SCIENCE CHINA
Life Sciences

•LETTER TO THE EDITOR•

September 2019 Vol.62 No.9: 1257–1260
https://doi.org/10.1007/s11427-019-9822-2

Update soybean Zhonghuang 13 genome to a golden reference

Yanting Shen[1,2†], Huilong Du[3,4†], Yucheng Liu[2,4], Lingbin Ni[2,4], Zheng Wang[2], Chengzhi Liang[3,4*] & Zhixi Tian[2,4*]

[1]*School of Pharmaceutical Sciences, Guangzhou University of Chinese Medicine, Guangzhou 510006, China;*
[2]*State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing 100101, China;*
[3]*State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing 100101, China;*
[4]*University of Chinese Academy of Sciences, Beijing 100039, China*

Received July 22, 2019; accepted August 19, 2019; published online August 21, 2019

Citation: 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. https://doi.org/10.1007/s11427-019-9822-2

Dear Editor,

Soybean is one of the most important crops worldwide. A high-quality reference genome will facilitate its functional analysis and molecular breeding (Wang and Tian, 2015). Previously, we *de novo* assembled a high-quality Chinese soybean genome Gmax_ZH13 (Shen et al., 2018, Yang and Huang, 2018). However, due to technical limitations at the time when we generated Gmax_ZH13, a large number of small contigs were not anchored onto chromosomes. Therefore, we here build a new golden reference genome for Zhonghuang 13 consisting of 20 nearly complete chromosomes by adding more single-molecule real time (SMRT) sequencing reads. Furthermore, we add large RNA-seq and smRNA-seq datasets for improving the annotation of its protein coding genes.

For genome assembly, we sequenced additional 40 Gb SMRT reads from 5 cells for this update. Therefore, a total of 120× PacBio reads, 365× Bionano optical maps marked by BssSI, 275× Bionano optical maps marked by BspQI, 45× HiSeq reads and 125× chromosome conformation capture sequencing (Hi-C) reads were used for the new genome assembly (Figure 1A). To make best of the sequencing data, we also adopted a new genome assembly pipeline by adding HERA (Du and Liang, 2018) to improve the sequence contiguity and reduce errors by accurately assembling the repetitive genome regions (Figure 1A). Briefly, the whole assembly process differed from previous pipeline at (1) using CANU (v1.7.1) to replace Smrtmake for assembling PacBio subreads to PacBio contigs; (2) using HERA to generate longer contigs; (3) using Juicer and 3D-DNA to replace HiC-Pro and LACHESIS to anchor the hybrid scaffolds into chromosomes with Hi-C reads.

With these improvements, we finally assembled a new version of genome with a length of 1,011,174,350 bp (named Gmax_ZH13_v2.0). Compared with the previous assembly (Gmax_ZH13), the contig N50 size of Gmax_ZH13_v2.0 increased 6.5 times (from 3.46 to 22.6 Mb), the gap number decreased 1.8 times (from 815 to 448) and the gap length decreased 8.8 times (from 20.49 to 2.33 Mb) (Table S1 in Supporting Information). Meanwhile, the un-anchored contig number decreased 17 times (from 549 to 36), resulting in the ratio of sequence that anchored to 20 chromosomes reaching to 98%. Moreover, we found that the mapping ratio of WGS HiSeq reads and RNA isoform sequencing (Iso-seq) reads reached to 99.89% and 99.81%, confirming the high completeness of Gmax_ZH13_v2.0. Besides nuclear chro-

†Contributed equally to this work
*Corresponding authors (Zhixi Tian, email: zxtian@genetics.ac.cn; Chengzhi Liang, email: cliang@genetics.ac.cn)

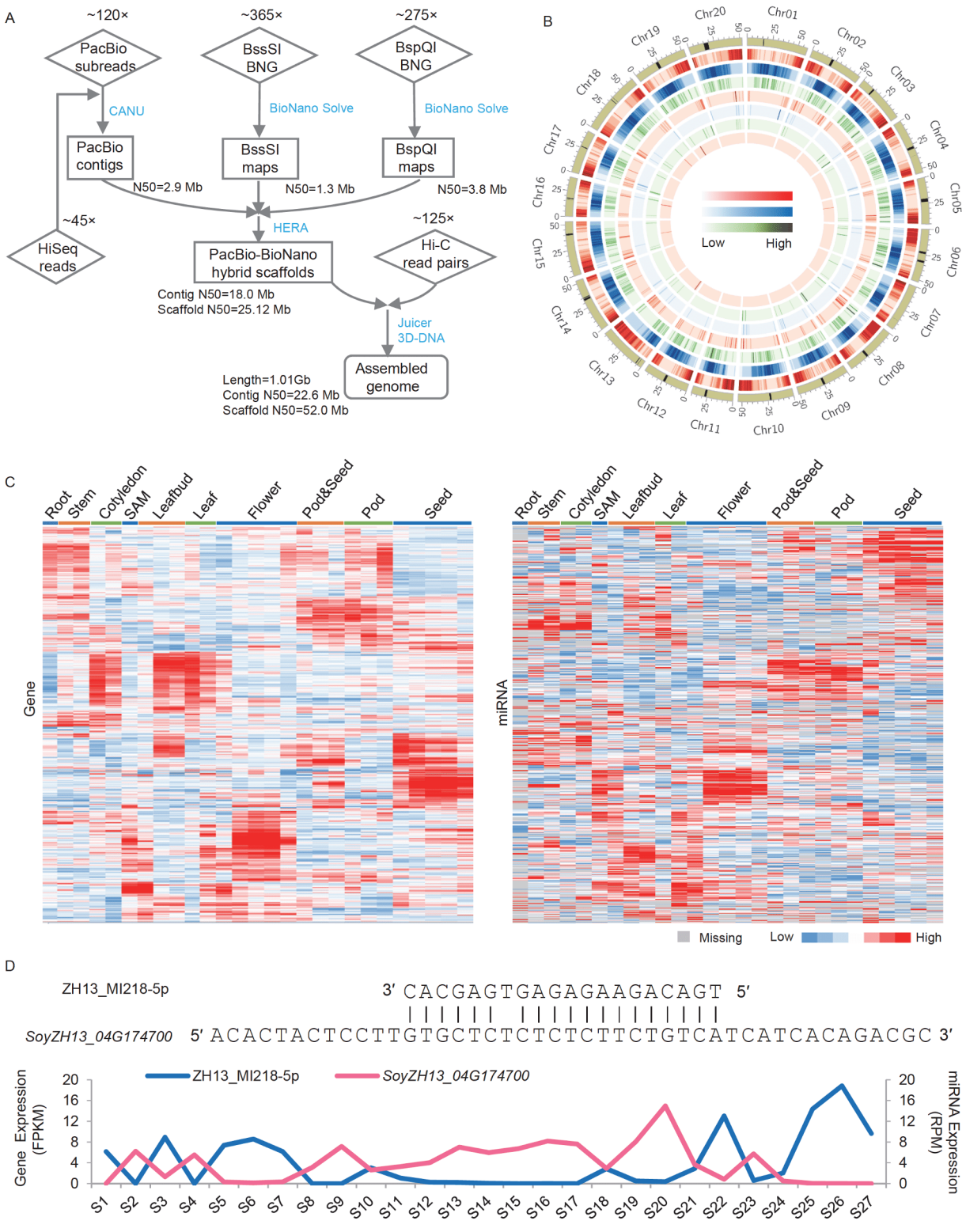

Figure 1 Update of Gmax_ZH13_v2.0 genome. A, Pipeline for genome assembly. B, Distribution of genome features. Tracks from outer to inner circles indicate chromosomes, and density of protein coding genes, repeat sequence, snoRNA, tRNA, miRNA, snRNA and rRNA, respectively. The black blocks on the outer circle indicate regions enriched of Cent91/92 (a soybean-specific centromeric repeat). C, Expression profiling of protein coding genes (left panel) and miRNAs (right panel) in 27 samples from different tissues of different development stages. D, An example of miRNA (top panel) to repress its target gene expression (bottom panel).

mosomes, we assembled the circular genomes of chloroplast and mitochondria with a length of 152,220 bp and 513,779 bp respectively.

To improve the accuracy of gene annotation, we performed RNA-seq for 27 Zhonghuang 13 samples, which were collected from different tissues at different developmental stages. Each sample was sequenced with two replications. In total, 353 Gb reads were produced. These RNA-seq data, together with the Iso-seq reads used for annotation in the last version, were used as expression sequence tag (EST) evidences to predict protein coding genes using MAKER (Cantarel et al., 2008), which was a professional pipeline popularly used in gene prediction. We annotated a total of 55,443 protein coding genes containing 96,366 mRNAs in the nuclear genome (Table S2 in Supporting Information). We found that 97% of the 1,440 single copy Embryophyta genes in BUSCO_v3 were completely assembled, confirming the high quality of our annotation. We found that 42,259 of the newly annotated genes matching to the genes in the last version, whose IDs were therefore inherited from Gmax_ZH13 (Table S2 in Supporting Information). In addition, we annotated 81 and 49 protein coding genes for chloroplast genome and mitochondrial genome respectively (Table S2 in Supporting Information). We also annotated non-coding genes, including 297 rRNA, 1,112 tRNA, 166 snRNA and 1,816 snoRNA (Table S2 in Supporting Information). Notably, we found that the rRNAs were not distributed evenly in each chromosome (Figure 1B). For instance, 94.49% (223/236) TSU rRNAs (5s/5.8s) were located at chromosome 19 from 15,393,361 to 15,465,926 bp (Table S2 in Supporting Information). In addition to genes, we identified 35,926 TEs with clear structural boundaries (Table S3 in Supporting Information). These TEs, together with their homologous truncated fragments and other repeat sequences, made up 53.06% of Gmax_ZH13_v2.0 genome (Tables S4 and S5 in Supporting Information). Moreover, we detected putative centromere regions for each chromosome by searching soybean specific centromere sequences CentGm-1 and CentGm-2 (Table S6 in Supporting Information). The results showed that the length of putative centromere regions for each chromosome exhibited significant variation and some chromosomes even had more than one region that is enriched with CentGm-1 and CentGm-2 (Figure 1B).

To annotate *MIRNA* genes, small RNAs from the same 27 samples used for protein coding gene annotation were sequenced. Using these reads, we annotated 331 *MIRNA* genes. Transcripts of these loci generated 349 pair of miRNAs (miRNA-3p and miRNA-5p) (Table S2 in Supporting Information). Using the RNA sequence and small RNA sequence data of 27 samples, we provided a detailed expression profiling for all the protein coding genes and miRNAs (Figure 1C, Tables S7 and S8 in Supporting Information). These expression profiling data will be helpful for soybean fundamental research, for instance, searching expression pattern of individual genes or choosing tissue specific expression genes. Moreover, the data can be used to investigate the relationship of miRNAs and their target genes because they came from the same sample sets. For example, our previous study indicated that *SoyZH13_04G174700* is a target of miRNA ZH13_MI218-5p (Liu et al., 2016). The expression profiling data demonstrated that the *SoyZH13_04G174700* and miRNA ZH13_MI218-5p exhibited opposite expression patterns, indicating the repression of ZH13_MI218-5p to *SoyZH13_04G174700* (Figure 1D).

In summary, we update the Gmax_ZH13 genome to a more complete and continuous platinum reference genome Gmax_ZH13_ v2.0 in this study. We also provide the expression profiling of mRNA and miRNA for different tissues from different developmental stages. The methods used for the construction of this high-quality genome sequence also can be used in soybean pan-genome projects in the future, which will greatly facilitate soybean fundamental research and molecular breeding.

All the sequencing data used for genome assembly and annotation have been deposited into the Genome Sequence Archive (GSA) database in BIG Data Center under Accession Number CRA001810. Information of Gmax_ZH13_ v2.0 genome and annotation was deposited into the Genome Warehouse (GWH) database in the BIG Data Center under Accession Number GWHAAEV00000000.1. All the assembly and annotation methods are detailed in Supplemental File 1.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Key Research & Development Program of China (2017YFD0101305), National Natural Science Foundation of China (31525018, 31788103), and the State Key Laboratory of Plant Cell and Chromosome Engineering (PCCE-KF-2019-05).*

References

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188–196.

Du, H., and Liang, C. (2018). Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. bioRxiv, http://dx.doi.org/10.1101/345983.

Liu, T., Fang, C., Ma, Y., Shen, Y., Li, C., Li, Q., Wang, M., Liu, S., Zhang, J., Zhou, Z., et al. (2016). Global investigation of the co-evolution of *MIRNA* genes and microRNA targets during soybean domestication. Plant J 85, 396–409.

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

Wang, Z., and Tian, Z.X. (2015). Genomics progress will facilitate molecular breeding in soybean. Sci China Life Sci 58, 813–815.

Yang, J., and Huang, X. (2018). A new high-quality genome sequence in soybean. Sci China Life Sci 61, 1604–1605.

SUPPORTING INFORMATION

Table S1 Comparison of Gmax_ZH13 and Gmax_ZH13_v2.0

Table S2 Genome annotation of protein coding genes and non-coding genes (gff3 format)

Table S3 Chromosome location of TEs with clear structural boundaries

Table S4 Transposable element and repeat sequence composition in the Gmax_ZH13_v2.0 genome

Table S5 Chromosome location of repeat sequences

Table S6 Locations of centromere region for each chromosome

Table S7 Expression pattern of protein coding genes in 27 different soybean samples

Table S8 Expression pattern of miRNA in 27 different soybean samples

Supplemental File 1 Materials and methods

The supporting information is available online at http://life.scichina.com and https://link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.