BMC Genomic Data

DATA NOTE

Open Access

De novo genome assembly of a high-protein soybean variety HJ117

Zhi Liu[1†], Qing Yang[1†], Bingqiang Liu[1†], Chenhui Li[1,2], Xiaolei Shi[1], Yu Wei[1], Yuefeng Guan[3], Chunyan Yang[1], Mengchen Zhang[1*] and Long Yan[1*]

Abstract

Objectives Soybean is an important feed and oil crop in the world due to its high protein and oil content. China has a collection of more than 43,000 soybean germplasm resources, which provides a rich genetic diversity for soybean breeding. However, the rich genetic diversity poses great challenges to the genetic improvement of soybean. This study reports on the de novo genome assembly of HJ117, a soybean variety with high protein content of 52.99%. These data will prove to be valuable resources for further soybean quality improvement research, and will aid in the elucidation of regulatory mechanisms underlying soybean protein content.

Data description We generated a contiguous reference genome of 1041.94 Mb for HJ117 using a combination of Illumina short reads (23.38 Gb) and PacBio long reads (25.58 Gb), with high-quality sequence coverage of approximately 22.44× and 24.55×, respectively. HJ117 was developed through backcross breeding, using Jidou 12 as the recurrent parent and Chamoshidou as the donor parent. The assembly was further assisted by 114.5 Gb Hi-C data (109.9×), resulting in a contig N50 of 19.32 Mb and scaffold N50 of 51.43 Mb. Notably, Core Eukaryotic Genes Mapping Approach (CEGMA) assessment and Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment results indicated that most core eukaryotic genes (97.18%) and genes in the BUSCO dataset (99.4%) were identified, and 96.44% of the genomic sequences were anchored onto twenty pseudochromosomes.

Keywords Soybean, De novo assembly, Genome feature, High protein content

†Zhi Liu, Qing Yang and Bingqiang Liu contributed equally to this work.

*Correspondence:
Mengchen Zhang
zhangmengchen@hotmail.com
Long Yan
dragonyan1979@163.com
[1]Hebei Key Laboratory of Crop Genetics and Breeding, Huang-Huai-Hai Key Laboratory of Biology and Genetic Improvement of Soybean, Ministry of Agriculture and Rural Affairs, Institute of Cereal and Oil Crops, National Soybean Improvement Center Shijiazhuang Sub- Center, Hebei Academy of Agricultural and Forestry Sciences, 050035 Shijiazhuang, Hebei, China
[2]College of Life Sciences, Hebei Agricultural University, 071001 Baoding, Hebei, China
[3]Guangdong Provincial Key Laboratory of Plant Adaptation and Molecular Design, Innovative Center of Molecular Genetics and Evolution, School of Life Sciences, Guangzhou University, 510006 Guangzhou, Guangdong, China

Objective

Soybean [*Glycine max* (L.) Merr.] is an important protein feed and vegetable oil crop worldwide. The cultivation of soy enables the production of various valuable products, including edible oils, biodiesel, and biofertilizers [1]. The main protein source in poultry and livestock feed is meal derived from soybean seeds. Commercial soybean cultivars generally have a seed protein content ranging from approximately 38–42% on a dry weight basis [2]. Only soybean grains with a protein content of 41.5% or higher on a dry weight basis can be used to produce meal with a protein content of 47.5% or higher [2]. Enhancing the amino acid content of soybean seeds would further increase the economic value of soybean. Soy protein

© The Author(s) 2024. **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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, 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 data made available in this article, unless otherwise stated in a credit line to the data.

content is influenced by complex factors such as genotype, environment, and genotype–environment interactions [3, 4]. Due to the strong negative correlations of soy protein content and oil content [4] with yield [5], it is quite difficult to increase soy protein content.

In the early stages of soybean breeding, farmers primarily relied on repeatedly selecting preferred seeds from cultivated populations [6]. Following that, artificial hybridization technology was introduced, and the initial artificially hybridized cultivated soybean was introduced in North America during the 1940s [7]. With the development and progress of molecular biology technology, marker-assisted selection (MAS) has been employed to expedite the breeding process [8]. The publication of the initial reference genome of soybean (cultivar Williams 82) in 2010 [9] signaled the commencement of the soybean functional genomics research era [10, 11]. The enhancement of sequencing technologies has significantly boosted the capacity to generate high-quality genome assemblies.

Data description

The *Glycine max* sample was collected from Shijiazhuang (37°6′25″N, 114°42′47″E). Genomic DNA and total RNA were isolated from leaf tissues. High-quality DNA was extracted using QIAGEN® Genomic kits. Three methods were used to quantify and check the extracted DNA, NanoDrop 2000 Spectrophotometer (Thermo Fischer Scientific), agarose gel electrophoresis and Qubit Fluorometer (Invitrogen). After the detection, the DNA was purified using AMPure PB beads (Pacbio 100-265-900), and the subsequent library construction utilized the final high-quality genomic DNA (gDNA). The size and concentration of the library fragments were assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). Qualified libraries were evenly loaded on SMRT Cell and sequenced for 30 h using Sequel II/IIe system (Pacific Biosciences, CA, USA).

Briefly, the DNA sample was initially fixed with formaldehyde and subsequently digested using *Hind*III restriction enzyme. Next, the DNA ends underwent repair and were labeled with biotin. Subsequently, T4 DNA ligase was used to ligate the interacting fragments to form a loop. After ligation, protease K was added for crosslinking, and then protein of ligated DNA fragments was digested to obtain purified DNA. Finally, the purified DNA was fragmented into sizes ranging from 300 to 500 base pairs. The biotin-labeled DNA fragments were then isolated using Dynabeads® M-280 Streptavidin (Life Technologies). Subsequently, the Hi-C library was constructed and sequenced on the Illumina NovaSeq6000 sequencing platform using paired-end reads of 150 base pairs.

To ensure the acquisition of high-quality data, the raw polymerase reads were subjected to quality control using the PacBio SMRT-Analysis package (https://www.pacb.com). This involved filtering out the following types of polymerase reads: (1) polymerase reads less than 50 bp in length, (2) Polymerase readings with a mass value below 0.8, (3) a polymerase read comprising an adaptor attached to itself and removing the adaptor sequence in the polymerase read. Then use SMRTLink 9.0 (parameter --min-passes=3 --min-rq=0.99) to generate CCS reads for subsequent assembly.

Hifiasm (https://github.com/chhylp123/hifiasm) was employed to assemble the HiFi reads, and the preliminarily assembled genome version (primary contigs) was obtained. To obtain chromosome level genome, we performed Hi-C assisted assembly. For the ~114.5 Gb raw reads (Data file 1 and Data file 2), preliminary quality control was performed using Fastp [14], and the resulting clean reads were subsequently aligned to primary contigs using hicup. Valid pair reads were utilized for further analysis. AllHIC was used for auxiliary assembly, and then Juicebox was used for fine-tune AllHIC clustering results. Finally, A genome was obtained with a contig N50 length of 19.32 Mb and a total contig length of 1041.94 Mb, as well as a scaffold N50 length of 51.43 Mb and a total scaffold length of 1041.95 Mb (Data file 3 and Data file 4).

To assess the quality of the assembly the self-written script was used to perform statistics on the number of single chromosome cluster scaffolds, chromosome sequence length, and genome mounting rate. According to the number of sequences assembled to the chromosome level and the number of sequences that were not assembled to the chromosome level, the Hi-C mounting rate was calculated. The chromosome-level genome was partitioned into 500 Kb bins of equal length. The number of Hi-C read pairs spanning any two bins was used as the intensity signal to represent the interaction between the respective bins. Heatmaps (Data file 5) were generated based on these signals. BUSCO (Benchmarking Universal Single-Copy Orthologs: http://busco.ezlab.org/) [18] was also applied to perform a quality assessment of the genome. The conserved genes (248 genes) existing in six eukaryotes were selected to construct the core gene library for CEGMA [19] evaluation. The evaluation results revealed that the majority of core eukaryotic genes (97.18%) and genes in the BUSCO dataset (99.4%) were successfully identified (Data file 6).

Repeatmasker [21] and repeatproteinmask (http://www.repeatmasker.org/) were employed to identify sequences that exhibit similarity to known repeat sequences. LTR_FINDER [22] was used to perform de novo prediction. Totally, 361,475,923 bp RepBase TEs and 453,714,080 bp de novo repetitive sequences were

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File type (file extension)	Data repository and identifier (DOI or accession number)
Data file1	Statistics on sequence data	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [12]
Data file2	Hi-C raw data	Fastq file (.fastq)	NGDC Genome Sequence Archive, https://ngdc.cncb.ac.cn/gsa/browse/CRA014073 [13]
Data file3	Assembly statistics of HJ117	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [15]
Data file4	genome.fa	Fasta file (.fasta)	NGDC Genome warehouse, https://ngdc.cncb.ac.cn/gwh/Assembly/83716/show [16]
Data file5	Hi-C interaction heatmap	Image file (.tif)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [17]
Data file6	Assessment results of CEGMA and BUSCO	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [20]
Data file7	Results of transposable element classification statistics	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [23]
Data file8	Results of gene structure prediction	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [25]
Data file9	Glycine.max.gene.gff	Gff file (.gff)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [26]
Data file10	Genome annotation of HJ117	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [27]
Data file11	Overview of the HJ117 reference genome	Image file (.tif)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [28]
Data file12	Statistics on non-coding RNA annotation results	Spreadsheet (.xls)	Figshare,https://doi.org/10.6084/m9.figshare.24865518figshare.com/s/6de11eca18b3ccef8314 [31]
Data file13	Raw RNA reads of leaf tissues	Fastq file (.fastq)	NGDC Genome Sequence Archive, https://ngdc.cncb.ac.cn/gsa/browse/CRA014073 [34]
Data file14	HiFi raw data	Fastq file (.fastq)	NGDC Genome Sequence Archive, https://ngdc.cncb.ac.cn/gsa/browse/CRA014073 [35]

identified, respectively (Data file 7). Structural prediction of genes was performed by using AUGUSTUS (http://bioinf.uni-greifswald.de/AUGUSTUS/) [24] (Data file 8 and Data file 9). Then, we used the protein databases NR (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/), SwissProt (http://www.uniprot.org/), KEGG (http://www.genome.jp/kegg/) and InterPro (https://www.ebi.ac.uk/interpro/) to annotate the gene set obtained from the gene structure annotation. A total of 57,151 genes were predicted, with 54,550 of these genes being functionally annotated in the database (Data file 10). The circular plot illustrates gene density, transposable element (TE) density, and GC density (Data file 11). The tRNAscan-SE [29] (http://lowelab.ucsc.edu/tRNAscan-SE/) was used to identify tRNA sequences within the genome. Blast [30] alignment was used to find the rRNA in the genome. The prediction of miRNA and snRNA sequences within the genome was performed using INFERNAL (http://infernal.janelia.org/). The copy number of miRNA, tRNA, rRNA and snRNA ranged from 68 to 5,116 (Data file 12) (See Table 1).

Limitations

Soybean is considered to have undergone an allotetraploidy event [9] that have resulted in 75% of its genes being present in multiple copies [32]. Repetitive DNA made up ~54.4% of each genome [33]. In this study, 23.38 Gb Illumina short reads (Data file 13) and 25.58 Gb of PacBio long reads (Data file 14) were obtained, providing approximately 22.44× and 24.55× sequence coverage. Although Hi-C sequencing obtained 114.5 Gb of data with a depth of 109.9×, the overall sequencing depth was relatively low, which may result in incomplete genomic information being obtained.

The contig N50 length of the de novo assembled HJ117 genome is 19.32 Mb, and the scaffold N50 reaches 51.43 Mb, indicating that the genome assembly level has achieved the average level of soybean genome assemblies during the same period. However, gaps still exist in the genome. To achieve accurate genome assembly, optical mapping technology could be incorporated, and HiFi sequencing depth could be increased in the later stages. Alternatively, HJ117 genome could be assembled to a telomere-to-telomere level using ONT Ultra-long technology to obtain more comprehensive genomic information for HJ117.

Abbreviations

CEGMA	Core Eukaryotic Genes Mapping Approach
BUSCO	Benchmarking Universal Single-Copy Orthologs
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
TE	Transposable Element
Hi-C	High-resolution Chromosome Conformation Capture

HiFi High-Fidelity Sequencing
HJ117 Ji HuiJiao No.117

Acknowledgements
Not applicable.

Author contributions
ZL data curation and writing-original draft; QY visualization of the work; BL project administration; CL and XS resources; YW data curation; YG, CY, MZ supervision; LY conceptualization and methodology.

Funding
This work was financially supported by the National Key R&D Project (2021YFD1201602), National Natural Science Foundation of China (31871652), and Natural Science Foundation of Hebei (C2020301020).

Data availability
Data files 2,13,14 described in this Data note can be freely and openly accessed on the Genome Sequence Archive in National Genomics Data Center China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences under GSA: CRA014073 (https://ngdc.cncb.ac.cn/gsa/browse/CRA014073) [13,34,35]. Data files 4 described in this Data note can be freely and openly accessed on the Genome warehouse in National Genomics Data Center China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences under GWH: GWHERCR00000000 (https://ngdc.cncb.ac.cn/gwh/Assembly/83716/show) [16]. Data files 1,3,5-12 are available on Figshare (https://doi.org/10.6084/m9.figshare.24865518) [12,15,17,20,23,25,26,27,28,31]. Please see Table 1 and references for details and links to the data.

Declarations

Ethics approval and consent to participate
The current study complies with relevant institutional, national, and international guidelines and legislation for experimental research and field studies on plants (either cultivated or wild), including the collection of plant material. Permissions were obtained to collect *Glycine max* samples. Sampling was conducted in Institute of Cereal and Oil Crops (ICOC), Hebei Academy of Agricultural and Forestry Sciences field plots and permission was granted by the ICOC to perform data collection.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 25 December 2023 / Accepted: 22 February 2024
Published online: 04 March 2024

References

1. Vianna GR, Cunha NB, Rech EL. Soybean seed protein storage vacuoles for expression of recombinant molecules. Curr Opin Plant Biol. 2023;71:102331. https://doi.org/10.1016/j.pbi.2022.102331.
2. Willis S. The use of soybean meal and full fat soybean meal by the animal feed industry. In: 12th Australian soybean conference. Soy Australia, Bundaberg. 2003.
3. Carver BF, Burton JW, Carter TE, Wilson RF. Response to environmental variation of soybean lines selected for altered unsaturated fatty acid composition. Crop Sci. 1986;26:1176–81. https://doi.org/10.2135/cropsci1986.0011183X002600060021x.
4. Chaudhary J, Patil GB, Sonah H, et al. Expanding Omics resources for improvement of soybean seed composition traits. Front Plant Sci. 2015;6:1021. https://doi.org/10.3389/fpls.2015.01021.
5. Kim M, Schultz S, Nelson RL, Diers BW. Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci. 2016;56:219–25. https://doi.org/10.2135/cropsci2015.06.0340.
6. Zhang M, Liu S, Wang Z, et al. Progress in soybean functional genomics over the past decade. Plant Biotechnol J. 2022;20(2):256–82. https://doi.org/10.1111/pbi.13682.
7. Rincker K, Nelson RL, Specht J, Sleper D, Cary T, Cianzio S, Casteel S, et al. Genetic improvement of U.S. soybean in maturity groups II, III, and IV. Crop Sci. 2014;54:1419–32. https://doi.org/10.2135/cropsci2013.10.0665.
8. Li MW, Wang Z, Jiang B, Kaga A, Wong FL, Zhang G, Han T, et al. Impacts of genomic research on soybean improvement in East Asia. Theor Appl Genet. 2020;133:1655–78. https://doi.org/10.1007/s00122-019-03462-6.
9. Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178–83. https://doi.org/10.1038/nature08670.
10. Li MW, Xin D, Gao Y, et al. Using genomic information to improve soybean adaptability to climate change. J Exp Bot. 2017;68(8):1823–34. https://doi.org/10.1093/jxb/erw348.
11. Wang Z, Tian Z. Genomics progress will facilitate molecular breeding in soybean. Sci China Life Sci. 2015;58(8):813–5. https://doi.org/10.1007/s11427-015-4908-2.
12. Data file 1.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
13. Data file 2.: De novo genome assembly of a high-protein soybean variety-HJ117. NGDC Genome Seq Archive. 2023. https://ngdc.cncb.ac.cn/gsa/browse/CRA014073.
14. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
15. Data file 3.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
16. Data file 4.: De novo genome assembly of a high-protein soybean variety-HJ117. NGDC Genome warehouse. 2023. https://ngdc.cncb.ac.cn/gwh/Assembly/83716/show.
17. Data file 5.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
18. 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):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
19. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7. https://doi.org/10.1093/bioinformatics/btm071.
20. Data file 6.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
21. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinf. 2004; Chap. 4. https://doi.org/10.1002/0471250953.bi0410s05
22. 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):W265-8. https://doi.org/10.1093/nar/gkm286
23. Data file 7.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
24. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44. https://doi.org/10.1093/bioinformatics/btn013.
25. Data file 8.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
26. Data file 9.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
27. Data file 10.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
28. Data file 11.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
29. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64. https://doi.org/10.1093/nar/25.5.955.
30. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32(Web Server issue):W20-5. https://doi.org/10.1093/nar/gkh435
31. Data file 12.: De novo genome assembly of a high-protein soybean variety-HJ117. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24865518.
32. Roulin A, Auer PL, Libault M, et al. The fate of duplicated genes in a polyploid plant genome. Plant J. 2013;73(1):143–53. https://doi.org/10.1111/tpj.12026.

33. Liu Y, Du H, Li P, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162–176e13. https://doi.org/10.1016/j.cell.2020.05.023
34. Data file 13.: De novo genome assembly of a high-protein soybean variety-HJ117. NGDC Genome Seq Archive. 2023. https://ngdc.cncb.ac.cn/gsa/browse/CRA014073.
35. Data file 14.: De novo genome assembly of a high-protein soybean variety-HJ117. NGDC Genome Seq Archive. 2023. https://ngdc.cncb.ac.cn/gsa/browse/CRA014073.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.