Genome assembly of the JD17 soybean provides a new reference genome for Comparative genomics

Xinxin Yi¹, Jing Liu^{1,2}, Shengcai Chen^{1,2}, Hao Wu¹, Min Liu¹, Qing Xu¹, Lingshan Lei¹, Seunghye Lee³, Bao Zhang², Dave Kudrna³, Wei Fan^{1,2}, Rod A. Wing³, Chunyan Yang⁴, Mengchen Zhang⁴, Jianwei Zhang^{1,3,*}, Xuelu Wang^{2,*} and Nansheng Chen^{1,5*}

¹National Key Laboratory of Crop Genetic Improvement, Center of Integrative Biology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China

² State Key Laboratory of Crop Stress Adaptation and Improvement, Henan University, Kaifeng 475004, Henan, China

³Arizona Genomics Institute and BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA

⁴Institute of Food and Oil Crops, Hebei Academy of Agricultural and Forestry Sciences, Shijiazhuang 050031, Hebei, China

⁵Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

*e-mails: jzhang@mail.hzau.edu.cn, xueluw@henu.edu.cn, chenn@qdio.ac.cn

Abstract

Cultivated soybean (*Glycine max*) is an important source for protein and oil. Many elite cultivars with different traits have been developed for different conditions. Each soybean strain has its own genetic diversity, and the availability of more high-quality soybean genomes can enhance comparative genomic analysis for identifying genetic underpinnings for its unique traits. In this study, we constructed a high-quality *de novo* assembly of an elite soybean cultivar Jidou 17 (JD17) with chromosome contiguity and high

29 accuracy. We annotated 52,840 gene models and reconstructed 74,054 high-
 30 quality full-length transcripts. We performed a genome-wide comparative
 31 analysis based on the reference genome of JD17 with three published
 32 soybeans (WM82, ZH13 and W05) , which identified five large inversions and
 33 two large translocations specific to JD17, 20,984 - 46,912 PAVs spanning
 34 13.1 - 46.9 Mb in size, and 5 - 53 large PAV clusters larger than 500kb.
 35 1,695,741 - 3,664,629 SNPs and 446,689 - 800,489 Indels were identified and
 36 annotated between JD17 and them. Symbiotic nitrogen fixation (SNF) genes
 37 were identified and the effects from these variants were further evaluated. It
 38 was found that the coding sequences of 9 nitrogen fixation-related genes
 39 were greatly affected. The high-quality genome assembly of JD17 can serve
 40 as a valuable reference for soybean functional genomics research.

41

Introduction

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 (Schmutz et al. 2010a) 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 (Zhou et al. 2015). Pan-genome study on wild and cultivated soybeans has uncovered numerous genetic differences among soybean strains (Liu et al. 2020), suggesting that a single reference genome is inadequate for representing the genetic richness of soybean lines.

The latest progress in sequencing technologies has greatly advanced the ability to construct high-quality genome assemblies with chromosome-level continuity with dramatically reduced cost (Risse et al. 2015; Deschamps et al. 2018). The application of Iso-Seq protocol has enhanced genome annotation (Jiao et al. 2017; Li et al. 2017; Jiang et al. 2017; Magrini et al. 2018).

Benefiting from improved technologies, more and more reference genomes of various soybean materials have been sequenced and reported, including the Chinese cultivar ZH13 (Shen et al. 2019), three reference genomes (*Glycine max* WM82v4, *Glycine max* Lee and *Glycine soja* PI 483463) (Valliyodan et al. 2019), the wild strain W05 genome (Xie et al. 2019), the pan-genome constructed from 26 different soybean species, and the recently published Korean Hwangkeum genome and eight soybean genomes (Chu et al. 2021).

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 (YanlixWilliams)] (paternal parent) (Qin et al. 2014),

and is famous for its lodging resistance, high yield, and strong adaptability (Zhao et al. 2013; Zhao et al. 2015). The goal of this project was to construct a high-quality genome assembly for JD17, and apply the whole genome sequence to identify genetic basis of important traits at the molecular level. In 2019, nodulation and symbiotic nitrogen fixation in legume were summarized and aggregated (Roy et al. 2020), but we still have many unknowns about the reasons for phenotypic differences in rhizobia number, morphology, and nitrogen fixation capacity in different strains of soybean.

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 mapping technologies to construct a high-quality soybean reference genome. After being annotated more completely, the JD17 pseudomolecules were used to identify structural differences in a comparative analysis with three published soybean reference genomes. We identified JD17-specific PAVs and a large number of SNPs and Indels, and also resolved the influence of these variants on the coding structure of nitrogen fixation-related genes.

Materials and methods

Plant and Sample preparation

Soybean seeds of *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 (Kereszt et al. 2007) 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 mins and watered with sterile Fahraeus solution (FÅHRAEUS 1957)

containing 2 mM KNO₃. Seeds were grown in a growth chamber (light) at 28°C and 8 h (dark) at 23°C with 60% humidity for 16 h.

RNA preparation and Sequencing

Underground tissues of inoculation and uninoculation from the 9 timepoints (1, 4, 6, 8, 10, 15, 20, 25 and 30 day post inoculation, dpi) were collected and used for RNA extraction respectively. After that, nine RNA samples of inoculation and uninoculation were mixed equally as a sample, respectively. 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-dpi root tissue of plants was used for SMRT whole-genome sequencing. Underground tissues were collected for genomic DNA preparation with modified CTAB method (Bergman and Quesneville 2007). Using 97 ug DNA, PacBio sequencing libraries were produced following manufacturers protocols as described for the Greater than 30kb-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 S1). At the same time, we also used a part of DNA samples for resequencing on the Illumina HiSeq 2500 platform for 150 bp paired end reads, with a sequencing depth of approximately 45X, for a total of

44.5Gb. 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 (Lieberman-Aiden et al. 2009). 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 150X, for a total of 152.0 Gb.

***De novo* genome assembly of JD17**

To perform *de novo* assembly of the JD17, we combined three different assemblers, including CANU (Koren et al. 2017) (v1.4), FALCON (<https://github.com/PacificBiosciences/FALCON-integrate>) and HGAP4 (SMRT Link v 5.0.1.9585) (Table S2). 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 (Figure S1). 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 S2. 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. Subsequently, the draft assembly was polished twice using Quiver (SMRT Link v 5.0.1.9585) and finally corrected using Illumina short reads with Pilon (Figure S2). Then non-soybean contigs were filtered by aligning them to NT library using blastn.

Continuation and connection of contigs

Due to the complementarity among three different assembly results from CANU (Koren et al. 2017), FALCON and HGAP4, we optimized our assembly results using the GPM (Zhang et al. 2016) pipeline to extend and connect contigs for better contiguity. Firstly, GPM loaded WM82 as a reference genome, and CANU assembly as the back-bone contigs. 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 (Camacho et al. 2009a), and got the potential overlapping relationships between CANU and FALCON contigs. Then CANU contigs were extended or connected by FALCON contigs as needed to produce the JD17 v0.2 assembly. Thirdly, we repeated the above steps with HGAP4 assembly and the JD17 v0.3 was generated. Fourthly, we loaded contigs from an assembly performed with the longest 70 Gb PacBio reads (extracted from the total 120 Gb sorted raw reads) using CANU and repeated the above editing steps to reach the JD17 v0.4. Finally, the contigs of chloroplast and mitochondrial sequences were identified and removed from JD17 v0.4 (Table S3). At editing steps, the JD17 v0.4 was re-polished using Quiver over twice iterations and corrected using Illumina short reads with Pilon (Walker et al. 2014). The finalized JD17 genome assembly (named as Glycine_max_JD17v1.0) was 995.0 Mb in size, with a contig N50 of 18.0 Mb (Table 1). The assembly processing details are 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 on kmer individuals is 0.029, and the corrected estimation 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 (Zhang et al. 2019) 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 (Durand et al. 2016). The genome assembly was finalized after this correction step (Table S3).

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 (Li 2014), 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 (McKenna et al. 2010; DePristo et al. 2011) 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%. 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 (Bergman and Quesneville 2007). MITEs (miniature inverted transposable elements) were collected by MITE-Hunter (Han and Wessler 2010) with all

default parameters. In order to get as much reliable LTR (long terminal repeat) retrotransposons information as possible, we used the LTR_retriever (Ou and Jiang 2018) analysis process, which integrates the output of LTR_FINDER (Xu and Wang 2007) and LTRharvest tools in GenomeTools (Gremme et al. 2013). 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 (Nawrocki and Eddy 2013) and comparing the similarity of secondary structure between the JD17 genome sequence and Rfam (Nawrocki et al. 2014) (v12.0) database.

Annotation of protein-coding genes

We performed gene calling analysis with Exonerate (Slater and Birney 2005), Trinity (Grabherr et al. 2011) and PASA (Haas et al. 2003), by using multi-sourced EST and protein sequences as evidences (including nonredundant soybean EST / Iso-seq ([SRX7016448](#)) (Chu et al. 2021) / protein sequences from NCBI, assembled JD17 RNA-Seq from mixed samples from underground samples, WM82 transcripts and proteins, *Lotus japonicus* and *Medicago truncatula* protein sequences), AUGUSTUS (Hoff and Stanke 2013) and GENEMARK (Bruna et al. 2020) as *ab initio* gene predictors, and a customized repeat library for RepeatMasker (Bergman and Quesneville 2007; Saha et al. 2008).

The RNA-Seq data was *de novo* assembled using Trinity 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. The WM82 protein sequence will be used for GENEMARK model training and prediction. Protein sequences of *Arabidopsis thaliana*, *Medicago sativa*, *Lotus corniculatus* and three different strains of soybean (WM82, ZH13, W05) were used in the Exonerate de-prediction protein-coding gene models. In order to obtain more accurate and complete annotation results, EVM was called to integrate the gene model prediction results from AUGUSTUS, GENEMARK, Exonerate and PASA. After all, the PASA software is used to update these annotation results.

Gene functions were inferred according to the best match of the alignments to the National Center for Biotechnology Information (NCBI) Swiss-Prot(Boeckmann et al. 2003) protein databases using BLASTP (ncbi blast v2.6.0+)(Altschul et al. 1997; Camacho et al. 2009b) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database(Kanehisa et al. 2012) with an E-value threshold of 1E-5. Gene Ontology (GO)(Ashburner et al. 2000) IDs for each gene were obtained from Blast2GO(Conesa and Gotz 2008).

Genome-wide rearrangement and SV detection

To identify large-scale synteny among the four soybean lines, we created a genome-wide alignment using the Mauve aligner with the progressiveMauve algorithm (Darling et al. 2010) with default parameters: default seed weights, determination of LCBs (minimum weights = default), and full alignment with iterative refinement (Chakraborty et al. 2021).

The PAV sequences, PAV clusters and PAV genes between JD17 and the three genomes (WM82, ZH13, and W05) were identified using the sliding window method described in the 2018 publication by Sun et al (Sun et al. 2018), with the same slide window size, alignment software and parameters as theirs.

We aligned WM82, ZH13 and W05 to JD17 using MUMmer 4.0 (NUCmer -maxmatch) (Marcais et al. 2018). MUMmer alignments were processed using

SyRI (Synteny and Rearrangement Identifier) (Goel et al. 2019) (<https://github.com/schneebergerlab/syri> commit 3f16e01), which identifies syntenic regions, chromosome rearrangement events such as inverson and translocation, and also identified SNPs, Indel, Copygains and Copylosses between chromosomes. The annotations of JD17 were indexed using the SnpEff tool for construction, and then SNPs and Indels identified genome-wide were annotated to finally find genes subject to large genetic variation.

Identification and structural variation analysis of SNF genes

Identification of SNF gene sequences based on the literature published by Roy et al. in 2020 (Roy et al. 2020). These sequences were first aligned to the protein sequences of JD17, WM82, ZH13 and W05 by blastp tool and filtered by coverage and identity greater than 40% to obtain possible nitrogen fixation related genes. These filtered genes were then subjected to clustering analysis by the OrthoFinder tool (Emms and Kelly 2019). The SNF genes in the four soybeans were finally identified by determining the gene cluster where the published SNF genes were located. To obtain the expansion and contraction of SNF genes, these single-copy SNF genes were further used to construct an evolutionary tree by RAxML tool (Stamatakis 2014) and finally combined with café (De Bie et al. 2006) for the analysis of expansion and contraction.

Then the SNPs and Indels loci between JD17 and them were combined to assess the protein encoding affected SNF genes. These genes (and gene sequences from the article) were then aligned using mafft(Nakamura et al. 2018) software, constructing an evolutionary tree based on fasttree(Price et al. 2010) tools, and manually checking the evolutionary tree manually to determine that these genes whose structure was affected were homologous (protein-coding genes on the same chromosome; on the same branch of the

evolutionary tree, and then expanded if genes of the species were not present on that branch).

Results and discussion

High quality genome assembly and annotation

Genome assembly using PacBio SMRT data and Hi-C data (detail in methods) resulted in a JD17 genome assembly with 411 contigs anchored to 20 pseudomolecules, with a total size of 965.8Mb and contigs N50 up to 18.0 Mb (Table 1), accounting for 97.1% of the total contigs (Figure 1). The genome accuracy rate was evaluated as 99.999% by using JD17 resequencing data and GATK tools (McKenna et al. 2010; DePristo et al. 2011) (See methods). BUSCO score of the assembly was 93.2% (Simão et al. 2015) (Table S4), indicating that the completeness of JD17 was higher than that of the WM82, ZH13, and W05 genomes (Shen et al. 2018; Xie et al. 2019).

In addition to the nuclear genome, we also constructed the full-length genomes of the soybean chloroplast (152.2 Kb) and the complete genome of *Bradyrhizobium japonicum* USDA 110 (9.1 Mb), which are consistent with previously reported results (Figure S4)(Saski et al. 2005; Kaneko et al. 2002).

We identified 52,840 protein-coding genes and 74,054 full-length transcripts in the JD17 genome (Table 2). The average length of mRNA, 5'UTR, CDS and 3'UTR were 4,465 bp, 302 bp, 1,183 bp and 487 bp, respectively. Our gene annotation is relatively in agreement with the other three genomes. After searching with existing databases and conserved structural domains for functional annotation, a total of 72,529 (97.94%) transcripts have known domains or functions, which suggested that high-confidence annotation of JD17 genome was performed.

Transposable elements

We identified 580.2Mb (58.30% of the JD17 genome) repeat elements in JD17 genome (Figure 2), which contain 1,138,787 intact transposable elements (TEs), including 734,061 class I RNA retrotransposons and 293,279 class II DNA transposons (Table S5).

Among the genomic sequences of JD17, the highest proportion of repetitive sequences is the long terminal repeat (LTR) retrotransposon, and a total of 631,056 LTR elements were identified, totaling 397.58 Mb, which is about 39.95% of the whole genome. Further by assessing the distribution and insertion time of LTRs, we found that the LTR enrichment was all in close proximity to the centromeric region, fewer in the gene enrichment region, and that the amplification of LTR retrotransposons in soybean occurred mainly within the past 1.5 million years and were all relatively young (Figure S5). When comparing the activity of LTR elements (Figure S6), Copia elements appear to be increasingly active and persistent over the past 3 million years. Similarly, Gypsy elements were relatively active during this period. In fact, between 1.4 and 3 million years, their activity levels do not differ much, but since 1.4 million years, the Gypsy element has become less active compared to the Copia element, while the unknown element has also become active over the last 3 million years, but at a much lower level compared to both the Gypsy and Copia elements. Based on the timing of the recent occurrence of WGD in soybean (~13 Mya)(Schmutz et al. 2010b), it appears that these LTR insertions and expansions are different from gene duplications, but occur mainly after WGD.

To compare the differences in the composition and distribution of repetitive sequences between them, we annotated the repetitive sequences of other three genomes using the same approach. The results indicate that the composition of their TEs is essentially the same, except that WM82 has slightly less LTR content. (Figure 2, Table S5 and Figure S7-9). Meanwhile,

we found some blank regions in the genomes of W82 and W05 in the distribution of TEs, both of which are N sequences in the genome sequence. To determine whether these blank regions are the centromeric regions, we identified the centromeric regions of the JD17 genome by using the soybean-specific centromeric satellite repeats sequences (CentGm-1 and CentGm-2)(Gill et al. 2009). The evidence suggests that the location of JD17 centromeric regions is consistent with that of the blank regions in WM82 and W05 (Figure 2, Figure S10). In general, the assemblies of JD17 and ZH13 are more complete in the centromeric regions, and the TEs of W05 has slightly richer components than the other strains.

Genomic rearrangements

To explore how many and how large scale discordant regions of genetic mapping exist between the JD17 reference genome and the three published genomes, we performed a genome-wide comparative analysis. When the pseudochromosomes of JD17 were aligned to the other pseudochromosomes of WM82, ZH13 and W05, a total of 3,727 syntenic blocks were identified (Figure S11), of which 1,368 were common syntenic blocks among the four strains (Table S6). Approximately 88.51% of the JD17 genome sequence shares the syntenic blocks with 88.12% of WM82, 86.10% of ZH13, and 81.88% of W05. We found a total of 251 syntenic blocks that exist between Wm82, ZH13 and W05, but not in JD17, with a size between 3.87-5.27Mb. Similarly, there are 113 such syntenic blocks between JD17, WM82 and ZH13, but not in W05, and the size is between 2.10-3.33Mb. The size of the specific sequence in each of the four strains (that is, the sequence that does not match any other strains) is between 82.58-119.37Mb, of which the wild-type W05 has the most specific sequences and the W82 has the least. This may be due to the loss of sequences in cultivars during the long domestication process of artificial selection.

The chromosomal differences between JD17 and the other three strains were further observed by comparative genome analysis. We observe that compared to WM82 and ZH13, JD17 has slightly more Copygains, 4.0Mb and 3.8Mb respectively, but W05 has more Copygains compared to JD17, with a combined total of about 7.1Mb (Table S7). We also found 56 inversions and 1,051 translocations events for the rearrangement events that occurred between JD17 and WM82 (Table S7). Similarly, 86 and 1,273 inversions and translocations occurred between JD17 and ZH13, and 93 and 2,996 inversions and translocations occurred between JD17 and W05, respectively. We observed five large inversions (>1Mb) on Chr04, Chr05, Chr06, Chr07, Chr19 and two large translocation events (>200kb) on Chr02 and Chr07 (Figure S12-S14, Table S8), respectively, that were present in JD17 compared to all other three species. Examination of the breakpoint loci of these variants by comparing the sequenced subreads to the genome showed that the assembly of JD17 was correct (Figure S15-S21). Although the breakpoint at JD17 (Chr06:32,732,757) is an N sequence (i.e. gaps) and the subreads at Chr07:27,723,877 are poorly supported, the corresponding breakpoint positions in the WM82 genome are both gaps. This suggests that these structural differences such as inversions and translocations are more likely to be true genetic variation in JD17 relative to the other three strains.

Identified PAVs by comparison with WM82, ZH13 and W05

Comparison of the genomes of JD17 and WM82 revealed 20,984 JD17-specific fragments (total length: 13.13 Mb) and 22,635 WM82-specific genomic fragments (total length: 30.81 Mb) (Table S9). Similarly, comparing the genomes of JD17 and ZH13 revealed 22,818 JD17-specific genomic fragments (13.68 Mb) and 23,456 ZH13-specific genomic fragments (33.27 Mb). Comparison of the JD17 and W05 genomes also revealed 36,658 JD17-specific genomic fragments (21.86 Mb) and 37,443 W05-specific genomic

fragments (46.92 Mb). We further merged PAV sequences within 100 kb from the physical coordinates to identify PAV clusters (Table S10). The majority of these PAV sequences (99.8%) were shorter than 5 kb. We identified 23 PAV sequences longer than 5 kb in JD17 relative to Wm82, respectively. These PAV sequences were unevenly distributed in the genome (Figure 1), with some located in clusters. The largest of these PAV cluster fragments is a 1.6-Mb JD17-specific fragment containing 5 predicted genes located between 14.8 Mb and 16.4 Mb on chromosome 1. However, there is a PAV cluster of less than 0.7 Mb that is very rich in genes, up to 73 coding genes, located between 45.3 Mb and 46.0 Mb on chromosome 8. After GO enrichment analysis, they were found to be associated with NADH dehydrogenase, aspartate-type endopeptidase activity, histidine, protein and sugar de metabolism. (Figure S22)

Based on the criterion that a gene can be designated as a PAV gene if its coding sequence is covered by $\geq 75\%$ of the PAV sequence (Sun et al. 2018), we compared JD17 with WM82, respectively, and we identified 100 JD17-specific and 75 WM82-specific PAV genes. Similarly, by comparing JD17 with ZH13, we identified a total of 80 JD17-specific and 156 ZH13-specific PAV genes, and between JD17 and W05, 101 JD17-specific and 67 W05-specific PAV genes were identified. It can be seen that there are large differences in genomic sequences between JD17 and WM82, ZH13 and W05.

By comparing these JD17-specific genes, it was found that only 20 JD17-specific genes were identified simultaneously (Figure S23), and most of the JD17-specific genes could still be found in different strains. The functional enrichment analysis of these 20 genes showed that these genes were related to signal transduction, response to stimulus, protein synthesis, and nitrogen metabolism. The results of functional annotation also showed that these

genes were associated with powdery mildew and TMV resistance protein synthesis.(Figure S24)

Identified SNPs and Indels by comparison with WM82, ZH13 and W05

To find SNPs and Indel between JD17 and the other three strains, we combined MUMmer and SyRI tools. The results showed that we identified 1,695,741 (2,675,463 in ZH13 and 3,664,629 in W05) SNP, 213,509 (324,109 in ZH13 and 391,977 in W05) insertions and 233,180 (391,977 in ZH13 and 408,512 in W05) deletions variations In JD17 genome by comparative genome-wide analysis. Combined with the results of structural variation and PAVs, both showed the smallest difference between JD17 and WM82 and the largest difference with wild-type soybean W05. This may be due to the fact that JD17 has the lineage of WM82, but it also supported that wild soybean can provide richer genetic diversity.

Annotation of these variants showed that they mainly affect intergenic regions, but there are still 4,947 variants between JD17 and WM82 that have a very large impact on 1,785 protein-coding genes. Enrichment analysis of these affected genes revealed that the primary function of these genes is associated with sulfate transmembrane transport, regulation of organ growth, regulation of developmental growth, carbohydrate binding. (Figure S25)

Similarly, a large effect of 7,363 variants on 2,567 protein-coding genes was observed between JD17 and ZH13, and a large effect of 6,674 variants on 3,611 protein-coding genes was observed between JD17 and W05.

Identification of SNF genes and differences among JD17, WM82, ZH13 and W05

To identify genes associated with SNF in the four genomes, we blast identified genes from published articles to these genomes. We identified 331, 360 and 285 and 346 genes potentially related to nitrogen fixation in JD17, WM82, ZH13 and W05 genomes, respectively (Table S11). Of these, 88.8%

of the SNF genes were identified in all four soybeans, and in JD17, 96% of the genes were identified, while fewer genes were identified in ZH13 (Figure S26). Subsequently, by comparative analysis of SNF genes from these four species, there are two expansion genes (*LjCLE-RS2* and *MtNAC969*), two contraction genes (*LjCLC1* and *MtCP6*), and five lost genes (*GmENOD93*, *GmRj2_GmRFG1*, *LjENOD40-1/ENOD40-2*, *LjHIP*, *MtCAS31*) (see Table 12 for details) in JD17 relative to their common ancestor (Table S12). By examining the structural variation of 331 genes potentially associated with nitrogen fixation in JD17, we finally found 18 genes with structural variation between WM82, ZH13 or W05 and affecting protein coding (Table S13). Further by constructing an evolutionary tree to manually confirm the effect of these genes undergoing structural variation on them, we finally observed that nine of them were not in the same branch of the evolutionary tree as the published sequence of WM82. For example, the *LjIGN1* genes (Glyma.10G291900 and Glyma.20G241200), which encodes the synthetic Ankyrin-Repeat Membrane Protein, which is important for symbiosis and nodulation (Kumagai et al. 2007). By evolutionary analysis of their homologous genes, we found that Glyma.10G291900 and Glyma.20G241200 are located in two branches, and the branch where Glyma.10G291900 is located has corresponding genes in all four species with essentially the same protein code (Figure S27). However, the branch of Glyma.20G241200 showed differences. By multiple sequence comparison, JD17020G0232500 in fact encodes 192 amino acids that are very different from several other proteins (Figure S28), and this genetic difference most likely affects symbiosis and nitrogen fixation and still need further studies to be verified.

Conclusions

Here, we constructed a high-quality de novo assembly of the soybean cultivar Jidou 17 (JD17) with high continuity, completeness, and accuracy,

with contigs N50 is close to 18 Mb. We predicted 52,840 gene models and reconstructed 74,054 high quality full-length transcripts. By performing a genome-wide comparative analysis of the reference genome of JD17 with three published soybean (WM82, ZH13 and W05), we identified five large inversions and two large translocations, and 20 PAV genes, which are specific to JD17. SNPs and Indels between JD17 and the three soybeans were also identified and annotated separately. In total, 1,785 genes in JD17 were found to have altered protein coding due to these structural variants, which may affect their function. Further identification of SNF genes and assessment of the impact by these variants revealed that the coding sequences of 9 SNF genes were significantly affected. We hope that our dataset will provide a valuable resource for comparative genomics and functional genomics of soybean.

Data availability

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.

Author information

Affiliations

National Key Laboratory of Crop Genetic Improvement, Center of Integrative Biology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China

Xinxin Yi, Jing Liu, Shengcai Chen, Hao Wu, Min Liu, Qing Xu, Lingshan Lei,
Bao Zhang, Wei Fan, Jianwei Zhang and Nansheng Chen

**State Key Laboratory of Crop Stress Adaptation and Improvement,
Henan University, Kaifeng 475004, Henan, China**

Jing Liu, Shengcai Chen, Bao Zhang, Wei Fan and Xuelu Wang
**Arizona Genomics Institute and BIO5 Institute, School of Plant Sciences,
University of Arizona, Tucson, AZ 85721, USA**

Seunghee Lee, Dave Kudrna, Rod Wing and Jianwei Zhang
**Institute of Food and Oil Crops, Hebei Academy of Agricultural and
Forestry Sciences, Shijiazhuang 050031, Hebei, China**

Chunyan Yang, Mengchen Zhang
**Department of Molecular Biology and Biochemistry, Simon Fraser
University, Burnaby, BC V5A 1S6, Canada**

Nansheng Chen

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.

Corresponding author

Correspondence to Jianwei Zhang, jzhang@mail.hzau.edu.cn; Xuelu
Wang, xueluw@henu.edu.cn; Nansheng Chen, chenn@qdio.ac.cn.

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.).

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17):3389-3402.
- Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25 (1):25-29.
- Bergman, C.M., and H. Quesneville, 2007 Discovering and detecting transposable elements in genome sequences. *Briefings in bioinformatics* 8 (6):382-392.
- Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher *et al.*, 2003 The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31 (1):365-370.
- Bruna, T., A. Lomsadze, and M. Borodovsky, 2020 GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* 2 (2):lqaa026.
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009a BLAST+: architecture and applications. *BMC bioinformatics* 10 (1):421.
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009b BLAST+: architecture and applications. *BMC bioinformatics* 10:421.
- Chakraborty, M., C.H. Chang, D.E. Khost, J. Vedanayagam, J.R. Adrion *et al.*, 2021 Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Res* 31 (3):380-396.
- Chu, J.S., B. Peng, K. Tang, X. Yi, H. Zhou *et al.*, 2021 Eight soybean reference genome resources from varying latitudes and agronomic traits. *Sci Data* 8 (1):164.
- Conesa, A., and S. Gotz, 2008 Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* 2008:619832.
- Darling, A.E., B. Mau, and N.T. Perna, 2010 progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS one* 5 (6):e11147.
- De Bie, T., N. Cristianini, J.P. Demuth, and M.W. Hahn, 2006 CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22 (10):1269-1271.
- DePristo, M.A., E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *43* (5):491.
- Deschamps, S., Y. Zhang, V. Llaca, L. Ye, A. Sanyal *et al.*, 2018 A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *9* (1):4844.
- Durand, N.C., J.T. Robinson, M.S. Shamim, I. Machol, J.P. Mesirov *et al.*, 2016 Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* 3 (1):99-101.
- Emms, D.M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20 (1):238.
- FÄHRÆUS, G.J.M., 1957 The infection of clover root hairs by nodule bacteria studied by a simple glass slide technique. *16* (2):374-381.
- Gill, N., S. Findley, J.G. Walling, C. Hans, J. Ma *et al.*, 2009 Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol* 151 (3):1167-1174.
- Goel, M., H. Sun, W.B. Jiao, and K. Schneeberger, 2019 SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* 20 (1):277.
- Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29 (7):644-652.

- Gremme, G., S. Steinbiss, and S. Kurtz, 2013 GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 10 (3):645-656.
- Haas, B.J., A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr *et al.*, 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. 31 (19):5654-5666.
- Han, Y., and S.R.J.N.a.r. Wessler, 2010 MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. 38 (22):e199-e199.
- Hoff, K.J., and M. Stanke, 2013 WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res* 41 (Web Server issue):W123-W128.
- Jiang, X., A.B. Hall, J.K. Biedler, and Z. Tu, 2017 Single molecule RNA sequencing uncovers trans-splicing and improves annotations in *Anopheles stephensi*. *Insect Mol Biol* 26 (3):298-307.
- Jiao, Y., P. Peluso, J. Shi, T. Liang, M.C. Stitzer *et al.*, 2017 Improved maize reference genome with single-molecule technologies. *Nature* 546 (7659):524-527.
- Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, 2012 KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40 (Database issue):D109-114.
- Kaneko, T., Y. Nakamura, S. Sato, K. Minamisawa, T. Uchiumi *et al.*, 2002 Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. 9 (6):189-197.
- Kereszt, A., D. Li, A. Indrasumunar, C.D. Nguyen, S. Nontachaiyapoom *et al.*, 2007 Agrobacterium rhizogenes-mediated transformation of soybean to study root biology. 2 (4):948.
- Koren, S., B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research:gr*. 215087.215116.
- Kumagai, H., T. Hakoyama, Y. Umehara, S. Sato, T. Kaneko *et al.*, 2007 A novel ankyrin-repeat membrane protein, IGN1, is required for persistence of nitrogen-fixing symbiosis in root nodules of *Lotus japonicus*. *Plant Physiol* 143 (3):1293-1305.
- Li, H., 2014 Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30 (20):2843-2851.
- Li, Y., W. Wei, J. Feng, H. Luo, M. Pi *et al.*, 2017 Genome re-annotation of the wild strawberry *Fragaria vesca* using extensive Illumina- and SMRT-based RNA-seq datasets. *DNA Res*.
- Lieberman-Aiden, E., N.L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy *et al.*, 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (5950):289-293.
- Liu, Y., H. Du, P. Li, Y. Shen, H. Peng *et al.*, 2020 Pan-genome of wild and cultivated soybeans. *Cell*.
- Magrini, V., X. Gao, B.A. Rosa, S. McGrath, X. Zhang *et al.*, 2018 Improving eukaryotic genome annotation using single molecule mRNA sequencing. *BMC Genomics* 19 (1):172.
- Marcais, G., A.L. Delcher, A.M. Phillippy, R. Coston, S.L. Salzberg *et al.*, 2018 MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* 14 (1):e1005944.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. 20 (9):1297-1303.

634 Nakamura, T., K.D. Yamada, K. Tomii, and K. Katoh, 2018 Parallelization of MAFFT for large-scale
635 multiple sequence alignments. *Bioinformatics* 34 (14):2490-2492.

636 Nawrocki, E.P., S.W. Burge, A. Bateman, J. Daub, R.Y. Eberhardt *et al.*, 2014 Rfam 12.0: updates to the
637 RNA families database. *Nucleic acids research* 43 (D1):D130-D137.

638 Nawrocki, E.P., and S.R. Eddy, 2013 Infernal 1.1: 100-fold faster RNA homology searches.
639 *Bioinformatics* 29 (22):2933-2935.

640 Ou, S., and N.J.P.p. Jiang, 2018 LTR_retriever: A highly accurate and sensitive program for
641 identification of long terminal repeat retrotransposons. 176 (2):1410-1422.

642 Price, M.N., P.S. Dehal, and A.P. Arkin, 2010 FastTree 2--approximately maximum-likelihood trees for
643 large alignments. *PloS one* 5 (3):e9490.

644 Qin, J., F. Wang, F. Gu, J. Wang, Q. Chen *et al.*, 2014 A genetic composition analysis of soybean sibling
645 varieties Jidou17 and Ji nf58. *Australian Journal of Crop Science* 8 (5):8.

646 Risse, J., M. Thomson, S. Patrick, G. Blakely, G. Koutsovoulos *et al.*, 2015 A single chromosome
647 assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing
648 data. 4 (1):60.

649 Roy, S., W. Liu, R.S. Nandety, A. Crook, K.S. Mysore *et al.*, 2020 Celebrating 20 Years of Genetic
650 Discoveries in Legume Nodulation and Symbiotic Nitrogen Fixation. *Plant Cell* 32 (1):15-41.

651 Saha, S., S. Bridges, Z.V. Magbanua, and D.G. Peterson, 2008 Empirical comparison of ab initio repeat
652 finding programs. *Nucleic acids research* 36 (7):2284-2294.

653 Saski, C., S.-B. Lee, H. Daniell, T.C. Wood, J. Tomkins *et al.*, 2005 Complete chloroplast genome
654 sequence of *Glycine max* and comparative analyses with other legume genomes. 59 (2):309-
655 322.

656 Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros *et al.*, 2010a Genome sequence of the
657 palaeopolyploid soybean. *Nature* 463 (7278):178.

658 Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros *et al.*, 2010b Genome sequence of the
659 palaeopolyploid soybean. *Nature* 463 (7278):178-183.

660 Shen, Y., H. Du, Y. Liu, L. Ni, Z. Wang *et al.*, 2019 Update soybean Zhonghuang 13 genome to a golden
661 reference. *Sci China Life Sci* 62 (9):1257-1260.

662 Shen, Y., J. Liu, H. Geng, J. Zhang, Y. Liu *et al.*, 2018 De novo assembly of a Chinese soybean genome.
663 *Sci China Life Sci* 61 (8):871-884.

664 Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, and E.M. Zdobnov, 2015 BUSCO:
665 assessing genome assembly and annotation completeness with single-copy orthologs.
666 *Bioinformatics* 31 (19):3210-3212.

667 Slater, G.S., and E. Birney, 2005 Automated generation of heuristics for biological sequence
668 comparison. *BMC bioinformatics* 6:31.

669 Stamatakis, A., 2014 RAXML version 8: a tool for phylogenetic analysis and post-analysis of large
670 phylogenies. *Bioinformatics* 30 (9):1312-1313.

671 Sun, S., Y. Zhou, J. Chen, J. Shi, H. Zhao *et al.*, 2018 Extensive intraspecific gene order and gene
672 structural variations between Mo17 and other maize genomes. *Nat Genet* 50 (9):1289-1295.

673 Valliyodan, B., S.B. Cannon, P.E. Bayer, S. Shu, A.V. Brown *et al.*, 2019 Construction and comparison of
674 three reference-quality genome assemblies for soybean. *Plant J* 100 (5):1066-1082.

- Walker, B.J., T. Abeel, T. Shea, M. Priest, A. Abouelliel *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* 9 (11):e112963.
- Xie, M., C.Y. Chung, M.W. Li, F.L. Wong, X. Wang *et al.*, 2019 A reference-grade wild soybean genome. *Nat Commun* 10 (1):1216.
- Xu, Z., and H.J.N.a.r. Wang, 2007 LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. 35 (suppl_2):W265-W268.
- Zhang, J., D. Kudrna, T. Mu, W. Li, D. Copetti *et al.*, 2016 Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. *Bioinformatics* 32 (20):3058-3064.
- Zhang, X., S. Zhang, Q. Zhao, R. Ming, and H. Tang, 2019 Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature plants* 5 (8):833-845.
- Zhao, Q., L. Yan, B. Liu, R. Di, X. Shi *et al.*, 2015 Breeding of High-yield Widespread and High-quality Soybean Cultivar Jidou 17. *Soybean Science* 34 (4):000736-000739.
- Zhao, S., X. Zhao, X. Tang, J. Zhang, Y. Xu *et al.*, 2013 High yield characteristics of summer sowing soybean varieties. *Soybean Science* (2013):168-175.
- Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu *et al.*, 2015 Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* 33 (4):408-414.

Figure legend

Figure 1 Overview of the JD17 reference genome

Tracks from outer to inner circles indicate: the chromosome of the genome; the gene density map; the repeat sequence density map; density distribution of SNPs between JD17 and WM82 ; density distribution of InDel between JD17 and WM82. PAV length distribution between JD17 and WM82; GC content of JD17.

Figure 2 Chromosome distribution map and percentage of different types of TEs.

708 Chromosomes were split into 200kb bins without overlap, and the percentage

709 of major types of TE elements (include DNA: DNA transposons, LINEs: long
 710 interspersed nuclear elements, Low complexity, LTR: long terminal repeat
 711 retrotransposons, NON LTR: Non-long terminal repeat retrotransposons,
 712 other: simple repeat, SINE: Short interspersed nuclear elements) in each
 713 bin was counted.

714

715 Tables

716 **Table 1 Assembly statistics of Glycine_max_JD17 (JD17), Glycine_max_v4.0 (WM82),**
 717 **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	9,200	1,528	1,870
Total size of contigs (Mb)	995.0	952.5	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)	18.0(PacBio)	0.4	CANU:2.9 +Bionno:18.0 +Hi-C:22.6	3.3
L50 contig count	21	649	66	58
Anchored contigs				
Number of chromosomes	20	20	20	20
Number of contigs	411	-	-	772
Total size (Mb)	965.8	978.4	1,011.2	1,013.2
Number of gaps	391	7,221	448	750

Table 2 Comparison of genome annotation of JD17, WM82, ZH13 and W05.

	JD17	WM82	ZH13	W05
Number of genes	52,840	52,872	55,573	47,201
Number of transcripts	74,054	86,256	96,496	69,277
Average number of transcripts per gene	1.4	1.6	1.7	1.5
Average length of transcript (bp)	4,465	4,889	5,230	5,198
Average exons number per transcript	5.9	6.5	6.5	6.6
Average length of 5' UTR (bp)	302	294	395	252
Average length of 3' UTR (bp)	487	448	562	336
Number of single exon mrna	10,057	12,065	8,466	7,803

Table S11 SNF genes identified in JD17, WM82, ZH13 and W05

Gene_name	JD17
GmACP	JD009G0233800,JD018G0230900
GmARF8a_GmARF8b	JD002G0227900,JD014G0210800
GmBCH1_GmBCH2,Gs52,LjLNP	JD010G0225300,JD020G0157500,J D016G0002000,JD016G0044000
GmBEHL1	JD001G0176800,JD011G0061400
GmbHLHm1	JD013G0231400
GmENOD93	
GmEXPB2	JD010G0121900
GmFWL1	JD009G0177500
GmGα	JD004G0053100,JD014G0096500,J D017G0224100
GmGβ1_GmGβ2_GmGβ3_GmGβ4	JD004G0011900,JD006G0011400,J D011G0114000,JD012G0040100
GmIFS1_GmIFS2,MtFNSII	JD007G0197300,JD013G0156000,J D013G0156100
GmINS1	JD002G0079900,JD007G0166000
GmmiR160	JD011G0248300
GmmiR167	
GmmiR172c	
GmNMNa	JD001G0238700
GmNNC1	JD011G0140000,JD012G0066800
GmPT5	JD003G0149000,JD019G0152000
GmPT7	JD010G0184700,JD020G0197500
GmRGS1,GmRGS2	JD011G0251200,JD018G0010200
GmRj2_GmRFG1	
GmUPS1	JD001G0056400,JD002G0108900
LjAMSH	JD007G0088600,JD009G0173400
LjAMT1.1	JD010G0132100,JD020G0083000
LjAPN1	JD008G0155500
LjARPC1	JD008G0294800,JD010G0211500,J D018G0093300
LjBRUSH	JD009G0160800,JD016G0186800
LjBZF	JD008G0285900,JD018G0104800
LjCASTOR,MtDMI1(LjPOLLUX)	JD012G0157800,JD016G0006700
LjCCaMK(GOF),MtCCaMK(LOF)	JD008G0212700,JD015G0219800
LjCCD7	JD001G0073900,JD011G0144800

LjCHIT5	JD014G0110600,JD017G0213700
LjCIP73	JD017G0102400
LjCKX3	JD009G0060800,JD015G0163200
LjCLC1	JD001G0178300,JD002G0217400,J D002G0245800,JD011G0060000,J D018G0023900
LjCLE-RS1	JD013G0339200
LjCLE-RS2	JD006G0254900,JD012G0113900
LjCLV2	JD009G0236900,JD018G0227600
LjCRE1(GOF),MtCRE1(LOF)	JD005G0234100,JD008G0046600
LjCYCLOPS(MtIPD3)	JD001G0148600,JD009G0203600
LjEIN2b	JD010G0055700,JD013G0126600
LjENOD40-1_ENOD40-2	
LjEPR3	JD008G0268100
LjERF1	JD020G0197200
LjGLB1	JD011G0117400
LjGLE1	JD011G0249700
LjHAR1	JD012G0036300
LjHIP	
LjIGN1	JD010G0282200,JD020G0232500
LjIPT3	JD010G0025800
LjKLV	JD008G0230100,JD018G0253200
LjLAN	JD004G0113300,JD009G0204700
LjLSK1	JD003G0015500,JD007G0073600
LjMATE1	JD009G0099000,JD015G0266500
LjMPK6	JD002G0127900,JD007G0200900
LjNAP1	JD020G0018600
LjNENA	JD013G0139500,JD001G0222000,J D011G0017200
LjNFR1_MtLYK3	JD002G0257500
LjNFR5	JD011G0060200
LjNFRe	JD020G0053500,JD002G0257400
LjNiCK4	JD006G0215300,JD012G0148400
LjNIN	JD004G0000600
LjNPF8.6	JD011G0202200,JD018G0061900
LjNPL	JD011G0250400,JD018G0010900

LjNRSYM1	JD010G0228100,JD020G0155000
LjnsRING	JD005G0230300,JD007G0140800,JD008G0042800
LjNUP133,LjNUP85	JD002G0291000,JD014G0008600
LjPHYB	JD009G0034300,JD015G0134500
LjPIR	JD003G0235000,JD019G0236700
LjRINRK1	JD006G0172200
LjROP6	JD001G0163100,JD002G0043800,JD004G0021900,JD006G0021100,JD011G0075400,JD011G0103700,JD012G0031400,JD016G0117600
LjSCARN	JD007G0256100,JD017G0011600
LjSEN1	JD005G0119500,JD005G0119600,JD008G0072200,JD008G0072300
LjSIE3	JD001G0095300,JD008G0257600
LjSIP1	JD004G0012900,JD006G0012300
LjSIP2	JD007G0003100,JD008G0209000
LjSUNER1	JD010G0047300,JD013G0118400
LjSUS1_LjSUS3,MtSUCS1	JD009G0071200,JD013G0096400,JD015G0173800,JD017G0044800
LjSYMRK	JD001G0019900,JD009G0193000
LjSYP71	JD003G0165600,JD010G0049900,JD013G0121000,JD019G0167500
LjTCO	JD003G0189900,JD019G0192100
LjTML(MtTML1)	JD016G0056100,JD019G0084900
MsNADH-GOGAT	JD004G0225300,JD006G0119500,JD014G0165400,JD019G0062300
MtARF16a	JD010G0206800
MtARP3	JD005G0170600,JD008G0123100
MtBG2	JD014G0122800,JD017G0197900
MtBHLH1	JD004G0091800
MtBHLH476	JD001G0196700
MtBRI1	JD004G0208200
MtCAS31	
MtCBS1	JD009G0124600

MtCBS1	JD016G0158100
MtCDC16	JD006G0213300,JD012G0144600
MtCDPK1	JD001G0165000,JD011G0073600
MtCHR	JD002G0293000,JD014G0006000
MtCHS	JD019G0097900
MtCNGC15a	JD010G0051900,JD013G0122800
MtCNGC15c	JD012G0165100,JD013G0296900
MtCOPT1	JD007G0133500,JD018G0179300

MtCP6	JD011G0169100
-------	---------------

MtCPK3	JD008G0298500,JD018G0084900
MtCRA2	JD004G0082900,JD006G0083900
MtCRN	JD008G0242500,JD018G0266300
MtCYP15a	JD014G0217700,JD017G0251800
MtDELLA1	JD011G0222500,JD018G0037900
MtDELLA2	JD005G0138700,JD008G0090700

MtDELLA3	JD020G0194300,JD010G0188300,JD004G0141800,JD006G0198400
----------	---

MtDGD1	JD003G0186300,JD019G0188800
MtDME	JD020G0181900
MtDNF2	JD002G0095800
MtEFD	JD007G0024600,JD008G0202900

MtERN1	JD016G0041200,JD019G0106000
--------	-----------------------------

MtERN2	JD002G0067700,JD016G0141900
MtEXO70H4	JD011G0138000,JD012G0068400
MtFER3	JD003G0042700
MtFLOT2,MtFLOT4	JD006G0060500
MtGA2ox10	JD012G0028200,JD011G0100800
MtHMGR1	JD001G0155000,JD011G0084400

MtKNAT3_MtKNAT5_MtKNAT9_MtKNAT10	JD006G0060200,JD014G0112400,JD013G0140900,JD017G0101300,JD009G0091200,JD015G0195400
MtLIN_LjLIN	JD010G0192500,JD012G0164600
MtLOG1	JD003G0167000,JD010G0051800,JD013G0122700,JD019G0169000
MtMMPL	JD002G0027300
MtMOT1.3	JD014G0122600,JD017G0198000,JD017G0198200
MtMTP2	JD011G0129500,JD012G0054000
MtN5	JD018G0179600,JD018G0179800
MtNAC969	JD001G0049700,JD007G0224000
MtNAD1	JD018G0269400
MtNF-YA1	JD002G0184900
MtNIP_MtLATD_MtNPF1.7	JD017G0185900
MtNLP1	JD004G0015900
MtNOOT	JD002G0178700
MtNPD1_MtNPD2_MtNPD3_MtNPD4_MtNPD5	JD015G0009000
MtNSP1	JD007G0038100,JD016G0010700
MtNSP2	JD006G0103600
MtPEN3-like	JD007G0014800,JD008G0187900,JD013G0326500,JD015G0018800
MtPHO2-like	JD013G0218300,JD015G0071300
MtPIN2_MtPIN3_MtPIN4	JD007G0159800,JD009G0113300,JD007G0096700,JD009G0166700,JD013G0084900,JD017G0055800
MtPLT1_MtPLT2_MtPLT3_MtPLT4	JD011G0126500,JD012G0051500,JD009G0233500,JD018G0231200,JD001G0022100,JD009G0190500
MtPNO1-like	JD018G0067600,JD018G0067700
MtRab7a1_MtRab7a2	JD011G0114300,JD012G0040300,JD007G0004200,JD008G0191600

MtRbohA	JD004G0066800,JD006G0067500
MtRDN1	JD002G0266200
MtRPG	JD010G0196300
MtRR5_MtRR9_MtRR11	JD011G0179100
MtRSD	JD007G0129800
MtSER6	JD004G0034100,JD006G0033600
MtSGR	JD001G0212400,JD011G0026100
MtSPK1	JD013G0076500,JD017G0064600
MtSymCRK	JD002G0296400,JD014G0001600
MtSymREM1	JD008G0012400
MtSYP132	JD012G0184500,JD013G0279900
MtSYT1_MtSYT2_MtSYT3	JD010G0206100,JD020G0175200,JD011G0103400,JD012G0031100,JD014G0221200,JD017G0255800
MtTIP1g	JD011G0136300,JD012G0060300
MtTML2	JD005G0075300
MtVAMP721d_MtVAMP721e	JD009G0019300,JD015G0121100,JD007G0217900,JD020G0021500
MtVPE	JD014G0091600,JD017G0227800
MtZIP6	JD007G0218200,JD020G0021900
MtZPT2	JD006G0042200
MtyECS	JD005G0203400,JD008G0013800
PvAGO5_GmAGO5	JD012G0076200
PvANN1	JD013G0178500,JD015G0233800
PvBECLIN1	JD004G0131200,JD011G0178000
PvNF-YC1	JD003G0222000,JD019G0223800
PvPI3K	JD004G0088100,JD006G0089300
PvPRAT3	JD004G0006800
PvRabA2	JD011G0129400,JD012G0053900
PvRACK1	JD005G0236500,JD008G0049000
PvRbohA	JD003G0219200,JD019G0221300
PvSIN1	JD007G0260400,JD017G0007400
PvTOR	JD001G0238300,JD011G0002500

WM82	ZH13
Glyma.09G248500,Glyma.18G244300	SoyZH13_09G225200,SoyZH13_18G216100
Glyma.02G239600,Glyma.14G208500	SoyZH13_02G219500,SoyZH13_14G191600
Glyma.20G162600,Glyma.10G231200,Glyma.16G043300	SoyZH13_10G213300,SoyZH13_20G148900,SoyZH13_16G039600
Glyma.11G064300,Glyma.01G178000	SoyZH13_01G161500,SoyZH13_11G062800
Glyma.13G253500	SoyZH13_13G230800
Glyma.06G247600	SoyZH13_06G228201,SoyZH13_16G087301,SoyZH13_18G165901
Glyma.10G122300	
Glyma.09G187000	
Glyma.04G056600,Glyma.14G098000,Glyma.17G226700	SoyZH13_04G054000,SoyZH13_17G215700
Glyma.06G013000,Glyma.11G118500,Glyma.12G043900,Glyma.04G013100	SoyZH13_04G012200,SoyZH13_06G011800,SoyZH13_11G115300,SoyZH13_12G040700
Glyma.13G173401,Glyma.07G202300,Glyma.13G173500	SoyZH13_07G187900,SoyZH13_13G154803,SoyZH13_13G155300
Glyma.02G086200,Glyma.07G172000	SoyZH13_02G080800,SoyZH13_07G159300
Glyma.11G242700	SoyZH13_11G240400
Glyma.10G030400,Glyma.02G143900	
Glyma.11G195600	SoyZH13_11G166802
Glyma.01G241900	SoyZH13_01G223400
Glyma.11G149825,Glyma.12G073300	SoyZH13_11G141800,SoyZH13_12G067200
Glyma.19G164300,Glyma.03G162800	SoyZH13_03G145800,SoyZH13_19G148900
Glyma.20G204000,Glyma.10G186500	SoyZH13_10G171400,SoyZH13_20G189200
Glyma.11G245800,Glyma.18G011400	SoyZH13_18G010202
Glyma.16G212300	SoyZH13_16G188500
Glyma.02G116300,Glyma.01G058500	SoyZH13_01G055600,SoyZH13_02G109100
Glyma.09G183100,Glyma.07G093100	SoyZH13_09G164800
Glyma.20G082450,Glyma.10G132300	SoyZH13_10G122500,SoyZH13_20G073500
Glyma.08G166200	SoyZH13_08G158800
Glyma.08G311700,Glyma.18G103800,Glyma.10G216300	SoyZH13_10G199400
Glyma.16G218000,Glyma.09G168500	SoyZH13_09G151300,SoyZH13_16G193200
Glyma.08G302500,Glyma.18G117100	SoyZH13_08G285400,SoyZH13_18G105800
Glyma.16G003300,Glyma.12G167300	SoyZH13_12G148700,SoyZH13_16G002700
Glyma.08G227200,Glyma.15G222300	SoyZH13_08G215600,SoyZH13_15G208700
Glyma.11G161947,Glyma.01G073200	SoyZH13_11G146600

Glyma.14G110700,Glyma.17G217000	SoyZH13_14G105300
Glyma.17G106000	SoyZH13_17G102300
Glyma.15G170300,Glyma.09G063900	
Glyma.01G179300,Glyma.02G228400,Glyma.11G062900,Glyma.11G231100,Glyma.14G195400,Glyma.18G026000,Glyma.02G258200,Glyma.14G058300	SoyZH13_01G162900,SoyZH13_02G208900,SoyZH13_02G237400,SoyZH13_11G061600,SoyZH13_11G228800,SoyZH13_14G052900,SoyZH13_14G179200,SoyZH13_18G024500
Glyma.13G368300	SoyZH13_13G339800
Glyma.06G284100	SoyZH13_06G257803
Glyma.18G240800,Glyma.09G251800	SoyZH13_09G228200,SoyZH13_18G212900
Glyma.08G049000,Glyma.05G241600	SoyZH13_05G226900,SoyZH13_08G046600
Glyma.09G213700,Glyma.01G149500	SoyZH13_01G132800,SoyZH13_09G193800
Glyma.10G058300,Glyma.13G145100	SoyZH13_10G055200,SoyZH13_13G125900
Glyma.01G028500	SoyZH13_01G028402
Glyma.08G283300	SoyZH13_08G267100
Glyma.20G203700	SoyZH13_20G188900
Glyma.11G121800	SoyZH13_11G118600
Glyma.11G244100	SoyZH13_11G241800
Glyma.12G040000,Glyma.11G114100	SoyZH13_12G036800,SoyZH13_11G111300
Glyma.05G051600	
Glyma.10G291900,Glyma.20G241200	SoyZH13_10G269500,SoyZH13_20G224800
Glyma.10G025300	SoyZH13_10G024900
Glyma.08G246100,Glyma.18G267000,Glyma.13G056200	SoyZH13_08G232900,SoyZH13_18G237000,SoyZH13_13G043000
Glyma.09G214900,Glyma.04G123300	SoyZH13_04G110300,SoyZH13_09G195100
Glyma.03G016500,Glyma.07G076700	SoyZH13_03G015200,SoyZH13_07G072100
Glyma.15G274600,Glyma.09G102800	
Glyma.02G138800,Glyma.07G206200	SoyZH13_07G191800
Glyma.20G019300	SoyZH13_20G018500
Glyma.13G155800,Glyma.01G224700,Glyma.11G018300	SoyZH13_01G206800,SoyZH13_11G017600
Glyma.02G270800	
Glyma.11G063100	SoyZH13_01G162600,SoyZH13_11G061800
Glyma.20G054500,Glyma.02G270700	SoyZH13_20G049300,SoyZH13_02G248700
Glyma.06G230800,Glyma.12G160500	SoyZH13_06G213900,SoyZH13_12G141600
Glyma.04G000600	SoyZH13_04G001000
Glyma.11G176800,Glyma.18G064900	SoyZH13_11G197700
Glyma.11G244900,Glyma.18G012200,Glyma.18G012267,Glyma.18G012333,Glyma.11G244800	SoyZH13_11G242600,SoyZH13_18G011000,SoyZH13_11G242500

Glyma.10G234100,Glyma.20G160200	SoyZH13_10G216100
Glyma.08G044700,Glyma.05G237500	SoyZH13_05G222900,SoyZH13_08G042900
Glyma.02G305200,Glyma.14G008200	SoyZH13_02G281800,SoyZH13_14G008300
Glyma.09G035500,Glyma.15G140000	SoyZH13_09G033900,SoyZH13_15G135400
Glyma.19G250600,Glyma.03G253000	SoyZH13_03G231700,SoyZH13_19G235102
Glyma.06G184400	SoyZH13_06G175500
Glyma.01G164200,Glyma.04G023300,Glyma.04G023400,Glyma.06G023300,Glyma.11G079100,Glyma.11G107700,Glyma.12G032700,Glyma.16G124400,Glyma.02G046100	SoyZH13_01G147500,SoyZH13_02G044000,SoyZH13_04G022101,SoyZH13_06G021700,SoyZH13_11G076900,SoyZH13_11G104700,SoyZH13_16G110400
Glyma.07G262000,Glyma.17G012100	SoyZH13_07G246100,SoyZH13_17G012000
Glyma.05G121500,Glyma.05G121600,Glyma.08G076300	SoyZH13_05G110900,SoyZH13_05G111000,SoyZH13_08G072502,SoyZH13_08G072503
Glyma.08G273300,Glyma.01G097700	SoyZH13_01G087200,SoyZH13_08G258600
Glyma.04G014100,Glyma.06G013900	SoyZH13_04G013200,SoyZH13_06G012700
Glyma.08G223400,Glyma.07G003200	SoyZH13_07G002900
Glyma.10G048300,Glyma.13G136100	SoyZH13_10G046700,SoyZH13_13G117300
Glyma.09G073600,Glyma.13G114000,Glyma.15G182600,Glyma.17G045800	SoyZH13_09G070100,SoyZH13_13G095400,SoyZH13_15G173700,SoyZH13_17G044200
Glyma.01G020100,Glyma.09G202300	SoyZH13_01G020302,SoyZH13_09G183400
Glyma.03G179500,Glyma.13G138900,Glyma.19G180200,Glyma.10G051400	SoyZH13_10G049500
Glyma.19G203400,Glyma.03G205900	SoyZH13_03G186000,SoyZH13_19G190600
Glyma.19G090600,Glyma.16G057300	SoyZH13_16G053000,SoyZH13_19G079100
Glyma.04G236900,Glyma.14G162300,Glyma.19G065600,Glyma.06G127400	SoyZH13_04G213100,SoyZH13_06G122600,SoyZH13_14G150200,SoyZH13_19G060300
Glyma.10G210600	SoyZH13_10G194000
Glyma.05G172600	SoyZH13_05G161200,SoyZH13_08G124600,SoyZH13_08G124801
Glyma.14G130300,Glyma.17G202900	SoyZH13_17G191800
Glyma.04G098400	SoyZH13_04G091800
Glyma.01G198100	SoyZH13_01G181700
Glyma.04G218300,Glyma.06G147600	SoyZH13_06G141800
Glyma.09G185500	SoyZH13_09G167200
Glyma.09G129700	

Glyma.16G177500	SoyZH13_16G158000
Glyma.12G156300,Glyma.06G228700	SoyZH13_12G138000
Glyma.01G166100,Glyma.11G077300	SoyZH13_01G149700,SoyZH13_11G075000
Glyma.14G005700,Glyma.02G307300	SoyZH13_02G283600,SoyZH13_14G005800
Glyma.19G105100	SoyZH13_19G091600
Glyma.10G053900,Glyma.13G141000	SoyZH13_10G051500,SoyZH13_13G121700
Glyma.13G324600,Glyma.12G175000	SoyZH13_12G156100,SoyZH13_13G297600
Glyma.07G141200,Glyma.18G191300	SoyZH13_07G132101,SoyZH13_18G166200
Glyma.06G278000,Glyma.06G278200,Glyma.06G279100,Glyma.06G279200,Glyma.06G279900,Glyma.06G283000,Glyma.06G283100,Glyma.11G144900,Glyma.12G077202,Glyma.12G123500,Glyma.12G123600,Glyma.12G127300,Glyma.12G208200,Glyma.15G225000,Glyma.U044800,Glyma.U045200,Glyma.12G127200	SoyZH13_06G255100
Glyma.18G096500,Glyma.08G316500	SoyZH13_08G298600,SoyZH13_18G087600
Glyma.04G088700,Glyma.06G090700	SoyZH13_04G083500,SoyZH13_06G085300
Glyma.08G257700,Glyma.18G282100	
Glyma.14G216300,Glyma.17G254900	
Glyma.11G216500,Glyma.18G040000	SoyZH13_11G214900,SoyZH13_18G039400
Glyma.05G140400,Glyma.08G095800	SoyZH13_05G129500,SoyZH13_08G090800
Glyma.20G200500,Glyma.10G190200,Glyma.04G150500,Glyma.06G213100	SoyZH13_10G175100,SoyZH13_04G132300,SoyZH13_06G200000
Glyma.03G202100,Glyma.19G200000	SoyZH13_19G187200
Glyma.20G188300	SoyZH13_20G173800
Glyma.02G102700	
Glyma.07G025800,Glyma.08G216600	SoyZH13_08G206400
Glyma.16G040000,Glyma.19G113100	SoyZH13_19G099200
Glyma.02G072800,Glyma.16G154100	SoyZH13_02G068800,SoyZH13_16G137600
Glyma.11G144600,Glyma.12G075200	SoyZH13_11G139700,SoyZH13_12G069000
Glyma.03G050100	
Glyma.04G064400,Glyma.06G065600	SoyZH13_06G061200
Glyma.12G029800,Glyma.11G104800	SoyZH13_12G028700,SoyZH13_11G101800
Glyma.11G088000,Glyma.01G156700	SoyZH13_01G139600,SoyZH13_11G085800

Glyma.04G064100,Glyma.06G065200,Glyma.14G112400,Glyma.13G157300,Glyma.17G104800,Glyma.09G095700,Glyma.15G202200	SoyZH13_15G189800
Glyma.10G194500	SoyZH13_10G179300
Glyma.10G053800,Glyma.13G140900,Glyma.19G182100,Glyma.03G181300	SoyZH13_03G163500,SoyZH13_10G051400,SoyZH13_13G121600,SoyZH13_19G167100
Glyma.02G028051,Glyma.02G028100,Glyma.02G028700	SoyZH13_02G027200
Glyma.14G130000	SoyZH13_14G115500,SoyZH13_17G191900,SoyZH13_17G192000
Glyma.12G059000,Glyma.11G135500	SoyZH13_11G131300,SoyZH13_12G054300
Glyma.18G191600,Glyma.18G191800	SoyZH13_18G166302,SoyZH13_18G166303
Glyma.07G229100	SoyZH13_07G214400
Glyma.18G285300,Glyma.02G096200	SoyZH13_02G089500
Glyma.02G195000	SoyZH13_02G177601,SoyZH13_12G217905
Glyma.17G189300	SoyZH13_17G179200
Glyma.04G017400	SoyZH13_04G016600
Glyma.02G188700	SoyZH13_02G172900
Glyma.15G008800	SoyZH13_15G008900
Glyma.07G039400,Glyma.16G008200	SoyZH13_07G038100,SoyZH13_16G006900
Glyma.06G110800	SoyZH13_06G105700
Glyma.07G015800,Glyma.08G201300,Glyma.15G019200,Glyma.13G355000	SoyZH13_07G014700,SoyZH13_08G192300,SoyZH13_13G326800,SoyZH13_15G018600
Glyma.13G239100,Glyma.15G074200	SoyZH13_13G217400,SoyZH13_15G071800
Glyma.07G164600,Glyma.09G117900,Glyma.09G176300,Glyma.07G102500,Glyma.13G101900,Glyma.17G057300	SoyZH13_09G158500,SoyZH13_13G083900
Glyma.11G131900,Glyma.12G056300,Glyma.18G244600,Glyma.09G248200,Glyma.01G022500,Glyma.09G199800	SoyZH13_18G216400,SoyZH13_01G022600
Glyma.18G072600,Glyma.18G072700	SoyZH13_18G069700
Glyma.11G118800,Glyma.12G044100,Glyma.07G004500,Glyma.08G205300	SoyZH13_07G004200,SoyZH13_08G196100

Glyma.06G072900,Glyma.04G071600	SoyZH13_04G067700,SoyZH13_06G068400
Glyma.02G279600	SoyZH13_02G257500
Glyma.10G198700	SoyZH13_10G182900
Glyma.11G155100	SoyZH13_11G177500
Glyma.07G135800	SoyZH13_07G127001
Glyma.06G036300,Glyma.04G036200	SoyZH13_04G034200,SoyZH13_06G034600
Glyma.11G027400,Glyma.01G214600	SoyZH13_01G197200,SoyZH13_11G026500
Glyma.17G066800,Glyma.13G093600	SoyZH13_17G064300
Glyma.14G001300,Glyma.02G311200	SoyZH13_02G287100
Glyma.08G012800	SoyZH13_08G012400
Glyma.12G194800,Glyma.13G307600	SoyZH13_12G174900,SoyZH13_13G281000
Glyma.20G180600,Glyma.10G210000,Glyma.11G107300,Glyma.12G032400,Glyma.14G220000,Glyma.17G259000	SoyZH13_10G193400,SoyZH13_20G166700,SoyZH13_11G104400,SoyZH13_12G031200,SoyZH13_14G202100,SoyZH13_17G246300
Glyma.11G143100	SoyZH13_11G138400,SoyZH13_12G060800
Glyma.05G077700	SoyZH13_05G075100
Glyma.15G126200,Glyma.09G019800,Glyma.07G223000,Glyma.20G022200	SoyZH13_09G019400,SoyZH13_15G121400,SoyZH13_07G208100,SoyZH13_20G021300
Glyma.17G230700,Glyma.14G092800	SoyZH13_14G088300,SoyZH13_17G219400
Glyma.07G223200,Glyma.20G022500	SoyZH13_07G208300,SoyZH13_20G021700
Glyma.06G045400	SoyZH13_06G042400
Glyma.08G014300,Glyma.05G207600	SoyZH13_05G193800,SoyZH13_08G013800
Glyma.12G083500	
Glyma.U031126,Glyma.13G199800	SoyZH13_13G178100,SoyZH13_15G222100
Glyma.04G141000,Glyma.11G153900	SoyZH13_04G124100,SoyZH13_11G176600
Glyma.19G236400,Glyma.03G239400	SoyZH13_03G218600,SoyZH13_19G221900
Glyma.04G094500,Glyma.06G096300	SoyZH13_04G088700,SoyZH13_06G091200
Glyma.04G007300	SoyZH13_04G007000
Glyma.12G058900,Glyma.11G135400	SoyZH13_11G131200,SoyZH13_12G054200
Glyma.08G051600,Glyma.05G243800	SoyZH13_05G228900,SoyZH13_08G049000
Glyma.19G233900,Glyma.03G236300	SoyZH13_03G215600,SoyZH13_19G219600
Glyma.17G007600,Glyma.07G266500	SoyZH13_07G250800
Glyma.11G002600,Glyma.01G241300	SoyZH13_11G002400

W05

rna-XM_028359879,rna-XM_028391935
rna-XM_028344772,rna-XM_028359715
rna-XM_028327365,rna-
XM_028364081,rna-XM_028352100
rna-XM_028334137,rna-XM_028386668
rna-XM_028332361.1

rna-XM_028328179.1
rna-XM_028392314.1
rna-XM_028345547,rna-
XM_028355718,rna-XM_028372054
rna-XM_028334610,rna-
XM_028338657,rna-XM_028371519,rna-
XM_028378767
rna-XM_028332021,rna-
XM_028332131,rna-XM_028385310
rna-XM_028344394,rna-XM_028384948
rna-XM_028341291.1

rna-XM_028368516.1
rna-XM_028333947,rna-XM_028338810
rna-XM_028361463,rna-XM_028369781
rna-XM_028326546,rna-XM_028365418
rna-XM_028340227,rna-XM_028358240
rna-XM_028349877,rna-
XM_028349881,rna-XM_028350802
rna-XM_028347473.1
rna-XM_028384056,rna-XM_028391607
rna-XM_028328729,rna-XM_028365968
rna-XM_028391367.1
rna-XM_028327002,rna-
XM_028358545,rna-XM_028390122
rna-XM_028325953,rna-XM_028350730
rna-XM_028356459,rna-XM_028390007
rna-XM_028338234,rna-XM_028351622
rna-XM_028346757,rna-XM_028389387
rna-XM_028333002,rna-XM_028377033

rna-XM_028352742.1
rna-XM_028353796.1
rna-XM_028324035,rna-XM_028347438
rna-XM_028335080,rna-
XM_028341568,rna-XM_028342859,rna-
XM_028344815,rna-XM_028358324,rna-
XM_028358825,rna-XM_028361398,rna-
XM_028386865
rna-XM_028331100.1

rna-XM_028356791,rna-XM_028392500
rna-XM_028378204,rna-XM_028387087
rna-XM_028383666,rna-XM_028392590
rna-XM_028328580,rna-XM_028341341

rna-XM_028391542.1

rna-XM_028333013.1
rna-XM_028341053.1
rna-XM_028338795,rna-XM_028334130
rna-XM_028375869.1
rna-XM_028327090,rna-XM_028364240
rna-XM_028328148.1
rna-XM_028357678,rna-
XM_028389202,rna-XM_028341170
rna-XM_028372859,rna-XM_028391650
rna-XM_028368290,rna-XM_028383850
rna-XM_028346596,rna-XM_028392751
rna-XM_028349943,rna-XM_028385348
rna-XM_028366848.1
rna-XM_028333438,rna-
XM_028331674,rna-XM_028392254
rna-XM_028362808.1
rna-XM_028336270,rna-XM_028386851
rna-XM_028366019,rna-XM_028362796
rna-XM_028339617,rna-XM_028381808
rna-XM_028371380.1
rna-XM_028331335,rna-XM_028358323
rna-XM_028340286,rna-
XM_028357086,rna-XM_028341342,rna-
XM_028357861

rna-XM_028327038,rna-XM_028365181
rna-XM_028378153,rna-XM_028387047
rna-XM_028343854,rna-XM_028366556
rna-XM_028324451,rna-XM_028349322
rna-XM_028363187,rna-XM_028370987
rna-XM_028381337.1
rna-XM_028333985,rna-
XM_028334396,rna-XM_028338498,rna-
XM_028349491,rna-XM_028371654,rna-
XM_028371655,rna-XM_028378881,rna-
XM_028385393
rna-XM_028355120,rna-XM_028386072
rna-XM_028376553,rna-
XM_028376554,rna-XM_028388095,rna-
XM_028391348
rna-XM_028379089,rna-XM_028389618
rna-XM_028371531,rna-XM_028378778
rna-XM_028382932,rna-XM_028388937
rna-XM_028330029,rna-XM_028339734
rna-XM_028324926,rna-
XM_028339921,rna-XM_028347793,rna-
XM_028355901
rna-XM_028326238,rna-XM_028372306
rna-XM_028326345,rna-
XM_028339894,rna-XM_028360186,rna-
XM_028370026
rna-XM_028361525,rna-XM_028370352
rna-XM_028349819,rna-XM_028361145
rna-XM_028344117,rna-
XM_028362432,rna-XM_028374202,rna-
XM_028380288
rna-XM_028329716.1
rna-XM_028377221,rna-
XM_028387902,rna-XM_028387903
rna-XM_028343276,rna-XM_028354420
rna-XM_028372643.1
rna-XM_028389247.1
rna-XM_028373932,rna-XM_028380567
rna-XM_028326094.1
rna-XM_028325647.1

rna-XM_028349645.1
rna-XM_028339299,rna-XM_028381784
rna-XM_028334279,rna-XM_028385608
rna-XM_028345136,rna-XM_028366811
rna-XM_028363044.1
rna-XM_028327510,rna-XM_028340379
rna-XM_028332914,rna-XM_028337844
rna-XM_028357786,rna-XM_028382843

rna-XM_028333642,rna-
XM_028336523,rna-XM_028337377,rna-
XM_028337391,rna-XM_028337400,rna-
XM_028338151,rna-XM_028338227,rna-
XM_028338229,rna-XM_028338315,rna-
XM_028382215,rna-XM_028382218,rna-
XM_028382219

rna-XM_028358295,rna-XM_028390169
rna-XM_028372511,rna-XM_028379792
rna-XM_028356187,rna-XM_028389350
rna-XM_028342841,rna-XM_028355527
rna-XM_028340010,rna-XM_028356957
rna-XM_028376791,rna-XM_028387543
rna-XM_028366287,rna-
XM_028329878,rna-XM_028373097,rna-
XM_028381669
rna-XM_028361231,rna-XM_028370304
rna-XM_028366222.1
rna-XM_028346127.1
rna-XM_028367399,rna-XM_028391275
rna-XM_028351094,rna-
XM_028361984,rna-XM_028377039
rna-XM_028343334,rna-XM_028350105
rna-XM_028336232,rna-XM_028336745
rna-XM_028368661.1
rna-XM_028379410.1
rna-XM_028333624,rna-XM_028336813
rna-XM_028330429,rna-XM_028384490

rna-XM_028372158,rna-
XM_028379406,rna-XM_028344570,rna-
XM_028334123,rna-XM_028354070,rna-
XM_028325063,rna-XM_028348047
rna-XM_028329692.1
rna-XM_028327390,rna-
XM_028340354,rna-XM_028361806,rna-
XM_028370047

rna-XM_028338418.1

rna-XM_028343316,rna-
XM_028353508,rna-XM_028353603
rna-XM_028331972,rna-XM_028339454
rna-XM_028359382.1

rna-XM_028355062.1
rna-XM_028353759.1
rna-XM_028371580.1
rna-XM_028354518.1

rna-XM_028346732.1

rna-XM_028383399,rna-XM_028350282
rna-XM_028380049.1
rna-XM_028383094,rna-
XM_028388693,rna-XM_028335750,rna-
XM_028348589
rna-XM_028334798,rna-XM_028347803
rna-XM_028367411,rna-
XM_028392418,rna-XM_028326025,rna-
XM_028384145,rna-XM_028340536,rna-
XM_028353643
rna-XM_028332770,rna-
XM_028338016,rna-XM_028323965,rna-
XM_028357144,rna-XM_028372615,rna-
XM_028392414
rna-XM_028359630,rna-
XM_028359871,rna-XM_028390427
rna-XM_028334692,rna-
XM_028338739,rna-XM_028382946,rna-
XM_028389487

rna-XM_028372275,rna-XM_028379524
rna-XM_028363775.1
rna-XM_028328282.1
rna-XM_028332401.1
rna-XM_028382790.1
rna-XM_028371809,rna-XM_028379053
rna-XM_028336188,rna-XM_028391106
rna-XM_028342430,rna-XM_028354294
rna-XM_028344847,rna-XM_028367191
rna-XM_028391498.1
rna-XM_028333851,rna-XM_028338421
rna-XM_028327343,rna-
XM_028364265,rna-XM_028333585,rna-
XM_028338835,rna-XM_028344688,rna-
XM_028354753
rna-XM_028334972,rna-XM_028336480
rna-XM_028376138.1

rna-XM_028365771,rna-XM_028385592

rna-XM_028343442,rna-XM_028353823
rna-XM_028366639,rna-XM_028385595
rna-XM_028379161.1
rna-XM_028377732,rna-XM_028386755
rna-XM_028337801.1
rna-XM_028334439,rna-XM_028348757
rna-XM_028334639,rna-XM_028373001
rna-XM_028362832,rna-XM_028370797
rna-XM_028372577,rna-XM_028379851
rna-XM_028371467.1
rna-XM_028334857,rna-XM_028339458
rna-XM_028378240,rna-XM_028387109
rna-XM_028363039,rna-XM_028370759
rna-XM_028355547,rna-XM_028386121
rna-XM_028325378,rna-XM_028332399