Plant Communications
Correspondence

CellPress
Partner Journal

A complete reference genome for the soybean cv. Jack

Dear Editors,

Soybeans are a global commodity for their edible protein and vegetable oil. The global population is predicted to be 9.7 billion by 2050 (UN, 2022), with a concomitant drastic increase in protein demand. With already 2.4 billion people suffering from food insecurity (FAO et al., 2023), there is an urgent need to meet future production demands for plant-based proteins.

High-quality soybean genomes will play a key role in future efforts to enhance seed quality through breeding and functional genomics. The first soybean genome assembled was from the cultivated accession Williams 82 (WM82) (Schmutz et al., 2010). In comparison to WM82, variations have been discovered between varieties/accessions/cultivars show the need for genomic information of frequently used soybean cultivars (Belzile, et al., 2022). Recent advances in CRISPR-Cas genome editing technologies emphasize that improving transformation efficiencies in soybeans remains the largest impediment to making functional genomics a reality in this important oil seed crop. Due to its amenability and quick regeneration, soybean cv. Jack has become a proof-of-concept cultivar when investigating seed traits in soybeans (Schmidt and Herman, 2008a, 2008b).

In this study, we assembled and validated a complete telomere-to-telomere (T2T), soybean cv. Jack reference genome. This cv. Jack genome expresses a seed-specific GFP reporter for glycinin-regulon–controlled protein accumulation (Schmidt and Herman 2008b). The Jack cultivar is used widely as a model to develop new seed traits with this GFP-Jack line providing a platform to identify and evaluate the role of *trans* or *cis* factors that function to mediate enhanced seed traits and the economic performance of soybeans. A GFP-seed-specific expressing Jack cultivar will be a valuable reference genome for the manipulation and subsequent characterization of protein and oil seed traits—the foundation of many soybean improvement programs (Supplemental Figure 1).

To elucidate the soybean cv. Jack genome, we obtained three single-molecule real-time (SMRT) cells of continuous long reads (CLR) data using CLR sequencing mode and one SMRT cell of high-fidelity data using the circular consensus sequencing (CCS) mode in PacBio Sequel II Systems (Supplemental Figure 2 and Supplemental Table 1).

The complete soybean genome was assembled in five stages (Supplemental Figure 3 and Supplemental Table 2). Because a single assembler program was unable to generate a high-quality genome, we used multiple datasets as inputs in different *de novo* assembler programs (Supplemental Table 3). The first two stages of the Jack genome assembly consisted of stage 1 using CLR data and stage 2 using CCS data both edited by using Genome Puzzle Master (Zhang et al., 2016). The resultant assembly contained only two gaps: one in chr1 and another in chr11 (Supplemental Figure 4 and Supplemental Tables 4 and 5). These two gaps were subsequently filled by using a novel cyclic elongation process called DEGAP (Huang et al., 2023). The stage 3 assembly was achieved after the two gaps were filled by DEGAP with two extended sequences: one in chr1 consisting of 1 076 216 bp (Supplemental Figure 4B) and the other in chr11 consisting of 885 638 bp (Supplemental Figure 4C and Supplemental Tables 6 and 7).

The stage 4 assembly involved the resolution of the triplet in chr10 (Supplemental Figure 5). The highly repetitive region in chr10 was both too long and too complex to be assembled by any available assembler programs, so this region remained unresolved in previous published soybean genomes (Schmutz et al., 2010). The chr10 triplet was validated by first designing four primer sets (Figure 1B, Supplemental Figure 6, and Supplemental Table 8) and then using the amplicons in a combined mixture as a probe in fluorescence *in situ* hybridization (FISH) experiments. This hybridization experiment showed three signals present only in chr10 (Figure 1B). Collectively, these results indicate there is a triple repeat region unique to chr10 (Supplemental Figure 6) in the cv. Jack soybean genome.

After polishing with CCS data, the finalized stage 5 assembly of the soybean cv. Jack genome (called Gmax-GtJack-RS1), consisting of 1 011 764 152 bp in total length, was obtained (Supplemental Table 9). A high level of both completeness and quality of this genome was confirmed by elevated mapping rates—for example, 99.03% of Illumina reads, 99.80% of the 110-Gbp longest CLR reads, and 99.99% of all high-fidelity reads (Supplemental Table 1). We captured 98.3% complete BUSCOs in the *embryphyta* reference gene set and 99.6% compete BUSCOs in the *eukaryote* reference gene set (Supplemental Table 10). Quality value scores, calculated by using Merqury with CCS and Illumina data, indicate an error rate lower than 0.03% for our soybean Jack genome assembly (Supplemental Table 11).

Centromere function is conserved across *eukaryote*, but the DNA sequences are highly divergent, even between closely related species. Centromeric DNA usually contains highly repetitive DNAs, such as satellite DNAs and retrotransposons. *CentGm* (soybean centromere-specific satellite repeats) sequences have been reported previously (Tek et al., 2010) to consist of two types of tandem repeat sequences, *CentGm-1* and *CentGm-4*. We mapped these two repetitive sequences to our Jack genome and, as expected, they occurred in every

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

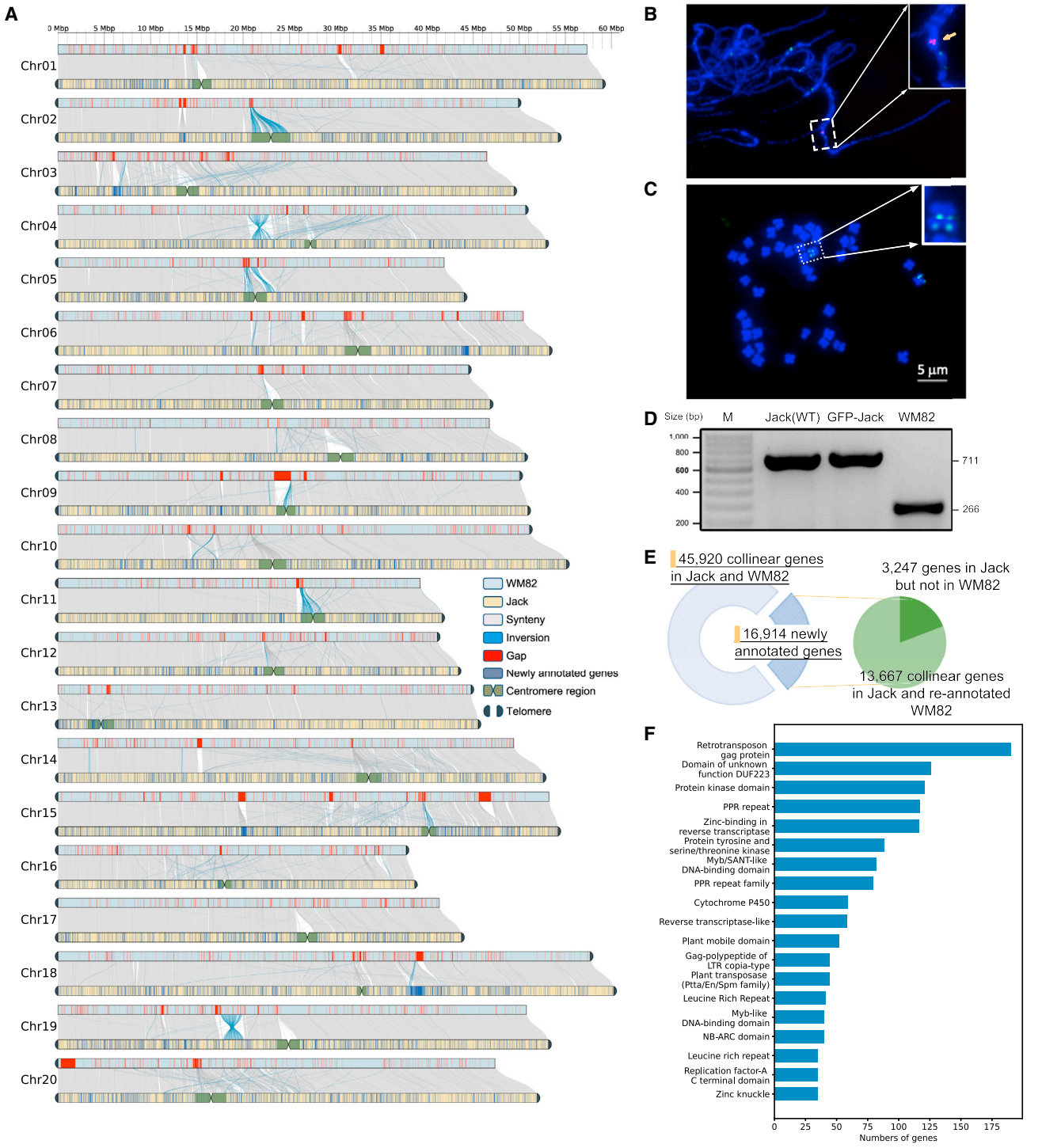

Figure 1. Collinearity analysis between the soybean Jack and WM82 genomes.

(A) Whole-genome comparison of the Jack complete genome to WM82 (Wm82.a4.v1). Linear comparison of the 20 individual chromosomes from the publicly available reference WM82 genome (top) compared to our Jack genome (bottom). Red segments denote missing sequences in WM82 that are resolved in Jack. Blue lines indicate regions of inversions and/or translocations. Gray areas indicate synteny alignment regions. Dark blue indicates density of 16 914 newly annotated genes in Jack when compared to the Wm82.a4.v1 annotation. Purple half-circles indicate telomeres found in the chromosomes. Retraction area in Jack indicates the centromere regions.

(B) FISH mapping of tandem repeat in chr10 of Jack.

(C) FISH mapping of GFP regions in the transgenic Jack genome. Two main GFP insertion sites were detected.

(D) PCR validation of one small insertion (SI1) between Jack (WT), GFP-Jack (this genome), and WM82.

(E) Collinear genes between Jack and WM82.

(F) GO enrichment analysis of newly annotated genes in Jack.

chromosome. Using the *CentGm* sequences as markers to denote centromeric regions, we estimated that 222 genes lie within this highly repetitive centromeric region that are now resolved in this complete T2T soybean genome (Supplemental Figure 1 and Supplemental Table 12). Similarly, we estimated that 23 genes are located in the highly repetitive telomeric regions (Supplemental Table 18).

The soybean cv. Jack (PI540556) was released in 1989 and developed by an F3-derived selection from the cross Fayette × Hardin (Nickell et al., 1990). The cv. Fayette was developed from the cross of WM82 with PI88788 (Bernard, 1988). The GFP-ER (endoplasmic reticulum [ER]–targeted GFP) transgenic cv. Jack soybean plant used in this genome project was determined to have two major GFP insertion locations in chr20, ∼27.9 kbp and ∼47.8 kbp (Figure 1C and Supplemental Figure 7A). A total of eight GFP insertions (i.e., five complete open reading frames and three partial open reading frames) exist in this GFP-ER transgenic cv. Jack genome (Supplemental Figure 7).

This complete assembly Gmax-GtJack-RS1 was compared to several known soybean genomes, including the latest Wm82.a4.v1 for WM82 (Schmutz et al., 2010), the recently released T2T assembly ASM3086415v1 for WM82-NJAU (Wang et al., 2023), JD17_chr_final for Jidou 17 (JD17) (Yi et al., 2022), and Gmax_ZH13_v2.0 for Zhonghuang 13 (ZH13) (Shen et al., 2019) (Figure 1A, Supplemental Figure 8, and Supplemental Table 13). Compared with the Jack genome, 52 and 38 inversions, 844 and 704 translocations, and 2097 and 1462 duplications were found in WM82 (Wm82.a4.v1) and WM82-NJAU (ASM3086415v1), respectively; 78 inversions, 1187 translocations, and 3100 duplications were found in JD17; and 72 inversions, 1110 translocations, and 3461 duplications were found in ZH13 (Supplemental Figure 9 and Supplemental Table 14). A comparison between Jack and WM82 genomes uncovered 6816 insertions and 4257 deletions in the cv. Jack; among them, 5461 insertions and 3123 deletions did not overlap with any WM82 gaps. We randomly selected six insertions/deletions (indels) that were located in putative genes and designed primer sets for genomic PCRs to validate our detections. The PCR results confirmed the authenticity of these indels found in comparing the genomes of Jack and WM82 (Supplemental Figure 10 and Supplemental Table 15).

Small indels may cause structural changes in encoded proteins. For example, *GmJack15g02718000* is predicted to encode for a cysteine-rich receptor-like protein kinase that is a single-pass membrane protein (Figure 1D and Supplemental Figure 10). A 445-bp insertion (SI1) in its intergenic region resulted in the disruption of one exon in Jack (Supplemental Figure 10B) and the resultant protein product containing only one transmembrane sequence, whereas the counterpart gene in WM82 encodes for three transmembrane sequences. As a direct result, there are distinct structural changes as the Jack encoded protein has fewer alpha helices and more beta turns than the protein encoded by WM82 (Supplemental Figure 11).

We analyzed potential sequences in Jack that may be located in WM82 gap regions by focusing on the shared collinear flanking sequences within the same chromosomes. As a result, 38.6 Mbp sequences within the Jack genome (Supplemental Table 16) were found to contain 92 genes and shared collinear flanking sequences with 702 gap regions in WM82. Among these Jack sequences, 3.6 Mbp (9.21%) can be mapped to WM82 unplaced contigs, and 0.6 Mbp (1.44%) are interspersed repeats and low-complexity DNA sequences selected by RepeatMasker (Supplemental Table 17). This result suggests that the high occurrence of repeats in a particular region may have led to gaps in the WM82 assembly (Supplemental Figure 12). These genes found in Jack that correspond to WM82 gap regions are considered newly uncovered due to the enhanced resolution of this cv. Jack genome.

Annotation of the soybean cv. Jack genome for transposable elements (TEs) and other repetitive sequences identified 266 033 TEs constituting 360 889 913 bp in total (Supplemental Table 19). Long terminal repeat/Gypsy retrotransposons were found to be enriched around centromeric regions (Supplemental Figure 1). Moreover, we found two rRNA gene-enriched regions, located in chr13 and chr19 (Supplemental Figure 13A). The 18S, 5.8S, and 28S rRNA genes are located in chr13 and display a high level of gene transcription (Supplemental Figure 13B). Similarly, sequence analysis between the 18S and 28S rRNA genes detected many promoter elements, including TATA and CAAT boxes, in this area (Supplemental Figures 13C and 14). The 5S rRNA genes were found in high abundance at ∼15.6 Mbp in chr19 (Supplemental Figure 13D).

There are 63 703 genes annotated in our complete soybean cv. Jack genome. In comparison to the published Wm82.a4.v1 annotation, 16 914 were considered to be newly annotated Jack genes because they had no homolog in the WM82 annotation (Figure 1E). Some of these genes reflect unique differences between these two cultivars. For example, we found 285 genes located only in the Jack genome during our presence/absence variant analysis (Supplemental Table 20). By using Gene Ontology (GO) resources, the newly annotated genes are predicted to consist of 5412 (32%) protein-coding genes, whereas the remaining 11 502 (68%) genes are of unknown function. Those genes with known GO annotations can be grouped into a few functional categories (Figure 1F), including reverse transcriptases and cytochrome P450s.

To make a pairwise comparison regarding the number of differential genes in our soybean cv. Jack genome, we reannotated the currently available WM82 assembly (Wm82.a4.v1) using the same annotation process for our cv. Jack genome. In so doing, 13 667 newly annotated Jack genes were found to have homologs in the reannotated WM82 genome, with 3247 genes in Jack having no WM82 homolog (Figure 1E). Noteworthy is this complete cv. Jack genome uncovered 337 genes that were previously unresolved in highly repetitive areas, including centromeric, telomeric, and gap regions in WM82 (Supplemental Tables 12, 16, and 18). In comparison to the recent T2T assembly WM82-NJAU, reannotated with the same process, we found that 2710 genes in Jack still had no homolog in WM82-NJAU.

In summary, we assembled a gap-free T2T soybean genome and provided details of the heterochromatic regions of all 20 chromosomes. The availability of this complete genome of a transformation-favored cultivar allowed us to uncover 337

genes that were previously unresolved, and will be a valuable resource to investigate current and emerging agricultural issues.

DATA AND CODE AVAILABILITY

All of the data needed to evaluate the conclusions in the paper are present in the paper and the supplementary materials. Genome sequences and RNA-seq data are in repository NCBI Bioproject PRJNA701655, and the genome assembly Gmax-GtJack-RS1 was deposited at DDBJ/ENA/GenBank: JAGXCU000000000. Additional data related to this paper may be requested from the authors.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This work was financially supported by the National Institute of Food and Agriculture award 2014-33522-22531(to M.A.S. and E.M.H.), and the Start-up Fund of the Huazhong Agricultural University (HZAU) (to J.Z.).

AUTHOR CONTRIBUTIONS

M.A.S. and J.Z. designed the research; Y.H., D.-H.K., and Y.M. conducted the experiments; Y.H., M.A.S., J.Z., and E.M.H. analyzed the data; M.A.S., J.Z., D.-H.K, and E.M.H. contributed the reagents/materials/analysis tools; and Y.H., M.A.S., and J.Z. wrote the paper.

ACKNOWLEDGMENTS

We thank the Arizona Genomics Institute (www.genome.arizona.edu) for generating the data used to assemble the genome, and the National Key Laboratory of Crop Genetic Improvement, HZAU for providing the bioinformatics computing platform in this study. The authors declare that they have no competing interests.

Received: October 16, 2023
Revised: November 8, 2023
Accepted: November 9, 2023
Published: November 14, 2023

Yicheng Huang (黄臆丞)[1], Dal-Hoe Koo[2], Yizhou Mao (毛一舟)[3], Eliot M. Herman[3], Jianwei Zhang (张建伟)[1,] and Monica A. Schmidt[3,*]*

[1]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China
[2]Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, KS 66506, USA
[3]BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
***Correspondence: Jianwei Zhang (jzhang@mail.hzau.edu.cn), Monica A. Schmidt (monicaschmidt@arizona.edu)**
https://doi.org/10.1016/j.xplc.2023.100765

REFERENCES

Belzile, F., Jean, M., Torkamaneh, D., Tardivel, A., Lemay, M.A., Boudhrioua, C., Arsenault-Labrecque, G., Dussault-Benoit, C., Lebreton, A., de Ronne, M., et al. (2022). The SoyaGen Project: Putting Genomics to Work for Soybean Breeders. Front. Plant Sci. **13**, 887553.

Bernard, R.L. (1988). Origins and Pedigrees of Public Soybean Varieties in the United States and Canada (US Department of Agriculture), p. 61. Technical Bulletin No. 1746.

FAO, I.F.A.D., and UNICEF, W.F.P.; WHO (2023). The State of Food Security and Nutrition in the World 2023. In Urbanization, Agrifood Systems Transformation and Healthy Diets across the Rural–Urban Continuum (Rome: FAO).

Huang, Y., Wang, Z., Schmidt, M., and Zhang, J. (2023). DEGAP: Dynamic Elongation of a Genome Assembly Path. Preprint at bioRxiv.

Nickell, C.D., Noel, G.R., Thomas, D.J., and Waller, R. (1990). Registration of 'Jack' soybean. Crop Sci. **30**:1365.

Schmidt, M.A., and Herman, E.M. (2008a). A RNAi knockdown of soybean 24 kda oleosin results in the formation of micro-oil bodies that aggregate to form large complexes of oil bodies and ER containing caleosin. Mol. Plant **1**:910–924.

Schmidt, M.A., and Herman, E.M. (2008b). The Collateral Protein Compensation Mechanism Can Be Exploited to Enhance Foreign Protein Accumulation In Soybean Seeds. Plant Biotechnol. J. **6**:832–842.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463**:178–183.

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

Tek, A.L., Kashihara, K., Murata, M., and Nagaki, K. (2010). Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon. Chromosome Res. **18**:337–347.

UN (United Nations). (2022). World Population Prospects 2022 (New York: United Nations). https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf.

Wang, L., Zhang, M., Li, M., Jiang, X., Jiao, W., and Song, Q. (2023). A telomere-to-telomere gap-free assembly of soybean genome. Mol. Plant **16**:1711–1714.

Yi, X., Liu, J., Chen, S., Wu, H., Liu, M., Xu, Q., Lei, L., Lee, S., Zhang, B., Kudrna, D., et al. (2022). Genome assembly of the JD17 soybean provides a new reference genome for comparative genomics. G3 (Bethesda). **12**:jkac017.

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