scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A telomere-to-telomere genome of wild soybean with resistance to soybean cyst nematode X12

Yun Lian, Yongkang Wu, Chun Li, He Wei, Pei Du, Jinying Li, Chenfang Lei, Haichao Li, Shiwei Wang, Hui Zhang, Jinshe Wang ✉ & Weiguo Lu ✉

The soybean cyst nematode (SCN) is one of the most destructive pests affecting soybean production worldwide. Wild soybean (*Glycine soja*) germplasm offers valuable genetic resources for developing SCN-resistant cultivars. Here, we present a telomere-to-telomere assembly of the wild soybean *Glycine soja* accession YSD56, which is resistant to the highly virulent SCN race X12. The assembled genome has a total length of 1,008.52 Mb, with a contig N50 of 51.97 Mb, and successfully resolves all 20 centromeres and 40 telomeres. The assembly is high quality, with 99.7% completeness based on conserved single-copy orthologs (BUSCO) and a base-level accuracy of QV 52.16. A total of 57,712 protein-coding genes were predicted, 98.79% of which were functionally annotated. This high-quality reference genome provides a valuable resource for investigating SCN resistance and accelerating soybean genetic improvement.

Background & Summary

Soybean is a globally important crop that serves as a major source of plant-based protein and oil. High-quality reference genomes of cultivated soybean (*Glycine max*), including Wm82 and ZH13, have facilitated the identification of genes associated with improvements in yield and quality[1,2]. However, the domestication process has markedly reduced genetic diversity in cultivated soybean (*Glycine max*), especially at stress-resistance loci, leading to the loss of numerous rare alleles[3,4]. In contrast, wild soybean (*Glycine soja*) maintains substantially greater genetic diversity, serving as a valuable reservoir of stress-adaptive alleles. Several genes associated with tolerance to abiotic (e.g., salinity, drought, and heat) and biotic stresses have already been identified in *G. soja*, highlighting its potential for improving stress resistance in cultivated soybean[4,5]. This underutilized germplasm thus offers critical genetic resources for broadening the genetic base of *G. max* and enhancing its resilience to environmental challenges.

The soybean cyst nematode (SCN, *Heterodera glycines*) is one of the most devastating pathogens affecting soybean production worldwide[6,7]. In recent years, the emergence of a highly virulent population, X12, has posed a serious challenge to existing resistance resources[8–10]. Although resistance loci such as *Rhg1* and *Rhg4*, derived from cultivated soybean accessions (e.g., PI 88788, PI 437654, PI 209332, PI 89772, PI 548316, and Peking), have been widely deployed in breeding programs[11–14], their effectiveness against the X12 population remains unverified, raising concerns regarding the durability of current resistance sources.

Here, we report the identification of a wild soybean accession, YSD56, which is strongly and stably resistant to the X12 population. We generated a telomere-to-telomere (T2T) reference genome of YSD56 using an integrated sequencing strategy combining PacBio HiFi, Oxford Nanopore (ONT) ultralong reads, Hi-C, and Illumina short reads. This high-quality T2T reference may facilitate the identification of novel SCN resistance genes and enrich the soybean pangenome, offering a foundation for future molecular breeding and genetic improvement of cultivated soybean (*G. max*).

[1]Institute of Crop Molecular Breeding, Henan Academy of Agricultural Sciences, Zhengzhou, China. ✉e-mail: wjs33314@126.com; 123bean@163.com

Varieties	Average cysts per plant #1	Female Index #2	Resistant level #3
Lee	228 ± 62.6 (181~320)	/	/
Peking	113 ± 95.4 (30~206)	49.4	MS
PI88788	196 ± 97.4 (127~339)	86	S
PI437654	61 ± 6.3 (52~66)	26.8	MR
PI209332	235 ± 189.5 (12~465)	102.9	S
PI89772	175 ± 73.8 (81~257)	76.8	S
PI548316	225 ± 49.2 (156~270)	98.4	S
YSD56	14 ± 3 (10~16)	5.9	R

Table 1. Evaluation YSD56 and some known resistant cultivars with soybean cyst nematode X12. Note: #1: The average number of cysts per plant ± standard deviation and minimum-maximum number of cysts per plant in parenthesis. #2: The female index (FI) is calculated as the mean number of cysts for each line relative to the mean number of cysts per plant of the susceptible control line Lee68, multiplied by 100. Bioassay was replicated three times with five plants per race testing for each line. #3: Resistance level definition: FI ≤ 10 = R (resistant, R); 10 < FI ≤ 30 = MR (moderately resistant, MR); 30 < FI ≤ 60 = MS (moderately sensitive, MS); FI > 60 = S (sensitive, S).

Methods

Materials and methods. *Sample collection and genome sequencing.* We systematically evaluated resistance to the highly virulent SCN population X12 across multiple soybean accessions, including six cultivated varieties (PI 88788, PI 437654, PI 209332, PI 89772, PI 548316, and Peking)[14] and the wild accession YSD56. All experiments were conducted in climate-controlled growth chambers at the Henan Academy of Agricultural Sciences, with the susceptible cultivar Lee used as a control. SCN resistance was assessed using the Female Index (FI), a standardized metric that quantifies the reproductive success of SCN females on test plants relative to the susceptible control[15,16]. Among the tested accessions, only the wild soybean YSD56 (*Glycine soja*) exhibited resistance to SCN race X12, with an FI value of 5.9. In contrast, the remaining accessions were moderately resistant or susceptible, with FI values ranging from 26.8–102.9 (Table 1). The YSD56 accession used in this study was originally collected from Xuchang City, Henan Province, China (34°16′77″ N, 113°47′98″ E) (Fig. 1A).

Fresh young leaves were harvested from two-week-old healthy seedlings, immediately frozen in liquid nitrogen, and stored at −80 °C for subsequent genomic DNA extraction and sequencing. Additionally, root, stem, leaf, and flower tissues from two-week-old seedlings were collected for transcriptome sequencing. High-quality genomic DNA was extracted from leaf tissue using the CTAB method for genome library construction and sequencing at Shanghai Personal Biotechnology Co., Ltd. (Shanghai, China). Short-read sequencing libraries with insert sizes of 300–350 bp were prepared, and subsequently sequenced on the Illumina X Plus platform (Illumina, San Diego, CA, USA). Long-read sequencing was performed on the PacBio Revio platform in circular consensus sequencing (CCS) mode. Ultralong-read libraries were prepared using an Oxford Nanopore SQK-LSK109 kit and sequenced on a PromethION flow cell (Oxford Nanopore Technologies, Oxford, UK). Chromosome conformation capture (Hi-C) libraries were constructed following a standard protocol using the DpnII restriction enzyme[17] and sequenced on an Illumina X plus platform. Total RNA was separately extracted from the root, stem, leaf, and flower tissues using a Qiagen RNeasy Mini Kit (Qiagen, Hilden, Germany), pooled in equal amounts, and used for Iso-Seq transcriptome library construction. The Iso-Seq cDNA library was prepared according to PacBio's standard protocol, with size selection performed using the BluePippin system (Sage Science, MA, USA), and sequenced on the PacBio Sequel II platform.

Karyotypic analysis. Karyotyping analysis was performed as described previously by Du, *et al.*[18]. Briefly, the dividing zones of the roots were collected and treated with 8-hydroxyquinoline in 2 ml microcentrifuge tubes for 3–5 h before being fixed in a mixture of 3:1 ethanol: acetic acid (v/v) for 2–3 d. The samples were then suspended in 45% acetic acid and squashed on a slide. After being frozen at −70 °C overnight, the slides that contained mitotic chromosomes were dehydrated in 100% ethanol, followed by 4′-,6-diamidino-2-phenylindole staining. The chromosomes were then observed and photographed under a Leica DM4 fluorescence microscope (Leica, Wetzlar, Germany) at an excitation wavelength of 450–490 nm. Chromosome classifications were made using the standardized nomenclature[19]. The results revealed that the wild soybean YSD56 has a chromosome number of 2n = 40 and a karyotype formula of 14 M/m + 4sm + 2st/t (Fig. 1B,C).

Genome assembly and telomere identification. We obtained 91.05 Gb (~91 × coverage) of Illumina short reads after quality control was performed using fastp (v0.23.1)[20] with the following parameters: "–length_required = 50 -e 20 -W 5–n_base_limit 5" (Table 2). Genome size, repeat content and heterozygosity of *Glycine. soja* YSD56 were estimated based on 19-mers using Jellyfish (v2.3.0)[21] and GenomeScope2[22]. The results revealed an estimated genome size of 1.14 Gb and a heterozygosity of 0.17% (Fig. 1D). PacBio subreads produced by the Revio platform were processed using CCS (v6.4.0)[23] (https://github.com/PacificBiosciences/ccs) with the following parameters "–min-passes 3–min-rq 0.99–min-length 500", yielding 44.19 Gb (~44 × coverage) of HiFi reads with an N50 of 17.47 kb (Table 2). Oxford Nanopore sequencing produced 240.85 Gb (~240 × coverage) of ultralong reads with an N50 of 44.98 kb (Table 2). For Hi-C sequencing, a total of 23.84 Gb (~123 × coverage) of data were generated and subjected to quality control using fastp (v0.23.1)[20] with the same parameters as above

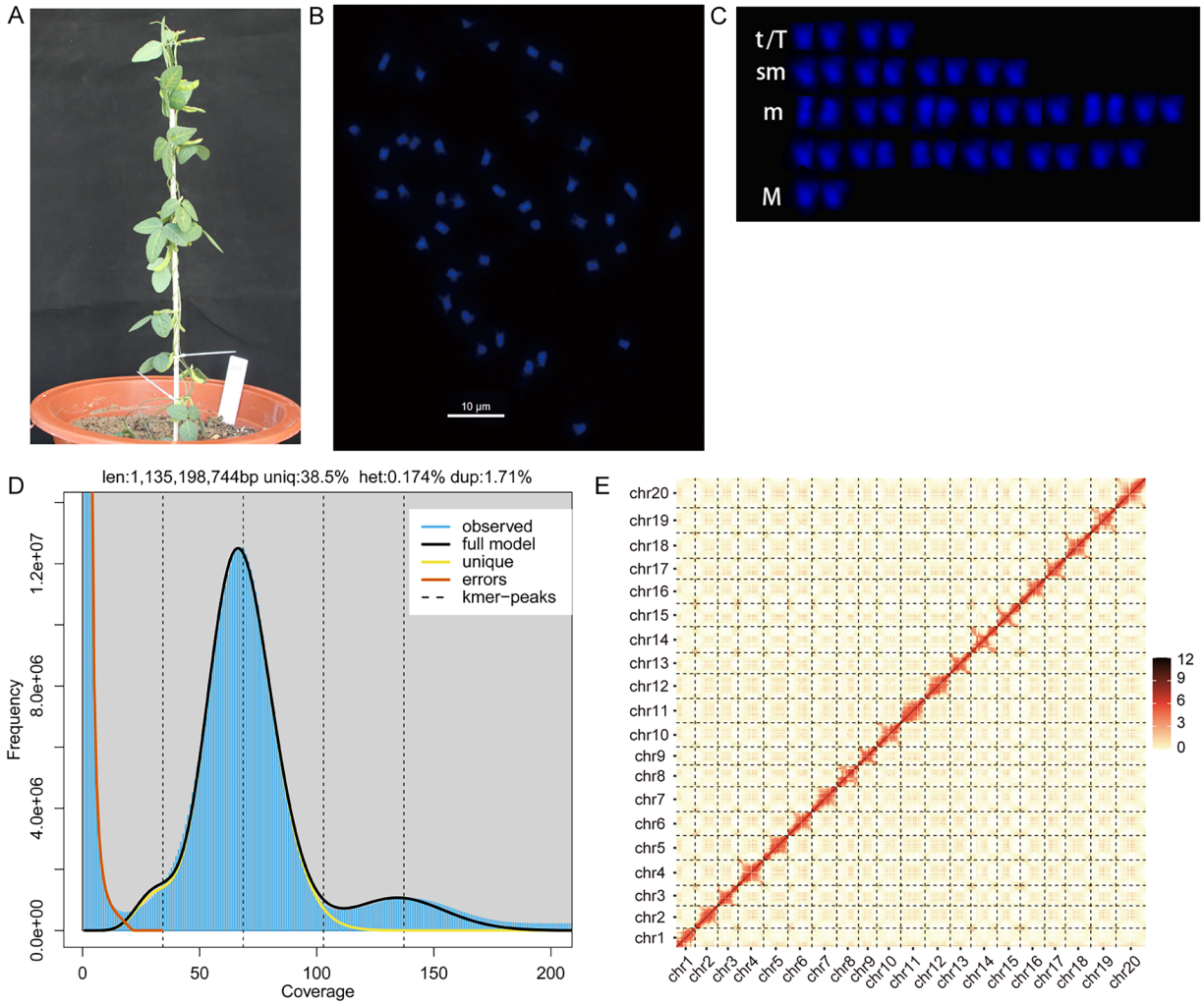

Fig. 1 The morphology, *k*-mer analysis, and karyotype of *Glycine soja* YSD56. (**A**) The plant sequenced in this study. (**B**) DAPI staining of chromosomes in YSD56. (**C**) The karyotype of YSD56. (**D**) 17-mer distribution was used for the estimation of genome size. (**E**) Hi-C chromatin interaction map of the *G. soja* YSD56 assembly.

Data Type	NGS	PacBio HiFi	ONT ultra-long	HiC	PacBio ISO-seq
Tissue	Leaf	Leaf	Leaf	Leaf	Mixture tissues (root, stem, leaf, and flower)
Plaform	Illumina	PacBio	PromethION	Illumina	PacBio
Total reads	301,486,205	2,617,224	13,527,546	412,815,679	45,210
Total bases (Gb)	91.05	44.19	240.85	123.84	—
Average length (bp)	150	16,885	17,804	150	1,582
Length N50 (bp)	—	17,474	44,976	—	1,716
Coverage ratio (%)	99.98	99.91	99.99	—	—
MeanDepth (X)	88.73	43.42	67.97	—	—

Table 2. Sequencing data yields of each library type for the *Glycine soja* YSD56.

(Table 2). PacBio Iso-Seq subreads were processed using SMRT Link (v11.1.0)[23] for classification, clustering, and consensus sequence generation, resulting in 45,210 full-length nonchimeric transcripts (Table 2).

The primary contig-level assembly was generated using Hifiasm (v0.18)[24] with PacBio HiFi reads, followed by the removal of redundant sequences using purge_dups (v1.2.5)[25]. Oxford Nanopore ultralong reads were assembled independently using NextDenovo (v2.5.0)[26] with the following parameters: "read_cutoff = 5k block-size = 2 g seed_cutoff = 89929 nextgraph_options = -a1". The Hifiasm-based assembly was selected as the preliminary assembly because of its superior continuity (higher contig N50) and higher completeness based on conserved single-copy orthologs (BUSCO (v5.4.5)[27], lineage: fabales_odb10) (Table 3).

	Hifiasm contig	Purge_dups contig	NextDenovo contig	HiC scaffold	TGS-GapCloser	Final assembly
Total number	675	39	41	32	20	20
Total length (bp)	1,044,147,026	1,008,801,473	1,010,190,352	1,009,023,018	1,007,794,034	1,008,523,555
N50 (bp)	49,649,712	49,649,712	35,720,038	51,918,298	51,918,298	51,965,791
N Numbers	0	0	0	1,107	0	0
GC (%)	35.11	35.01	35.28	35.01	35.01	35.01
BUSCO	C:97.9%[S:48.1%,D:49.8%]vv,F:0.2%,M:1.9%	C:97.9%[S:47.9%,D:50.0%],F:0.2%,M:1.9%	C:97.8%[S:39.3%,D:58.5%],F:0.3%,M:0.9%		—	C:99.7%[S:36.2%,D:63.5%],F:0.2%,M:0.1%

Table 3. Statistics of the YSD56 assembly of each step.

The Hi-C clean reads were then mapped to the preliminary assembly using Bowtie2 (v2.4.4)[28] with the following parameters: "-end-to-end -very-sensitive -L 30". Invalid Hi-C read pairs, including dangling ends, self-circles, re-ligations, and dumped pairs, were filtered out using the HiC-Pro pipeline[29]. The valid interactions were used to cluster, order, and orient contigs into pseudochromosomes using Juicer (v1.5)[30] and 3D-DNA (v180922)[31]. Juicebox (v1.11.08)[32] was used for manual inspection and refinement of the Hi-C contact maps. The final pseudochromosome orientation and nomenclature were manually curated based on the reference genome of cultivated soybean Wm82. A Hi-C contact map was constructed by binning the genome into 100-Kb windows and calculating the number of interactions between each pair of bins, which was visualized as a heatmap (Fig. 1E). Approximately 99.87% of the contigs were anchored to 20 chromosomes. The assembled genome was further refined by three rounds of error correction using Racon (v1.3.1)[33] with HiFi long reads, followed by two rounds of error correction using Pilon (v1.22)[34] with Illumina short reads. Finally, TGS-gapclose (v1.0.1)[35] was run with ONT long reads (–min_idy 0.3) to fill the remaining gaps, resulting in a gap-free assembly (Table 3).

A four-step strategy was employed to identify and validate telomeric regions in the *Glycine soja* YSD56 genome. First, ultralong Oxford Nanopore Technologies (ONT) reads were mapped to the gap-free genome assembly using Winnowmap (v2.03)[36] with "k = 15 -MD", retaining only those reads that uniquely aligned within the terminal 1,000 bp of each chromosome. Second, the retained reads were screened for telomeric 7-mer motifs (CCCTAA/TTAGGG) using tidk (v0.2.0)[37]. The reads with the highest motif frequencies were designated as telomeric references, whereas the remaining reads served as telomeric queries. Third, consensus sequences were generated from both telomeric references and queries using the medaka_consensus command in Medaka (v1.7.3)[38]. These consensus sequences were then aligned to the assembled chromosomes using MUMmer (v4.0.0)[39] to map telomere locations. Alignments with < 90% identity or those located > 20 kb from chromosome ends were discarded. High-confidence telomeric alignments were used to replace the corresponding terminal regions in the pseudochromosomes. A total of 40 telomeres ranging in length from 10,807 to 44,380 bp were identified (Fig. 2, Table 4). Finally, two rounds of error correction were performed using Racon (v1.3.1)[33] with PacBio HiFi reads to refine low-quality regions potentially introduced during the gap-filling and telomere correction steps. This process resulted in a telomere-to-telomere, gap-free reference genome of *G. soja* YSD56, with a total size of 1,008.52 Mb and a contig N50 of 51.97 Mb (Table 3).

Repeat annotation and centromere identification. Repetitive elements in the YSD56 genome were annotated using a combination of homology-based and *de novo* approaches. The homology-based method was performed using RepeatMasker (v4.1.4)[40] and RepeatProteinMask (v4.1.4)[40] with the Repbase database[41]. An ab initio approach was conducted using EDTA (v2.2.2)[42], which integrates structure-based detection of LTR retrotransposons, DNA transposons, and helitrons, and incorporates RepeatModeler (v2.0.4)[43] to annotate residual non-LTR elements. The results from both approaches were merged into a nonredundant repeat annotation set based on coordinate overlaps. A total of 563.47 Mb of repetitive sequences were identified, accounting for 55.87% of the entire genome, including 409.92 Mb (40.65%) of retroelements and 65.41 Mb (6.49%) of DNA transposons. The retroelements consisted of 15.40 Mb of long interspersed nuclear elements (LINEs, 1.53%), 430.64 kb of short interspersed nuclear elements (SINEs, 0.04%), and 394.09 Mb of long terminal repeat elements (LTRs, 39.08%).

Centromeric regions in the YSD56 genome were identified using a previously described pipeline (https://github.com/Immortal2333/Telomeres_and_Centromeres)[44]. Briefly, Tandem Repeat Finder (TRF v4.09.1)[45] was employed to scan satellite sequences from 30–500 bp across the genome with the parameters: "2 7 7 80 10 50 500 -f -d -m". The top seven repeat units with the highest period values on each chromosome were manually examined (filtering criteria: period ≥ 60, number of copies ≥ 10) to infer candidate centromeric regions. Among these families, repeat families with lengths of 91 and 92 bp, which have been previously reported as soybean-specific centromeric markers[46], were frequently detected. Based on the repeat annotation from EDTA (v2.2.2)[42], repeat family information within the candidate centromeric regions was extracted by keyword matching (e.g., Copia, Gypsy, Helitron). These candidate regions were further validated through visual inspection using the Integrative Genomics Viewer (IGV)[47,48]. The chromosomal density distributions of three major transposable element (TE) families (Copia, Gypsy, and Helitron) and the top seven tandem repeat units (trf91, trf92, trf182, trf273, trf276, trf183, and trf184) identified by Tandem Repeat Finder (TRF) are presented (Supplementary Fig. 1). Regions enriched with trf91 and trf92 correspond to the putative centromeric regions that were validated in this study. A total of twenty centromeres were identified (Fig. 2), with an average length of 2.67 Mb, ranging from 1.68 Mb (chr16) to 3.77 Mb (chr3) (Table 4). To further validate the identified centromeric regions, quarTeT (v1.1.5)[49]

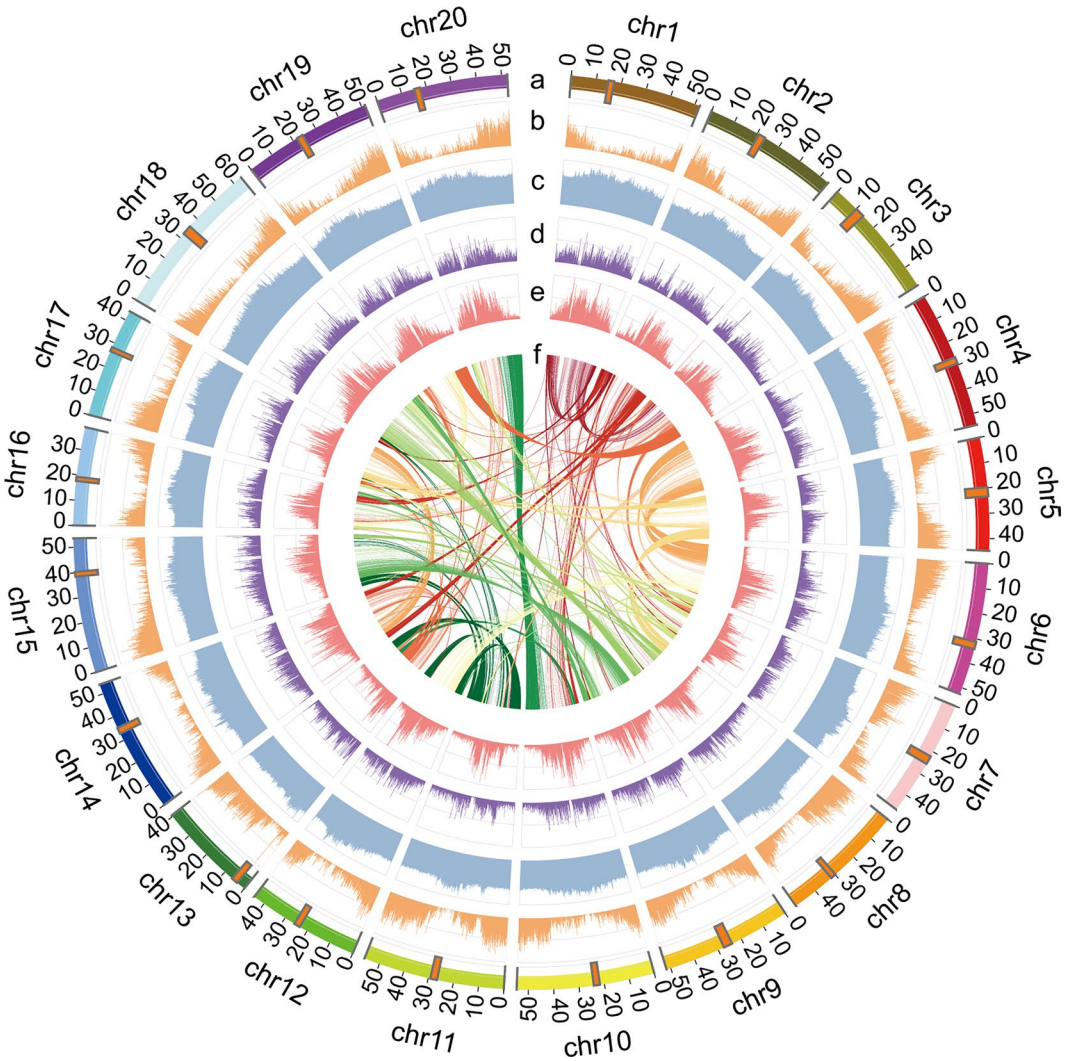

Fig. 2 Characteristics of the *Glycine soja* YSD56 genome. (**a**) Pseudochromosome length, telomeres and centromere of the *G. soja* YSD56 genome, (**b**) gene density, (**c**) GC density (**d**) LTR/Copia transposon density, (**e**) LTR/Gypsy transposon density, and (**f**) paralogous genes among different chromosomes; b-e were drawn in 100-Kb sliding windows.

was also applied. The centromeric coordinates predicted by quarTeT were highly consistent with those obtained from our TRF-EDTA-IGV pipeline across all chromosomes (Table 4).

Gene prediction and annotation. Protein-coding gene prediction in the YSD56 genome was performed by integrating homology-based annotation, ab initio prediction, and transcriptome-based evidence. For homolog-based annotation, the protein sequences of *Arabidopsis thaliana*, *G. max* Wm82, *G. max* ZH13, *G. soja* (W01, W02, W03), *Medicago truncatula* and *Phaseolus vulgaris* were downloaded from public databases (Supplementary Table 1). Protein sequences from the selected species were aligned to the YSD56 genome using Exonerate (v2.4.0)[50] with the parameters: "-model protein2genome–showalignment false -showtargetgff true -percent 50 -bestn 10–minintron 10–maxintron 130000". For each species, the best-scoring alignment at each genomic locus was retained for subsequent analysis.

For ab initio prediction, 2,000 full-length genes with intact structures, including canonical start/stop codons and complete intron–exon boundaries, were selected from the homology-based results to train AUGUSTUS (v3.5.0)[51] and SNAP (2006-07-28)[52].

For transcriptome-based annotation, short-read RNA-seq data from the roots, stems, leaves, flowers, and seeds of *G. soja* (W01, W02 and W03) at different developmental stages were retrieved from the Genome Warehouse (GWH) database (https://bigd.big.ac.cn/gsa/index.jsp) under accessions CRR108975 to CRR109207 (Supplementary Table 2). The reads were aligned to the YSD56 genome using HISAT2 (v2.2.1)[53], and the transcripts were assembled with StringTie (v2.2.1)[54]. Additionally, Iso-Seq reads were aligned to the genome using Gmap (20241010)[55] and Blat (v37x1)[56]. The resulting alignments were integrated and validated using PASA (v2.5.2)[57] (https://github.com/PASApipeline/PASApipeline).

Chromosome		Telomere length (bp)		Previous pipeline	quarTeT pipeline
Number	Length (bp)	Left	Right	Centromere boundary (Left : Right)	Centromere boundary (Left : Right)
chr1	51,965,791	24,290	22,852	14,725,354:16,959,149	11779045:17703825
chr2	53,953,519	44,380	15,930	20,784,311:23,518,192	21729768:26792144
chr3	48,555,363	12,600	31,504	10,089,016:13,862,687	12037366:13146700
chr4	54,718,854	30,870	14,835	26,778,277:29,326,900	25892053:34746321
chr5	44,815,473	22,540	13,584	19,752,904:23,351,587	15746595:24468211
chr6	52,557,790	30,030	11,941	31,111,143:34,013,213	29162865:35385731
chr7	46,925,628	15,400	40,959	22,027,357:25,376,614	21293333:26946605
chr8	49,694,999	31,080	11,870	29,808,479:31,963,646	29845484:31414723
chr9	51,749,905	28,630	31,876	24,062,836:27,584,044	23116513:26951771
chr10	54,554,946	13,720	10,807	21,723,247:24,146,385	21096178:23912573
chr11	56,189,710	15,960	13,871	26,489,161:29,197,312	26070765:30030836
chr12	44,692,882	18,060	34,563	23,246,419:25,694,235	23854546:25869936
chr13	40,529,669	30,660	34,810	3,664,961:6,232,852	5707146:6276172
chr14	53,859,623	16,660	14,294	32,945,992:35,448,361	31706818:37592794
chr15	53,557,297	15,680	16,608	38,450,810:40,355,155	38354933:39996911
chr16	38,978,300	13,300	19,101	17,517,473:19,195,023	14824680:20004147
chr17	43,795,261	17,080	29,232	26,366,000:28,089,958	24607136:29313212
chr18	63,645,620	17,500	15,481	32,075,316:35,847,483	30702294:38447477
chr19	51,506,071	28,350	14,072	21,733,521:24,296,767	14871548:23963214
chr20	52,276,854	16,450	13,945	15,273,420:17,594,385	13356987:16826554

Table 4. The statistics for each chromosome the YSD56.

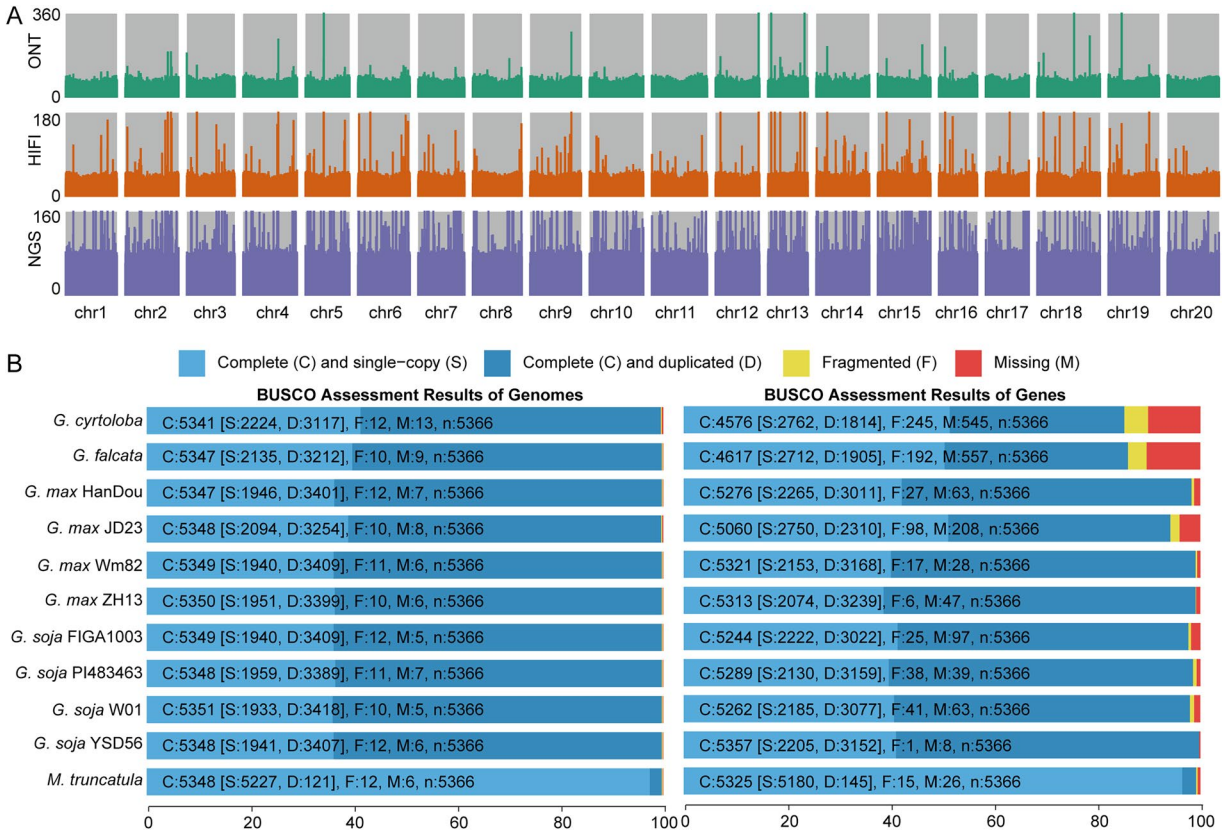

Fig. 3 Sequencing coverage and assembly validation of the T2T-YSD56. (**A**) Whole-genome coverage of mapped ONT, HiFi and Illumina reads is shown with primary alignments. Ultra-long ONT reads longer than 100 Kb and HiFi reads longer than 1 Kb were used for coverage analysis. (**B**) The relative proportion of complete and single copy (light blue), complete and duplicated (dark blue), fragmented (yellow), and missing (red) Benchmark Universal Single Copy Ortholog (BUSCO) genes identified for the assembly (left) and gene annotation set (right) of the 11 Fabaceae species.

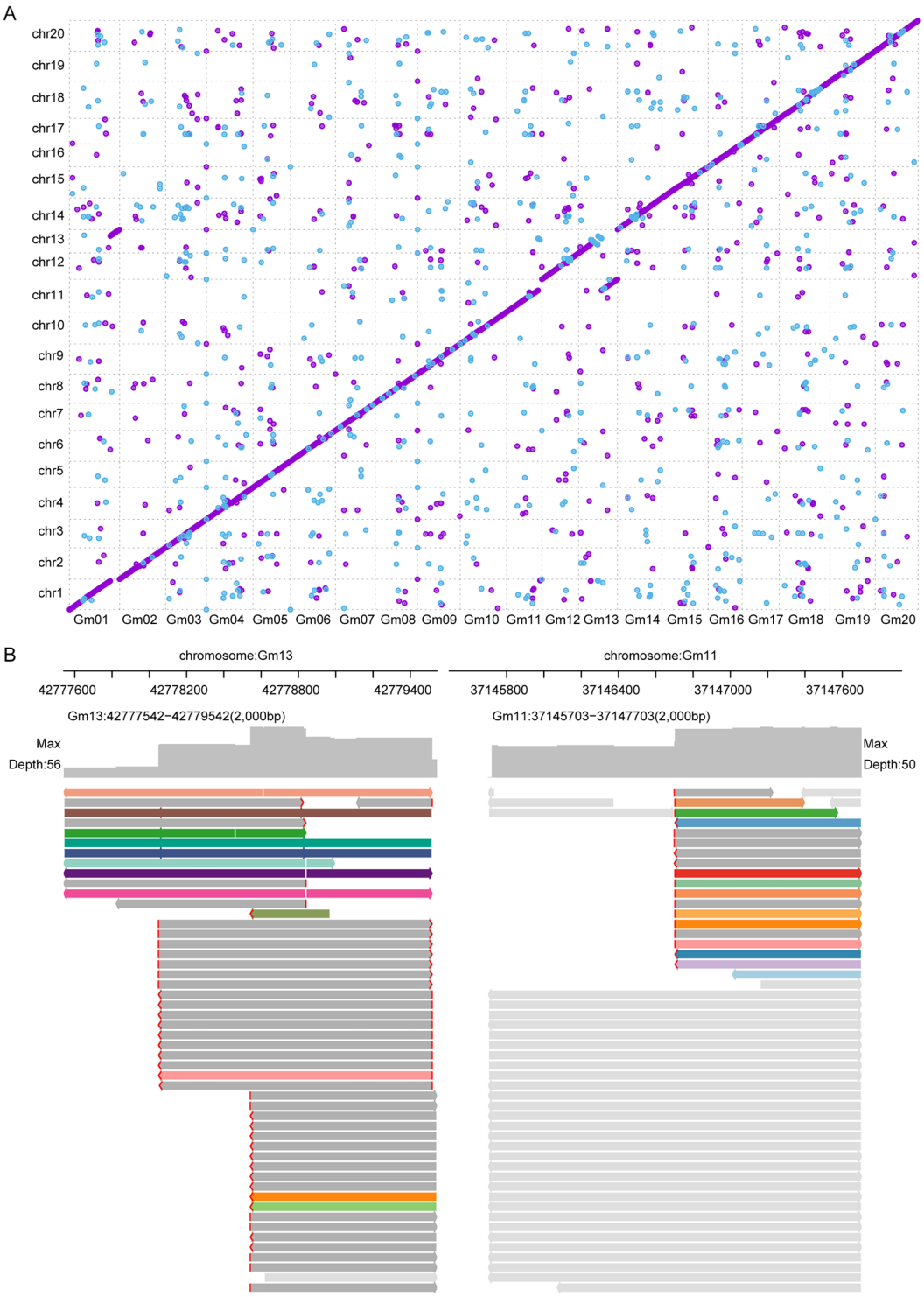

Fig. 4 Validation of reciprocal translocation between chromosomes 11 and 13 in wild soybean YSD56 compared to cultivated soybean Wm82. (**A**) Whole-genome alignment dot plot. MUMmer pairwise comparison between YSD56 and Wm82 reveals a balanced translocation between Gm11 and Gm13. Purple dots represent forward alignments (YSD56 vs Wm82), while light blue dots indicate reverse alignments. (**B**) Split-read mapping validation. YSD56 HiFi reads from BAM alignments to the Wm82 genome support a translocation breakpoint at Gm11: 37,145,703 and Gm13: 42,777,542, overlapping one predicted by Sniffles (Table 5). Color-filled bars indicate reads spanning the breakpoint regions.

The final set of protein-coding genes was generated using MAKER2 (v3.01.03)[58], which integrates ab initio prediction with the transcriptome and homology-based evidence. A total number of 57,712 protein-coding genes were identified, with an average gene length of 4337.50 bp and an average coding sequence (CDS) length of 1103.60 bp.

Breakpoint_chr_A	Breakpoint_pos_A	Breakpoint_chr_B	Breakpoint_pos_B	Support_Reads	VAF	Strand
Gm13	261184	Gm11	37001379	7	0.175	+−
Gm13	557275	Gm11	21592961	29	1	+−
Gm13	915903	Gm11	32213808	11	0.256	+−
Gm13	1300272	Gm11	32554877	12	0.522	+−
Gm13	1300272	Gm11	33727998	5	0.217	+−
Gm13	1330966	Gm11	31374628	7	0.206	+−
Gm13	2093150	Gm11	11047727	8	0.571	+−
Gm13	2403523	Gm11	34900849	14	0.452	+−
Gm13	4129248	Gm11	5634819	5	0.106	+−
Gm13	5792597	Gm11	22256454	6	0.545	+−
Gm13	5808072	Gm11	22258275	8	1	+−
Gm13	8174925	Gm11	29377238	7	0.189	+−
Gm13	10416405	Gm11	33557763	11	0.125	+−
Gm13	10419309	Gm11	18495549	7	0.212	+−
Gm13	12494505	Gm11	13766477	10	0.323	+−
Gm13	14473275	Gm11	39959617	10	0.263	+−
Gm13	14473275	Gm11	39968312	19	0.5	+−
Gm13	14723107	Gm11	18778209	14	0.298	+−
Gm13	18032729	Gm11	9827061	9	0.231	+−
Gm13	22515582	Gm11	39972295	41	0.719	+−
Gm13	25572601	Gm11	23791119	23	0.793	+−
Gm13	34233311	Gm11	24183800	25	0.272	+−
Gm13	34251746	Gm11	24191090	24	0.533	+−
Gm13	40461056	Gm11	17553512	8	0.154	+−
Gm13	42778542	Gm11	37146703	5	0.06	+−

Table 5. Translocation Breakpoints Identified from YSD56 HiFi Reads Aligned to the Wm82 Genome.

The average exon number per gene was 4.90, and the average exon length was 225.34 bp (Supplementary Table 3). The intragenomic synteny within YSD56 was identified using MCScan from the JCVI utility library[59] with the parameters: "–minspan = 20–minsize = 20" and visualized using Circos (v0.69-8)[60] (Fig. 2).

Gene function annotation was applied using sequence similarity and domain-based approaches. The predicted proteins were aligned to the SwissProt (20240327)[61], Trembl (20240327)[61] and NR (20240206)[62] databases using DIAMOND (v2.0.14)[63] with an E-value threshold of 1e-5. KEGG annotations were obtained using EggNOG-mapper (v2.0)[64]. Protein domains were identified using InterProScan (v5.71-102.0)[65], which utilizes information from the SUPERFAMILY[66], NCBIFAM[67], PRINTS[68], Pfam[69], SMART[70], ProSiteProfiles[71] and ProSitePatterns[71] databases. Overall, 98.79% of the protein-coding genes were annotated at least once in the above databases, including 77.50% with InterPro entries (44,727), 49.55% with GO terms (28,599), 97.23% with TrEMBL (56,114), and 98.68% with NR (56,950).

Data Records

All the raw sequencing data utilized in this study were submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under BioProject accession number PRJNA1095640. The Iso-Seq, ONT, Hi-C, HiFi and Illumina data were archived under the accession number SRP502474[72]. The genome assembly was submitted to GenBank with the accession number GCA_040083835.1[73], and the genome assembly and annotation are available on Figshare (https://doi.org/10.6084/m9.figshare.26004370.v2)[74].

Technical Validation

Evaluation of the genome assembly. The quality of the YSD56 genome assembly was evaluated to check its continuity, completeness and accuracy. For continuity evaluation, the NGS short reads were aligned to the assembly using BWA (v0.7.17)[75], and the PacBio HiFi and ONT ultralong reads (>100 kb) were aligned to the assembly using Minimap2 (v2.17)[76]. The assembly achieved mapping coverages of 99.98% (NGS, mean depth of 88.73×), 99.91% (PacBio HiFi, mean depth of 43.42×), and 99.99% (ONT, mean depth of 67.97×), indicating high continuity (Fig. 3A). Completeness was assessed using BUSCO (v5.4.5)[27] with the fabales_odb10 lineage (5,366 genes), revealing 5,348 (99.7%) complete BUSCOs, including 1,941 (36.2%) single-copy and 3,407 (63.5%) duplicated genes (Fig. 3B). The assembly accuracy was further evaluated using Merqury (v1.3)[77] with a *k*-mer size of 19, which yielded an average quality value (QV) of 52.16. These results demonstrate that the YSD56 genome has high continuity, completeness and accuracy.

A translocation between chromosomes 11 and 13 in *G. soja* relative to *G. max* Wm82 was previously inferred and validated using a fluorescence *in situ* hybridization (FISH)-based karyotyping system[78]. In this study, whole-genome alignments were performed between the YSD56 assembly and the W01, PI483463,

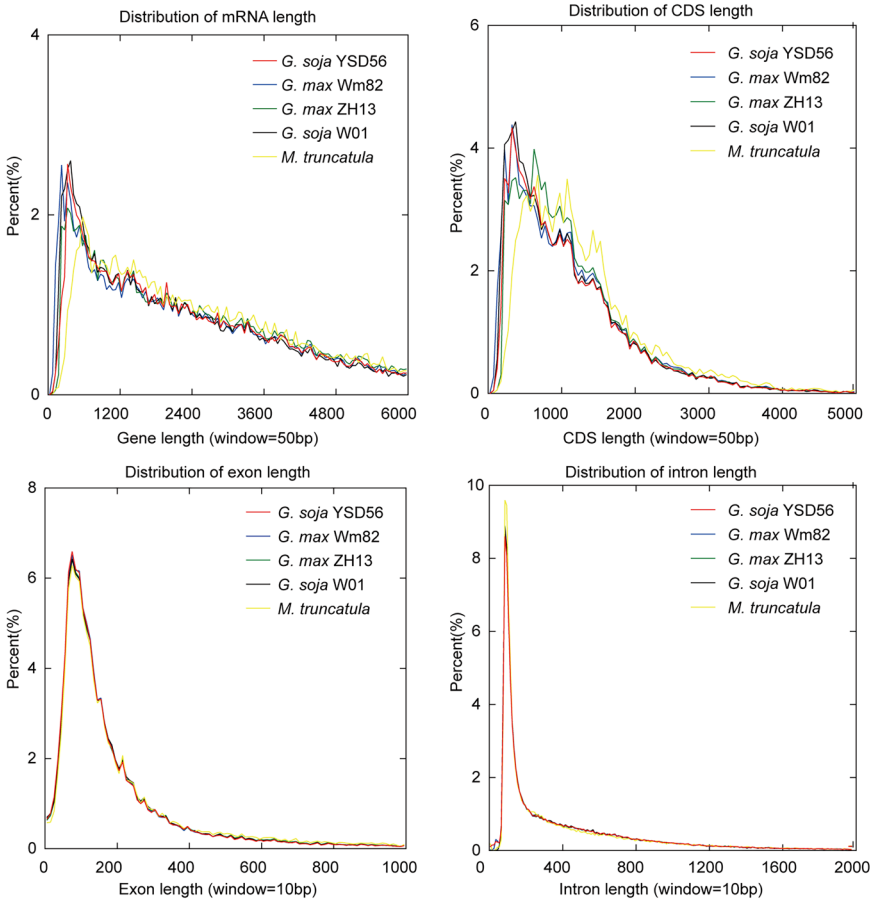

Fig. 5 Comparison of gene features among five Fabaceae gene sets.

and Wm82 genomes using MUMmer (v4.0.0)[39], followed by filtering with delta-filter to identify large-scale inter-chromosomal rearrangements (Fig. 4A, Supplementary Figs. 2A, 3A). Moreover, PacBio HiFi reads from YSD56 were aligned to these genomes using Minimap2 (v2.17)[76], and structural variants (SVs) were detected using Sniffles (v2.6.1)[79]. Translocations between chromosomes 11 and 13 were extracted using BCFtools (v1.13)[80], and breakpoint coordinates along with supporting read evidence were summarized (Table 5, Supplementary Tables 4, 5). Representative events were visualized using SVhawkeye[81] to confirm split-read support (Fig. 4B, Supplementary Figs. 2B, 3B). These analyses confirmed the occurrence of inter-chromosomal translocations in YSD56 relative to the W01, PI483463, and Wm82 genomes.

Evaluation of the gene annotation. The completeness of the YSD56 genome annotation was first evaluated using BUSCO (v5.4.5)[27] with the fabales_odb10 lineage, revealing that 99.8% of the conserved orthologous genes were found in the predicted protein-coding genes, including 41.1% single-copy and 58.7% duplicated genes (Fig. 3B). Compared with the two other T2T genomes of Wm82 and ZH13, the BUSCO score of the YSD56 gene set was greater (Fig. 3B). Moreover, the length distributions of genes, CDSs, exons, and introns were compared among *G. soja* YSD56, *G. max* Wm82, *G. max* ZH13, *G. soja* W01 and *M. truncatula*, indicating that YSD56 has similar distribution patterns to those observed across the Fabaceae family (Fig. 5, Supplementary Table 3).

Code availability

All data processing commands and pipelines were carried out in accordance with the instructions and guidelines provided by the relevant bioinformatic software. This study does not involve custom scripts or code.

Received: 11 June 2024; Accepted: 1 August 2025;
Published online: 13 August 2025

References

1. Wang, L. *et al.* A telomere-to-telomere gap-free assembly of soybean genome. *Molecular Plant* **16**, 1711–1714 (2023).
2. Zhang, C. *et al.* The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. *Molecular Plant* **16**, 1715–1718 (2023).
3. Fang, Y., Lu, L., Chen, K. & Wang, X. Tradeoffs among root functional traits for phosphorus acquisition in 13 soybean genotypes contrasting in mycorrhizal colonization. *Annals of Botany* **134**, 179–190 (2024).
4. Zhuang, Y. *et al.* Phylogenomics of the genus Glycine sheds light on polyploid evolution and life-strategy transition. *Nature Plants* **8**, 233–244 (2022).

5. Xie, M. *et al*. A reference-grade wild soybean genome. *Nature communications* **10**, 1216 (2019).
6. Peng, D., Jiang, R., Peng, H. & Liu, S. Soybean cyst nematodes: a destructive threat to soybean production in China. *Phytopathology Research* **3**, 19 (2021).
7. Miraeiz, E. *et al*. Early transcriptional responses to soybean cyst nematode HG Type 0 show genetic differences among resistant and susceptible soybeans. *Theoretical and Applied Genetics* **133**, 87–102 (2020).
8. Lian, Y. *et al*. Chromosome-level reference genome of X12, a highly virulent race of the soybean cyst nematode Heterodera glycines. *Molecular ecology resources* **19**, 1637–1646 (2019).
9. Lian, Y. *et al*. The spatial distribution and genetic diversity of the soybean cyst nematode, Heterodera glycines, in China: It is time to take measures to control soybean cyst nematode. *Frontiers in Plant Science* **13**, 927773 (2022).
10. Lian, Y. *et al*. A new race (X12) of soybean cyst nematode in China. *Journal of nematology* **49**, 321 (2017).
11. Cook, D. E. *et al*. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *science* **338**, 1206–1209 (2012).
12. Liu, S. *et al*. A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. *Nature* **492**, 256–260 (2012).
13. Liu, S. *et al*. The soybean GmSNAP18 gene underlies two types of resistance to soybean cyst nematode. *Nature communications* **8**, 14822 (2017).
14. Niblack, T. L. *et al*. A revised classification scheme for genetically diverse populations of Heterodera glycines. *Journal of nematology* **34**, 279 (2002).
15. Usovsky, M. *et al*. Classification Methods and Identification of Reniform Nematode Resistance in Known Soybean Cyst Nematode-Resistant Soybean Genotypes. *Plant Disease* **106**, 382–389 (2022).
16. Bent, A. F. Exploring soybean resistance to soybean cyst nematode. *Annual Review of Phytopathology* **60**, 379–409 (2022).
17. Belton, J.-M. *et al*. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
18. Pei, D. U. *et al*. Chromosome painting of telomeric repeats reveals new evidence for genome evolution in peanut. *Journal of Integrative Agriculture* **15**, 2488–2496 (2016).
19. Levan, A., Fredga, K. & Sandberg, A. A. Nomenclature for centromeric position on chromosomes. *Hereditas* **52**, 201–220 (2009).
20. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
21. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
22. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications* **11**, 1432 (2020).
23. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**, 563–569 (2013).
24. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170–175 (2021).
25. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
26. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology* **25**, 107 (2024).
27. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
28. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357–359 (2012).
29. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**, 1–11 (2015).
30. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95–98 (2016).
31. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
32. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101 (2016).
33. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737–746 (2017).
34. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
35. Xu, M. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, giaa094, https://doi.org/10.1093/gigascience/giaa094 (2020).
36. Jain, C. *et al*. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
37. Brown, M. R., Gonzalez de La Rosa, P. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics*, btaf049 (2025).
38. Luan, T. *et al*. Benchmarking short and long read polishing tools for nanopore assemblies: achieving near-perfect genomes for outbreak isolates. *BMC genomics* **25**, 679 (2024).
39. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* **14**, e1005944 (2018).
40. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4–10 (2004).
41. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **6**, 1–6 (2015).
42. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology* **20**, 1–18 (2019).
43. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
44. Shi, X. *et al*. The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. *Horticulture Research* **10**, uhad061 (2023).
45. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573–580 (1999).
46. Gill, N. *et al*. Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant physiology* **151**, 1167–1174 (2009).
47. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178–192 (2013).
48. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
49. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture research* **10**, uhad127 (2023).
50. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* **6**, 1–11 (2005).
51. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435–W439 (2006).
52. Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 1–9 (2004).
53. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907–915 (2019).
54. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295 (2015).

55. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
56. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656–664 (2002).
57. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654–5666 (2003).
58. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* **12**, 1–14 (2011).
59. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211 (2024).
60. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639–1645 (2009).
61. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* **31**, 365–370 (2003).
62. Wheeler, D. L. *et al*. Database resources of the national center for biotechnology information. *Nucleic acids research* **36**, D13–D21 (2007).
63. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).
64. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* **38**, 5825–5829 (2021).
65. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
66. Wilson, D., Madera, M., Vogel, C., Chothia, C. & Gough, J. The SUPERFAMILY database in 2007: families and functions. *Nucleic acids research* **35**, D308–D313 (2007).
67. Li, W. *et al*. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic acids research* **49**, D1020–D1028 (2021).
68. Attwood, T. K. *et al*. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Research* **28**, 225–227 (2000).
69. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res* **42**, D222–230, https://doi.org/10.1093/nar/gkt1223 (2014).
70. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res* **37**, D229–232, https://doi.org/10.1093/nar/gkn808 (2009).
71. Hulo, N. *et al*. The PROSITE database. *Nucleic Acids Res* **34**, D227–230, https://doi.org/10.1093/nar/gkj063 (2006).
72. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP502474 (2024).
73. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_040083835.1 (2024).
74. Lian, Y., Li, C., Wang, J. & Lu, W. A telomere-to-telomere genome of wild soybean with resistance to soybean cyst nematode X12. *figshare* https://doi.org/10.6084/m9.figshare.26004370.v2 (2024).
75. Jung, Y. & Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **38**, 2404–2413, https://doi.org/10.1093/bioinformatics/btac137 (2022).
76. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
77. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
78. Findley, S. D. *et al*. A fluorescence *in situ* hybridization system for karyotyping soybean. *Genetics* **185**, 727–744 (2010).
79. Smolka, M. *et al*. Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology* **42**, 1571–1580 (2024).
80. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
81. Xiao, Y., Yu, T., Liang, F. & Hou, T. SVhawkeye: an ultra-fast software for user-friendly visualization of targeted structural fragments from BAM files. *Frontiers in Genetics* **15**, 1352443 (2024).

Acknowledgements

This research was funded by Joint Fund for S&T Research and Development Plan of Henan in China (222301420028).

Author contributions

Y.L. and W.L. conceived and designed the study. Y.L. collected the samples. C.L. and J.W. performed the data analysis. P.D. processed the Karyotypic analysis. Y.L. wrote the manuscript. H.W., C.L., Y.W., J.L., C.L., H.L., S.W., H.Z., J.W. and W.L. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05741-y.

Correspondence and requests for materials should be addressed to J.W. or W.L.