satellites (%)	0.12	0.11	0.35	0.1	0.11	0.32	0.11	0.43
	Simple repeats (%)	1.23	1.2	0.67	0.98	1.03	1.51	0.92	1.01
Total (%)	50.2	49.75	48.76	49.78	49.94	50.74	49.6	50.32	

Table 4. Genome annotation and BUSCO assessment of the 8 soybean accessions.

Accession	IGA1001	IGA1002	IGA1003	IGA1004	IGA1005	IGA1006	IGA1007
SNPs							
Total SNPs	2429031	1860802	4275544	2321059	1938310	1970083	2121229
Indels							
Total Indels	563950	462627	922735	565841	478827	441705	495694
Structural Variations							
Large SV indels	14747	11750	25335	14083	11933	12340	12980
Translocations	1213	726	3006	1142	706	950	885
Inversions	290	223	413	393	200	210	208
PAVs							
PV events	2004	1593	3575	1869	1659	1504	1780
AV events	1905	1515	3330	1851	1586	1469	1692

Table 5. Genomic variation of 7 soybean accessions compared with IGA1008.

performed utilizing *ab initio*-, homology-, RNA-sequencing-, and Iso-seq-based methods. Augustus v.3.3⁸⁰ and Glimmer v.3.0.4⁸¹ were used for *ab initio* gene prediction. For homology-based annotation, protein sequences from *Glycine max* Williams 82, *Glycine soja*, *Arabidopsis thaliana*, *Arachis duranensis*, and *Cajanus cajan* were obtained from NCBI and aligned to each of the 8 genomes using TBLASTN⁸². Exonerate was used to build gene structure based on the Blast results. For RNA-seq based gene prediction, short reads were mapped to their respective genome using TopHat v2.1.1⁸³ and the gene structure was predicted using Cufflinks v2.2.1⁸⁴. For Iso-seq based gene prediction, reads were mapped to IGA1008 assembly using GMAP v2016-09-14⁸⁵ and TransDecoder v4.1.0 was used to filter for high quality gene models. Lastly, a consensus gene set was generated by integrating gene annotations from each source using MAKER⁸⁶. With this process, we identified 57,286 to 58,392 protein coding genes (Table 4).

Non-coding RNA (ncRNA) annotation was performed using INFERNAL⁸⁷ based on Rfam v.14.1 models. tRNA was annotated using tRNAscan-SE v.1.3.1⁸⁸. rRNA annotation was performed using BLASTN with known soybean rRNA sequences.

BUSCO evaluation of genome and annotations. BUSCO (Benchmarking Universal Single-Copy Orthologs) evaluation was performed on the genome assembly and the gene annotations using BUSCO v.3.0.2⁸⁹ with embryophyta_odb9 data set.

Variation detection. Whole genome alignment based variation detection was performed by aligning each genome to IGA1008 using MUMer4⁹⁰ with parameters `-l 40 -c 90`. The MUMer alignment was filtered using delta-filter with parameters `--l`. SNVs, small Indels were called using show-snps with parameters `-rlTHC`. Large SVs were called using custom scripts MumSV (<https://github.com/jeff-sc-chu/MumSV>) based on output from show-coords with parameters `-THrcl`. Alignment gaps associated with assembly gaps were filtered away. Presence-absence variations (PAVs) were assessed based on large insertions and deletions (>100 bp) using BLASTN against the reference genome. A deletion or insertion is considered a PAV if BLASTN alignment to the reference genome is <50% query coverage or <90% PID. Accession specific SNP, Indel, and SVs were identified by selecting variants not overlapping any variants of the same class in all other genomes. Using the above process, we identified 1.86–4.27 million SNVs, 0.44–0.92 million small indels, 11.75–25.33 thousand large indels (>100 bp), 706–3006 translocation events, and 200–413 inversion events (Table 5).

Title	Description
Annotated peptide sequence for each soybean genome	Peptide sequence of each annotated coding gene for each genome
Annotated CDS sequences for each soybean genome	Coding sequences of each annotated gene for each genome
Gene annotations (GFF)	Gene annotation of each genome in GFF format
SNVs for each genome compared to IGA1008	SNV location detected by whole genome MUMer alignment of each genome using IGA1008 as a reference.
Indels between soybean genome assemblies and IGA1008	Indel location detected by whole genome MUMer alignment of each genome using IGA1008 as a reference.
PAVs of between soybean genomes compared with IGA1008	PAV location detected by whole genome MUMer alignment of each genome using IGA1008 as a reference.
Gene orthology correspondence	Gene orthology correspondence between IGA1008 and Williams 82 v4 genome.

Table 6. Description of file hosted on the Figshare record (<https://doi.org/10.6084/m9.figshare.c.5106161.v4>) accompanying this paper.

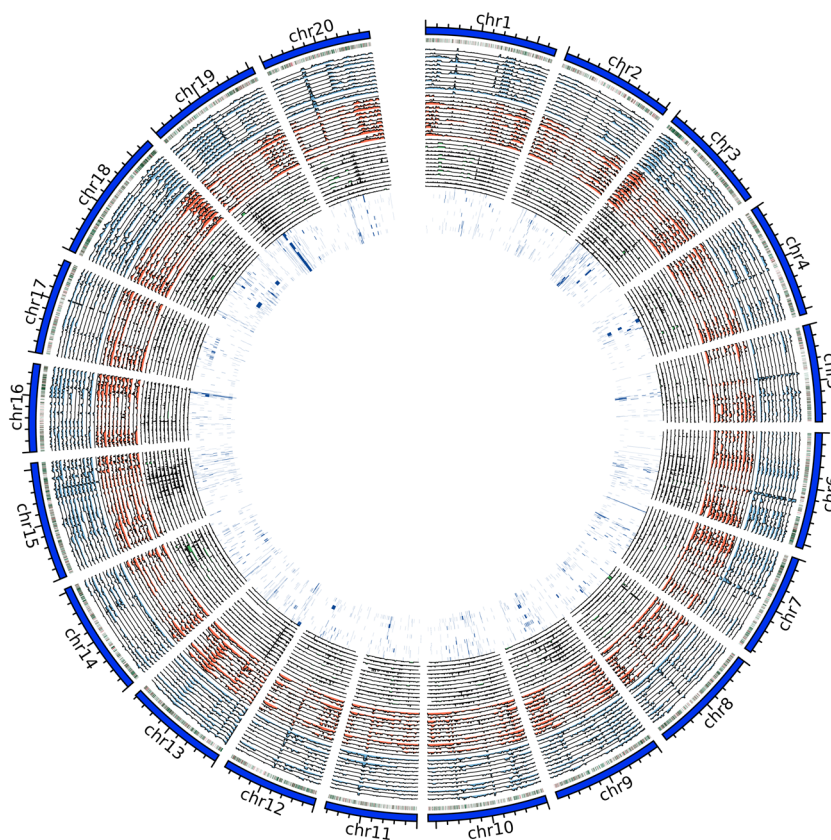


Fig. 2 Genomic variations of 12 soybeans compared to IGA1008 genome. The rings are in the following order from outer-most to inner-most: IGA1008 chromosomes (navy blue), PAV regions with dark red representing PV and dark green representing AV, a set of blue histograms represent SNP density in 400Kb window, the set of orange histograms represent small indel density in 400Kb window, the set of green histograms represent the large indel sizes, the inner-most set of dark blue highlights represent inversions. Each set depicts the variation of 12 soybeans ordered from outside to inside: IGA1001, IGA1002, IGA1003, IGA1004, IGA1005, IGA1006, IGA1007, ZH13, W05, Lee, PI483463, W82_v4.

Data Records

The sequencing data, genome assembly, and genome annotation of the 8 soybeans have been deposited in NCBI under the BioProject PRJNA561626. This includes whole genome sequencing data from the PacBio Sequel platform^{29–36} and from the Illumina platform^{37–44}, RNA sequencing data from the Illumina platform^{45–52}; Hi-C library sequencing data from the Illumina platform^{54–61}; and genome assembly sequences and annotations^{68–75}. The individual BioProject, BioSample, and SRA accession IDs are also listed in Supplementary Table 2. Gene annotations in GFF3 format, one-to-one gene correspondences between IGA1008 and W82_v4, and the coordinates of SNPs, Indels, and PAVs are available on Figshare with <https://doi.org/10.6084/m9.figshare.c.5106161.v4>⁹¹. We have also detailed the resources available on Figshare in Table 6. The plant materials used in this study are available from the authors upon request.

Technical Validation

Genome assembly quality assessment. The final assemblies of the 8 genomes ranged between 986.1 Mb and 1001.3 Mb with more than 98.7% of the genome anchored to 20 chromosomes (Table 3). The genome assembly size is similar to its respective k-mer estimation (Table 3) and also comparable to soybean genomes, such as W82_v4³, ZH13²⁸, and W05²⁶, which were assembled using similar sequencing technologies. The contig N50 were between 1.4 Mb–6.1 Mb and the scaffold N50 were between 48.8 Mb–50.7 Mb (Table 3)^{68–75}. The number of gaps ranged between 830 and 2922 (Table 3). Each unspanned gap in the assembly was arbitrarily set with 500 bp of Ns. As a comparison, the current soybean reference Williams 82 version 4 genome has a genome size of 978 Mb with contig N50 of 0.41 Mb, scaffold N50 of 20.44 Mb, and 8920 gaps³.

We further assessed the quality of our genome assemblies in a number of ways. First, the presence of centromeric repeats CentGm-1/2 were found in all 20 chromosomes for all genome assemblies. We assessed the completion and accuracy of the assemblies using BUSCO⁸⁹ and re-mapping of Illumina short read data. We observed an average of 96.98% complete BUSCO alignment and an average 99.75% remapping rate. Next, a high accuracy genome assembly would expect a very low level of homozygous SNVs from the remapping analysis and as expected we see an average of 0.00037% homozygous SNVs and 0.0073% heterozygous SNVs indicating the low error rates in these assemblies (Table 3).

Genome comparison with soybean reference genome Williams 82. IGA1008 is a soybean cultivar derived from Williams 82. Between IGA1008 and the W82_v4 assemblies, we identified 0.38 million SNVs, 0.14 million small indels, 3,203 large indels (>100 bp), 255 translocation-like and 135 inversion-like events. These events do not border or cross assembly gaps and may represent genomic variations between the two W82 lines.

Gene set completeness assessment. The numbers of protein-coding genes annotated for these cultivars were highly similar, ranging between 57,286 and 58,392 genes. We used BUSCO to evaluate the completion of our gene annotation and found that over 96% of the 1440 genes were found completely (Table 4).

Structural variation assessment. IGV⁹² was used to visually inspect many structural variations. We also examined known structural variations for their presence in the 8 genomes and our data confirmed the 40Kb Williams 82-specific insertion on Chromosome 15 (Supplementary Fig. 1)^{7,26}, the I-locus inversion event (Supplementary Fig. 2)²⁶ and the large deletion (~15Kb) affecting the E3 gene (Supplementary Fig. 3)²⁷.

Usage Notes

Identifying variations in known genes. The resources generated in this dataset allows one to search for genomic variations in genes of interest. For instance, we were able to identify nonsense mutations in the E2 gene as well as frameshift indels in J and FT1b (Supplementary Fig. 3). These resources, such as the indel found in FT1b, has not been characterized before and provide candidate alleles for further soybean research.

PAV gene screening. As more soybean genomic resources become available, one can take a pan-genome approach to identify genomic variations that are unique or shared between different soybean accessions. As a demonstration, we included 5 additional soybean genomes (ZH13²⁸, W05²⁶, Lee³, PI483463³, W82³) to identify genomic variations that are shared in a subset of the genomes (Fig. 2). These 5 genomes were chosen for their genome assembly quality, which were improved using sequencing technologies with longer reads. We identified 60 genes that were found in 3 wild soybeans but missing or truncated in the other 10 cultivated soybeans and 185 genes found missing or truncated in all 3 wild soybeans (Supplementary Table 3). Analyses such as this could generate additional resources to further soybean research and crop improvement.

Code availability

The versions and parameters of published software used in this study were described in the Methods. MumSV is a set of custom scripts to call large SVs based on MUMer alignments and it can be accessed at <https://github.com/jeff-sc-chu/MumSV>.

Received: 2 November 2020; Accepted: 30 April 2021;

Published online: 01 July 2021

References

- Dashiell, K. SOYBEANS: Improvement, Production, and Uses. Third Edition. *Agricultural Systems* vol. 83 110–111 (2005).
- Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
- Valliyodan, B. *et al.* Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J.* **100**, 1066–1082 (2019).
- Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **7**, 13390 (2016).
- Da Silva, C. *et al.* The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* **25**, 4777–4788 (2013).
- Li, M. *et al.* Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res.* **27**, 865–874 (2017).
- 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).
- Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol* **2**, 1940–1955 (2018).
- Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
- Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**, 419–423 (2011).

11. Read, B. A. *et al.* Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **499**, 209–213 (2013).
12. Yu, J. *et al.* Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol. J.* **17**, 881–892 (2019).
13. Zhang, Q.-J. *et al.* Rapid diversification of five *Oryza* AA genomes associated with rice adaptation. *Proc. Natl. Acad. Sci. USA* **111**, E4954–62 (2014).
14. Neafsey, D. E. *et al.* Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347**, 1258522 (2015).
15. Hu, Z. *et al.* Novel sequences, structural variations and gene presence variations of Asian cultivated rice. *Sci Data* **5**, 180079 (2018).
16. Wang, W. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
17. Stein, J. C. *et al.* Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
18. Yao, W. *et al.* Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* **16**, 187 (2015).
19. Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
20. Song, J.-M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants* **6**, 34–45 (2020).
21. Valliyodan, B. *et al.* Landscape of genomic diversity and trait discovery in soybean. *Sci. Rep.* **6**, 23598 (2016).
22. Shen, Y. *et al.* De novo assembly of a Chinese soybean genome. *Sci. China Life Sci.* **61**, 871–884 (2018).
23. Shimomura, M. *et al.* The *Glycine max* cv. Enrei Genome for Improvement of Japanese Soybean Cultivars. *Int. J. Genomics Proteomics* **2015**, 358127 (2015).
24. 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).
25. Qi, X. *et al.* Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nat. Commun.* **5**, 4340 (2014).
26. Xie, M. *et al.* A reference-grade wild soybean genome. *Nat. Commun.* **10**, 1216 (2019).
27. Liu, Y. *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162–176.e13 (2020).
28. Shen, Y. *et al.* Update soybean *Zhonghuang 13* genome to a golden reference. *Sci. China Life Sci.* **62**, 1257–1260 (2019).
29. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7016449> (2020).
30. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010599> (2020).
31. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010600> (2020).
32. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7026392> (2020).
33. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7016855> (2020).
34. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010602> (2020).
35. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010601> (2020).
36. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010603> (2020).
37. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7009432> (2020).
38. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX6983681> (2020).
39. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7009429> (2020).
40. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7009431> (2020).
41. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX6825670> (2020).
42. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7009428> (2020).
43. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7009433> (2020).
44. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7009430> (2020).
45. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010783> (2020).
46. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010779> (2020).
47. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010782> (2020).
48. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7016199> (2020).
49. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010787> (2020).
50. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010786> (2020).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010784> (2020).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010785> (2020).
53. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
54. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010335> (2020).
55. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010338> (2020).
56. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010337> (2020).
57. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7026256> (2020).
58. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010339> (2020).
59. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010348> (2020).
60. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010349> (2020).
61. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRX7010336> (2020).
62. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
63. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
64. Zhang, J. *et al.* Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. *Bioinformatics* **32**, 3058–3064 (2016).
65. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
66. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
67. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
68. Feng, X. *Glycine Max Cultivar Hefeng 25, Whole Genome Shotgun Sequencing Project.* *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIWX000000000> (2021).
69. Feng, X. *Glycine Max Cultivar Huaxia 3, Whole Genome Shotgun Sequencing Project.* *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIXC000000000> (2021).
70. Feng, X. *Glycine Max Cultivar Jinyuan, Whole Genome Shotgun Sequencing Project.* *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIXB000000000> (2021).
71. Feng, X. *Glycine Max Cultivar Wenfeng 7, Whole Genome Shotgun Sequencing Project.* *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIYW000000000> (2021).
72. Feng, X. *Glycine Max Cultivar Williams 82, Whole Genome Shotgun Sequencing Project.* *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIWW000000000> (2021).

73. Feng, X. Glycine Max Cultivar Zhonghuang 13, Whole Genome Shotgun Sequencing Project. *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIXA000000000> (2021).
74. Feng, X. Glycine Max Cultivar Zhonghuang 35, Whole Genome Shotgun Sequencing Project. *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIWZ000000000> (2021).
75. Feng, X. Glycine Soja Cultivar F, Whole Genome Shotgun Sequencing Project. *GenBank* <https://www.ncbi.nlm.nih.gov/nucleotide/WIXD000000000> (2021).
76. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
77. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–8 (2007).
78. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
79. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4, Unit 4.10 (2009).
80. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–9 (2006).
81. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
82. Altschul, S. B. Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
83. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
84. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
85. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
86. Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).
87. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
88. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol. Biol.* **1–14**, 2019 (1962).
89. Seppy, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
90. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
91. Chu, J. Datasets for Eight soybean reference genome resources from varying latitudes and agronomic traits. *figshare* <https://doi.org/10.6084/m9.figshare.c.5106161.v4> (2021).
92. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).

Acknowledgements

This work was funded by National Key Research and Development Project from the Ministry of Science and Technology of China (Grant number: 2016YFD0101900); Key Research Program of Chinese Academy of Sciences (Grant number: 593-ZDRW-ZS-2019-2-02); CAS Pioneer Hundred Talents Program (to NC); and Taishan Scholar Project Special Fund (to NC). The funding bodies were not involved in the design, data collection, and analysis for this study.

Author contributions

X.F. and N.C. conceived and designed the research project. K.T. and J.L. prepared the soybean materials and measured growth traits. H.W., B.P. and X.Y. performed sequencing, genome assembly, and genome annotation. J.S.C., G.L. and H.Z. analyzed the data. J.S.C., N.C. and X.F. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-021-00947-2>.

Correspondence and requests for materials should be addressed to N.C. or X.F.

Reprints and permissions information is available at www.nature.com/reprints.

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 Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2021