Theoretical and Applied Genetics (2025) 138:119
https://doi.org/10.1007/s00122-025-04881-4

ORIGINAL ARTICLE

# Chromosome-level genome assembly of a high-yield Chinese soybean variety Mengdou1137 unlocks genetic potential of disease and lodging resistance

Rujian Sun[1,2] · Bincheng Sun[2] · Zihao Zheng[3] · Qi Zhang[2] · Xingguo Hu[2] · Rongqi Guo[2] · Lei Feng[2] · Shen Chai[2] · Jingshun Wang[2] · Ping Qiu[2] · Ping Yu[2] · Ying Liu[2] · Wei Song[2] · Yinghui Li[1] · Lijuan Qiu[1]

Received: 12 September 2024 / Accepted: 11 March 2025


## Abstract

***Key message*** **We assembled the genome of Mengdou1137 with high quality and revealed the specific disease resistance genes and a large number of genomic variations related to agronomic traits.**

**Abstract** As a cornerstone in the global agricultural landscape, soybean stands as a pivotal oilseed crop, underpinning both nutritional and industrial applications. The burgeoning development of novel soybean varieties significantly propels the crop's industrialization, offering enhanced traits that cater to diverse agricultural and commercial needs. In this study, we present the de novo assembly of the genome a high-yield Chinese soybean variety Mengdou1137, employing an integrated approach of both long-read and short-read sequencing technologies to achieve comprehensive genomic insights. Achieving a notable assembly with a genome size of 999.99 Mb, our work features a contig N50 of 14.92 Mb and a scaffold N50 of 50.26 Mb, successfully anchoring 98.24% of sequences across the 20 chromosomes. Through meticulous comparative analysis with existing soybean genomes, our research unveiled 115 Mengdou1137-specific disease resistance genes alongside a substantial array of agronomical trait-associated genomic variants. Among the salient genomic features, we identified a favorable haplotype of the dwarf gene *PH13*, a critical determinant of plant stature, underscoring its potential for breeding compact soybean varieties with lodging resistance. This high-quality assembly of the Mengdou1137 genome not only enriches the repository of soybean genetic resources but also paves the way for future innovations in soybean breeding and trait improvement, offering valuable insights for the enhancement of this crucial agricultural commodity.

Communicated by Henry T. Nguyen.

Rujian Sun, Bincheng Sun and Zihao Zheng have contributed equally to the work.

✉ Yinghui Li
liyinghui@caas.cn

✉ Lijuan Qiu
qiulijuan@caas.cn

[1] State Key Laboratory of Crop Gene Resources and Breeding/ the National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI)/Key Laboratory of Grain Crop Genetic Resources Evaluation and Utilization (MARA), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2] Hulunbuir Institute of Agriculture and Animal Husbandry, Hulunbuir 021000, Inner Mongolia, China

[3] Department of Agronomy, Iowa State University, Ames, IA 50011-1051, USA

## Abbreviations

| | |
|---|---|
| TE | Transposable elements |
| SV | Structural variation |
| GO | Gene Ontology |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |

## Introduction

Soybean (*Glycine max* [L.] Merr.), a cornerstone of global agriculture, stands at the forefront of economic and food crops, providing essential protein and oil to humans and animals alike (Leamy et al. 2017; Zhang et al. 2022). Over the past few decades, the continuous emergence of new elite varieties has contributed to increase soybean yields (Chu et al. 2021; Shen et al. 2018). Despite these advancements, the scope of available soybean genomic data is still restricted, revealing a substantial gap in our comprehension of the genetic underpinnings of crucial agronomic traits, pivotal in modern breeding endeavors.

Published online: 15 May 2025

The progress in soybean genetic breeding is deeply rooted in the exploration and utilization of diverse germplasm resources, focusing on the precise decoding of their genetic makeup. This endeavor is instrumental in tracking genomic variations, uncovering novel genes and identifying new molecular markers (Chu et al. 2021; Liu et al. 2020). These essential genetic analyses are inherently reliant on a series of high-quality reference genomes that encompass a wide array of genetic resources. The rapid advancements in high-throughput sequencing and assembly technologies have paved the way for extensive genomic applications across various soybean varieties.

The first soybean reference genome was the American cultivated soybean variety Williams 82 developed in the 1980s (Schmutz et al. 2010). With the rapid development of whole-genome sequencing technologies, particularly long-read sequencing, an increasing number of individual soybean genomes have been published, reflecting diverse germplasm sources. Early assemblies included the Japanese variety Enrei (Shimomura et al. 2015), followed by Zhonghuang 13 (ZH13) from central China, which also facilitated the construction of a comprehensive gene co-expression network using public RNA-seq datasets (Shen et al. 2018, 2019). Subsequent efforts focused on the American southern cultivar Lee (Valliyodan et al. 2019), the wild soybean accession W05 (Xie et al. 2019) and the Korean variety Hwangkeum (Kim et al. 2021). More recently, a high-quality genome of the modern Chinese cultivar Nongdadou2 has been published (Zhang et al. 2024b). Collectively, these assemblies have significantly expanded our understanding of soybean genomics over the past decade (Zhang et al. 2022), and the development of soybean pan-genomes from genetically diverse panels (Liu et al. 2020; Torkamaneh et al. 2021) has further enriched the genomic landscape and facilitated deeper insights into soybean genetics. Jack cultivar and JD17 cultivar were genome-sequenced by Huang et al. (2024) and Yi et al. (2022). Highly contiguous, nearly gapless, genome assemblies (Wm82.a5) were developed for two economically important soybean cultivars (Williams82 and Lee) (Garg et al. 2023). Two genome assemblies including the Wm82.a6 and Fiskeby III with their comparative analysis were presented (Espina et al. 2024). In addition, recent efforts have generated telomere-to-telomere (T2T) genome assemblies of the soybean cultivars Wm82 and ZH13, fully closing all gaps and providing complete telomeric and centromeric regions, thereby greatly advancing our ability to dissect soybean genomic complexity and evolutionary dynamics (Wang et al. 2023; Zhang et al. 2024b).

Mengdou1137, a novel Chinese soybean cultivar developed via pedigree selection and sexual hybridization, demonstrates potential for widespread adoption across China's Northeastern region, owing to its consistently high yields and exceptional resistance to adverse conditions (Fig. 1). Here, we report a highly contiguous, chromosome-level

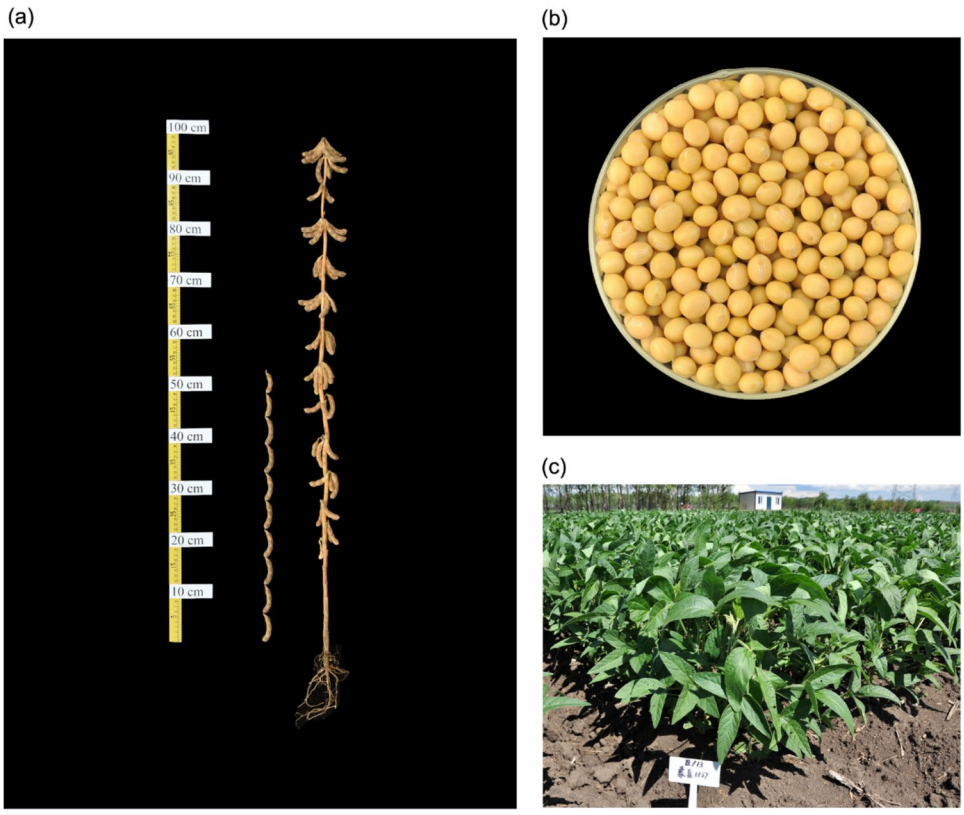


**Fig. 1** Morphology and agronomic traits of Mengdou1137: plant (**a**), grain (**b**) and field performance (**c**) of Mengdou1137

assembly for Mengdou1137, with contig and scaffold N50 values of 14.92 Mb and 50.26 Mb, respectively. Through comparative analysis with a wide collection of published soybean genomes, we illuminated Mengdou1137-specific disease resistance genes and structural variations related to key agronomic traits. This genome assembly is poised to enrich soybean genomic research and facilitate the refinement of elite soybean varieties.

## Materials and methods

### Genome sequencing, assembly and anchoring

The genomic DNA was extracted from Mengdou1137 leaves. Libraries for Illumina short-read (NovaSeq 6000) and Oxford Nanopore (ONT PromethION platform) long-read sequencing were prepared in accordance with the manufacturer's guidelines. Low-quality ONT reads ($q$-value < 7) were filtered out, and the remaining high-quality data were used to assemble the draft genome using NextDenovo with the following parameters: 'sort_options = – m 4 g – t 8 – k 40 minimap2_options_cns = – x ava-ont – t 8 – k17 – w17' (Hu et al. 2024). NextPolish was applied to correct assembly errors combined with Illumina and ONT reads (Hu et al. 2020). The Nanopore assembly results were combined with Illumina data for error correction, ensuring contig assembly accuracy of more than 99.9% (Jain et al. 2018). With the reference of Wm82 (Wm82.a4.v1), we used the RAGOO software and RBSA (https://github.com/likui345/rbsa) to attach contig genome to chromosome level (Alonge et al. 2019). These two methods use the most common software minimap2 and Mummer, respectively. Finally, the results of the two methods are combined to generate chromosomal genomes. The method does not change the contig sequence, only clustered, ordered and oriented the contigs by sequence similarity, which ensure the accuracy of the assembly sequence. In the end, we used BUSCO for evaluation and reached 99.5%, which further illustrates the accuracy of the assembly results. Finally, we successfully get a chromosomal-level genome for Mengdou1137.

### Genome annotation

For repeat sequence annotation, we applied the software Repeatmasker based on Repbase repetitive sequences database (Jurka 2000) as well as de novo prediction by LTRFinder (Xu and Wang 2007).

For protein-coding gene annotations, we combined three strategies, including homolog-based, de novo and transcriptome-based predictions. In detail, protein sequences from Arabidopsis thaliana, *Glycine max* (Wm82), *Medicago truncatula*, *Oryza sativa*, *Phaseolus vulgaris* and *Solanum lycopersicum* were collected to predict gene models using Genewise (Birney et al. 2004). Five distinct algorithms were used to predict de novo gene models, including Augustus (Stanke et al. 2008), Genscan (Burge and Karlin 1997), GlimmerHMM (Burge and Karlin 1997), Geneid (Guigó 1998) and SNAP (Korf 2004). We downloaded and aligned publicly available RNA-seq datasets from SoyC05, which included nine distinct tissue samples and developmental stages, as follows: (A) root at growth stage V1, (B) stem at V1, (C) young leaf at V1, (D) mature leaf at R1, (E) old leaf at R4, (F) flower at R1, (G) pod and seed before 4 weeks, (H) seed at 6 weeks and (I) seed at 8 weeks. These datasets were obtained from the Genome Sequence Archive (GSA) and Genome Warehouse (GWH) in the BIG Data Center (https://bigd.big.ac.cn/gsa/index.jsp) under Accession Number PRJCA002030. The predicted gene models were performed using Cufflinks (Trapnell et al. 2010) and PASA (Haas et al. 2003). Finally, all predicted gene models were integrated by EVidenceModeler (Haas et al. 2008).We predicted the function of each gene against four databases, including Swiss-Prot (http://www.uniprot.org/), InterPro (https://www.ebi.ac.uk/interpro/), KEGG (http://www.genome.jp/kegg/) and Non-Redundant Protein Sequence (NR, http://www.ncbi.nlm.nih.gov/protein).

### Identification of disease resistance genes

We employed the integrative RGAugury pipeline (Li et al. 2016b) for the identification of disease resistance genes in Mengdou1137. This process involved the use of resistance gene analogs, such as NBS-encoding proteins (Kim et al. 2012), receptor-like protein kinases (RLKs) and receptor-like proteins (RLPs) (Böhm et al. 2014; Li et al. 2016a) as benchmarks to ascertain the conserved domains and motifs within the protein sequences. To pinpoint Mengdou1137-specific disease resistance genes, a gene family analysis was conducted comparing Mengdou1137 against 26 publicly accessible soybean genomes (Liu et al. 2020) (Supplementary Table S1 and S2) utilizing OrthoMCL (Li et al. 2003). This facilitated the isolation of Mengdou1137-specific genes, which were then cross-referenced with the initially predicted disease resistance genes to identify unique disease resistance genes in Mengdou1137. To further elucidate the metabolic and signaling pathways associated with Mengdou1137-specific disease resistance genes, we performed KEGG pathway enrichment analysis by mapping these genes to the Kyoto Encyclopedia of Genes and Genomes database (http://www.genome.jp/kegg/) and identifying significantly enriched pathways through a hypergeometric test with false discovery rate (FDR) correction (FDR < 0.05).

### Identification of structural variants

The sequence of Mengdou1137 and aforementioned 26 soybean genomes was aligned to Wm82 using MUMmer (Delcher et al. 2002). Structural variants, including insertions, deletions, translocations and inversions, were identified using software SyRI with default parameters (Goel et al. 2019). The resulting structural variants across the 27 soybean genomes were then consolidated and refined using Jasmine (Kirsche et al. 2023). Subsequently, single-nucleotide polymorphisms (SNPs) were extracted based on whole-genome alignment.

### Comparative analysis of agronomical trait-associated SNPs in Mengdou1137

A set of SNPs were obtained by aligning Mengdou1137 assembly use nucmer software (Kurtz et al. 2004) (version 3.1, nucmer –mum -c 1000 -l 1000) with the soybean reference genome Wm82 (Wm82.a4.v1). The results were filtered by delta filter (-i 95 -l 100 −1). After obtaining the SNPs using dnadiff, these SNPs were compared with the previously identified 3,986 QTL (Supplementary Table S3) and trait-associated SNPs for important agronomical traits including yield, seed quality, disease resistance and adaptability from GWAS in the Soybase database (https://www.soybase.org/) and determined how many of these SNPs in Mengdou1137 carried favorable alleles and performed the same calculation for a set of 598 additional resequenced lines (Li et al. 2023). By comparing Mengdou1137's favorable allele counts against this panel, we established its relative ranking and demonstrated its genetic potential in key agronomic traits (Supplementary Table S4).

In the validation of *sss1* gene haplotypes (Zhu et al. 2022), upon maturation of the soybeans, 20–30 plants were randomly selected from each test plot for sampling. The investigation focused on two key metrics: Firstly, the number of nodes from the cotyledon node to the main stem tip was counted and averaged ($P = 2.7 \times 10^{-16}$); and secondly, the number of pods per plant was assessed by counting the grains per plant, including immature, diseased and worm-affected grains, while excluding dark grains, with averages calculated for each plot. This assessment was conducted using the Wilcoxon rank-sum test to ensure statistical rigor.

## Results

### Genome sequencing and assembly

A total of 115.93 Gb Oxford Nanopore long reads and 52.28 Gb whole-genome shotgun sequencing data were produced to assemble a draft genome of 999.99 Mb for Mengdou1137 (Table 1), which is close to 1,010 Mb of the cultivated soybean reference genome of Wm82 (Garg et al. 2023; Wang et al. 2023). This preliminary assembly contains 283 contigs and 117 scaffolds, with an N50 length of 14.92 Mb and 50.26 Mb, respectively (Table 1). The evaluation of BUSCO completeness (99.50%), LTR assembly index (10.20%) and RNA-seq mapping ratio (90.79%, Supplementary Fig. S1, Supplementary Fig. S2, Supplementary Table S5) all supported the high quality of our assembly (Table 1). By aligning this draft assembly to Wm82, 982.41 Mb sequences (98.24%) were successfully anchored to 20 pseudo-chromosomes ($2n = 40$; Fig. 2a, b; Supplementary Table S6). This high-quality chromosomal-level genome was thus employed for further analysis.

### Genome annotation

A total of 522.70 Mb repetitive elements were annotated, representing 52.27% of the Mengdou1137 genome (Table 1; Supplementary Table S7, Supplementary Fig. S3). Among those, long terminal repeat (LTR) retrotransposons are the predominant classified repetitive elements, occupied 44.49% (444.86 Mb) of the Mengdou1137 genome, which were mainly attributed to the possible expansion of subclass Gypsy (31.31% or 313.16 Mb; Supplementary Table S7). A relatively small fraction of the Mengdou1137 genome harbored DNA transposons (5.70% or 57.02 Mb) and unclassified repeats (4.42% or 44.22 Mb; Supplementary Table S7).

**Table 1** Summary of assembly and annotation

| Items | Number |
|---|---|
| Total size of genome assembly | 999.99 Mb |
| Chromosome number (2*n*) | 40 |
| Number of scaffolds | 117 |
| Longest scaffold | 61.67 Mb |
| Scaffold N50 length | 50.26 Mb |
| Scaffold L50 | 10 |
| Number of contigs | 283 |
| Longest contig | 34.99 Mb |
| Contig N50 length | 14.92 Mb |
| Contig L50 | 24 |
| GC content (%) | 35.00 |
| BUSCO evaluation (%completeness) | 99.50 |
| LTR assembly index (%) | 10.20 |
| RNA mapping ratio (%) | 90.79 |
| Repetitive elements (%) | 52.27 |
| Number of genes | 57,002 |
| Number of transfer RNAs | 1064 |
| Number of ribosomal RNAs | 345 |
| Number of small nuclear RNAs | 853 |
| Number of microRNAs | 833 |

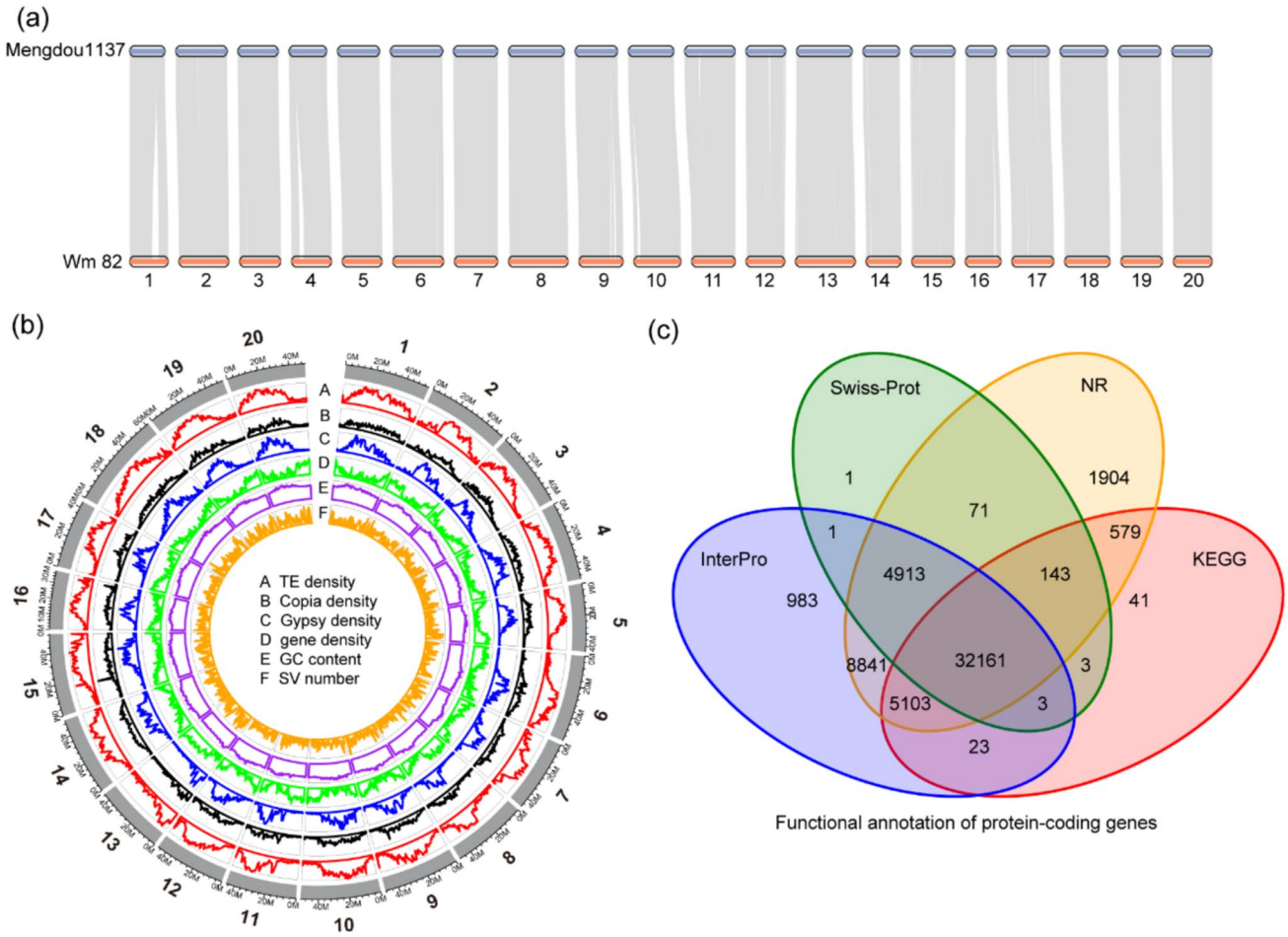


**Fig. 2** Assembly and annotation: **a** synteny blocks between Mengdou1137 and Wm82 were used to anchor scaffolds to each chromosome. **b** A physical map illustrating the 20 chromosomes for Mengdou1137, featuring: (A) TE density, (B) Copia density, (C) Gypsy density, (D) gene density, (E) GC content and (F) SV number. **c** Results of functional annotations of protein-coding genes using four distinct databases

Utilizing a comprehensive approach that combines homology-based, transcriptome-based and de novo predictions, we successfully annotated a total of 57,002 protein-coding genes in the Mengdou1137 genome (Table 1; Supplementary Table S8; Supplementary Fig. S4). The average gene length is 3,623 bp, with a coding region length of 1,069 bp, and ~ 4.81 exons per gene (Supplementary Table S8). Out of these genes, 54,770 (96.10%) received functional annotations in at least one of the NR (http://www.ncbi.nlm.nih.gov/protein), KEGG (http://www.genome.jp/kegg/), Swiss-Prot (http://www.uniprot.org/) and InterPro (https://www.ebi.ac.uk/interpro/) databases (Fig. 2c, Supplementary Table S9). Additionally, noncoding RNAs were also identified, encompassing 1,064 transfer RNAs, 345 ribosomal RNAs, 853 small nuclear RNAs and 833 microRNAs (Table 1; Supplementary Table S10).

## Identification of disease resistance in Mengdou1137

Deciphering disease resistance genes holds paramount importance for enhancing growth and yield in soybean breeding programs. To achieve this objective, we conducted a comprehensive prediction of 2,739 disease resistance genes in the Mengdou1137 genome (Supplementary Table S1). This set includes 336 nucleotide-binding site (NBS)-encoding proteins, 413 RLKs, 161 RLPs, 1,353 transmembrane coiled coils (TM-CCs) and 476 genes with uncertain functions (Fig. 3a).

Through a comparative analysis between Mengdou1137 and 26 other soybean genomes (Liu et al. 2020), we identified a set of 7,484 gene families (comprising 7,588 genes) exclusive to the Mengdou1137 genome (Fig. 3b). Among these Mengdou1137-specific genes, 115 are related to disease resistance (Fig. 3b, c, Supplementary Table S11), with a

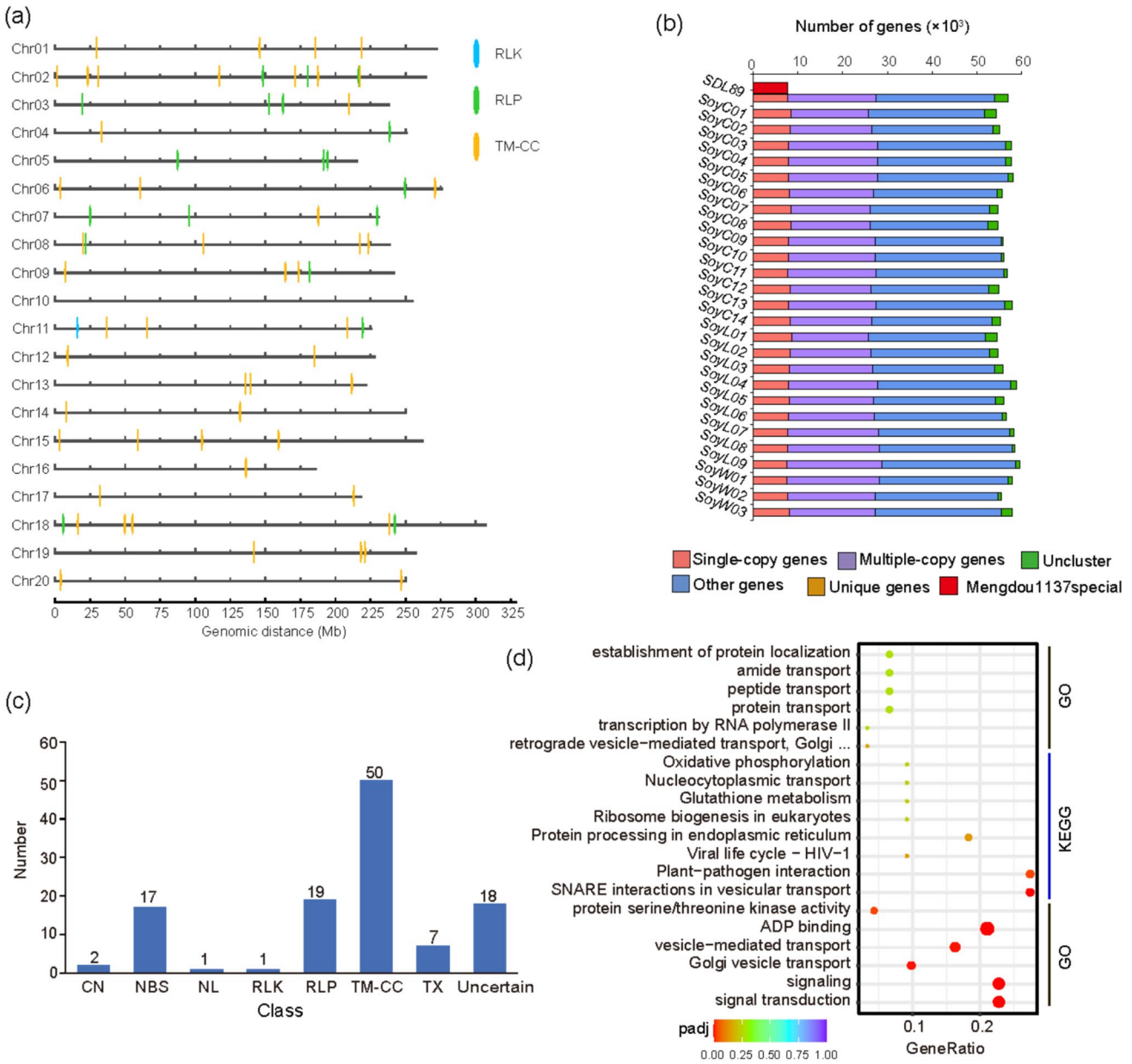


**Fig. 3** Identification of disease resistance genes in Mengdou1137: **a** genomic distribution of disease resistance genes in Mengdou1137. **b** Gene family clustering among Mengdou1137 and 26 soybeans genomes. **c** Summary of Mengdou1137-specific disease resistance genes. **d** GO and KEGG enrichment of Mengdou1137-specific disease resistance genes

notable enrichment in ADP binding, signaling and transport functions (Fig. 3d, Supplementary Table S12). These findings suggest that the unique genetic composition of Mengdou1137 may confer specialized disease resistance capabilities, offering promising avenues for the genetic improvement of soybean cultivars in breeding programs.

## Allele mining for agronomical trait improvement in Mengdou1137

By aligning Mengdou1137 to the Wm82 reference genome, we identified a large set of genomic variants, including 40,938 deletions, 53,409 insertions, 139 inversions, 644

translocation and the most abundant of 7,107,093 SNPs (Supplementary Table S13). Based on the collected information of 3,986 SNP loci associated with important agronomical traits, including disease resistance, seed quality, adaptability and yield (i.e., GWAS hits and QTLs in Soybase database) (Li et al. 2023), we examined the information regarding these trait-associated SNP loci with positive effect in Mengdou1137 and performed statistical analysis in conjunction with other sequenced cultivated soybean varieties ($n = 598$) (Supplementary Table S4) (Li et al. 2023). As expected, we found outstanding phenotypes in Mengdou1137, typically, disease resistance (ranking 71st, Fig. 4a), quality (ranking 89th, Fig. 4b), adaptability (ranking 212th, Fig. 4c) and yield (ranking 190th, Fig. 4d).

Non-synonymous mutation in *sss1* gene was essential for controlling soybean seed weight (Zhu et al. 2022). Here we discovered a Mengdou1137-prominent SNP ('G', chr19: 45,846,917) located in coding regions of *sss1* gene, which is significantly associated with the number of sections and pods per plant (Fig. 4e, f). In addition, a total of 17 non-coding insertions and deletions may also influence *sss1* (Fig. 4g). To further validate the potential breeding value of *sss1*, we conducted haplotype analysis for trait section number and number pods per plant using the 2-kb regions upstream and downstream of this gene. The results indicated the haplotypes representing superior trait values occupied a higher proportion in upstream ($n = 85$, $P < 2.2 \times 10^{-16}$ for section number, $P = 3.1 \times 10^{-16}$ for number pods of per plant, Wilcoxon rank-sum test) and downstream ($n = 91$, $P < 2.2 \times 10^{-16}$ for section number, $P = 2.7 \times 10^{-16}$ for number pods of per plant, Wilcoxon rank-sum test).

### The favorable haplotype of *PH13* in Mengdou1137

Utilizing the variation loci information identified in Wm82, we conducted a comparative analysis between Mengdou1137 and Wm82 to determine the presence of beneficial trait loci in Mengdou1137. Notably, we discovered that Mengdou1137 harbors the *PH13* gene, which is associated with the dwarf trait. A collinear analysis of the *PH13* gene between Wm82 and Mengdou1137 revealed an insertion of approximately 5.4-kb sequence at 52 bp into the fifth exon, aligning with findings from previous research (Qin et al. 2023).This insertion is likely to result in a truncated PH13 protein, comprised of 742 amino acids and missing the terminal part of the WD40 domain. The identified haplotype, named $PH13^{H3}$, was observed to significantly reduce plant height in $PH13^{H3}$ accessions when compared to those with alternate haplotypes. Given the agronomic significance of the dwarf trait, Mengdou1137 emerges as a valuable donor for enhancing soybean germplasm through the introduction of the dwarf gene.

## Discussion

Soybean, being one of the world's most economically significant crops, has seen a rapid development in genomic research in recent years. Specifically, many soybean cultivar genomes with extensive genetic diversity have been successfully assembled. For instance, Liu and colleagues constructed a pan-genome of 26 representative wild soybean and cultivated soybean varieties, offering a promising platform for soybean evolution and functional genomics research (Liu et al. 2020). Recent telomere-to-telomere (T2T) assemblies of soybean classic cultivars Wm82 and ZH13 have fully closed genomic gaps and enhanced the resolution of complex regions, significantly advancing our understanding of soybean genomic architecture and diversity (Wang et al. 2023; Zhang et al. 2024b).

However, the existing soybean genomic database falls short in effectively cataloging the dominant loci and genetic insights of the rapidly expanding range of emerging soybean varieties. The current soybean reference genome (Wm82) was derived from an American domesticated cultivar, potentially overlooking crucial genetic details specific to the majority of Chinese soybean varieties. While significant progress has been made with the publication of T2T soybean genomes, such as that of the Chinese variety ZH13 (Zhang et al. 2024a) and the availability of multiple high-quality reference assemblies covering portions of Chinese germplasm diversity (Liu et al. 2020), there remains a pressing need for comprehensive genomic resources that capture the genetic innovations of newly developed, widely cultivated varieties with superior yield, disease resistance and lodging resistance. Establishing such high-quality reference genomes will not only refine our understanding of soybean's genetic landscape but also enhance breeding strategies, ultimately driving the improvement of agronomically important traits.

This study presents a high-quality reference genome for the Chinese soybean cultivar Mengdou1137. Our assembled genome boasts an impressive contig N50 length of 14.92 Mb and a scaffold N50 length of 50.26 Mb, surpassing the contiguity of Wm82 (contig N50 = 419.3 kb and scaffold N50 = 20.4 Mb). Our comparison with other soybean genomes led to the identification of 115 novel disease resistance genes, which potentially contribute to Mengdou1137's specific resilience to adverse conditions. Further, we uncovered a large set of genomic variants related to important agronomic traits, such as yield, quality and resistance.

Mengdou1137 contains *PH13* gene which can lead to dwarf traits. Besides having excellent agronomic traits such as lodging resistance (Qin et al. 2023), Mengdou1137 can also be used as dwarf trait donor in the process of

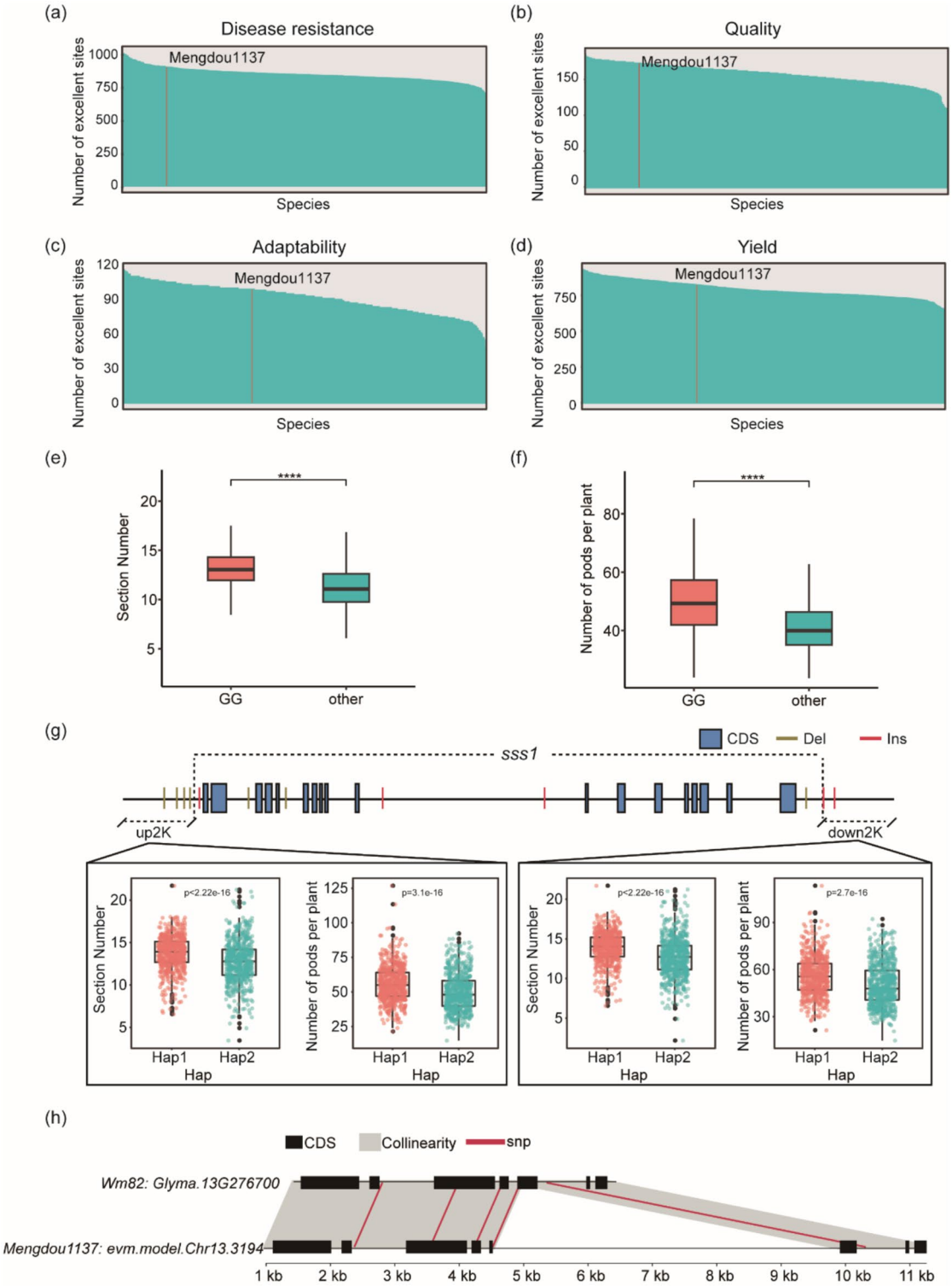


**Fig. 4** Excellent agronomic traits in Mengdou1137: **a**–**d** ranks in Mengdou1137 among 598 materials for SNP number related to disease resistance (**a**), quality (**b**), adaptability (**c**) and yield (**d**). **e**, **f** Comparison between 'GG' and other genotypes for SNP (chr19: 45,846,917) in coding regions of *sss1* for the number of sections (**e**) and pods per plant (**f**). **g** Haplotype analysis for 2-kb regions of upstream (left) and downstream (right) of *sss1*. 'Hap1' and 'Hap2' represent the superior phenotypic haplotypes. **h** Identification of *PH13* characteristics of Mengdou1137 with Wm82

soybean genetic improvement. This mutant exhibits a decrease in plant height and early maturity, making it suitable for planting at high latitudes. Soybean planting with high density tolerance has great potential for high yield in production, and density tolerance is a hot and difficult topic in soybean breeding (Yang et al. 2021). It is of great significance to cultivate new varieties of soybean with high density tolerance by utilizing the excavated excellent new germplasm.

In summary, this new genome represents a valuable source of theoretical knowledge for advancing soybean genomics research and enhancing yield.

## Conclusion

This research introduced the chromosome-level de novo assembly of Mengdou1137, a high-yield Chinese soybean variety. Through genome annotation and a comparative genomics analysis with 26 previously published soybean genomes, we uncovered Mengdou1137's potential as a donor of disease resistance genes for breeding purposes. Further, allele mining for SNPs associated with agronomic traits, along with a comparative analysis against resequenced soybean lines, underscored Mengdou1137's superior genetic qualities. This was exemplified by the discovery of functional variations in the *sss1* gene and favorable haplotypes in the *PH13* gene, further emphasizing Mengdou1137's excellence as a variety.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00122-025-04881-4.

**Author Contribution statement** LQ and YL designed the study; RS performed the experiments and wrote original draft; BS provided soybean accessions and performed the experiments; ZZ analyzed the data; QZ, XH, RG, LF and SC analyzed the data; JW, PQ, PY, YL and WS performed the experiments; and RS, LQ and YL interpreted the results and wrote the manuscript. All authors read and approved the final manuscript.

**Funding** This work was supported by the 'select the best candidates to undertake key research projects' of Inner Mongolia Autonomous Region (2023JBGS0006).

**Data availability** All sequencing data, genome sequences and annotation results have been uploaded to the National Genomics Data Center under accession number PRJCA033726 (https://ngdc.cncb.ac.cn/search/all?&q=PRJCA033726). The data are not yet public; here is the reviewer link: https://ngdc.cncb.ac.cn/gwh/Assembly/reviewersPage/qKDULIdbmuqfNJdYfZPkVlFTKTpjxjERKQBqUysWhWoKMJEHApygLhwzdotyVXgS. These data will be made publicly available upon acceptance of the manuscript.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.



## References

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol 20:224. https://doi.org/10.1186/s13059-019-1829-6

Birney E, Clamp M, Durbin R (2004) GeneWise and genomewise. Genome Res 14:988–995. https://doi.org/10.1101/gr.1865504

Böhm H, Albert I, Fan L, Reinhard A, Nürnberger T (2014) Immune receptor complexes at the plant cell surface. Curr Opin Plant Biol 20:47–54. https://doi.org/10.1016/j.pbi.2014.04.007

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94. https://doi.org/10.1006/jmbi.1997.0951

Chu JS-C, Peng B, Tang K, Yi X, Zhou H, Wang H, Li G, Leng J, Chen N, Feng X (2021) Eight soybean reference genome resources from varying latitudes and agronomic traits. Sci Data 8:164. https://doi.org/10.1038/s41597-021-00947-2

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30:2478–2483. https://doi.org/10.1093/nar/30.11.2478

Espina MJC, Lovell JT, Jenkins J, Shu S, Sreedasyam A, Jordan BD, Webber J, Boston L, Brůna T, Talag J, Goodstein D, Grimwood J, Stacey G, Cannon SB, Lorenz AJ, Schmutz J, Stupar RM (2024) Assembly, comparative analysis, and utilization of a single haplotype reference genome for soybean. Plant J 120:1221–1235. https://doi.org/10.1111/tpj.17026

Garg V, Khan AW, Fengler K, Llaca V, Yuan Y, Vuong TD, Harris C, Chan TF, Lam HM, Varshney RK, Nguyen HT (2023) Near-gapless genome assemblies of Williams 82 and Lee cultivars for accelerating global soybean research. Plant Genome 16:e20382. https://doi.org/10.1002/tpg2.20382

Goel M, Sun H, Jiao W-B, Schneeberger K (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 20:277. https://doi.org/10.1186/s13059-019-1911-0

Guigó R (1998) Assembling genes from predicted exons in linear time with dynamic programming. J Comput Biol 5:681–702. https://doi.org/10.1089/cmb.1998.5.681

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666. https://doi.org/10.1093/nar/gkg770

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to

assemble spliced alignments. Genome Biol 9:R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hu J, Fan J, Sun Z, Liu S (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36:2253–2255. https://doi.org/10.1093/bioinformatics/btz891

Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, Ruan J, Xiao C-L, Wang D, Wu D-D, Wang S (2024) NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol 25:107. https://doi.org/10.1186/s13059-024-03252-4

Huang Y, Koo D-H, Mao Y, Herman EM, Zhang J, Schmidt MA (2024) A complete reference genome for the soybean cv. Jack. Plant Commun 5:100765. https://doi.org/10.1016/j.xplc.2023.100765

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338–345. https://doi.org/10.1038/nbt.4060

Jurka J (2000) Repbase Update: a database and an electronic journal of repetitive elements. Trends Genet 16:418–420. https://doi.org/10.1016/s0168-9525(00)02093-x

Kim J, Lim CJ, Lee B-W, Choi J-P, Oh S-K, Ahmad R, Kwon S-Y, Ahn J, Hur C-G (2012) A genome-wide comparison of NB-LRR type of resistance gene analogs (RGA) in the plant kingdom. Mol Cells 33:385–392. https://doi.org/10.1007/s10059-012-0003-8

Kim M-S, Lee T, Baek J, Kim JH, Kim C, Jeong S-C (2021) Genome assembly of the popular Korean soybean cultivar Hwangkeum. G3 (Bethesda) 11:jkab272. https://doi.org/10.1093/g3journal/jkab272

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC (2023) Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods 20:408–417. https://doi.org/10.1038/s41592-022-01753-3

Korf I (2004) Gene finding in novel genomes. BMC Bioinform 5:59. https://doi.org/10.1186/1471-2105-5-59

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome bio 5:1–9. https://doi.org/10.1186/gb-2004-5-2-r12

Leamy LJ, Zhang H, Li C, Chen CY, Song B-H (2017) A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC Genomics 18:18. https://doi.org/10.1186/s12864-016-3397-4

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503

Li L, Yu Y, Zhou Z, Zhou J-M (2016a) Plant pattern-recognition receptors controlling innate immunity. Sci China Life Sci 59:878–888. https://doi.org/10.1007/s11427-016-0115-2

Li P, Quan X, Jia G, Xiao J, Cloutier S, You FM (2016b) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. BMC Genomics 17:852. https://doi.org/10.1186/s12864-016-3197-x

Li Y-H, Qin C, Wang L, Jiao C, Hong H, Tian Y, Li Y, Xing G, Wang J, Gu Y, Gao X, Li D, Li H, Liu Z, Jing X, Feng B, Zhao T, Guan R, Guo Y, Liu J, Yan Z, Zhang L, Ge T, Li X, Wang X, Qiu H, Zhang W, Luan X, Han Y, Han D, Chang R, Guo Y, Reif JC, Jackson SA, Liu B, Tian S, Qiu L-J (2023) Genome-wide signatures of the geographic expansion and breeding of soybean. Sci China Life Sci 66:350–365. https://doi.org/10.1007/s11427-022-2158-7

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020) Pan-genome of wild and cultivated soybeans. Cell 182:162-176.e113. https://doi.org/10.1016/j.cell.2020.05.023

Qin C, Li Y-H, Li D, Zhang X, Kong L, Zhou Y, Lyu X, Ji R, Wei X, Cheng Q, Jia Z, Li X, Wang Q, Wang Y, Huang W, Yang C, Liu L, Wang X, Xing G, Hu G, Shan Z, Wang R, Li H, Li H, Zhao T, Liu J, Lu Y, Hu X, Kong F, Qiu L-J, Liu B (2023) PH13 improves soybean shade traits and enhances yield for high-density planting at high latitudes. Nat Commun 14:6813. https://doi.org/10.1038/s41467-023-42608-5

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. https://doi.org/10.1038/nature08670

Shen Y, Liu J, Geng H, Zhang J, Liu Y, Zhang H, Xing S, Du J, Ma S, Tian Z (2018) De novo assembly of a Chinese soybean genome. Sci China Life Sci 61:871–884. https://doi.org/10.1007/s11427-018-9360-0

Shen Y, Du H, Liu Y, Ni L, Wang Z, Liang C, Tian Z (2019) Update soybean Zhonghuang 13 genome to a golden reference. Sci China Life Sci 62:1257–1260. https://doi.org/10.1007/s11427-019-9822-2

Shimomura M, Kanamori H, Komatsu S, Namiki N, Mukai Y, Kurita K, Kamatsuki K, Ikawa H, Yano R, Ishimoto M, Kaga A, Katayose Y (2015) The glycine max cv. enrei genome for improvement of Japanese soybean cultivars. Int J Genomics. https://doi.org/10.1155/2015/358127

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24:637–644. https://doi.org/10.1093/bioinformatics/btn013

Torkamaneh D, Lemay M-A, Belzile F (2021) The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. Plant Biotechnol J 19:1852–1862. https://doi.org/10.1111/pbi.13600

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621

Valliyodan B, Cannon SB, Bayer PE, Shu S, Brown AV, Ren L, Jenkins J, Chung CYL, Chan TF, Daum CG, Plott C, Hastie A, Baruch K, Barry KW, Huang W, Patil G, Varshney RK, Hu H, Batley J, Yuan Y, Song Q, Stupar RM, Goodstein DM, Stacey G, Lam HM, Jackson SA, Schmutz J, Grimwood J, Edwards D, Nguyen HT (2019) Construction and comparison of three reference-quality genome assemblies for soybean. Plant J 100:1066–1082. https://doi.org/10.1111/tpj.14500

Wang L, Zhang M, Li M, Jiang X, Jiao W, Song Q (2023) A telomere-to-telomere gap-free assembly of soybean genome. Mol Plant 16:1711–1714. https://doi.org/10.1016/j.molp.2023.08.012

Xie M, Chung CY-L, Li M-W, Wong F-L, Wang X, Liu A, Wang Z, Leung AK-Y, Wong T-H, Tong S-W, Xiao Z, Fan K, Ng M-S, Qi X, Yang L, Deng T, He L, Chen L, Fu A, Ding Q, He J, Chung G, Isobe S, Tanabata T, Valliyodan B, Nguyen HT, Cannon SB, Foyer CH, Chan T-F, Lam H-M (2019) A reference-grade wild soybean genome. Nat Commun 10:1216. https://doi.org/10.1038/s41467-019-09142-9

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268. https://doi.org/10.1093/nar/gkm286

Yang Q, Lin G, Lv H, Wang C, Yang Y, Liao H (2021) Environmental and genetic regulation of plant height in soybean. BMC Plant Biol 21:63. https://doi.org/10.1186/s12870-021-02836-7

Yi X, Liu J, Chen S, Wu H, Liu M, Xu Q, Lei L, Lee S, Zhang B, Kudrna D, Fan W, Wing RA, Wang X, Zhang M, Zhang J, Yang C, Chen N (2022) Genome assembly of the JD17 soybean provides a new reference genome for comparative genomics. G3. https://doi.org/10.1093/g3journal/jkac017

Zhang M, Liu S, Wang Z, Yuan Y, Zhang Z, Liang Q, Yang X, Duan Z, Liu Y, Kong F, Liu B, Ren B, Tian Z (2022) Progress in soybean functional genomics over the past decade. Plant Biotechnol J 20:256–282. https://doi.org/10.1111/pbi.13682

Zhang A, Kong T, Sun B, Qiu S, Guo J, Ruan S, Guo Y, Guo J, Zhang Z, Liu Y, Hu Z, Jiang T, Liu Y, Cao S, Sun S, Wu T, Hong H, Jiang B, Yang M, Yao X, Hu Y, Liu B, Han T, Wang Y (2024a) A telomere-to-telomere genome assembly of Zhonghuang 13, a widely-grown soybean variety from the original center of Glycine max. Crop J 12:142–153. https://doi.org/10.1016/j.cj.2023.10.003

Zhang C, Shao Z, Kong Y, Du H, Li W, Yang Z, Li X, Ke H, Sun Z, Shao J, Chen S, Zhang H, Chu J, Xing X, Tian R, Qin N, Li J, Huang M, Sun Y, Huo X, Meng C, Wang G, Liu Y, Ma Z, Tian S, Li X (2024b) High-quality genome of a modern soybean cultivar and resequencing of 547 accessions provide insights into the role of structural variation. Nat Genet 56:2247–2258. https://doi.org/10.1038/s41588-024-01901-9

Zhu W, Yang C, Yong B, Wang Y, Li B, Gu Y, Wei S, An Z, Sun W, Qiu L, He C (2022) An enhancing effect attributed to a nonsynonymous mutation in *SOYBEAN SEED SIZE 1*, a *SPINDLY*-like gene, is exploited in soybean domestication and improvement. New Phytol 236:1375–1392. https://doi.org/10.1111/nph.18461