SCIENCE CHINA
Life Sciences

•COVER ARTICLE•

August 2018 Vol.61 No.8: 871–884
https://doi.org/10.1007/s11427-018-9360-0

De novo assembly of a Chinese soybean genome

Yanting Shen[1,5], Jing Liu[2], Haiying Geng[3], Jixiang Zhang[1,5], Yucheng Liu[1,5], Haikuan Zhang[4], Shilai Xing[4], Jianchang Du[2*], Shisong Ma[3*] & Zhixi Tian[1,5*]

[1]*State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*Provincial Key Laboratory of Agrobiology, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China;*
[3]*School of Life Sciences, University of Science and Technology of China, Hefei 230027, China;*
[4]*Berry Genomics Corporation, Beijing 100015, China;*
[5]*University of Chinese Academy of Sciences, Beijing 100039, China*

Received June 15, 2018; accepted July 5, 2018; published online July 27, 2018

Soybean was domesticated in China and has become one of the most important oilseed crops. Due to bottlenecks in their introduction and dissemination, soybeans from different geographic areas exhibit extensive genetic diversity. Asia is the largest soybean market; therefore, a high-quality soybean reference genome from this area is critical for soybean research and breeding. Here, we report the *de novo* assembly and sequence analysis of a Chinese soybean genome for "Zhonghuang 13" by a combination of SMRT, Hi-C and optical mapping data. The assembled genome size is 1.025 Gb with a contig N50 of 3.46 Mb and a scaffold N50 of 51.87 Mb. Comparisons between this genome and the previously reported reference genome (*cv. Williams 82*) uncovered more than 250,000 structure variations. A total of 52,051 protein coding genes and 36,429 transposable elements were annotated for this genome, and a gene co-expression network including 39,967 genes was also established. This high quality Chinese soybean genome and its sequence analysis will provide valuable information for soybean improvement in the future.

***de novo* soybean genome, Zhonghuang 13, Gmax_ZH13, structure variation, gene co-expression network**

Citation: 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. https://doi.org/10.1007/s11427-018-9360-0

INTRODUCTION

Soybean (*Glycine max* [L.] Merr.) is one of the most important crops, providing more than half of global oilseed production and more than a quarter of the world's protein for food and animal feed (Wilson, 2008). Studies have indicated that the cultivated soybean was domesticated from its annual wild relative (*Glycine soja* (Sieb. and Zucc.)) in China approximately 5,000 years ago (Carter et al., 2004). After domestication, cultivated soybeans were introduced to Korea and Japan approximately 2,000 years ago, then to North America in 1765 and to Central and South America during the first half of the last century (Wilson, 2008). Currently, cultivated soybeans are planted extensively worldwide, including in Asia (China, Japan, Korea, and India), North America (United States of America and Canada), and South America (Brazil, Argentina, and Paraguay).

To meet the ever-increasing consumer demand, soybean breeders have made considerable efforts to develop elite varieties. Soybean yield has increased dramatically in the last several decades. However, it is predicted that the rate of yield increase in soybean need to be further accelerated, from the current rate 1.3% to at least 2.4% per year (Ray et al., 2013)

*Corresponding authors (Zhixi Tian, email: zxtian@genetics.ac.cn; Shisong Ma, email: sma@ustc.edu.cn; Jianchang Du, email: dujianchang@hotmail.com)

to meet the food consumption demands of an increasing world population (Foley et al., 2011). A better understanding of the underlying genetic bases of important agronomical traits will expedite the progress of marker-assisted breeding programs for soybean (Wang and Tian, 2015).

The first soybean genome was sequenced for Williams 82 (Glycine_max_v.1.0), a cultivated soybean developed in the 1980s (Schmutz et al., 2010). This reference genome opened the door to soybean functional genomics (Chan et al., 2012; Wang and Tian, 2015). Although more than 45,000 accessions had been developed during the long history of soybean cultivation (Carter et al., 2004), less than 0.02% of landraces (approximately 80 accessions) were used as progenitors for cultivars development in North American (Gizlice et al., 1994), which resulted in a diversity-reduced introduction bottleneck (Hyten et al., 2006). Intergenomic comparisons demonstrated that due to the genetic bottleneck during soybean domestication, wild soybeans and cultivated soybeans exhibited quite a number of lineage-specific genes and genes with copy number variation (CNV) or large-effect mutations (Lam et al., 2010; Li et al., 2013; Li et al., 2014). Further investigation suggested that even the cultivated soybeans from different geographic areas exhibited extensive genetic diversity (Li et al., 2010; Zhou et al., 2015). A recent comparison between Chinese and American soybeans revealed that the genetic basis of Chinese soybeans is distinct from that of those in the USA (Liu et al., 2017). Genetic variations among different accessions, particularly the presence/absence variations and CNVs, are highly associated with agronomic traits (Li et al., 2014; Zhou et al., 2015; Wang et al., 2016; Wei and Cao, 2016). Therefore, a reference genome from one cultivated soybean may not fully represent the genetic diversities, particularly for soybeans from Asia with large genetic variations.

"Zhonghuang 13" is a soybean cultivar bred by Chinese scientists in 2001. This accession was derived from cultivar accessions "Yudou 18" and "Zhongzuo 90052-76" by pedigree selection and exhibits high yield capacity and high stress tolerance. Here, we *de novo* assembled the genome of "Zhonghuang 13" (Gmax_ZH13) in high-quality. Furthermore, we established a comprehensive gene co-expression network using public available soybean RNA-seq expression datasets and used this network to help to mine candidate key genes in controlling agronomically important traits. This genome will facilitate soybean genomics research and elite cultivar improvement.

RESULTS

Genome sequencing, assembly and annotation

We sequenced "Zhonghuang 13" genomic DNA using different approaches, including single-molecule real-time (SMRT) sequencing, optical mapping, chromosome conformation capture sequencing (Hi-C) and next generation sequencing (HiSeq). In total, we generated 80.67 Gb (~80×) SMRT sequences, 45.03 Gb paired HiSeq reads with a length of 250 bp, 638,970 Mb BioNano single-molecular maps (>150 kb) and 125.84 Gb (~125×) Hi-C reads (Table S1 in Supporting Information). We conducted the assembly in a stepwise fashion following a previously reported approach (Bickhart et al., 2017). The initial assembly using PacBio SMRT data alone generated 1,559 contigs with an N50 length of 2.6 Mb. These contigs were corrected with HiSeq sequences, after which they were scaffolded using BioNano optical mapping data and further clustered using Hi-C data. Finally, we assembled the genome (Gmax_ZH13) into 836 contigs with an N50 length of 3.46 Mb (Table 1), of which 287 were assembled into 21 scaffolds with an N50 length of 51.87 Mb. The assembled genome size was 1.025 Gb and the 21 scaffolds constituted approximately 97% of the whole genome. These scaffolds were named chromosome 1–20 and chloroplast following previously reported order based on synteny analysis (Schmutz et al., 2010). We assessed the completeness of our assembled genome by remapping the HiSeq and isoform sequencing (Iso-Seq) reads, and found that more than 99.87% of HiSeq reads and 99.98% of Iso-Seq reads could align properly, indicating high completeness of the assembled genome.

To analyze repetitive sequences, we searched the genome sequence via a combination approach of *de novo* structure-based analysis and homology-based comparisons referring to previous methods (Schmutz et al., 2010). A total of 36,429 transposable elements (TEs) with clear structural boundaries were identified (Table 2, Supplemental File 1 in Supporting Information). As found in other plant genomes, long terminal repeat (LTR)-retrotransposons were the most abundant elements, including 12,641 *Copia*-like, 17,935 *Gypsy*-like and 100 unclassified LTR elements, representing 84.2% of all the identified TEs. The ratios of intact LTRs to solo LTRs for *Copia*-like elements (5,089 vs. 7,552) and *Gypsy*-like elements (9,214 vs. 8,721) were quite similar. We also identified a total of 330 long interspersed nuclear element (LINE)-retrotransposons. In addition to class Ⅰ retrotransposons, 5,423 class Ⅱ DNA transposons were identified, including 7 Tc1/Mariners, 42 hATs, 2,242 Mutators, 71 PIF/Harbingers, 10 Pongs, 59 CACTAs, 2,918 MITEs, and 74 Helitrons. These TEs, together with abundant truncated elements and other repetitive fragments, made up 52.75% of the Gmax_ZH13 genome.

To predict protein-coding genes, we sequenced the transcriptomes of roots, stems, leaves, flowers and seeds of "Zhonghuang 13" using Iso-Seq. In total, ~8.93 Gb Iso-Seq reads were generated (Table S1 in Supporting Information). Combining *ab initio* prediction, protein-based homology searches and transcript evidence gathered from Iso-Seq se-

Table 1 Assembly statistics of the soybean Gmax_ZH13 genome

Assembly[a)]	Contigs[b)]	Scaffolds	Unplaced contigs[c)]	Contig N50 (Mb)[d)]	Scaffold N50 (Mb)[d)]	Assembly size (Gb)	Assembly in scaffolds (%)
PacBio	1,559	–	–	2.6	–	1.007	–
BioNano-BspQI	–	518	–	–	3.79	1.012	–
BioNano-BssSI	–	1,181	–	–	1.3	1.031	–
PacBio+BioNano	826	59	717	3.46	25.12	1.025	96.85
PacBio+BioNano +Hi-C	836	21	549	3.46	51.87	1.025	97

a) Assemblies are listed as the steps of combining different sequence data types for genome assembly. b) The number of continuous stretches of sequence within the scaffold without gaps >3 bases in length of at least 100 bases. c) Unplaced contigs are defined as input contigs that were not placed by the optical map or Hi-C in a scaffold. d) All N50 values are based on the Gmax_ZH13 assembled size.

Table 2 Transposable element and repeat sequence composition in the Gmax_ZH13 genome

Repeat type		Classification		Intact/Solo number[a)]	DNA content (bp)	DNA content (%)
Class I: Retrotransposon	LTR-Retrotransposon	Ty1/copia		12,641	106,803,505	10.42%
		Ty3/gypsy		17,935	331,019,054	32.29%
		Others		100	4,012,789	0.39%
	Non-LTR Retrotransposon	LINE		330	9,300,199	0.91%
		SINE			399,615	0.04%
Class II: DNA Transposon	Subclass I:	Tc1/Mariner		7	144,681	0.01%
		hAT		42	219,006	0.02%
		Mutator		2,242	22,699,372	2.22%
		PIF/Harbinger		71	1,169,142	0.11%
		Pong		10	352,522	0.03%
		CACTA		59	6,404,368	0.62%
		MITE	Tourist	1,356	1,212,517	0.11%
			Stowaway	1,562	1,110,643	0.11%
	Subclass II:	Helitron		74	2,905,865	0.29%
	Tandem Repeat				10,546,540	1.03%
	Unknown				42,453,694	4.14%
	Total			36,429	540,753,512	52.75%

a) Number of transposable elements with clear boundaries and signatures of insertion sites.

quences (Chen et al., 2017), we identified 52,051 protein-coding genes (Supplemental File 2 in Supporting Information). Compared with 1,440 single copy orthologs from Embryophyta in Plantae BUSCO v2 dataset (Simão et al., 2015), 97.01% of these genes were completely assembled and annotated in our genome, 0.83% of these genes were incomplete, and only 2.15% were not assembled or annotated, suggesting that the gene annotation in this study has high completeness.

A comparison between the annotated genes from Gmax_ZH13 and Glycine_max_v2.0 showed that most of them were shared (Table S2 in Supporting Information). Using the criterion of minimum alignment identity of 90%, 45,068 (86.58%) genes from Gmax_ZH13 corresponded to 45,882 (81.77%) genes from Glycine_max_v2.0. Further investigation classified these corresponding gene pairs into four classes: class 1, identical in both length and sequence, which contained 24,685 Gmax_ZH13 genes and 24,683 Glycine_max_v2.0 genes; class 2, having the same length but with SNPs or small indels, which contained 8,132 Gmax_ZH13 genes and 8,134 Glycine_max_v2.0 genes; class 3, having minimum alignment coverage of 80%, which contained 11,036 Gmax_ZH13 genes and 11,576 Glycine_max_v2.0 genes; and class 4, with alignment coverage less than 80%, which contained 1,733 Gmax_ZH13 genes and 1,774 Glycine_max_v2.0 genes (Table S2 in Supporting Information). Besides these corresponding gene pairs, we also identified 6,983 and 10,162 lineage-specific genes in Gmax_ZH13 and Glycine_max_v2.0 respectively, which may result from the sequence variations between the two genomes and different annotation pipelines.

In addition to the nuclear genome, we assembled a chloroplast genome using the SMRT and HiSeq sequences. The total length of Gmax_ZH13 chloroplast genome is 152.22

kb. The chloroplast genome contains 127 genes, including 90 protein coding genes, 29 tRNA genes and 8 rRNA genes (Supplemental File 2 in Supporting Information). The chloroplast genome size and gene composition were similar to those from a previous assemble (Saski et al., 2005) and other plant species (Du et al., 2017; Guo et al., 2017).

Genome comparison to the Williams 82 reference genome

The Williams 82 soybean genome was first released in 2010 (Schmutz et al., 2010) and was updated in 2015 (Glycine_max_v2.0). In addition to Williams 82, two additional genomes from soybean cultivars Enrei (Glycine max_Enrei_2.0) (Shimomura et al., 2015) and Lee (Glyma.Lee.gnm1) were released in 2015 and 2018, respectively. Comparing to these three previously released soybean genomes, Gmax_ZH13 had the longest total sequence length, highest contig N50 and scaffold N50 and fewest contig number (Table S3 in Supporting Information). Detailed investigation showed that Gmax_ZH13 had only 815 gaps in the chromosomes and 549 unplaced scaffolds, whereas Glycine_max_v2.0 had 12,761 gaps and 1,169 unplaced scaffolds (Table S4 in Supporting Information, Figure 1).

Synteny analysis demonstrated that although Gmax_ZH13 and Glycine_max_v2.0 exhibited high chromosome-level similarity, with 867.26 Mb (84.60%) of Gmax_ZH13 sequences aligned to 868.99 Mb (88.77%) of Glycine_max_v2.0 sequences, these two genomes also showed a certain number of significant structure variations (SVs), including translocations, inversions and presence variations (Figure 1). In total, we identified 1,404 translocations including 637 inter-translocations (occurred between chromosomes; approximately 3.82 Mb) and 767 intra-translocations (occurred within chromosomes; approximately 17.17 Mb) (Table S5 in Supporting Information). The largest translocation was 2.22 Mb locating from 13.32 to 15.56 Mb on chromosome 5 in Gmax_ZH13, which was anchored to Glycine_max_v2.0 from 18.33 to 20.52 Mb on chromosome 5. We also identified 161 inversion events (approximately 8.57 Mb) within chromosomes (Table S6 in Supporting Information). The most distinct inversion occurred at chromosome 11 from 27.78 to 30.00 Mb in Gmax_ZH13, which assembled as 22.23 to 24.6 Mb in Glycine_max_v2.0 in an opposite direction. In addition to these translocation and inversion events, there were also some regions not only translocated but also inverted (translocation & inversion) between Gmax_ZH13 and Glycine_max_v2.0 genomes. We detected 528 and 705 of these translocation & inversion events within chromosomes (inter-translocation & inversion) and between chromosomes (intra-translocation & inversion), respectively (Table S7 in Supporting Information). The inter-translocation & inversion events had a total length of 5.82 Mb, and the intra-translocation & inversion events had a total length of 19.54 Mb. The largest one was 0.91 Mb in length from 16.36 to 17.29 Mb on chromosome 1 in Gmax_ZH13, which aligned to Glycine_max_v2.0 from 15.00 to 14.10 Mb on chromosome 1.

Comparing to the Glycine_max_v2.0 genome, the Gmax_ZH13 sequence was 46.13 Mb longer, and this length was mainly occupied by repeat sequences. In addition to these repeat sequences, we found 12,170 presence variations (PVs, only fragments >100 bp were counted) in Gmax_ZH13 (Table S8 in Supporting Information) and 5,239 PVs in Glycine_max_v2.0 (Table S9 in Supporting Information). The 12,170 PVs from Gmax_ZH13 accounted for 12.07 Mb and contained 1,365 genes; and the 5,239 PVs from Glycine_max_v2.0 accounted for 3.44 Mb and contained 641 genes. In addition to the longer PVs, we also detected 255,971 small insertions (1–99 bp) in the Gmax_ZH13 (accounting for 1.20 Mb) and 249,535 small insertions in the Glycine_max_v2.0 genome (accounting for 1.16 Mb) (Table S10 in Supporting Information). To check whether these SVs were real or incorrect assembly, we randomly picked up seven SVs and performed validation by PCR. Our results demonstrated that all of the SVs could be detected, confirming the correct assembly of the Gmax_ZH13 genome (Figure S1 in Supporting Information).

We found that some of the genetic variations were associated with phenotypic changes in these two cultivars. The *F3'5'H* (*SoyZH13_13G057600*) gene has been reported to control soybean flower color and a substitution of 65 bp tandem repeat by 12 bp in the third exon of *F3'5'H* was responsible for a change from purple flower to white flower (Zabala and Vodkin, 2007). Zhonghuang 13 has purple flower and Williams 82 has white flowers. This polymorphism was identified at the corresponding locus when the sequences from Zhonghuang 13 and Williams 82 were compared (Figure S2 in Supporting Information).

Gene co-expression network assists mining of important agronomic genes

Gene co-expression network is a popular approach to explore gene regulatory relationships (Oldham et al., 2006; Wolfe et al., 2005; Rhee and Mutwil, 2014). The genes in the same module have similar expression patterns and a tendency toward the same biological function (Serin et al., 2016). Gene co-expression networks can be used to predict gene function (Ma et al., 2007; Childs et al., 2011) and to identify key genes in biological pathways (Krouk et al., 2010; Le et al., 2010; Windram et al., 2012; Wei et al., 2013). Graphical Gaussian model (GGM), which employs partial correlation coefficient to measure direct correlation between genes, is a robust method for co-expression network analysis, and GGM gene

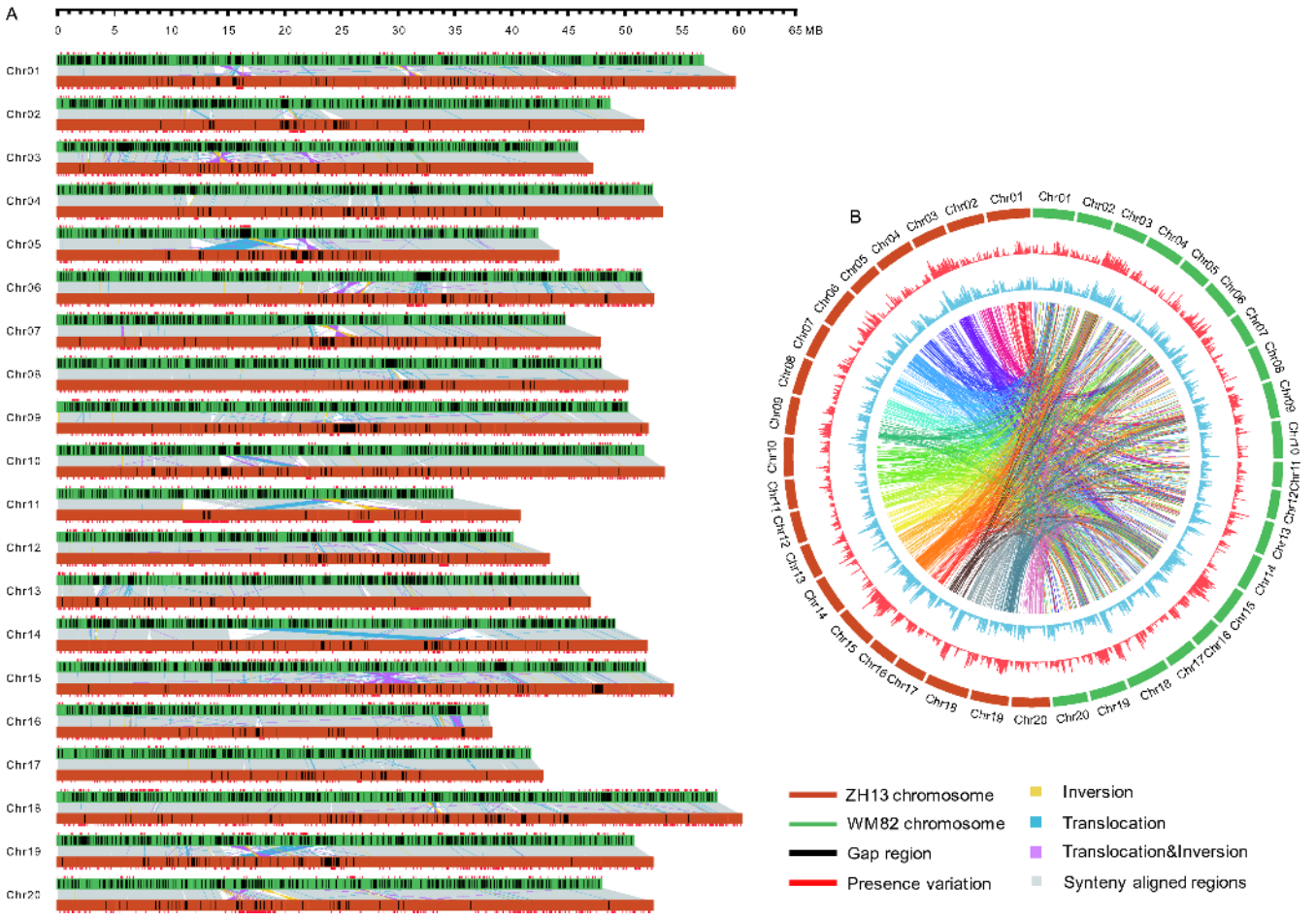


Figure 1 Whole-genome comparison between Gmax_ZH13 and Glycine_max_v2.0. A, Intra-chromosome comparisons. Gaps in assembled chromosomes, specifically presence regions, synteny aligned regions, inversion regions, translocation regions and translocation & inversion regions are included. B, Inter-chromosome comparisons. Tracks from outer to inner circles indicate SNP number and small insertion number, lines between each chromosome show translocation or translocation & inversion events.

networks have been built previously for *Arabidopsis* and maize (Schäfer and Strimmer, 2005; Ma et al., 2007; Ma et al., 2015; Ma et al., 2017). To promote practical utility of the Gmax_ZH13 genome, we constructed a soybean GGM gene co-expression network based on the genes from Gmax_ZH13 using transcriptome datasets from 1,978 soybean RNA-seq runs deposited in the NCBI Sequence Read Archive (SRA), which span more than 25 different tissues at different developmental stages. The established Gmax_ZH13 GGM network contained 39,967 (76.78%) genes and 330,864 co-expressed gene pairs (Supplemental File 3 in Supporting Information).

Quantitative trait locus (QTL) and genome wide association study (GWAS) are commonly used to explore the genes controlling agronomic traits, but they usually result in large candidate regions that make it hard to mine the causal genes. A combination of QTL/GWAS and gene co-expression network techniques may help to identify causal genes in a specific QTL/GWAS region. We tested this approach by exploring new genes controlling soybean flowering time and linoleic acid content.

Flowering time is a complex agronomic trait, which is controlled by multiple QTLs (Zhang et al., 2017). To date, at least 41 GWAS and 92 QTL regions have been detected (Keim et al., 1990; Mansur et al., 1993; Mansur et al., 1996; Orf et al., 1999; Yamanaka et al., 2000; Tasma et al., 2001; Yamanaka, 2001; Zhang et al., 2004; Funatsuki et al., 2005; Pooprompan et al., 2006; Reinprecht et al., 2006; Gai et al., 2007; Githiri et al., 2007; Komatsu et al., 2007; Khan et al., 2008; Palomeque et al., 2009; Oyoo et al., 2010; Kuroda et al., 2013; Jun et al., 2014; Contreras-Soto et al., 2017; Fang et al., 2017). These reported QTL/GWAS regions contained 7,971 genes in the Gmax_ZH13 genome, with an average of approximately 60 genes for each region. To mine the causal gene from 60 candidates is challenging. So far, nine genes have been functionally validated to control flowering time in soybean, including *E1* (*SoyZH13_06G195900*)(Xia et al., 2012), *E2* (*SoyZH13_10G204600*)(Watanabe et al., 2011), *E3* (*SoyZH13_19G210400*)(Kong et al., 2010), *E4* (*SoyZH13_20G079700*)(Kong et al., 2010), *E9*

(*SoyZH13_16G134200*)(Kong et al., 2014; Zhao et al., 2016), *E10* (*SoyZH13_08G341400*)(Samanfar et al., 2017), *J* (*SoyZH13_04G047800*)(Yue et al., 2017; Lu et al., 2017), *GmFT2b* (*SoyZH13_16G134600*)(Kong et al., 2010) and *GmFT5a* (*SoyZH13_16G040400*)(Kong et al., 2010). In the gene co-expression network we established, 152 genes were directly connected to these 9 reported genes controlling flowering time (Figure 2A, Table S11 in Supporting Information), among which 26 genes were located in the flowering time QTL/GWAS regions. Moreover, 4 of these 26 genes were homologous to genes controlling flowering time in *Arabidopsis* (Figure 2B). Among these, we investigated *SoyZH13_16G177400* in detail to see if it was a candidate causal gene for flowering time in soybean.

SoyZH13_16G177400 was located in QTL regions reported by two independent studies (Pooprompan et al., 2006; Mao et al., 2017). Its homologous gene in *Arabidopsis* encodes a MADS-box transcription factor negatively regulating the FLC/MAF clade genes and positively regulating *FT* (Koo et al., 2010). In the gene co-expression network, *SoyZH13_16G177400* was connected to three reported genes: *GmFT2b*, *GmFT5a* and *E9* (Figure 2A). We detected five nonsynonymous mutation sites for *SoyZH13_16G177400* in a natural population we previously re-sequenced (Fang et al., 2017), which classifed this gene into six haplotypes (Figure 3A). When the days to flowering of each accession were evaluated, we found that different haplotypes were significantly different from each other, with haplotype H1 having the shortest flowering time and haplotype H6 having the longest flowering time. This phenotypic variation in the flowering times of different haplotypes indicated that *SoyZH13_16G177400* might be a gene responsible for soybean flowering time in the natural population. We illustrated the phylogenetic relations among haplotypes from 809 previously re-sequenced accessions (Fang et al., 2017) for *SoyZH13_16G177400*. The results showed that accessions from high-latitude areas mainly contained haplotype H1, which relates to shorter days to flowering, while accessions from low-latitude areas contained higher proportion of the other five haplotypes, which relates to longer days to flowering (Figure 3B).

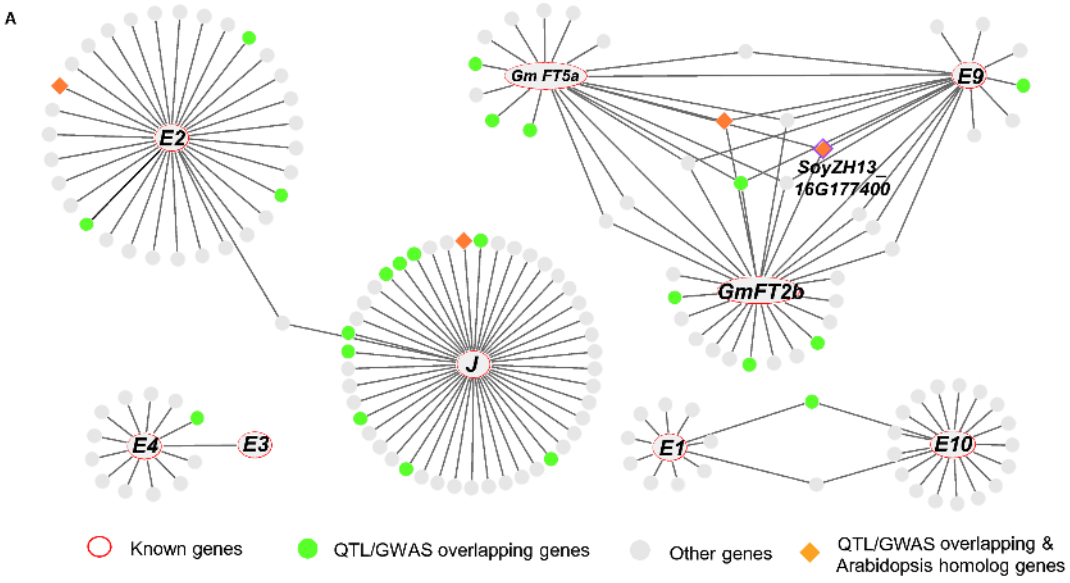
Linoleic acid and linolenic acid are two important fatty acid compounds. The ratio of linoleic acid to linolenic acid is important for the quality of soybean oil. *FAD3A* (*SoyZH13_14G178100*) was found to be an important gene involved in transforming linoleic acid to linolenic acid in soybean, and mutation of *FAD3A* resulted in higher ratio of linoleic acid to linolenic acid (Byrum et al., 1995). In the gene co-expression network we constructed, three genes were found to directly connect to *FAD3A* (Figure 4A). Among them, *SoyZH13_02G207800* was located in a linoleic acid content QTL reported by Kim et al. in 2010 (Kim et al., 2010) and its homologous gene in *Arabidopsis* was annotated as fatty acid desaturase. In the same natural population that we used in the flowering time analysis, only one nonsynonymous mutation located on the second exon of *SoyZH13_02G207800* was detected, which classified the accessions into two haplotypes. The accessions of each haplotype had significant differences in linoleic acid content (Figure 4B). Therefore, *SoyZH13_02G207800* might be another candidate gene controlling transformation from linoleic acid to linolenic acid in soybean.

DISCUSSION

A high-quality reference genome is crucial for functional analysis of a species. As an increasing number of reference genomes have been assembled (VanBuren et al., 2015; Hoshino et al., 2016; Badouin et al., 2017; Clavijo et al., 2017; Schmidt et al., 2017; Raymond et al., 2018), genetic diversity between different populations, ethnic groups, varieties and individuals has been revealed. For instance, a surprisingly large number of SVs were identified between two *indica* rice varieties, Zhenshan 97 and Minghui 63 (Zhang et al., 2016), and between maize Kill, W22 and B73 (Jiao et al., 2017). Sometimes, SVs have large effects on phenotype determination (Lupski et al., 1991; Dooner and He, 2008; Studer et al., 2011; Lv et al., 2018). Therefore, one reference genome can not represent the overall genetic information of a species, and more assembled genomes from different accessions/individuals are required. For important species, such as human, *Arabidopsis*, rice and maize, more than one genome have been assembled (Shi et al., 2016; Hirsch et al., 2016; Kawakatsu et al., 2016; Seo et al., 2016; Zhang et al., 2016; Du et al., 2017).

As one of the most economically important crops, soybean has undergone strict genetic bottlenecks during cultivation, resulting in accessions from different geographic areas possibly exhibiting high genetic diversity. The current soybean reference genome was sequenced from Williams 82, which is a cultivar domesticated in America. It is necessary to assemble another genome from an Asian soybean accession because Asia is one of the largest soybean planting area. In this study, we report a high-quality reference genome for a Chinese soybean cultivar “Zhonghuang 13”. With a contig N50 length of 3.46 Mb and a scaffold N50 length of 51.87 Mb, our assembled genome is more contiguous than those of most reported plant genomes, including *Oropetium thomaeum* (contig N50=2.39 Mb) (VanBuren et al., 2015), *indica* rice (Zhenshan 97 contig N50=2.3 Mb, Minghui 63 contig N50=3.1 Mb) (Zhang et al., 2016), *Zea mays* B73 (contig N50=1.2 Mb) (Jiao et al., 2017), the Japanese morning glory *Ipomoea nil* (contig N50=1.87 Mb) (Hoshino et al., 2016) and *Chenopodium quinoa* (scaffold N50=3.85 Mb) (Jarvis et al., 2017). To the best of our knowledge, it is

A



B

Gene_ID	QTL/GWAS/Homolog	Gene_ID	QTL/GWAS/Homolog
SoyZH13_02G122900	First_flower_16-2	*SoyZH13_10G204500*	Ref1,First_flower_24-4
SoyZH13_02G127700	First_flower_16-2	*SoyZH13_11G114800*	First_flower_11-1
SoyZH13_02G136600	First_flower_13-2	*SoyZH13_12G067700*	Ref1,First_flower_25-2
SoyZH13_02G137400	First_flower_13-2	*SoyZH13_15G091000*	First_flower_12-3
SoyZH13_02G139100	First_flower_13-2	*SoyZH13_15G145400*	First_flower_12-3
SoyZH13_02G140500	First_flower_13-2	*SoyZH13_16G111400*	First_flower_9-3
SoyZH13_02G143300	First_flower_13-2	*SoyZH13_16G177400*	First_flower_13-8,*AGL6*
SoyZH13_02G204100	First_flower_15-1	*SoyZH13_18G065500*	First_flower_9-2
SoyZH13_06G245300	First_flower_2-1	*SoyZH13_19G139600*	First_flower_15-2,First_flower_2-3,First_flower_4-3
SoyZH13_06G186500	First_flower_18-2,First_flower_26-2/5/6/7/8/9/12, *TOC1*	*SoyZH13_19G189400*	First_flower_22-3,First_flower_3-3
SoyZH13_06G194000	Ref1,First_flower_12-2,First_flower_5-1,First_flower_26-9/12/17, First_flower_4-1,*GmFULa*	*SoyZH13_19G204700*	First_flower_13-9,First_flower_16-4
SoyZH13_06G194400	Ref1,First_flower_12-2,First_flower_5-1,First_flower_26-9/12/17,First_flower_4-1	*SoyZH13_19G209600*	First_flower_16-4,First_flower_20-2
SoyZH13_07G031600	First_flower_4-2	*SoyZH13_19G244500*	First_flower_16-4,First_flower_5-3,First_flower_8-3,*APRR5*

Figure 2 Combining gene co-expression network and QTL/GWAS regions to predict soybean flowering time related genes. A, Genes co-expressed with 9 known soybean flowering time related genes at the first level. Nodes represent genes, and edges represent connections between genes. Edge width correlates to the connected genes' expression pattern similarity; the thicker the edge, the higher the expression correlation its connected genes have. B, 26 soybean flowering time related genes predicted by GWAS and/or QTL regions appearing in (A). Ref1 is Fang et al., 2017.

the fourth contiguous plant genome reported to date, together with rice R498 (Du et al., 2017), *Arabidopsis* TAIR 10 (https://www.arabidopsis.org/) and rose (Raymond et al., 2018) genomes. Comparing our assembled genome with the previously reported soybean reference genome, a large number of SVs and accession specific genes were identified, revealing the genetic diversities between accessions from different geographies. Previous study indicated that assembly errors in the previous reference genome may affect gene identification (Fang et al., 2017). We found that some assembly errors could be resolved by the higher quality genome sequences of Gmax_ZH13 (Figure S3 in Supporting Information). In addition, we established a GGM gene co-expression network based on the annotated genes of Gmax_ZH13, which will facilitate the identification of candidate genes controlling specific traits in combination with QTL/GWAS and homologs search.

In summary, a high quality Chinese soybean genome Gmax_ZH13 was assembled and annotated in this study. This new genome will facilitate legume genomics research

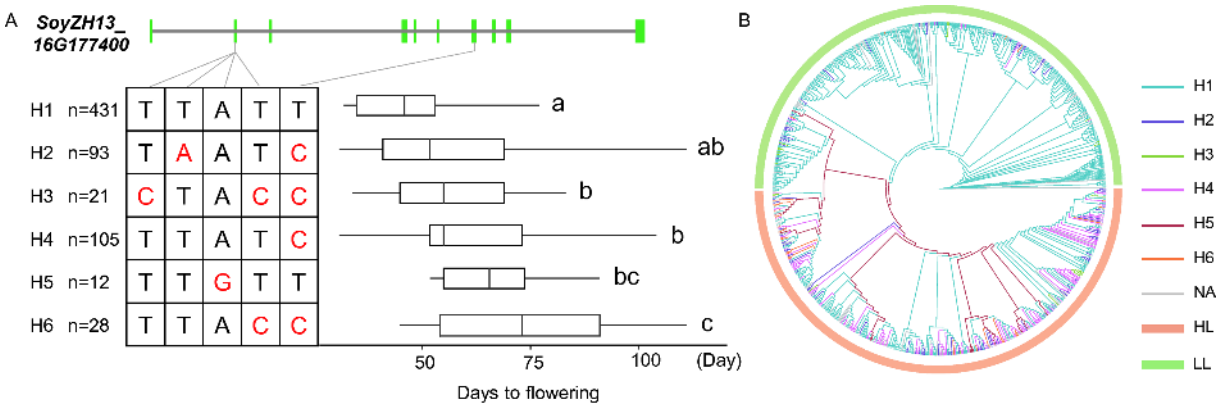


Figure 3 *SoyZH13_16G177400* is a gene controlling soybean flowering time. A, Different haplotypes of *SoyZH13_16G177400* show significantly different flowering times. Green blocks indicate the gene's CDS region and "H" is an abbreviation for "haplotype". Nucleotides marked in red are the mutant forms compared to Gmax_ZH13 genome. The different letters to the right of each column indicate significant differences by ANOVA test (P<0.01). B, Geographic distribution of accessions roughly in accordance with their haplotypes of *SoyZH13_16G177400*. The phylogenetic tree is modified from Figure 1b of the reference (Fang et al., 2017). "HL" is an abbreviation for "high latitude", and "LL" is an abbreviation for "low latitude".

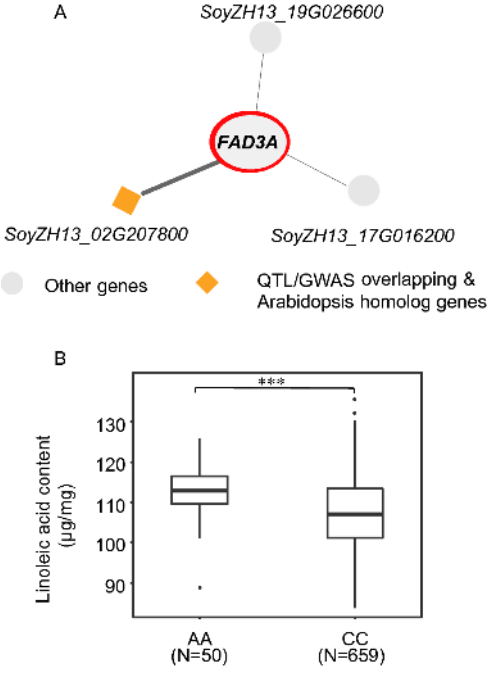


Figure 4 Combining gene co-expression network and QTL/GWAS regions to predict linoleic acid content related genes. A, Genes co-expressed with *FAD3A* at the first level. Nodes represent genes and edges represent connections between genes. Edge width correlates to the connected genes' expression pattern similarity; the thicker the edge, the higher the expression correlation its connected genes have. B, Linoleic acid content shows a significant difference between accessions with two different haplotypes in *SoyZH13_02G207800*. *** denotes *t*-test P<0.001.

and soybean crop improvement.

MATERIALS AND METHODS

Plant materials and sequencing

For SMRT, DNA was isolated using the Blood&Cell Culture DNA Midi Kit (Qiagen Inc., Valencia, CA, USA) according to a user-developed protocol (Isolation of genomic DNA from plants and filamentous fungi using the QIAGEN® Genomic-tip) provided by Qiagen. A 20 kb library was constructed and sequenced on 17 cells using the Sequel™ Sequencing Plate 1.2 on the Pacific Biosciences Sequel platform at Berry Genomic Corporation, Ltd (Beijing, China).

For HiSeq, DNA was isolated using the Plant Genomic DNA kit (Tiangen, Beijing, China) according to the manufacturer's protocol. A library with 450 bp small-insertion was prepared and sequenced on two lanes of the Illumina HiSeq2500 platform for 250 bp paired-end reads.

For optical mapping, young yellow leaves after dark treatment for three days were sent to Berry Genomic Corporation as DNA samples. High-molecular-weight DNA was isolated and labeled using the single-stranded nicking endonuclease Nt.BspQI and Nt.BssSI independently. These two labeled DNA samples were separately and automatically imaged using BioNano Irys system, and only molecules longer than 150 kb were used for further analysis.

For Hi-C, leaves fixed in 1% (vol/vol) formaldehyde were used for library construction. Cell lysis, chromatin digestion, proximity-ligation treatments, DNA recovery and subsequent DNA manipulations were performed as previously described (Lieberman-Aiden, 2009). MboI was used as the restriction enzyme in chromatin digestion. The Hi-C library was sequenced on the Illumina HiSeq X Ten platform for 150 bp paired-end reads.

For Iso-Seq, samples of leaf, flower and seed were collected separately at early, middle and late developmental stages from the same DNA sequenced "Zhonghuang 13" plant, while the samples for root and stem were collected from another "Zhonghuang 13" plant at two weeks after germination. The total RNA of all 11 samples was isolated

and mixed together at the same concentration. Five cDNA libraries (<1 kb, 1–2 kb, 2–3 kb, 3–6 kb and 5–10 kb) were prepared using this mixed RNA and sequenced on 7 SMRT cells (two cells for 1–2 kb and 2–3 kb library separately, one cell for each of the other libraries) using P6-C4 chemistry on the Pacific Biosciences RSII platform.

Genome assembly

De novo assembly was conducted with SMRT long reads using the smrtmake assembly pipeline (https://github.com/PacificBiosciences/smrtmake). In this draft assembly, the longest 50× subreads were selected for error correction and then the longest 20× error corrected subreads were set as seed reads for overlapping detection. Arrow was used to polish this draft assembly and Pilon v1.20 (Walker et al., 2014) was used to further error correction when adding Hi-Seq reads.

BioNano optical maps labeled by different endonucleases were assembled into consensus physical maps separately. Then, PacBio-BioNano hybrid scaffolds were generated by combining these maps and the PacBio SMRT draft assembly produced in the previous step. For these steps, BioNano Solve™ v3.0.1 (Solve_06082017Rel) was used.

To anchor hybrid scaffolds into chromosome, the Hi-C sequencing data were aligned into scaffolds by Bowtie2 in HiC-Pro_2.9.0 (Servant et al., 2015). According to the orders and orientations provided by the alignment, those scaffolds were clustered into chromosomes by LACHESIS (Burton et al., 2013) with parameters CLUSTER_N=20, RE_SITES=292, CLUSTER_MAX_LINK_DENSITY=3.48, CLUSTER_NONINFORMATIVE_RATIO=3, ORDER_MIN_N_RES_IN_TRUNK=50000, ORDER_MIN_N_RES_IN_SHREDS=500. As original LACHESIS results always have mis-ordered and mis-orientated scaffolds in some groups, manual correction and validation were also performed by drawing contact maps with HiCPlotter (Akdemir and Chin, 2015). The genome assembly was finalized after this correction.

SMRT long reads were aligned to the chloroplasts genome (NCBI accession: NC_007942.1) assembled previously (Saski et al., 2005) using BLASR (Chaisson and Tesler, 2012) to extract the subreads coming from the chloroplast, and Canu (v1.5) (Koren et al., 2017) was then used for assembly. The final assembled chloroplast genome was polished and redundant sequences were removed.

Repeat analysis and gene annotation

A combination of structure-based and homology-based approaches were employed to identify TEs. These approaches include the following: (1) LTR_STRUC was employed to identify LTR retrotransposons (McCarthy and McDonald, 2003); (2) elements with clear insertion sites deposited in SoyTEdb were mapped to the new genome sequence using element or junction sequences (100 bp each on both sites) (Du et al., 2010); (3) TE conserved domains were searched in the genome sequence (Holligan et al., 2006); (4) CrossMatch was used to identify new elements via known TEs. Repeat regions were detected by RepeatMasker (version open-4.0.7) (http://www.repeatmasker.org/) using annotated TEs as a library. If a region belonged to different TE types, it was defined with the following priorities: *Gypsy*>DNA/Stowaway>DNA/Tc1>DNA/CACTA>*Copia*>LTR_unclassified>DNA/Helitron>DNA/Mutator>DNA/hAT>LINE>DNA/PONG>DNA/PIF>DNA/Tourist>SINE>tandem_repeat>unknown.

A comprehensive strategy combining *ab initio* gene finding, homology-based gene prediction and Iso-Seq reads was used for annotation of protein-coding genes on chromosomes. Augustus (v3.0.3) (Stanke and Morgenstern, 2005), SNAP (v2013-02-16) (Korf, 2004) and GeneMark-ET (v4.21) (Besemer and Borodovsky, 2005) were used in *ab initio* gene finding. cDNA sequences of *Arabidopsis thaliana* (167_TAIR10), *Glycine max* (275_Wm82.a2.v1) and *Oryza sativa* (323_v7.0) downloading from Phytozome v12 (https://phytozome.jgi.doe.gov/pz/portal.html) were used to predict homologous genes by performing GeMoMa-1.4.2 (Keilwagen et al., 2018). Iso-Seq reads from 5 libraries were analyzed by smrtanalysis_2.3.0 (https://www1.smartanalysis.com) to obtain individual full-length transcripts. All full-length transcripts were merged together, and the redundancies were removed. All the gene structures predicted by the above methods were combined into consensus gene models using EVidenceModeler (EVM) (Haas et al., 2008), and genes that were predicted only in the *ab intio* step or that shorter than 150 bp were removed. The left gene models were then updated by PASA (r20140417) (Haas, 2003). The genes in the chloroplast genome were predicted by an online database CpGAVAS (Liu et al., 2012) which focuses on chloroplast gene analysis.

Genome comparisons and SV identification

Genome comparisons between Gmax_ZH13 and Glycine_max_v2.0 were performed via whole-genome alignment mainly by tools from MUMmer (ver 3.0) (Kurtz et al., 2004). Nucmer was used to align the two genomes (–g 2000) and then delta-filter was used to filter the alignment blocks in one-to-one alignment mode (-1). Blocks longer than 200 bp were used for further SV detections.

SNPs and indels were identified by performing show-snp (-ClrTo) on the alignment blocks. Translocations and inversions were identified from the aligned blocks by manual checking, and neighboring blocks belonging to the same event were merged together. Blocks belonging to both the

inversion and translocation categories were named translocation & inversion events. Blocks that did not belong to either translocations or inversions were defined as synteny blocks.

PVs were identified for the two genomes separately. Unaligned blocks were extracted depending on all alignment blocks for each genome and then mapped to the other genome using Blast+ (Camacho et al., 2009). Sequence with alignments identity >90% and length >100 bp were filtered and all other un-aligned sequences were defined as genome specific PVs.

Gene co-expression network establishment

The gene co-expression network was constructed following a procedure described previously (Ma et al., 2017). Briefly, publicly available soybean RNA-seq datasets were downloaded from the NCBI SRA database as raw sra files and converted to fastq files. The RNA-seq reads were trimmed using Trimmomatic (v0.36), and mapped to the Gmax_ZH13 genome via STAR (v2.5.3a) (Dobin et al., 2013; Bolger et al. 2014). After removing the RNA-seq runs with less than 70% reads mapped as well as those for small RNA sequencing, 1,978 high quality runs remained and were used for gene expression quantification via RSEM v1.3.0 (Li and Dewey, 2011). After discarding low expressed genes with <10 runs having expression values (TPM) ≥5, the remaining genes' expression values (TPM) were assembled into a gene expression matrix with 42,169 rows (genes) and 1,978 columns (runs). The matrix was used for partial correlation coefficient (pcor) calculation via a random sampling approach (Ma et al., 2017; Ma et al., 2007). The procedure consisted of 30,000 rounds, with 2,000 genes randomly selected in each round for pcor calculation via the GeneNet package in R (Schäfer and Strimmer, 2005). In total, each gene pair was selected ~66 times with ~66 pcors calculated, and the pcor with the lowest absolute value was selected as its final pcor. The Pearson correlation coefficient (r) between genes was also calculated. Finally, 330,864 gene pairs with pcor≥0.035 and r≥0.35 were selected for GGM gene co-expression network construction, which included 39,967 genes in total.

Gene and region conversions between different genomes

Gene correspondences between Gmax_ZH13 and Glycine_max_v2.0 were determined at the gene level (blastn), mRNA level (blastn) and protein level (blastp) using Blast+ (Camacho et al., 2009). Gene/mRNA/protein sequences from Glycine_max_v2.0 were mapped to the same type sequences from Gmax_ZH13 and those with alignment identity >90% were retained. Only genes with correspondence at all three levels were kept as corresponding gene pairs.

Previously, reported QTLs/GWAS related to soybean flowering time were detected based on the Glycine_max_v1.0 or Glycine_max_v2.0 genome. We located these regions in the Gmax_ZH13 genome by performing nucmer alignment with MUMmer (ver 3.0)(Kurtz et al., 2004).

Data availability

All the sequencing data used in the genome assembly have been deposited into the Genome Sequence Archive (GSA) database in BIG Data Center under Accession Number CRA001007. Information for the assembled genome Gmax_ZH13 was deposited both into the Genome Warehouse (GWH) (GWHAAEV00000000) database in the BIG Data Center and DDBJ/ENA/GenBank under the accession QKRT00000000.

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

Acknowledgements *This work was supported by the National Natural Science Foundation of China (91531304, 31525018, 31370266, and 31788103), the "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA08000000), and the State Key Laboratory of Plant Cell and Chromosome Engineering (PCCE-KF-2017-03).*

Akdemir, K.C., and Chin, L. (2015). HiCPlotter integrates genomic data with interaction matrices. Genome Biol 16, 198.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière, C., Owens, G.L., Carrère, S., Mayjonade, B., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546, 148–152.

Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33, W451–W454.

Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J., Lam, E.T., Liachko, I., Sullivan, S.T., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nat Genet 49, 643–650.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol 31, 1119–1125.

Byrum, J. R., Kinney, A. J., Shoemaker, R. C., and Diers, B. W. (1995). Mapping of the microsomal and plastid omega-3 fatty acid desaturases in soybean [*Glycine max* (L.) Merr.]. Soybean Genet Newslett 22, 181–184.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC BioInf 10, 421.

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

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC BioInf 13, 238.

Chan, C., Qi, X., Li, M.W., Wong, F.L., and Lam, H.M. (2012). Recent developments of genomic research in soybean. J Genets Genomics 39, 317–324.

Chen, G., Shi, T., and Shi, L. (2017). Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60, 116–125.
Childs, K.L., Davidson, R.M., and Buell, C.R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS ONE 6, e22196.
Clavijo, B.J., Venturini, L., Schudoma, C., Accinelli, G.G., Kaithakottil, G., Wright, J., Borrill, P., Kettleborough, G., Heavens, D., Chapman, H., et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27, 885–896.
Contreras-Soto, R.I., Mora, F., Lazzari, F., de Oliveira, M.A.R., Scapim, C. A., and Schuster, I. (2017). Genome-wide association mapping for flowering and maturity in tropical soybean: implications for breeding strategies. Breed Sci 67, 435–449.
Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., Ma, B., Qi, M., Li, Y., Zhao, X., et al. (2017). Sequencing and *de novo* assembly of a near complete *indica* rice genome. Nat Commun 8, 15324.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.
Dooner, H.K., and He, L. (2008). Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. Plant Cell 20, 249–258.
Du, J., Grant, D., Tian, Z., Nelson, R.T., Zhu, L., Shoemaker, R.C., and Ma, J. (2010). SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11, 113.
Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol 18, 161.
Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C., et al. (2011). Solutions for a cultivated planet. Nature 478, 337–342.
Funatsuki, H., Kawaguchi, K., Matsuba, S., Sato, Y., and Ishimoto, M. (2005). Mapping of QTL associated with chilling tolerance during reproductive growth in soybean. Theor Appl Genet 111, 851–861.
Gai, J., Wang, Y., Wu, X., and Chen, S. (2007). A comparative study on segregation analysis and QTL mapping of quantitative traits in plants—with a case in soybean. Front Agric China 1, 1–7.
Githiri, S.M., Yang, D., Khan, N.A., Xu, D., Komatsuda, T., and Takahashi, R. (2007). QTL analysis of low temperature induced browning in soybean seed coats. J Heredity 98, 360–366.
Gizlice, Z., Carter, T.E., and Burton, J.W. (1994). Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci 34, 1143–1151.
Guo, H., Liu, J., Luo, L., Wei, X., Zhang, J., Qi, Y ., Zhang, B., Liu, H., and Xiao, P. (2017). Complete chloroplast genome sequences of *Schisandra chinensis*: genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. Sci China Life Sci 60, 1–5.
Haas, B.J. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654–5666.
Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9, R7.
Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., et al. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell 28, 2700–2714.
Holligan, D., Zhang, X., Jiang, N., Pritham, E.J., and Wessler, S.R. (2006). The transposable element landscape of the model legume *Lotus japonicus*. Genetics 174, 2215–2228.
Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., Shin-I, T., Minakuchi, Y., Koda, Y., Nagano, A.J., et al. (2016). Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. Nat Commun 7, 13295.
Hyten, D.L., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., and Cregan, P.B. (2006). Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103, 16666–16671.
Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., Schmöckel, S.M., Li, B., Borm, T.J. A., Ohyanagi, H., Mineta, K., Michell, C.T., Saber, N., et al. (2017). The genome of *Chenopodium quinoa*. Nature 542, 307–312.
Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M. S., Stein, J.C., Wei, X., and Chin, C.S. (2017). Improved maize reference genome with single-molecule technologies. Nature 546, 524-527.
Jun, T.H., Freewalt, K., Michel, A.P., and Mian, R. (2014). Identification of novel QTL for leaf traits in soybean. Plant Breed 133, 61-66.
Kawakatsu, T., Huang, S.S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M. A., Castanon, R., Nery, J.R., Barragan, C., He, Y., et al. (2016). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell 166, 492–505.
Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC BioInf 19, 189.
Keim, P., Diers, B.W., Olson, T.C., and Shoemaker, R.C. (1990). RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126, 735-742.
Khan, N.A., Githiri, S.M., Benitez, E.R., Abe, J., Kawasaki, S., Hayashi, T., and Takahashi, R. (2008). QTL analysis of cleistogamy in soybean. Theor Appl Genet 117, 479–487.
Kim, H.K., Kim, Y.C., Kim, S.T., Son, B.G., Choi, Y.W., Kang, J.S., Park, Y.H., Cho, Y.S., and Choi, I.S. (2010). Analysis of quantitative trait loci (QTLs) for seed size and fatty acid composition using recombinant inbred lines in soybean. J Life Sci 20, 1186–1192.
Komatsu, K., Okuda, S., Takahashi, M., Matsunaga, R., and Nakazawa, Y. (2007). Quantitative trait loci mapping of pubescence density and flowering time of insect-resistant soybean (*Glycine max* L. Merr.). Genet Mol Biol 30, 635–639.
Kong, F., Liu, B., Xia, Z., Sato, S., Kim, B.M., Watanabe, S., Yamada, T., Tabata, S., Kanazawa, A., Harada, K., et al. (2010). Two coordinately regulated homologs of *FLOWERING LOCUS T* are involved in the control of photoperiodic flowering in soybean. Plant Physiol 154, 1220–1231.
Kong, F., Nan, H., Cao, D., Li, Y., Wu, F., Wang, J., Lu, S., Yuan, X., Cober, E.R., Abe, J., et al. (2014). A new dominant gene conditions early flowering and maturity in soybean. Crop Sci 54, 2529–2535.
Koo, S.C., Bracko, O., Park, M.S., Schwab, R., Chun, H.J., Park, K.M., Seo, J.S., Grbic, V., Balasubramanian, S., Schmid, M., et al. (2010). Control of lateral organ development and flowering time by the *Arabidopsis thaliana* MADS-box Gene *AGAMOUS-LIKE6*. Plant J 62, 807–816.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive*k*-mer weighting and repeat separation. Genome Res 27, 722–736.
Korf, I. (2004). Gene finding in novel genomes. BMC BioInf 5, 59.
Krouk, G., Mirowski, P., LeCun, Y., Shasha, D.E., and Coruzzi, G.M. (2010). Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. Genome Biol 11, R123.
Kuroda, Y., Kaga, A., Tomooka, N., Yano, H., Takada, Y., Kato, S., and Vaughan, D. (2013). QTL affecting fitness of hybrids between wild and cultivated soybeans in experimental fields. Ecol Evol 3, 2150–2168.
Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes.. Genome Biol 5, R12.
Lam, H.M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.L., Li, M.W., He, W., Qin, N., Wang, B., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42, 1053–1059.
Le, B.H., Cheng, C., Bui, A.Q., Wagmaister, J.A., Henry, K.F., Pelletier, J.,

Kwong, L., Belmonte, M., Kirkbride, R., Horvath, S., et al. (2010). Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci USA 107, 8063–8070.
Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC BioInf 12, 323.
Li, Y.H., Li, W., Zhang, C., Yang, L., Chang, R.Z., Gaut, B.S., and Qiu, L.J. (2010). Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytologist 188, 242–253.
Li, Y., Zhao, S., Ma, J., Li, D., Yan, L., Li, J., Qi, X., Guo, X., Zhang, L., He, W., et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14, 579.
Li, Y., Zhou, G., Ma, J., Jiang, W., Jin, L., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32, 1045–1052.
Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293.
Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., and Guan, X. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13, 715.
Liu, Z.X., Li, H.H., Wen, Z.X., Fan, X.H., Li, Y.H., Guan, R.X., Guo, Y., Wang, S.M., Wang, D.C., and Qiu, L.J. (2017). Comparison of genetic diversity between Chinese and American soybean (*Glycine max* (L.)) accessions revealed by high-density SNPs. Front Plant Sci 8, 2014.
Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A., et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66, 219–232.
Lu, S., Zhao, X., Hu, Y., Liu, S., Nan, H., Li, X., Fang, C., Cao, D., Shi, X., Kong, L., et al. (2017). Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. Nat Genet 49, 773–779.
Lv, S., Wu, W., Wang, M., Meyer, R.S., Ndjiondjop, M.N., Tan, L., Zhou, H., Zhang, J., Fu, Y., Cai, H., et al. (2018). Genetic control of seed shattering during African rice domestication. Nat Plants 4, 331–337.
Ma, S.S., Bohnert, H.J., and Dinesh-Kumar, S.P. (2015). AtGGM2014, an Arabidopsis gene co-expression network for functional studies. Sci China Life Sci 58, 276–286.
Ma, S., Ding, Z., and Li, P. (2017). Maize network analysis revealed gene modules involved in development, nutrients utilization, metabolism, and stress response. BMC Plant Biol 17, 131.
Ma, S., Gong, Q., and Bohnert, H.J. (2007). An *Arabidopsis* gene network based on the graphical Gaussian model. Genome Res 17, 1614–1625.
Mansur, L., Lark, K., Kross, H., and Oliveira, A. (1993). Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). Theor Appl Genet 86, 907-913.
Mansur, L.M., Orf, J.H., Chase, K., Jarvik, T., Cregan, P.B., and Lark, K.G. (1996). Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36, 1327–1336.
Mao, T., Li, J., Wen, Z., Wu, T., Wu, C., Sun, S., Jiang, B., Hou, W., Li, W., Song, Q., et al. (2017). Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. BMC Genomics 18, 415.
McCarthy, E.M., and McDonald, J.F. (2003). LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19, 362–367.
Oldham, M.C., Horvath, S., and Geschwind, D.H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA 103, 17973–17978.
Orf, J., Chase, K., Jarvik, T., Mansur, L., Cregan, P., Adler, F., and Lark, K. (1999). Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Sci 39, 1642-1651.
Oyoo, M.E., Githiri, S.M., Benitez, E.R., and Takahashi, R. (2010). QTL analysis of net-like cracking in soybean seed coats. Breed Sci 60, 28–33.
Palomeque, L., Li-Jun, L., Li, W., Hedges, B., Cober, E.R., and Rajcan, I. (2009). QTL in mega-environments: II. Agronomic trait QTL co-localized with seed yield QTL detected in a population derived from a cross of high-yielding adapted × high-yielding exotic soybean lines. Theor Appl Genet 119, 429-436.
Pooprompan, P., Wasee, S., Toojinda, T., Abe, J., Chanprame, S., and Srinives, P. (2006). Molecular marker analysis of days to flowering in vegetable soybean (*Glycine max* (L.) Merrill). Kasetsart Journal 40, 573-581.
Ray, D.K., Mueller, N.D., West, P.C., and Foley, J.A. (2013). Yield trends are insufficient to double global crop production by 2050. PLoS ONE 8, e66428.
Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., et al. (2018). The Rosa genome provides new insights into the domestication of modern roses. Nat Genet 50, 772–777.
Reinprecht, Y., Poysa, V.W., Yu, K., Rajcan, I., Ablett, G.R., and Pauls, K.P. (2006). Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. Genome 49, 1510–1527.
Rhee, S.Y., and Mutwil, M. (2014). Towards revealing the functions of all genes in plants. Trends Plant Sci 19, 212–221.
Samanfar, B., Molnar, S.J., Charette, M., Schoenrock, A., Dehne, F., Golshani, A., Belzile, F., and Cober, E.R. (2017). Mapping and identification of a potential candidate gene for a novel maturity locus, *E10*, in soybean. Theor Appl Genet 130, 377–390.
Saski, C., Lee, S.B., Daniell, H., Wood, T.C., Tomkins, J., Kim, H.G., and Jansen, R.K. (2005). Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol Biol 59, 309–322.
Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4, Article32.
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.
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.
Seo, J.S., Rhie, A., Kim, J., Lee, S., Sohn, M.H., Kim, C.U., Hastie, A., Cao, H., Yun, J.Y., Kim, J., et al. (2016). *De novo* assembly and phasing of a Korean human genome. Nature 538, 243–247.
Serin, E.A.R., Nijveen, H., Hilhorst, H.W.M., and Ligterink, W. (2016). Learning from co-expression networks: possibilities and challenges. Front Plant Sci 7, 444.
Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol 16, 259.
Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., et al. (2016). Long-read sequencing and *de novo* assembly of a Chinese genome. Nat Commun 7, 12065.
Shimomura, M., Kanamori, H., Komatsu, S., Namiki, N., Mukai, Y., Kurita, K., Kamatsuki, K., Ikawa, H., Yano, R., and Ishimoto, M. (2015). The *Glycine max* cv. Enrei genome for improvement of Japanese soybean cultivars. Int J Genomics 2015, 358127.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.
Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. N-

ucleic Acids Res 33, W465–W467.
Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nat Genet 43, 1160–1163.
Tasma, I.M., Lorenzen, L.L., Green, D.E., and Shoemaker, R.C. (2001). Mapping genetic loci for flowering time, maturity, and photoperiod insensitivity in soybean. Mol Breeding 8, 25–35.
VanBuren, R., Bryant, D., Edger, P.P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., Lyons, E., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature 527, 508–511.
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963.
Wang, K., Huang, G., and Zhu, Y. (2016). Transposable elements play an important role during cotton genome evolution and fiber cell development. Sci China Life Sci 59, 112–121.
Wang, Z., and Tian, Z.X. (2015). Genomics progress will facilitate molecular breeding in soybean. Sci China Life Sci 58, 813–815.
Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., Takahashi, R., Anai, T., Tabata, S., Kitamura, K., et al. (2011). A map-based cloning strategy employing a residual heterozygous line reveals that the*GIGANTEA* gene is involved in soybean maturity and flowering. Genetics 188, 395–407.
Wei, H., Yordanov, Y.S., Georgieva, T., Li, X., and Busov, V. (2013). Nitrogen deprivation promotes*Populus* root growth through global transcriptome reprogramming and activation of hierarchical genetic networks. New Phytol 200, 483–497.
Wei, L., and Cao, X. (2016). The effect of transposable elements on phenotypic variation: insights from plants to humans. Sci China Life Sci 59, 24–37.
Wilson, R.F. (2008). Soybean: Market Driven Research Needs in Genetics and Genomics of Soybean, G. Stacey, ed. (New York: Springer), pp. 3-16.
Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E., Jenkins, D.J., Penfold, C.A., Baxter, L., Breeze, E., et al. (2012). *Arabidopsis* defense against *Botrytis cinerea*: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. Plant Cell 24, 3530–3557.
Wolfe, C.J., Kohane, I.S., and Butte, A.J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC BioInf 6, 227.
Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., Anai, T., Sato, S., Yamazaki, T., Lü, S., et al. (2012). Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. Proc Natl Acad Sci USA 109, E2155–E2164.
Yamanaka, N., Nagamura, Y., Tsubokura, Y., Yamamoto, K., Takahashi, R., Kouchi, H., Yano, M., Sasaki, T., and Harada, K. (2000). Quantitative trait locus analysis of flowering time in soybean using a RFLP linkage map.. Breed Sci 50, 109–115.
Yamanaka, N. (2001). An informative linkage map of soybean reveals QTLs for flowering time, leaflet morphology and regions of segregation distortion. DNA Res 8, 61–72.
Yue, Y., Liu, N., Jiang, B., Li, M., Wang, H., Jiang, Z., Pan, H., Xia, Q., Ma, Q., Han, T., et al. (2017). A single nucleotide deletion in J encoding gmelf3 confers long juvenility and is associated with adaption of tropic soybean. Mol Plant 10, 656–658.
Zabala, G., and Vodkin, L.O. (2007). A rearrangement resulting in small tandem repeats in the *F3'5'H* gene of white flower genotypes is associated with the soybean locus. Crop Sci 47, S-113.
Zhang, J., Chen, L.L., Xing, F., Kudrna, D.A., Yao, W., Copetti, D., Mu, T., Li, W., Song, J.M., Xie, W., et al. (2016). Extensive sequence divergence between the reference genomes of two elite*indica* rice varieties Zhenshan 97 and Minghui 63. Proc Natl Acad Sci USA 113, E5163–E5171.
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.
Zhang, W.K., Wang, Y.J., Luo, G.Z., Zhang, J.S., He, C.Y., Wu, X.L., Gai, J.Y., and Chen, S.Y. (2004). QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. Theor Appl Genet 108, 1131–1139.
Zhao, C., Takeshima, R., Zhu, J., Xu, M., Sato, M., Watanabe, S., Kanazawa, A., Liu, B., Kong, F., Yamada, T., et al. (2016). A recessive allele for delayed flowering at the soybean maturity locus *E9* is a leaky allele of *FT2a*, a *FLOWERING LOCUS T* ortholog. BMC Plant Biol 16, 20.
Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33, 408–414.

SUPPORTING INFORMATION

Figure S1 PCR validation for insertion or deletions regions in Gmax_ZH13.

Figure S2 Gene controlling soybean flower color exists allelic difference between Zhonghuang 13 and Williams 82.

Figure S3 High quality ZH13 genome sequence can improve gene identification, an example from GWAS analysis.

Table S1 Data statistics for different sequence types

Table S2 Correspondence between Gmax_ZH13 and Glycine_max_v2.0 annotation genes

Table S3 Assembly comparison between Gmax_ZH13 and other previously released three soybean genomes

Table S4 Detailed chromosome comparison between Gmax_ZH13 and Glycine_max_v2.0

Table S5 Translocation events between Gmax_ZH13 and Glycine_max_v2.0 genome

Table S6 Inversion events between Gmax_ZH13 and Glycine_max_v2.0 genome

Table S7 Translocation & inversion events between Gmax_ZH13 and Glycine_max_v2.0 genome

Table S8 Specifically presence genome regions in Gmax_ZH13 genome

Table S9 Specifically presence genome regions in Glycine_max_v2.0 genome

Table S10 Small insertion regions in Gmax_ZH13 and Glycine_max_v2.0 genome

Table S11 Genes co-expressed with 9 known soybean flower time related genes

Supplemental File 1

Supplemental File 2

Supplemental File 3

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.