Platinum-grade soybean reference genome facilitates characterization of genetic underpinnings of traits

Xinxin Yi
Huazhong Agriculture University https://orcid.org/0000-0003-4865-3855
Jing Liu
Huazhong Agriculture University
Shengcai Chen
Huazhong Agriculture University
Hao Wu
Huazhong Agriculture University
Min Liu
Huazhong Agriculture University
Qing Xu
Huazhong Agriculture University
Lingshan Lei
Huazhong Agriculture University
Seunghee Lee
University of Arizona
Bao Zhang
Huazhong Agriculture University
Dave Kudrna
University of Arizona
Wei Fan
Huazhong Agriculture University
Rod Wing
University of Arizona
Chunyan Yang
Hebei Academy of Agriculture and Forestry Sciences
Mengchen Zhang
Hebei Academy of Agriculture and Forestry Sciences
Jianwei Zhang
Huazhong Agriculture University
Xuelu Wang
Henan University
Nansheng Chen (chenn@sfu.ca)
College of Life Science and Technology https://orcid.org/0000-0002-6361-964X

Research article

Keywords: contiguity, completeness, accuracy

DOI: https://doi.org/10.21203/rs.3.rs-57915/v1

Abstract

Background

Cultivated soybean (Glycine max) is an important source for protein and oil. Each soybean strain has its own genetic diversity, and the availability of more soybean genomes may enhance comparative genomic analysis of soybean.

Results

In this study, we constructed a high-quality *de novo* assembly of an elite soybean cultivar Jidou 17 (JD17) with high contiguity, completeness, and accuracy. We annotated 59,629 gene models and reconstructed 235,109 high-quality full-length transcripts. We have molecularly characterized the genotypes of some important agronomic traits of JD17 by taking advantage of these newly established genomic resources.

Conclusions

We reported a high-quality genome and annotations of a wide range of cultivars, and used them to analyze the genotypes of genes related to important agronomic traits of soybean in JD17. We have demonstrated that high-quality genome assembly can serve as a valuable reference for soybean genomics and breeding research community.

Background

Soybean (*Glycine max*) is an important crop for protein and dietary oil and is ranked the fourth largest crop in production in the world. The current soybean reference genome, which was based on the Williams 82 (Wm82) line [1] has greatly enhanced the identification of genes underlying important traits and facilitated research on the function and expression of soybean genes.

Recent studies using high-throughput sequencing has revealed extensive genetic diversities in soybean [2]. Pan-genome study on wild and cultivated soybeans has uncovered numerous genetic differences among soybean strains [3], suggesting that a single reference genome is inadequate for representing the genetic richness of soybean lines.

Recent progress in genome technologies has greatly advanced the ability to construct high-quality genome assemblies with chromosome-level continuity with dramatically reduced cost [4, 5]. The application of Iso-Seq protocol has enhanced genome annotation [6-9]. Such progresses make it possible to construct multiple reference genomes based on different lines.

JD17 is a major soybean cultivar in the Huang-Huai-Hai region of China, and it is also the main soybean variety recognized by the Ministry of Agriculture since 2010. It is the offspring of Hobbit (maternal parent) and Zao 5241 [7476 × 7527-1-1 (Yanli×Williams)] (paternal parent) [10], and is famous for its lodging resistance, high yield, and strong adaptability [11, 12]. The goal of this project was to construct a high-quality genome assembly for JD17, and apply the whole genome sequence to molecularly identify genetic bases of important traits. Especially in terms of important agronomic traits, such as regulating flower color, cultural cycle, regulating seed size, weight, oil content and protein content and other important trait genes, helping to determine their location and genotype in JD17, and then applying them to genetic improvement and Breeding.

In this study, we extracted genome DNA from developing underground tissues of JD17, and used PacBio single-molecule real-time (SMRT) sequencing and Hi-C technologies to construct a platinum-grade soybean reference genome annotated more completely, and we used it to identify the genotypes of important agronomic trait-related genes of JD17.

Results

Construction of a platinum-grade soybean reference genome

Genome assembly using PacBio SMRT data and Hi-C data resulted in a JD17 genome assembly with 411 contigs anchored to 20 pseudomolecules, with a total size of 965.8 Mb (Table 1), accounting for 97.1% of the total contigs (Fig. 1). The genome

accuracy rate was evaluated as 99.999% by using resequencing data and GATK tools [13, 14] (See methods). BUSCO score of the assembly was 93.2% [15] (Table 2), indicating that the completeness of JD17 was higher than that of the Wm82, ZH13, and W05 genomes [16, 17].

Table 1
Assembly statistics of Glycine_max_JD17 (JD17), Glycine_max_v2.0 (Wm82), Gmax_ZH13 (ZH13) and W05.

Assembly Feature	**JD17**	**Wm82**	**ZH13**	**W05**
Estimated genome size (by K-mer analysis) (Mb)	1,109	1,115	-	-
Number of contigs	446	16,311	1,528	1,870
Total size of contigs (Mb)	995.0	955.1	1,007	998.6
Longest contig (Mb)	31.8	-	-	-
Number of contigs > 1 Mb	97	-	-	-
Number of contigs > 10 Mb	39	-	-	-
N50 contig length (Mb)	17.998(PacBio)	0.189	3.46(PacBio + BioNano + Hi-C)	3.3
L50 contig count	21	1,492	66	58
Anchored contigs				
Number of chromosomes	20	20	20	20
Number of contigs	139	-	-	772
Total size (Mb)	971	937.3	1,025	1,013.2
Number of gaps	119	12,785	815	750

Table 2
BUSCO analysis for JD17, Wm82, ZH13 and W05.

	JD17	W82	ZH13	W05
Genome				
Complete BUSCOs (C)	1,342	1,340	1,280	1,340
Complete and single-copy BUSCOs (S)	669	690	631	673
Complete and duplicated BUSCOs (D)	673	650	649	667
Fragmented BUSCOs (F)	16	15	29	16
Missing BUSCOs (M)	82	85	131	84
Transcript				
Complete BUSCOs (C)	1,400	1,393	1,393	1,401
Complete and single-copy BUSCOs (S)	199	341	533	384
Complete and duplicated BUSCOs (D)	1,201	1,052	860	1,017
Fragmented BUSCOs (F)	9	13	11	6
Missing BUSCOs (M)	31	34	36	33
Protein				
Complete BUSCOs (C)	1,404	1,401	1,397	1,404
Complete and single-copy BUSCOs (S)	205	341	533	405
Complete and duplicated BUSCOs (D)	1,199	1,060	864	999
Fragmented BUSCOs (F)	5	8	12	7
Missing BUSCOs (M)	31	31	31	29
Total BUSCO groups searched	1440			

Annotation of JD17 genome

We identified 616.6 Mb (62.0% of the JD17 genome) repeat elements in JD17, which contain 1,138,787 intact transposable elements (TEs), including 864,771 class I RNA retrotransposons and 274,016 class II DNA transposons (**Table S1**).

We identified 58,738 protein-coding genes and 253,172 full-length transcripts in the JD17 genome. The average length of transcripts (2,525 bp) and 3'UTR (707 bp) in JD17 genome are comparatively longer than that in Wm82 (1,628 bp, 374 bp), ZH13 (1,512 bp, 338 bp), and W05 (1,650 bp, 405 bp) genome (Table 3).

Table 3
Comparison of genome annotation of JD17, Wm82, ZH13 and W05.

	JD17	Wm82	ZH13	W05
Number of genes	58,738	56,044	52,178	55,539
Number of transcripts	253,172	88,647	58,017	89,477
Average number of transcripts per gene	4.3	1.58	1.11	1.61
Average length of transcript* (bp)	2,525	1,628	1,512	1,650
Average exons number per transcript*	8.4	5.33	5.62	5.36
Average length of 5' UTR* (bp)	591	388	183	252
Average length of 3' UTR* (bp)	707	374	338	405
Numbers of genes with gap	0	1,063	12	0
* mean based on transcript with the longest CDS.				

Table 4
Genes related to important agronomic traits.

	Gene Name	Wm82.a2.v1 Gene id	JD17 Gene ID	Genotype in JD17	Speculated phenotype	Reference
Flowering related genes	*FLS1/gmfls1*	Glyma.13G082300	GmJD_13G0078400	*WmWm*	purple flower	(Takahashi et al., 2007)
	F3'5'H	Glyma.13G072100	GmJD_13G0068800	*W1*	Light purple flowers	(Takahashi et al., 2010)
	DFR1	Glyma.14G072700	GmJD_14G0066900	*W3*	Purple flower	(Park et al., 2015)
	DFR2	Glyma.17G252200	GmJD_17G0274600	*W4*	Purple flower	(Yan et al., 2014)
	FT2c	Glyma.02G069200 Glyma.02G069500	GmJD_02G0063800	*GmFT2c*	Days About 32 in short day (SD) and 46 in long day (LD) to flowering	(Wu et al., 2017)
	J gene	Glyma.04G050200	GmJD_04G0047400	*J*	Promoting flowering under short days	(Lu et al., 2017)
	E1	Glyma.06G207800	GmJD_06G0201400	*e1-as*	Early-Flowering	(Xia et al., 2012)
Seed related genes	*GmIRCHS*	Glyma.08G110300	GmJD_08G0106200	*I/I*	Inhibiting seed coat pigmentation	(Kasai et al., 2007)
	GmHs1-1	Glyma.02G269500	GmJD_02G0283800	*Gmhs1-1*	Permeable seed coat	(Sun et al., 2015)
	SoyWRKY15a	Glyma.05G096500	GmJD_05G0121200	*GmWRKY15a* (haplotypes H1)	Heavier seeds	(Gu et al., 2017)
	FATB1a	Glyma.05G012300	GmJD_05G0012300	No deletion, homozygous *FATB1a*	Relatively high palmitic acid levels of seed	(Goettel et al., 2016)
	ZF351	Glyma.06G290100	GmJD_06G0292100	*GmZF351*	Relatively high lipid content of seed	(Li, QT et al., 2017)
	GmDREBL	Glyma.12G103100	GmJD_12G0097600	variation type 2	Relatively higher oil content of seed	(Zhang, YQ et al., 2016)
	GA20OX	Glyma.07G081700	GmJD_07G0078300	cultivated *GmGA20OX*	Relatively heavy 100 seeds weight of seed	(Lu et al., 2016)
	NFYA	Glyma.02G303800	GmJD_02G0316700	*GmNFYA*	Relatively higher oil content of seed	(Lu et al., 2016)

	Gene Name	Wm82.a2.v1 Gene id	JD17 Gene ID	Genotype in JD17	Speculated phenotype	Reference
	GmSWEET10a	Glyma.15G049200	GmJD_15G0049100	Haplotypes H_III-3	Relatively heavy 100-seed weight, higher fatty acid content and lower protein content of seed	(Shoudong et al., 2020)
	GmSWEET10b	Glyma.08G183500	GmJD_08G0177700	Haplotypes H_III-1	Relatively heavy 100-seed weight, higher fatty acid content and lower protein content of seed	(Shoudong et al., 2020)
Other important traits related genes	*GmphyA2*	Glyma.10G141400	GmJD_10G0172700	*E4*	Normal phytochrome A gene (phyA)	(Liu et al., 2008)
	Dt2	Glyma.18G273600	GmJD_18G0289500	*dt2/dt2*	indeterminate	(Ping et al., 2014)
	SHAT1-5	Glyma.16G019400	GmJD_16G0022500	*GsSHAT1-5*	Thinner FCC secondary walls	(Dong et al., 2014)

Evolutionary relationship between JD17 and other soybean lines

We compared the genome sequencing results of the JD17 genome and 68 lines (including 10 wild soybean lines from different regions) [2]. SNVs and small Indels were obtained (**Table S2**) through aligning the Illumina sequencing results of 68 soybean strains to the JD17 reference genome by using BAW [18], followed by calling variations using GATK [19]. In order to determine the evolutionary status of JD17, all these variations were merged, obtaining a total of 18,156,237 SNV sites and 2,307,757 small Indel (≤ 10 bp) sites. After quality filtering, 5,544,995 SNV sites were used to construct a neighbor-joining tree. The results showed that JD17 was closely related to Williams (PI548631). This result is consistent with that Williams was the parental ancestor of JD17 [12]. JD17 was also closely related to the cultivated varieties Fen Dou No.85, Cang Dou-11 and Jin Da No.52, which were all cultivars in the Huang-Huai-Hai region (Fig. 2).

Characterization of flower color related genes in JD17

In order to determine the genes of important agronomic traits in JD17, we investigated a lot of literature and found 140 genes in total (**Table S3**). Among them, 25 genes related to soybean flowering were discovered (**Table S3**). Six of them control and regulate soybean flower color, including *W1, W2, W3, W4*, *Wm* and *Wp* [20]. Cloning results showed that these six locus were *F3'5'H* [21], MYB [22], *DFR1* [23],*DFR2* [24] *FLS1*/*gmfls1* [25] and *F3H* [26], respectively (Fig. 1). The flower color of soybeans with dominant *W1W3W4* genotype is dark purple, the flower color of *W1W3w4* genotype is light purple, the flower color of *W1w3W4* genotype is purple, and *W1w3w4* has nearly white flowers [27]. Subsequent research showed that *w1w3W4* has white flowers [23]. The F3'5'H gene was cloned in Clark and B09121 strains [21]. The flower colors of the two soybean lines are purple (*W1*) and light purple (*w1-lp*). Compared with the wild-type *W1*, the difference between the mutant-type *w1-lp* is that a single base G is replaced with A at the 653 bp after the initial codon, while the mutant-type *w1* has a 65 bp insertion (Fig. 3a). At the *W1* locus, the sequence of JD17 is exactly the same as Clark-*w1* (white flower), and there is such a 65 bp sequence insertion, indicating that the genotype of JD17 is *w1w1*, which is consistent with the white flower phenotype of JD17. For the two genes that regulate flower color, *W3* (*DFR1*) and *W4* (*DFR2*), the genotype of *w3w4* is characterized by nearly white flowers, while the

genotype of *W3w4* is characterized by purple throat flowers [23]. The DFR1 gene was first cloned in L70-4422, L68-1774, Harosoy and Williams 82 [23]. Comparative analysis found that the genotype of JD17 was consistent with Wm82 except that there was no loss of 32 bp (Fig. 3b). Although they have this difference in sequence, their flower color is still the same, indicating that the 32 bp deletion has no effect on flower color. The *W4* (*DFR2*) gene was cloned in Clark and kw4, Bay and E30-D-1 [24]. The comparative analysis found that the key site of JD17 gene sequence is exactly the same as that of Bay (*W4*) (Fig. 3c), which means that the genotype of JD17 is also *W4*. Therefore, we determined that the gene of JD17 is *w1w3W4* with white flowers, which is the same as Wm82 [23].

The *Wm* (*FLS*/*gmfls1*) gene was first cloned in Harosoy (*WmWm*) and Harosoy-*wm* (*wmwm*). The homozygous wild-type Harosoy (*WmWm*) has purple flowers, while the homozygous mutant Harosoy-*wm* (*wmwm*) is magenta [25]. The change in flowers color here is due to the deletion of a single base G from *FLS1* in Harosoy. In the Wm82 strain, the *FLS1* gene is exactly the same as Harosoy, and they are all homozygous *WmWm* without single-base G deletion [25]. The gene sequence of JD17 is consistent with Wm82 and Harosoy, except for a single base A insertion in the intron region (Fig. 3d). It is inferred that JD17 should also be a white flower, which is the same as Wm82.

Characterization of flowering time-related genes in JD17

Among the genes related to flowering (**Table S3**), there are 19 genes related to the flowering cycle, including 3 genes *FT2c* [28], *J* gene [29], and *E1* gene [30] DNA mutation sites were determined at the molecular level (Fig. 1). The comparison found that the *FT2c* gene in the JD17 strain is the same as the allele carrying wild-type *GsFT2c* (Fig. 4a). The flowering time of plants with wild-type *GsFT2c* was about 2.8 days earlier than that of Wm82 under short-day light, and about 2 days later under long-day light [28]. The mutations at this locus speculate that JD17 should have a short flowering cycle under short-day light, about 29 days, and about 45 days under long-day light.

The *J* gene was cloned in PI 159925, BR121 and Harosoy [29]. In these strains, PI 159925(*j*) has a single base C deletion at 1,352 bp after the start codon compared with Harosoy(*J*), which leading to the early termination of encoding (Fig. 4b). However, BR121(*j*) compared with Harosoy(*J*) has a 10 bp deletion at 587–596 bp after the initial codon, which resulting in a frameshift mutation. The variation in both cases delayed the flowering time under short-day sunlight, from about 32 days to about 45 days [29]. The sequence of this gene in JD17 is consistent with Wm82 and Harosoy (*J*), so it can be speculated that JD17 should behave to promote flowering under short-day light.

The *E1* gene was first cloned in the Harosoy-*E1* strain [30]. Compared with Wm82, the coding sequence of *E1* gene in the Harosoy-*E1* strain is G at 44 bp after the initial codon (Fig. 4c), while the mutant *e1-as* genotype is shown here as a missense mutation caused by a C base substitution, the genotype of *e1-fs* at 49 bp is a frameshift mutation caused by A base insertion, and the mutant *e1-as* genotype is shown here as a missense mutation caused by a C base substitution [30]. The comparative analysis showed that the genotype in the JD17 genome was consistent with *e1-as*, which suggested that JD17 should have a shorter flowering cycle.

Characterization of seed traits-related genes in JD17

Soybean seeds have many traits, including seed coat color, seed size, grain weight, oil content, protein content, etc., which are all very important agronomic traits.

The *GmIRCHS* gene is an important gene that controls the color of the seed coat. Compared with the TM strain, the study found that THM (A pigmented seed coat mutant Toyohomare) mainly caused the pigmentation of the seed coat through a 3.3-kb region deleted (Fig. 5a), from normal yellow to brown [31]. The result of sequence alignment revealed that the 3.3Kb was not lost in JD17, but about 56 kb was inserted between the adjacent J and CHS3, and the seed coat color of JD17 was still yellow [11].

The *GmHs1-1* gene controls the cracking of the seed coat and is closely related to water penetration. The genotype is distinguished based on mutations at 7 sites in the promoter region of the gene [32]. The sequence comparison evidence of

JD17 showed that it was completely consistent with the Wm82 sequence (Fig. 5b), the genotype was *Gmhs1-1*, and the phenotype was permeable seed coat.

In the seed weight study, the *SoyWRKY15a* gene was identified in cultivated soybean Suinong14 (SN14) and wild soybean ZYD00006. Through 6 polymorphic loci, 4 genotypes were defined, H1, H2, H3 and H4. The indel at -61 bp in the 5'UTR is a major and independent cis-motif variation (Fig. 5c) [33]. JD17 has the same genotype as Wm82, and can be determined to be H1. According to studies, the phenotype of this genotype has large seeds and relatively large leaves.

The *FATB1a* gene is related to the palmitic acid content of seeds. Compared with the Jack strain, there is a 254-kb deletion in the mutant N0304-303-3 (Fig. 5d), which directly affects the low-level palmitic acid content of soybean seeds, about 5% [34]. Through the result of sequence alignment, we found that there is such a sequence in JD17 (GmJD_05G0012300). The genotype is No deletion, Homozygous *FATB1a*, and it is speculated that the palmitic acid content phenotype of JD17 may be around 10%.

The *ZF351* gene is closely related to the oil content of soybean seeds, which is mainly determined by the A/T SNP at 767 bp upstream of the gene start codon (Fig. 5e), and the expression of the *GmZF351* allele with A at -767 bp is significantly higher than that of the *GsZF351* allele with T at -767 bp, and promotes the increase of oil content [35]. The result of sequence alignment shows that the genotype of JD17 is GmZF351, which is consistent with Wm82, and it is speculated that the oil content is higher. *GmDREBL* is another gene related to oil content, and its genotype is determined by the 7 sites upstream of the gene [36]. The sequence of JD17 shows that its genotype is consistent with Wm82 (variation type 2) (Fig. 6a). Since the two genes closely related to oil content in JD17 are the same as Wm82, it is inferred that the oil content of JD17 is also the same as Wm82, which is 18.2%.

GA20OX mainly affects the 100-seed weight of seeds, and its genotype is mainly determined by 29 sites in the promoter region (Fig. 6b) [37]. Sequence alignment showed that the genotype in JD17 was completely consistent with Wm82, and it was inferred that its 100-seed weight was 15–20 g, which is consistent with the phenotype [12].

NFYA gene is also related to oil content, and the main difference in its genotype comes from whether there is an insertion of ~ 1,500 bp in the promoter region (Fig. 6c) [36]. While the sequence of JD17showed that there is such an insertion, it is speculated that the oil content of JD17 is also similar to Wm82, which is 18–22% [12].

The *GmSWEET10a* and *GmSWEET10b* genes are related to the seed 100-kernel weight, fatty acid content, and protein content. They were identified by GWAS association analysis of more than 800 soybean lines [38]. Among them, the *GmSWEET10a* gene distinguishes different genotypes by 5 sites upstream of the gene and 5 sites of the gene body (Fig. 6d). The sequence of JD17 is showed be H_III-3 type through aligned with the region of Wm82. *GmSWEET10b* mainly defines different genotypes based on the differences of 17 sites (Fig. 6e). According to the sequence characteristics of JD17, its genotype was identified as H_III-1 type. According to the phenotypic statistics in the article [38], we can roughly speculate that the seed weight of JD17 is about 18 g, the fatty acid content is about 20%, and the protein content is about 36%[12].

Characterization of genes underlying other important traits

There are also genes related to other important traits, such as the gene *GmphyA2* related to pigment synthesis [39], the gene *Dt2* related to stem growth [40], the gene *SHAT1-5* (Fig. 1), which is closely related to fruit pod dehiscence [41]. And the genotypes and related phenotypes of JD17 were identified and predicted respectively based on the characteristics of these published genotypes.

In TK780 and wild soybean Hidaka 4 (H4), the insertion of the repetitive sequence GTTTT changed the *GmphyA2* genotype from *E4* to *e4* (Fig. 7a), resulting in yellow flowers appearing on plants under continuous far-red light [39]. By determining that the genotype of JD17 is consistent with the wild type H4, it is inferred that its phenotype is also consistent with that of the wild type H4.

The genotype of *Dt2* is mainly determined by the differences in the three sites within the three genes (Fig. 7b) [40]. Although there are some differences between the sequence of JD17 and Wm82 in these regions, these three key sites are consistent. It is concluded that the genotype of JD17 is consistent with Wm82 as *dt2/dt2*, and the predicted stem growth phenotype is indeterminate.

SHAT1-5 gene is a gene related to the control of pod dehiscence identified in cultivated soybean HEINONG44 and wild soybean ZYD00755, mainly by promoting the thickening of the secondary wall of fibrous cap cells to inhibit pod dehiscence [41]. The genotype of JD17 was inferred to be *GmSHAT1-5* based on the deletion of its sequence (Fig. 7c). It is speculated that the pod phenotype of JD17 is the same as that of cultivated soybean HEINONG44, and it is easy to crack.

Discussion

In this study, we successfully constructed a platinum-grade *de novo* assembly and annotation of JD17 genome. The contig N50 of the JD17 genome assembly reached approximately 18 Mb (**Figure S1**), taking the advantage of long PacBio reads, deep coverage (110X). The chromosomal continuity was obtained through Hi-C analysis. The high-quality assembly of JD17 led to the reduced number of gaps within genes and between genes (Table 3). The JD17 genome assembly, whose contig N50, scaffold N50, and accuracy have reached 18.0 Mb, 50.6 Mb, and 99.999%, respectively, suggesting that it has reached platinum-grade quality.

The high quality JD17 was further annotated using Iso-Seq long reads, which significantly improved gene annotation with increased alternative-splicing isoforms, more complete 3'UTRs and transcripts (Table 3). Therefore, the JD17 genome could be widely used as a high-quality reference genome for soybean biology and molecular breeding.

Taking advantage of this platinum-grade genome assembly, we successfully identified the molecular sequences that explain important traits of JD17. By aligning the sequence of cloned genes related to important agronomic traits to the genome sequence of JD17, the genotypes of these genes in JD17 were determined. And based on previous studies, it is consistent with the phenotype, such as white flowers, flowering cycles, yellow seed coats, 100-seeds weight, oil and protein content and so on [11, 12]. And combined with the results of SNPs, it is determined that there are few SNPs near these genes, especially no SNPs exist at the key sites for determining the genotype. High-quality assembly and more precise annotation will help gene location and cloning, and provide accurate locus and mutation information for its genotype. I believe that high-quality genomes will provide a good foundation for molecular cloning, GWAS, and genetic breeding research, especially the discovery of QTLs for important traits in the genetic background of JD17. It will also become an important foundation for soybean genetic diversity research

Conclusions

In summary, we have integrated a platinum-grade genome through PacBio and Hi-C technology, and obtained a more complete genome-wide annotation. Our precise genome and annotation will help accurately mapping the trait-related genes, important agronomic traits and their genotypes and genetic analysis. It will also provide important resources for soybean diversity research and pan-genome research.

Methods

Plant and Sample preparation

Soybean seeds (*Glycine max* cv. JD17) used in this study were from Hebei Academy of Agricultural and Forestry Sciences. The seeds are planted and extracted in Xuelu Wang's Laboratory in Huazhong Agricultural University. The seeds were sterilized with chlorine gas (5 ml of 32% (w/w) HCl to 100 ml 4–5% (wt/vol) sodium hypochlorite in a beaker) for 15 h [42] and then left in a sterile hood for 2 h. The sterilized seeds were sown in growth bottles filled with sand after being soaked in sterile Milli-Q water

for 30 min and watered with sterile Fahraeus solution [43] containing 2 mM KNO_3. Seeds were grown in a growth chamber (light) at 28℃ and 8 h (dark) at 23℃ with 60% humidity for 16 h.

RNA preparation and Sequencing

1, 4, 6, 8, 10, 15, 20, 25 and 30 dpi seedings were selected to carry out PacBio isoform sequencing. Underground tissues of inoculation and uninoculation from the 9 timepoints were collected and used for RNA extraction respectively. After that, nine RNA samples of inoculation and uninoculation were mixed equally as a sample, respectively. The two RNA samples were prepared for PacBio isoform sequencing (Iso-Seq). In addition, we also selected different tissues including root, nodule, stem, leaf, pod, seed and flower for mixed RNA-SEq. All the RNA was extracted by TRIzol reagent (Invitrogen 15596026). we performed RNA-Seq on illumina platform and produced approximately 10 Gb raw data with 150 bp pair-end reads.

Whole-genome sequencing using SMRT technology

10-day post inoculation (dpi) root tissue of plants was used for SMRT whole-genome sequencing. Underground tissues were collected for genomic DNA preparation with modified CTAB method [44]. Using 97 ug DNA, PacBio sequencing libraries were produced following manufacturers protocols as described for the Greater than 30 kb-SMRTbell Libraries Needle Shearing (SMRTbell Template Prep Kit 1.0) with Blue Pippin size selections (Sage Science, http://www.sagescience.com/), and the SMRTbell libraries were constructed through Pacific Biosciences SMRTbell Template Prep Kit 1.0 (http://www.pacb.com/). SMRT sequencing was performed on a PacBio RSII instrument using P6/C4 sequencing chemistry (DNA/Polymerase Binding Kit P6) and 6hr movies. We used a total of 118 SMRT cells and produced 127.3 Gb of raw data with an average subread length of 15 kb (**Table S**4). At the same time, we also used a part of DNA samples for resequencing with a sequencing depth of approximately 60X. These data are mainly used for the evaluation of genome size and its heterozygosity, post-assembly error correction and genome quality assessment.

Hi-C library construction and sequencing

For samples used for Hi-C assisted assembly, leaves fixed in 1% (volume / volume) formaldehyde were used for library construction. Cell lysis, chromatin digestion, proximity ligation treatment, DNA recovery and subsequent DNA manipulation were performed as previously described [45]. The restriction enzyme used in chromatin digestion is Mbol. Finally, the Hi-C library was sequenced on the Illumina HiSeq X 10 platform for 150 bp paired end reads, with a sequencing depth of approximately 100X.

De novo **genome assembly of JD17**

To perform *de novo* assembly of the JD17, we combined three different assemblers, including CANU [46] (v1.4), FALCON (v1.8.7 23 Mar 2017, https://github.com/PacificBiosciences/FALCON-integrate) and HGAP4 (SMRT Link v 5.0.1.9585) (**Table S5**). The main assembly was performed on whole SMRT sequenced long reads. All assembly softwares were performed with a presumed 1-Gb genome size. If not specified, all programs in our study were run with default parameters. CANU was run with 'errorRate = 0.013', and FALCON was run with 'length_cutoff = -1' for initial mapping of seed reads for the error-correction phase. For a better FALCON assembly, we additionally optimized parameters as 'DBsplit = -x500, –s400, pa_HPCdaligner = -v -B128 -t16 -e.70 -l1000 -s1000 -T8 -M24, ovlp_HPCdaligner = -v -B128 -t32 -h60 -e.96 -l500 -s1000 -T8 -M24 and overlap_filtering = –max_diff 60 –max_cov 60 –min_cov 2'. The stats of three initial assemblies were show in **Table S5**. We subsequently used the CANU assembly as the working set because it was better able to phase the diploid genome as well as indicating better contiguity from N50 evaluations. Subsequently, the draft assembly was polished using Quiver (SMRT Link v 5.0.1.9585) over twice iterations and finally corrected using Illumina short reads with Pilon (**Figure S3**). Then filtering contigs that were not part of soybeans were aligned to NT library using blastn.

Continuation and connection of contigs

Comparing three different assembly results from CANU [46], FALCON and HGAP4, we found that they have some complementarity, so we continue to further extend and connect the CANU contigs to optimize our assembly results. The extended and connected contigs were performed on CANU contigs using the GPM [47] pipeline. Firstly, GPM will loading Wm82

as reference genome, and loading CANU assembly as input. These contigs were ordered and located on chromosome based on Wm82 (Glycine_max_v2.0) using blastn. This step produced a draft chromosome assembly, as JD17 v0.1. Secondly, we loaded FALCON assembly result and aligned with CANU assembly using blastn [48], then we can get the relationships between CANU and FALCON contigs. So, the CANU contigs can be extended or connected by FALCON contigs, which is the JD17 v0.2. Thirdly, we loaded HGAP4 assembly and repeat the above steps, as the JD17 v0.3. Fourthly, we loaded parts of contigs from assembly was performed on 70 Gb PacBio long reads using CANU and repeat the above steps, as the JD17 v0.4. The whole 120 Gb PacBio long reads are sort by length of reads, and extract 70 Gb reads according the length of reads. In whole process, these extended or connected contigs are supported by PacBio subreads to avoid introducing new errors. Finally, the contigs belong to chloroplast and mitochondrial sequences were removed from JD17 v0.4 (**Table S6**). At every steps, we merged these assemblies by GPM [47], polished them with Quiver and Pilon [49].

But the JD17 v0.4 was also polished using Quiver over twice iterations and corrected using Illumina short reads with Pilon [49]. The JD17 genome assembly (Glycine_max_JD17v1.0) was 995.0 Mb in size, with a contig N50 of 18.0 Mb (Table 1). The detailed assembly process is shown in the **Figure S3**.

The assessment of genomic heterozygosity and size is using the Genomic Character Estimator program, gce v 1.0.0 (ftp://ftp.genomics.org.cn/pub/gce), and the heterozygous ratio based in kmer individuals is 0.029, and the corrected estimate of genome size is about 1.11 Gb.

To anchor hybrid contigs into chromosome, the Hi-C sequencing data were aligned into contigs using bwa. According to the orders and orientations provided by the alignment, those contigs were clustered into chromosomes by ALLHiC v0.9.8 [50] with default params. According to the ALLHiC groups and assembly results create hic files, manual correction and validation were also performed by drawing contact maps with juicerbox [51]. The genome assembly was finalized after this correction (**Table S6**).

Quality assessment of JD17 genome assembly

To assess the quality of the Glycine_max_JD17v1.0 assembly, we used our 65 Gb resequencing data. First, by aligning all reads to the assembly with BWA-MEM in BWA [52], the mapping rate is over 98.8% and the coverage was over 99.65%, which shows the consistency between the assembly and reads. By using the GATK tools [13, 14] for SNPs calling with JD17 resequencing data, we found 78,033 SNP, of which only 8,000 were homozygous, indicating that the JD17 genome has an accuracy of over 99.999% (**Table S2**). The completeness of the assembly was estimated by BUSCO with default parameters.

Annotation of TE and ncRNA sequences

To investigate the JD17 genome sequence features, we identified transposable elements (TEs) and other repetitive elements by RepeatMasker [44]. MITEs (miniature inverted transposable elements) were collected by MITE-Hunter [53] with all default parameters. In order to get as much reliable LTR (long terminal repeat) retrotransposons information as possible, we used the LTR_retriever [54] analysis process, which integrates the output of LTR_FINDER [55] and LTRharverst tools in GenomeTools [56]. Masking sequence with RepeatMasker (version 4.0.8) (http://www.repeatmasker.org/) based on MITEs and LTR library that has been identified. The other tandem repeats were identified by constructing a *de novo* repeat library using Repeatmodeler (version 1.0.11) (http://www.repeatmasker.org/). RepeatMasker was run against the genome assembly again, with all above library as the query library.

Non-coding RNAs were predicted by the Infernal program using default parameters [57] and comparing the similarity of secondary structure between the JD17 genome sequence and Rfam [58] (v12.0) database.

Annotation of protein-coding genes

We performed gene calling analysis with MAKER [59, 60], by using multi-sourced EST and protein sequences as evidences (including nonredundant soybean EST/protein sequences from NCBI, assembled JD17 RNA-Seq from mixed samples and Iso-Seq data from underground samples, Wm82 transcripts and proteins, *Lotus japonicus* and *Medicago truncatula* protein

sequences), AUGUSTUS [61] and SNAP [62] as *ab initial* gene predictors, and a customized repeat library for RepeatMasker [44, 63].

The RNA-Seq data was *de novo* assembled using Trinity [64] to obtain the assembled cDNA sequence, and annotated with the PASA tool along with the non-redundant isoforms sequence from Iso-SEq. The annotation results will be used for AUGUSTUS model training and prediction. At the same time, these Iso-Seq de-redundant sequences are also used by MAKER for preliminary prediction, and the results are used for de novo sequencing of SNAP. In order to obtain more accurate and complete annotation results, we will use the AUGUSTUS trained model, the SNAP hmm model, RNA-Seq and Iso-Seq evidence, ESTs from NCBI, William 82 transcripts, RepeatMasker annotated repeated sequence results, And protein sequences from multiple species are combined as MAKER's input evidence to make predictions, and EVM is called to integrate the prediction results, and finally the PASA software [65] is used to update these annotation results. The detailed process is shown in the **Figure S4**.

Finally, in order to obtain more complete alternative splicing information, we manually added the dropped Iso-forms. These transcripts are from the full-length transcript sequence of Iso-seq, and the results are generated by using GMAP [66] alignment to the genome.

Function annotation of protein sequence were predicted by interproscan software with InterPro databases [67]. Nucleotide sequences were aligned to the ‘nr’ database by BLASTX [48] (version 2.2.28+). Gene Ontology terms for all genes were annotated by Blast2GO [68].

Blastn were used to search highest hit of genes between two genomes (identity > = 90%). For the multiple genes which corresponded to one gene, we only retained the adjacent genes that mapped to different positions within one gene.

Evolutionary analysis

Data source of evolution analysis: The 68 strains used for evolution analysis are all from BIG Data Center (http://gsa.big.ac.cn/index.jsp) under Accession Number PRJCA000205 [69]. The selection of these strains is mainly based on different regions and higher sequencing depth. The detailed strains are shown in the **Table S2.**

By comparing these 68 strains with JD17, using the bwa tool to compare to the JD17 reference genome, and then using the GATK tool for SNP calling analysis, merge these result for evolutionary analysis, set the filter value to the filter value is "SNP: QD = 2.0, FS = 60.0, MQ = 40.0, MQRankSum = -12.5, ReadPosRankSum_snp = -8.0, ReadPosRankSum_indel = -20.0; INDE: QD = 2.0, FS = 200.0", where Indel sets the length filter value to $\leq$ 10 bp. Then use TreeBeST software to use the SNP set of all samples to construct a neighbor-joining tree.

Abbreviations

JD17: Jidou 17

Wm82: Williams 82

ZH13: Zhonghuang 13

SMRT: single-molecule real-time

GATK: Genome Analysis Toolkit

bp: Base pair

BUSCO: Benchmarking Universal Single-Copy Orthologs

UTR: untranslated region

CDS: Coding sequence

GPM: Genome puzzle master

SNVs: single-nuclotide variations

SNPs: single-nuclotide polymorphisms

Indel: Insertion and deletion

GWAS: Genome-wide association study

QTLs: Guantitative trait loci

CTAB: Cetyl trimethylammonium bromide

transposable elements (TEs)

MITEs (miniature inverted transposable elements)

LINEs:Long interspersed elements

LTRs:Long terminal repeats

SINEs: Short interspersed elements

Iso-seq: Full-length cDNA Sequencing of Alternatively Spliced Isoforms

RNA-seq: High-throughput messenger RNA sequencing

NCBI: National Center for Biotechnology Information

Declarations

Availability of data and materials

The sequencing data (PacBio Whole Genome Sequencing data for assembly, resequencing data, Hi-C data and Mixed-Sample RNA-seq data for Annotation) used in this study have been deposited into National Center for Biotechnology Information under BioProject Number PRJNA412346 with accession number SRR12416523 - SRR12416632, SRR12416444, SRR12416750 and SRR9643849 - SRR9643851. The final assembly had been deposited at GenBank JACKXA000000000.

Funding

The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. All costs of this work was supported by grant 2016YFD0100700 of The National Key Research and Development Program of China (to B. Z.) and 2015CB910200 of The National Key Basic Research Foundation of China (to X.W.), grants 31870257, 91535104 and 31430046 of The National Natural Science Foundation of China (to X.W.), and grant 31471522 of The National Natural Science Foundation of China (to M. Z.).

Contributions

NC, JZ and XW coordinated the research. XY, SC, HW, LL, BA and WF prepared the DNA and RNA samples for sequencing. SL, DK and RW performed PacBio sequencing. XY, JL, HW, ML, QU and LL performed assembly analysis and annotation of the genome and transcriptome. CY and MZ provided the plant seed from Hebei Academy of Agricultural and Forestry Sciences. XY, JZ, NC and XW wrote the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

All experiments, including experiments involving soybean planting, tissue collection, and sequencing, comply with all applicable laws and NIH guidelines.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J: **Genome sequence of the palaeopolyploid soybean**. *nature* 2010, **463**(7278):178.
2. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33(4):408–14.
3. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M: **Pan-genome of wild and cultivated soybeans**. *Cell* 2020.
4. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, Watson MJG. **A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data**. 2015, 4(1):60.
5. Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, May G, Lin HJNc. **A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping**. 2018, 9(1):4844.
6. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(7659):524–7.
7. Li Y, Wei W, Feng J, Luo H, Pi M, Liu Z, Kang C. **Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets**. *DNA Res* 2017.
8. Jiang X, Hall AB, Biedler JK, Tu Z. Single molecule RNA sequencing uncovers trans-splicing and improves annotations in Anopheles stephensi. Insect Mol Biol. 2017;26(3):298–307.
9. Magrini V, Gao X, Rosa BA, McGrath S, Zhang X, Hallsworth-Pepin K, Martin J, Hawdon J, Wilson RK, Mitreva M. Improving eukaryotic genome annotation using single molecule mRNA sequencing. BMC Genom. 2018;19(1):172.
10. Qin J, Wang F, Gu F, Wang J, Chen Q, Zhang M: **A genetic composition analysis of soybean sibling varieties Jidou17 and Jinf58**. *Australian Journal of Crop Science* 2014, **8**(5):8.
11. Zhao S, Zhao X, Tang X, Zhang J, Xu Y, Feng Y, Zhang M. High yield characteristics of summer sowing soybean varieties. Soybean Science. 2013;2013:168–75.
12. Zhao Q, Yan L, Liu B, Di R, Shi X, Zhao S, Zhang M, Yang C. Breeding of High-yield Widespread and High-quality Soybean Cultivar Jidou 17. Soybean Science. 2015;34(4):000736–9.
13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S. Daly MJGr: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. 2010, 20(9):1297–1303.
14. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA. Hanna MJNg: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. 2011, 43(5):491.
15. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
16. Shen Y, Liu J, Geng H, Zhang J, Liu Y, Zhang H, Xing S, Du J, Ma S, Tian Z. De novo assembly of a Chinese soybean genome. Sci China Life Sci. 2018;61(8):871–84.

17. Xie M, Chung CY, Li MW, Wong FL, Wang X, Liu A, Wang Z, Leung AK, Wong TH, Tong SW, et al. A reference-grade wild soybean genome. Nat Commun. 2019;10(1):1216.
18. Li H. **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. *arXiv preprint arXiv:13033997* 2013.
19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
20. Palmer RG, Pfeiffer TW, Buss GR, Kilen TC. **Qualitative genetics**. *Soybeans: improvement, production, and uses* 2004, 16:137–233.
21. Takahashi R, Dubouzet JG, Matsumura H, Yasuda K, Iwashina T. A new allele of flower color gene W1 encoding flavonoid 3'5'-hydroxylase is responsible for light purple flowers in wild soybean Glycine soja. BMC Plant Biol. 2010;10:155.
22. Takahashi R, Yamagishi N, Yoshikawa N. A MYB transcription factor controls flower color in soybean. J Hered. 2013;104(1):149–53.
23. Park GT, Sundaramoorthy J, Lee JD, Kim JH, Seo HS, Song JT. Elucidation of Molecular Identity of the W3 Locus and Its Implication in Determination of Flower Colors in Soybean. PLoS One. 2015;10(11):e0142643.
24. Yan F, Di S, Rojas Rodas F, Rodriguez Torrico T, Murai Y, Iwashina T, Anai T, Takahashi R. Allelic variation of soybean flower color gene W4 encoding dihydroflavonol 4-reductase 2. Bmc Plant Biology. 2014;14(1):58.
25. Takahashi R, Githiri SM, Hatayama K, Dubouzet EG, Shimada N, Aoki T, Ayabe S, Iwashina T, Toda K, Matsumura H. A single-base deletion in soybean flavonol synthase gene is associated with magenta flower color. Plant Mol Biol. 2007;63(1):125–35.
26. Zabala G, Vodkin LO. The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell. 2005;17(10):2619–32.
27. Hartwig EE, Hinson K. Inheritance of Flower Color of Soybeans 1. Crop Sci. 1962;2(2):152–3.
28. Wu F, Sedivy EJ, Price WB, Haider W, Hanzawa Y. Evolutionary trajectories of duplicated FT homologues and their roles in soybean domestication. Plant J. 2017;90(5):941–53.
29. Lu S, Zhao X, Hu Y, Liu S, Nan H, Li X, Fang C, Cao D, Shi X, Kong L, et al. Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. Nat Genet. 2017;49(5):773–9.
30. Xia Z, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, Anai T, Sato S, Yamazaki T, Lu S, et al. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc Natl Acad Sci U S A. 2012;109(32):E2155–64.
31. Kasai A, Kasai K, Yumoto S, Senda M. Structural features of GmIRCHS, candidate of the I gene inhibiting seed coat pigmentation in soybean: implications for inducing endogenous RNA silencing of chalcone synthase genes. Plant Mol Biol. 2007;64(4):467–79.
32. Sun L, Miao Z, Cai C, Zhang D, Zhao M, Wu Y, Zhang X, Swarm SA, Zhou L, Zhang ZJ, et al. GmHs1-1, encoding a calcineurin-like protein, controls hard-seededness in soybean. Nat Genet. 2015;47(8):939–43.
33. Gu Y, Li W, Jiang H, Wang Y, Gao H, Liu M, Chen Q, Lai Y, He C. Differential expression of a WRKY gene between wild and cultivated soybeans correlates to seed size. J Exp Bot. 2017;68(11):2717–29.
34. Goettel W, Ramirez M, Upchurch RG, An YQ. Identification and characterization of large DNA deletions affecting oil quality traits in soybean seeds through transcriptome sequencing analysis. Theor Appl Genet. 2016;129(8):1577–93.
35. Li QT, Lu X, Song QX, Chen HW, Wei W, Tao JJ, Bian XH, Shen M, Ma B, Zhang WK, et al. Selection for a Zinc-Finger Protein Contributes to Seed Oil Increase during Soybean Domestication. Plant Physiol. 2017;173(4):2208–24.
36. Zhang YQ, Lu X, Zhao FY, Li QT, Niu SL, Wei W, Zhang WK, Ma B, Chen SY, Zhang JS. Soybean GmDREBL Increases Lipid Content in Seeds of Transgenic Arabidopsis. Sci Rep. 2016;6:34307.
37. Lu X, Li QT, Xiong Q, Li W, Bi YD, Lai YC, Liu XL, Man WQ, Zhang WK, Ma B, et al. The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. Plant J.

2016;86(6):530–44.

38. Shoudong W, Shulin L, Jie W, Kengo Y, Bin Z, Ya-Chi Y, Zhi L, B FW, Feng MJ, Li-Qing C. **Simultaneous changes in seed size, oil content, and protein content driven by selection of SWEET homologues during soybean domestication**. *National Science Review* 2020.
39. Liu B, Kanazawa A, Matsumura H, Takahashi R, Harada K, Abe J. Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics. 2008;180(2):995–1007.
40. Ping J, Liu Y, Sun L, Zhao M, Li Y, She M, Sui Y, Lin F, Liu X, Tang Z, et al. Dt2 is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell. 2014;26(7):2831–42.
41. Dong Y, Yang X, Liu J, Wang BH, Liu BL, Wang YZ. Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nat Commun. 2014;5:3352.
42. Kereszt A, Li D, Indrasumunar A, Nguyen CD, Nontachaiyapoom S, Kinkema M, Gresshoff PMJNP. **Agrobacterium rhizogenes-mediated transformation of soybean to study root biology**. 2007, 2(4):948.
43. FÅHRAEUS GJM. **The infection of clover root hairs by nodule bacteria studied by a simple glass slide technique**. 1957, 16(2):374–381.
44. Bergman CM, Quesneville H. Discovering and detecting transposable elements in genome sequences. Brief Bioinform. 2007;8(6):382–92.
45. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *science* 2009, **326**(5950):289–293.
46. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. *Genome research* 2017:gr. 215087.215116.
47. Zhang J, Kudrna D, Mu T, Li W, Copetti D, Yu Y, Goicoechea JL, Lei Y, Wing RA. Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. Bioinformatics. 2016;32(20):3058–64.
48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform. 2009;10(1):421.
49. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one. 2014;9(11):e112963.
50. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nature plants. 2019;5(8):833–45.
51. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell systems. 2016;3(1):99–101.
52. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.
53. Han Y, Wessler, SRJNar. **MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences**. 2010, 38(22):e199-e199.
54. Ou S, Jiang NJPp. **LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons**. 2018, 176(2):1410–1422.
55. Xu Z, Wang HJNar. **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**. 2007, 35(suppl_2):W265-W268.
56. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinf. 2013;10(3):645–56.
57. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.
58. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J. Rfam 12.0: updates to the RNA families database. Nucleic acids research. 2014;43(D1):D130–7.

59. Holt C, Yandell, MJBb. **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects**. 2011, 12(1):491.
60. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D. Lawrence CJJPp: **MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations**. 2014, 164(2):513–524.
61. Hoff KJ, Stanke M. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic acids research. 2013;41:W123–8. (Web Server issue).
62. Korf IJBb. **Gene finding in novel genomes**. 2004, 5(1):59.
63. Saha S, Bridges S, Magbanua ZV, Peterson DG. Empirical comparison of ab initio repeat finding programs. Nucleic acids research. 2008;36(7):2284–94.
64. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29(7):644–52.
65. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB. Town CDJNar: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies**. 2003, 31(19):5654–5666.
66. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.
67. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang H-Y, Dosztányi Z, El-Gebali S, Fraser M. InterPro in 2017—beyond protein family and domain annotations. Nucleic acids research. 2016;45(D1):D190–9.
68. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.
69. Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, Hu G, Zhou Z, Yu H, Zhang M, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. 2017;18(1):161.

Figures

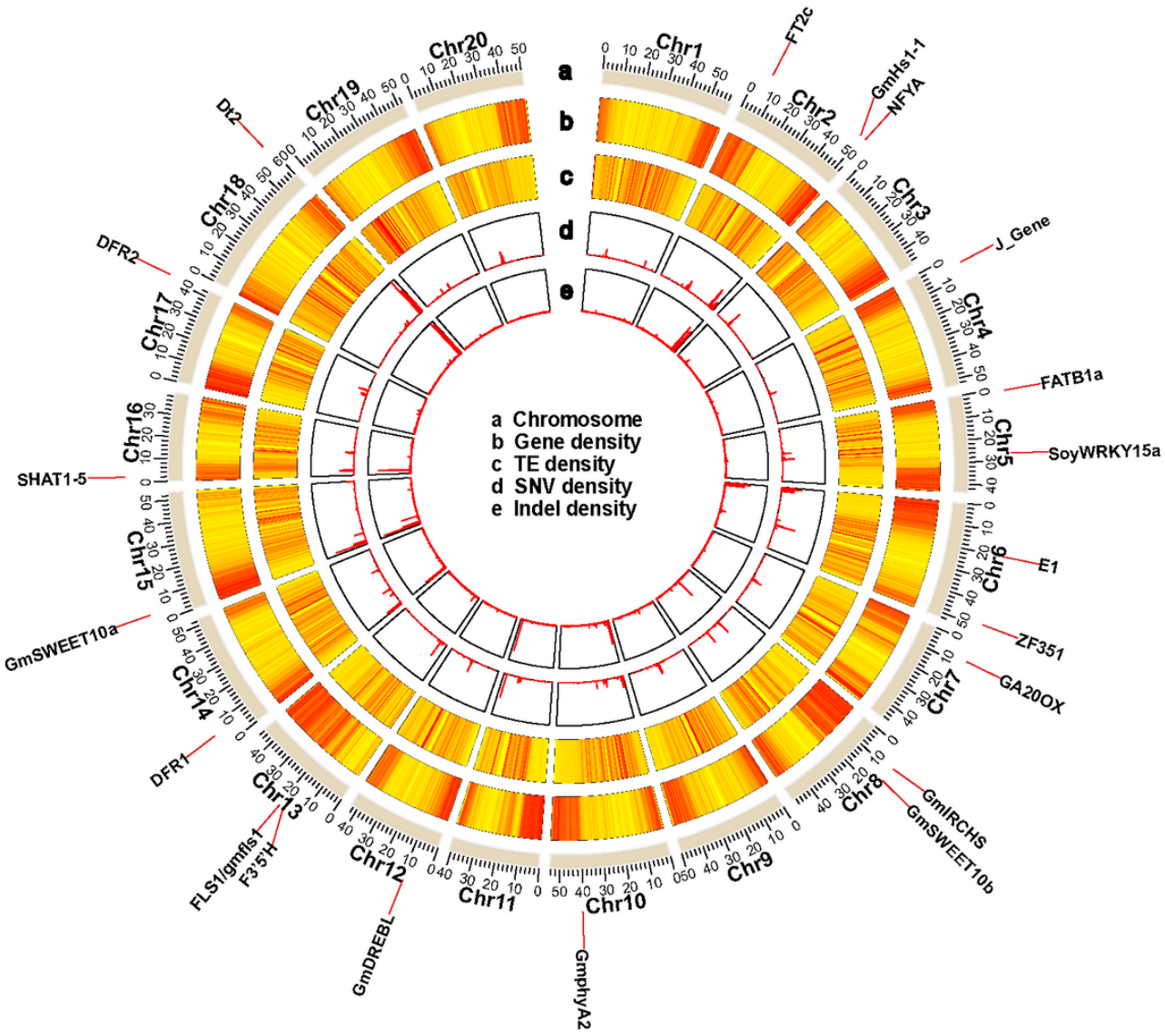

Figure 1

Figure 1

Overview of the JD17 reference genome Tracks from outer to inner circles indicate: a) the chromosome of the genome; b) the gene density map; c) the repeat sequence density map; d) density distribution diagram of SNP sites; e) density distribution diagram of InDel. The reddish color indicates greater density, the yellower color indicates lower density.

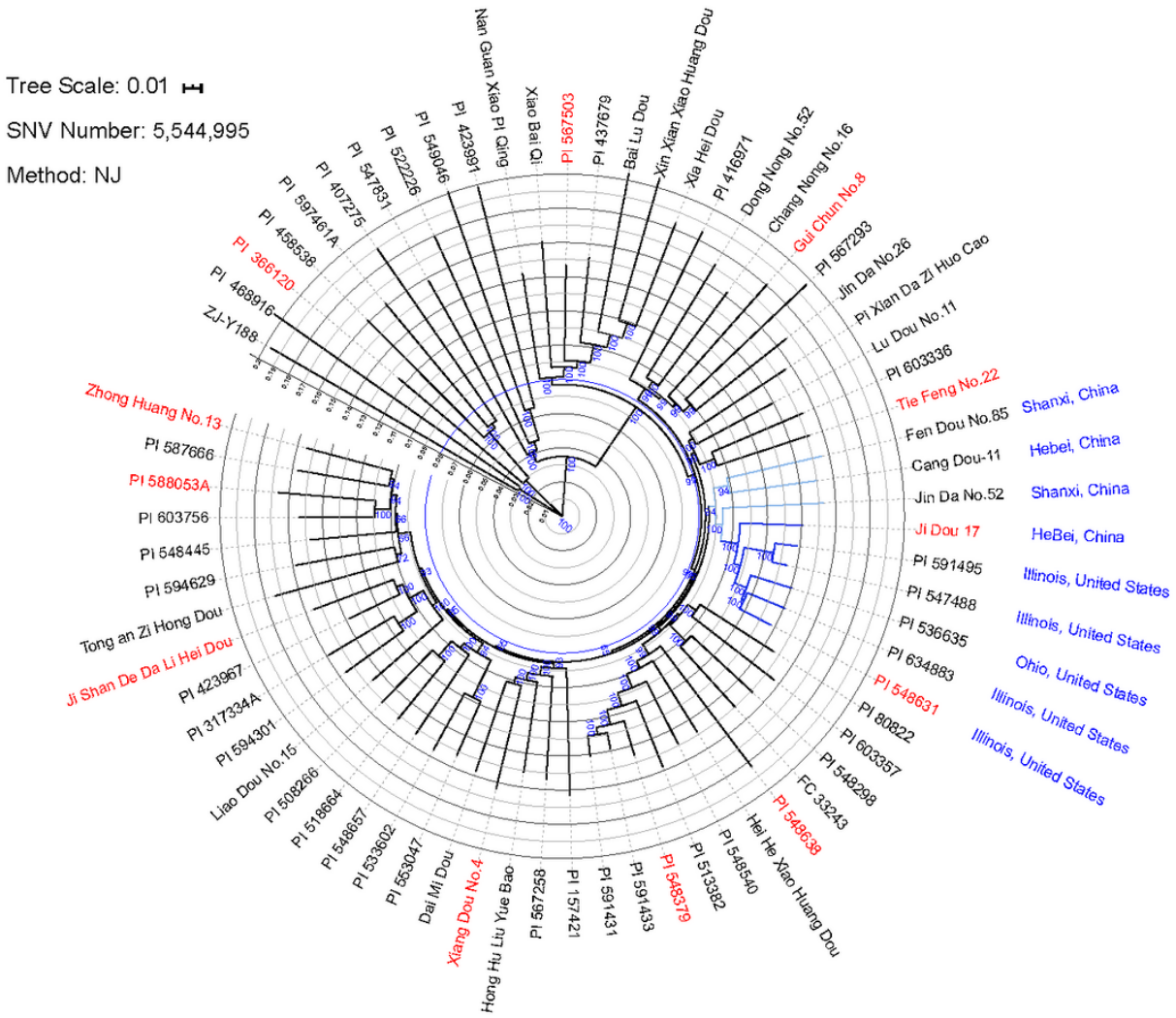

Figure 2

Figure 2

Evolutionary relationship between JD17 and other soybean lines The lines marked in red are the representative strains of each area, the blue lines are the branches where JD17 is located, and the blue text is the area to which the strains belong.

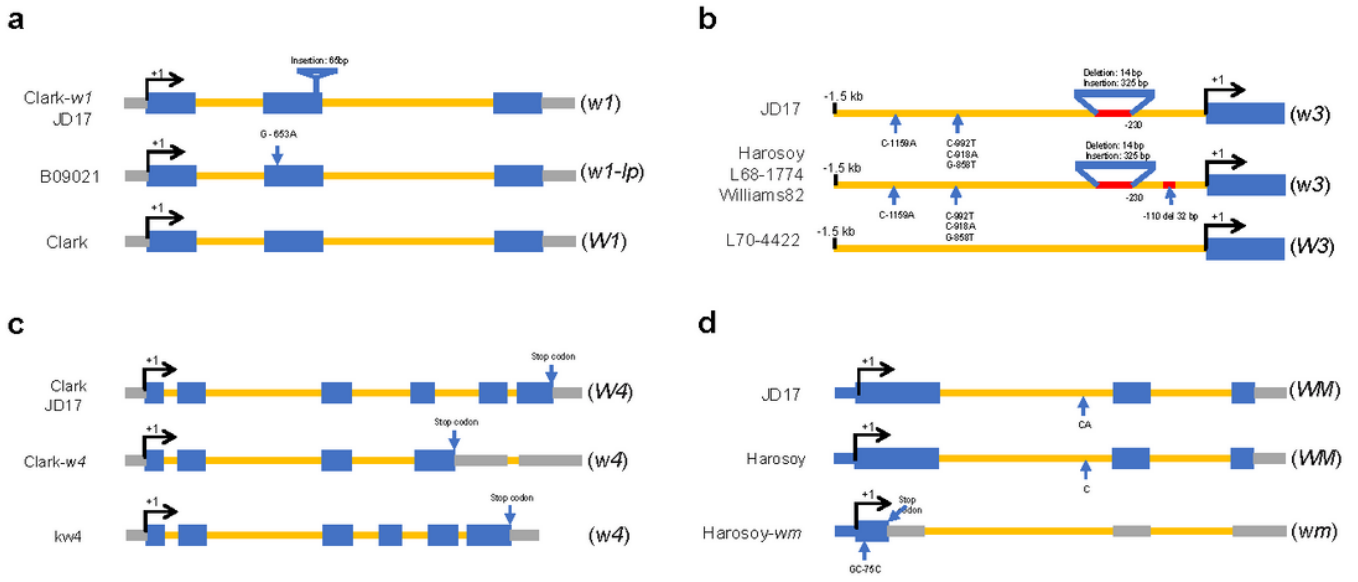

Figure 3

Figure 3

Gene models that regulate flower color a) Gene model of F3’5’H. b) Gene model of DRF1. c) Gene model of DRF2. d) Gene model of FLS1.

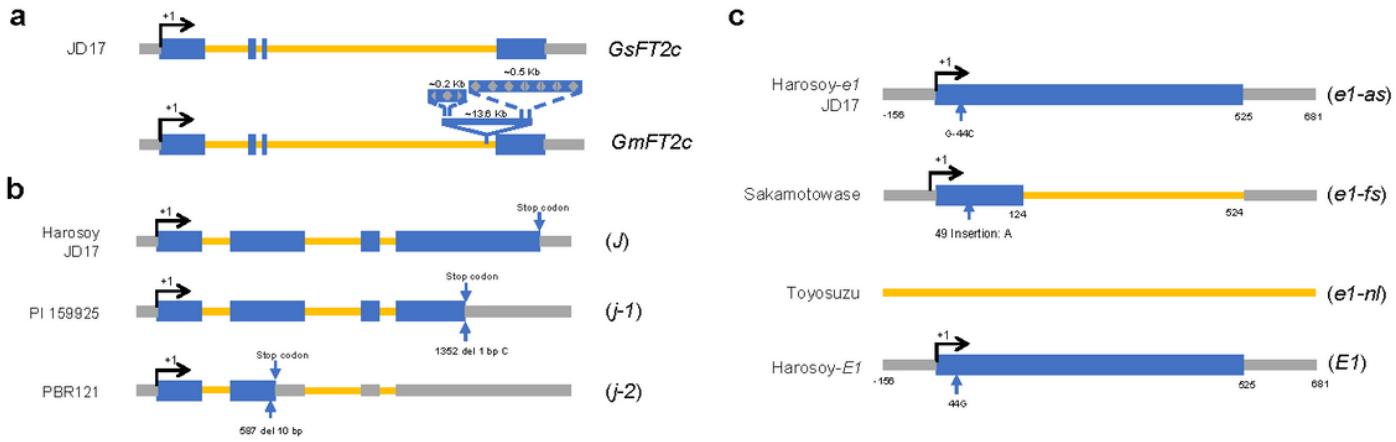

Figure 4

Figure 4

Gene models that regulate flowering time a) Gene model of FT2c. b) Gene model of J gene. c) Gene model of E1 gene.

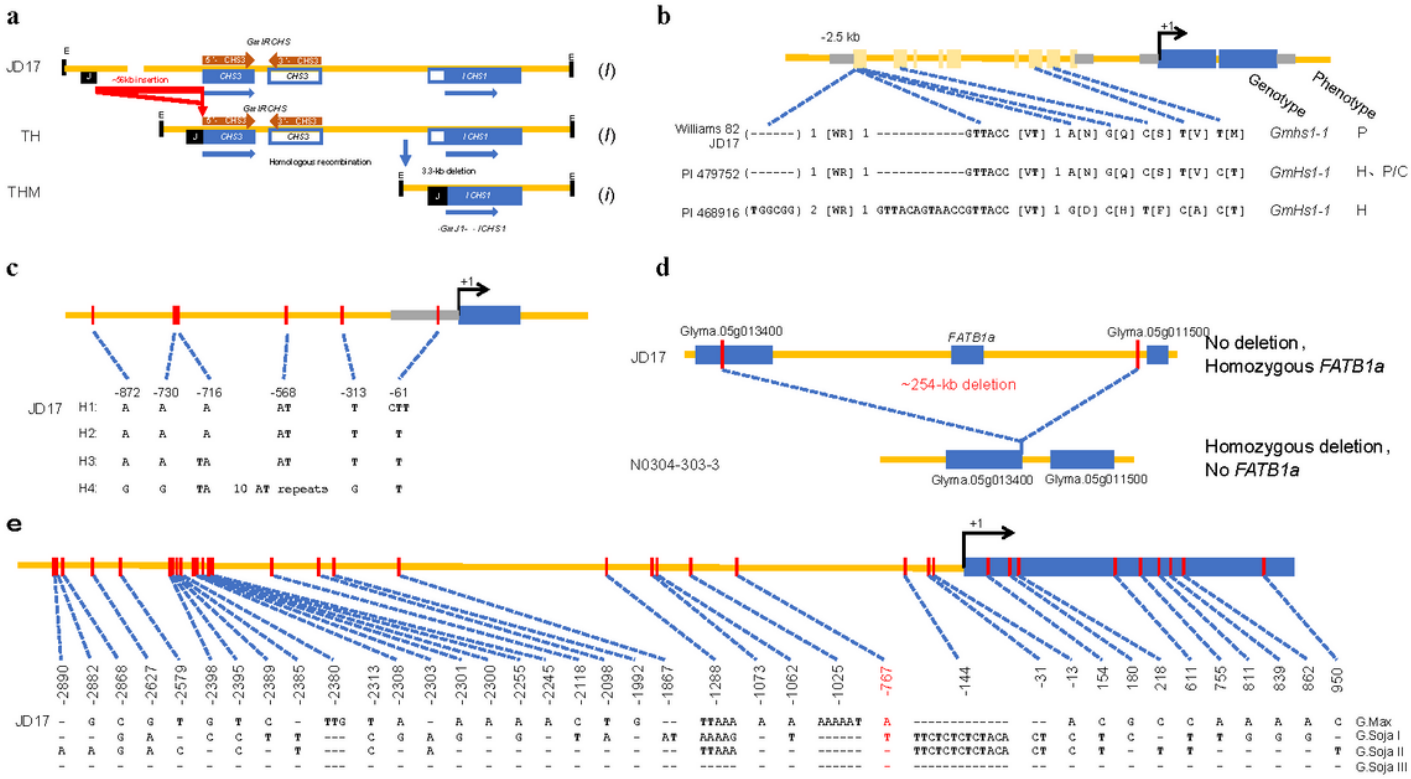

Figure 5

Figure 5

Gene models that regulate seed traits a) Gene model of GmIRCHS, which regulates the color of the seed coat. b) Gene model of GmHs1-1, which is related to the cracking of the seed coat. c) Gene model of SoyWRKY15a, which is related to the size of seed. d) Gene model of FATB1a, which is related to palmitic acid content of seed. e) Gene model of GmZF351, which is which is closely related to the oil content of seeds.

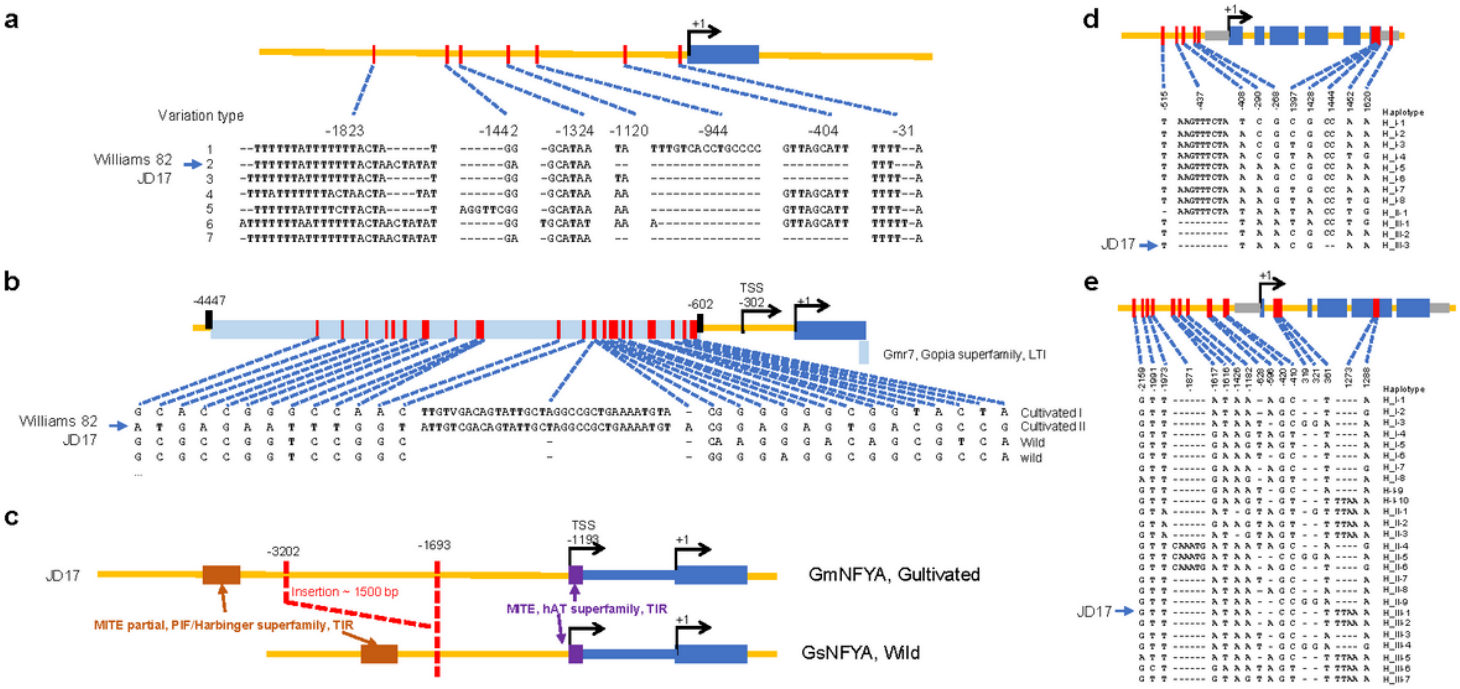

Figure 6

Figure 6

Gene models related to seed traits a) Gene model of GmDREBL, which is related to the oil content of seeds. b) Gene model of GA20OX, which is related to the 100-seed weight of seeds. c) Gene model of NFYA, which is related to the oil content of seeds. d) and e) Gene models of GmSWEET10a and GmSWEET10b, which are related to 100-seed weight, fatty acid content and protein content of seeds.

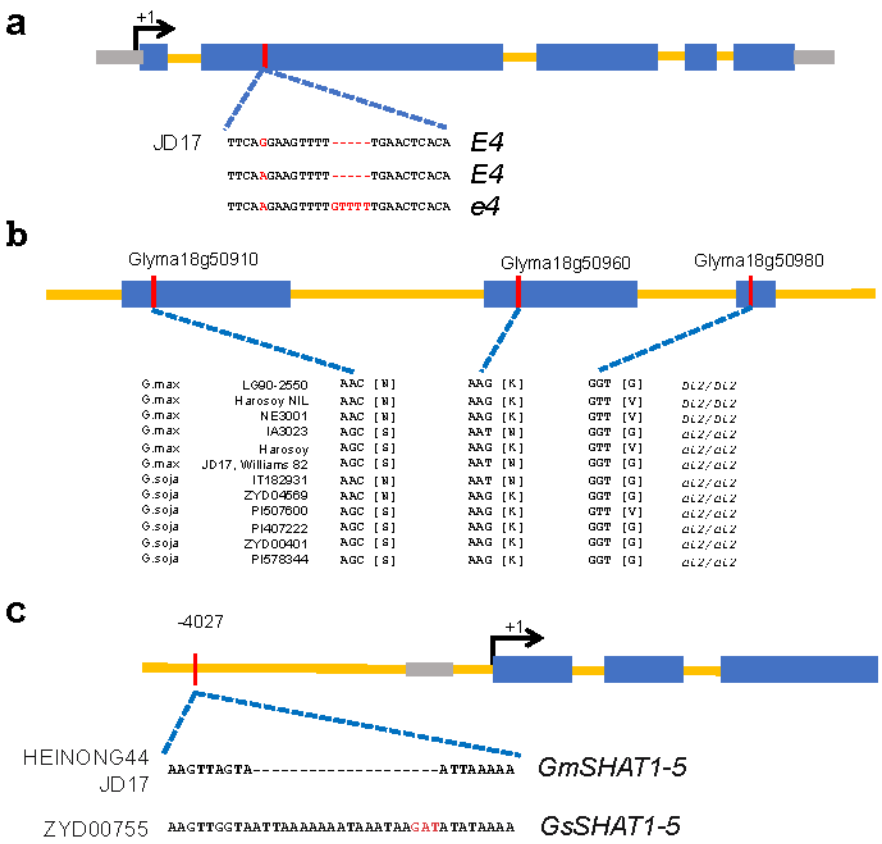

Figure 7

Figure 7

Gene models for other important agronomic traits a) Gene model of GmphyA2, which is related to pigment synthesis. b) Gene model of Dt2, which is related to stem growth. c) Gene model of SHAT1-5, which is related to fruit pod dehiscence.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureSsubmit.pdf
- SupplementaryTablessubmit.xlsx