SCIENCE CHINA
Life Sciences

•RESEARCH HIGHLIGHT•

December 2018 Vol.61 No.12: 1604–1605
https://doi.org/10.1007/s11427-018-9431-8

# A new high-quality genome sequence in soybean

Jun Yang[1*] & Xuehui Huang[2*]

[1]*Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Plant Science Research Center, Chinese Academy of Sciences, Shanghai Chenshan Botanical Garden, Shanghai 201602, China;*
[2]*College of life sciences, Shanghai Normal University, Shanghai 200234, China*

Received October 16, 2018; accepted October 22, 2018; published online November 22, 2018

**Citation:** Yang, J., and Huang, X. (2018). A new high-quality genome sequence in soybean. Sci China Life Sci 61, 1604–1605. https://doi.org/10.1007/s11427-018-9431-8

In the 1940s, Nikolai Ivanovich Vavilov (Russian: Николай Иванович Вавилов) published his manuscript "The theory of origins of cultivated plants after Darwin (Russian: Учение о происхождении культурных растений после Дарвина)". Vavilov believed that among various crops, soybean (*Glycine max*) originated from China. Since then, the cultivated soybean has been demonstrated to be originally domesticated in China approximately 5,000 years ago in the eastern half of North China (Fukuda, 1933; Carter et al., 2004). Northeastern China formerly served as the center of soybean production. In 1908, the Japanese firms achieved the first several soybean shipments to England from Manchuria. In 2018, the US soybean cargo Peak Pegasus docked at Dalian's Beiliang port, Liaoning province, China, to deliver 70,000 tons of US soybeans. An additional 6 million US dollars have been paid for the tariff as Peak Pegasus missed the deadline. Meanwhile, the 70,000 tons of soybeans accounts for less than 1% of the annual soybean import in China. In the US, the rate of on-farm soybean yield improvement from 1924 to 2012 is 23.3 kg per hectare per year. The yield improvement in the US is mainly contributed by the rapid breeding of high-yielding soybean cultivars, particularly the genetically modified ones. In the future, soybean breeding for higher yield and oil content will possibly rely on the information from molecular genetic studies, in which the genome sequence data bear importance. Combining Sanger sequencing and multiple genetic maps, the soybean genome has been decoded several years ago in the cultivar "Williams 82", a widely used US modern cultivar. The released genome has accelerated the basic soybean research and related production.

Considering the advantages of sequencing technology advances, Chinese scientists from Tian Zhixi, Ma Shisong, and Du Jianchang's groups have presented a better reference genome of the Chinese cultivar "Zhonghuang 13" (Shen et al., 2018). A comprehensive comparison between the currently available soybean genomes, "Williams 82" and "Zhonghuang 13," not only identified considerable differences, including structural variations and presence/absence variations, between these cultivars but also pinpointed the sequence differences in the specific gene *F3′5′H*, the genetic basis of the purple flower in "Zhonghuang 13" and the white flower in "Williams 82." This flower color, as an example, demonstrated the power of comparative genome studies in elucidating the genetic basis behind the phenotypic differences among the various cultivars. Such work may accelerate genome re-sequencing studies on soybean. The genome selection and genome editing technologies applied in soybean breeding will benefit the improvement of on-farm soybean yield in China.

The repetitive sequences hindered the efficiency of the plant genome project and connectivity of the reference genome. The PacBio long-read sequencing applied in "Zhonghuang 13" and Oxford Nanopore long-read sequencing in wild tomato (Schmidt et al., 2017) demonstrated that,

*Corresponding authors (Jun Yang, email: shanghai.junyang@gmail.com; Xuehui Huang, email: xhhuang@shnu.edu.cn) to a certain extent, plant scientists can now handle "junk and selfish DNA" at the whole-genome level to elucidate their effects on plant genome evolution and agronomic trials. In conclusion, future studies may provide more insights and more precise dating of the origin and dispersal of cultivated soybean.

Normally, several RNA-seq data sets are sufficient to predict most of protein-coding genes given that highly connected genome sequences are already available (Chen et al., 2017). Shen et al. have predicted the protein-coding genes in "Zhonghuang 13" using Iso-Seq, which features the capacity to catch the full-length transcripts. However, they mixed samples together, thereby lowering the values of the data set. On the other hand, a more comprehensive transcriptome analysis in "Zhonghuang 13" could be presented. Notably, Shen et al. considered the advantages of large-scale RNA-seq data sets during gene function analysis using co-expression network based on a graphical Gaussian model. Combined with some prior knowledge (Zhang et al., 2017), they have shown the straightforward pipeline to narrow down the candidate genes controlling the flowering time and linoleic acid content from high-resolution genetic mapping. The correlation between the identified haplotypes of the candidate genes and latitude distribution among the natural population further accelerated the demand for functional validation studies, particularly regarding the possible mechanisms and evolution routines of different haplotypes in "Zhonghuang 13" and other cultivars.

In conclusion, the high quality genome sequence of Chinese soybean "Zhonghuang 13" has provided a foundational resource for genetic studies on soybean (Figure 1). Considering that soybean cultivars and its wild relatives grow in diverse ecogeographic areas and feature a high level of genetic diversity, the *de novo* construction of the complete genome sequences for multiple diverse soybean accessions could be performed to generate the pangenome data sets in soybean. As applied in other crops (Hirsch et al., 2014; Zhao et al., 2018), the soybean pangenome, once available, may further facilitate the mining of allelic variations and promote the genetic studies on soybean.

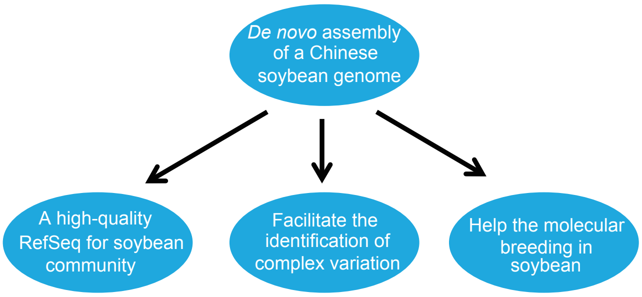
*De novo* assembly of a Chinese soybean genome
A high-quality RefSeq for soybean community
Facilitate the identification of complex variation
Help the molecular breeding in soybean

**Figure 1** (Color online) Utility of a new high-quality reference genome sequence in soybean using PacBio-based sequencing and *de novo* assembly of a Chinese cultivar "Zhonghuang 13."

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

Carter, T.E., Nelson, R., Sneller, C.H., and Cui, Z. (2004). Soybeans: Improvement, Production and Uses, Third edition (agronomy) (Madison, Wisconsin, USA).

Chen, G., Shi, T., and Shi, L. (2017). Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60, 116–125.

Fukuda, Y. (1933). Cytogenetical studies on the wild and cultivated Manchurian soybeans (*Glycine L*). Jpn J Bot 6, 489-506.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121–135.

Schmidt, M.H.W., Vogel, A., Denton, A.K., Istace, B., Wormit, A., van de Geest, H., Bolger, M.E., Alseekh, S., Maß, J., Pfaff, C., et al. (2017). *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. Plant Cell 29, 2336–2348.

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.

Zhang, S.R., Wang, H., Wang, Z., Ren, Y., Niu, L., Liu, J., and Liu, B. (2017). Photoperiodism dynamics during the domestication and improvement of soybean. Sci China Life Sci 60, 1416–1427.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50, 278–284.