Journal of Genetics and Genomics 53 (2026) 458–466

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Original Research

Assembly of a high-quality reference genome and characterization of a chemical-mutagenized library of an elite soybean cultivar Tianlong 1

Yinghua Sheng [a, b, 1], Yicheng Huang [c, 1, †], Zilun Jin [b, 1], Xuyan Wang [b], Chenghui Liu [e], Jingwen Zhang [b], Zhipeng Zhou [d], Chuang Ma [e], Jianwei Zhang [c, *], Min Chen [a, b, *]

[a] Sanya Institute of Henan University, Sanya, Hainan 572025, China
[b] State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, Henan University, Kaifeng, Henan 475004, China
[c] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei, 430070, China
[d] College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430074, China
[e] State Key Laboratory of Crop Stress Resistance and High-Efficiency Production, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Yangling, Shaanxi 712100, China

ARTICLE INFO

Article history:
Received 1 June 2025
Received in revised form
12 August 2025
Accepted 13 August 2025
Available online 21 August 2025

Keywords:
Soybean
EMS mutant library
Tianlong 1
SNPs
InDels
High-quality genome

ABSTRACT

Soybean (*Glycine max* L.) is a globally vital crop for oil production and food security. High-quality genomic resources are instrumental for both functional genomics and breeding. Here, we report a near-complete, high-quality genome assembly of the elite cultivar Tianlong 1 (TL1), featuring fully resolved telomeres and centromeres, as well as a gap-free assembly of 14 of its 20 chromosomes. On the basis of the genome assembly, we generate an ethyl methanesulfonate (EMS)-mutagenized population comprising 2555 M_7 plants. Whole-genome resequencing of 288 EMS mutants uncovers 1,163,869 high-confidence single-nucleotide polymorphisms (SNPs) and 542,709 insertions/deletions (InDels), achieving 91.89% coverage of predicted protein-coding genes. Phenotypic screening demonstrates robust genotype–phenotype associations, with two nonsynonymous mutants displaying pronounced defects in seed and leaf development. Collectively, the chromosome-scale TL1 genome assembly and the extensively characterized mutant population establish valuable resources for functional genomics and precision breeding in soybean and related legume species.

Introduction

Soybean has become an indispensable global crop, serving as a critical source of plant-based protein, edible oil, and industrial raw materials to meet the demands of a growing population (Tian et al., 2025). To address the pressing need for increased soybean production, genetic improvement of yield-related traits through targeted breeding strategies is essential. While recent advances in genomics have enabled the identification of key genes regulating important agronomic traits, such as seed composition, stress tolerance, and yield potential (Goettel et al., 2022; Liang et al., 2024; Yuan et al., 2024; Zhong et al., 2024; Zhang et al., 2025), the functional characterization of soybean genes remains largely incomplete. A systematic approach to uncovering gene-trait relationships is therefore crucial for accelerating soybean breeding.

Forward genetics has proven highly effective in linking genotypes to phenotypes, particularly when combined with chemical mutagenesis and high-throughput sequencing. EMS-induced mutagenesis, in particular, offers a powerful means to generate high-density mutations for functional gene discovery. This approach has been successfully implemented in major crops, including rice (Takagi et al., 2015), maize (Zhou et al., 2023), watermelon (Deng et al., 2022), and Chinese cabbage (Sun et al., 2022), where EMS libraries have enabled rapid gene identification through whole-genome

* Corresponding authors.
E-mail addresses: jzhang@mail.hzau.edu.cn (J. Zhang), chenmin@henu.edu.cn (M. Chen).
† Current address: Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China.
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jgg.2025.08.006

sequencing. In soybean, EMS-mutagenized populations have been developed in various genetic backgrounds, such as Williams 82 (Wm82, Zhang et al., 2022), Zhonghuang 13 (ZH13, Ge et al., 2018), and Zhongpin 661 (Li et al., 2017), facilitating the discovery of genes controlling diverse traits. However, the lack of a high-quality reference genome for many elite cultivars limits the precision of mutation mapping and functional studies.

To enable accurate mutation detection, we generated a near-complete, high-quality genome assembly of Tianlong 1 (TL1), a high-yielding soybean cultivar extensively cultivated in Huang-Huai-Hai region of China. We then performed whole-genome resequencing on 288 M_7 mutants, identifying a comprehensive set of SNPs and InDels with high genome-wide coverage. Phenotypic screening revealed strong genotype-phenotype correlations, highlighting the utility of this resource for functional genomics. Our study provides a robust platform for gene discovery and molecular breeding in soybean, offering opportunities to elucidate the genetic basis of agronomic traits.

Results

Chromosome-scale assembly of the TL1 high-quality genome

Through PacBio long-read sequencing combined with high-fidelity (HiFi) sequence polishing (Figs. 1A, 1B, S1), we generated a high-quality genome assembly for the elite soybean cultivar TL1 (TL1-HQ). The final 1017 Mb assembly represents a 0.4% reduction from the previous TL1 genome (TL1-PRE, 1021 Mb; NCBI: GCA_015227745.1) (Jia et al., 2020), primarily due to improved gap closure and error correction (Fig. 1C and 1D). The assembly exhibits exceptional contiguity, with only 26 contigs (N50 = 50.3 Mb) and 6 residual gaps localized to chromosomes 2, 11, 14, 15, 18, and 20 (Fig. S2; Table S1).

A hallmark of this assembly is its near-complete chromosomal reconstruction, with all 40 telomeres and 20 centromeres successfully resolved (Fig. 1D; Table S1). Rigorous quality assessments including a Genome Continuity Inspector (GCI) score of 40.93, an R-AQI of 96.89, and a S-AQI of 97.67 (Table S1), collectively confirmed the high contiguity and precision of the assembly. Transposable element (TE) analysis identified a genomic content of 35.56%, primarily composed of retroelements (31.32%) and DNA transposons (4.21%). This TE content was significantly lower than that reported in other soybean cultivars, including Zhonghuang 13 (53.06%) (Zhang et al., 2023) and Williams 82 (Wm82-NJAU, 55.73%) (Wang et al., 2023b), but comparable to the Jack cultivar (35.67%) (Huang et al., 2024), suggesting cultivar-specific TE dynamics in soybean genomes (Table S2).

Gene annotation and comparative genomics reveal structural improvements

Comparative genomic analysis revealed substantial conservation with existing soybean references, with 78.4% (60,403 genes) and 83.3% (64,150 genes) showing one-to-one orthology to the Wm82 a4.v1 (Valliyodan et al., 2019) and a6.v1 (Espina et al., 2024) assemblies, respectively (Table S3). The completeness of the assembly was further validated by BUSCO analysis, which identified 99.89% of the conserved *Fabales* orthologs, demonstrating nearly complete representation of the gene space. Comparative genomics revealed substantial improvements over TL1-PRE, including 130 previously unknown genomic regions (>1 Kb) and corrections to 45 translocations and 213 inversions (Figs. 1D and S3; Table S4).

Whole-genome alignments further highlighted extensive structural divergence between TL1-HQ and existing soybean references, including 517 translocations and 134 inversions relative to Wm82-NJAU (Figs. 1D, S3, S4), and 319 translocations and 71 inversions compared to ZH13 (Fig. S5). These findings underscore both the technical superiority of TL1-HQ and the natural genomic diversity among soybean cultivars. With its complete chromosomal architecture and accurate annotation, the TL1-HQ genome provides a gold-standard reference for both intra-species comparisons among soybean cultivars and inter-species legume genomics studies.

Construction and phenotypic characterization of the TL1 EMS mutant population

For EMS mutant library construction, the TL1 genetic background was strategically selected for mutagenesis due to its dual value for soybean improvement: (1) direct applicability for developing breeding-ready improved variants, and (2) long-term utility for forward genetic studies of agronomic trait loci. We established an EMS-mutagenized population by treating TL1 seeds with 0.6% EMS, achieving an optimal M_0 germination rate of ~50% that balances mutation density with population viability. Through 7 generations of field cultivation across multiple locations, we developed a stable mutant resource including 7300 M_2 plants in Wuhan (2018), 6000 M_3 and 5000 M_4 plants in Sanya and Wuhan (2019), 3500 M_5 and 3452 M_6 plants in Sanya and Kaifeng (2020), culminating in 2555 genetically stable M_7 lines from Kaifeng (2021) (Fig. 2A).

Phenotypic characterization of the M_4 population revealed extensive morphological variation throughout development. Vegetative stage mutants displayed altered leaf architecture (including trifoliate-to-quadra/quinquefoliate transitions), modified plant stature, and diverse branching patterns (Fig. 2B and 2C). Mature plants exhibited significant diversity in seed characteristics, including size, color, and seed coat morphology (Fig. 2D). This mutant collection represents a comprehensive genetic resource that captures substantial diversity within the TL1 background, which serves as a functional genomics platform for gene discovery, providing important alleles for direct breeding applications. The population's combination of genetic stability (M_7 generation) and phenotypic breadth makes it particularly valuable for dissecting the genetic architecture of complex agronomic traits.

Genome-wide characterization of 288 M_7 mutants through whole-genome resequencing

Whole-genome resequencing of mutant populations serves as a powerful tool for functional genomics, enabling precise genotype-phenotype correlations. To comprehensively characterize the mutational landscape of our population, we deep-sequenced 288 M_7 individuals (representing both visible mutants and phenotypically normal lines) selected from the 2555 mutant lines (Table S5). This generated 5513 Gb of high-quality sequencing data, achieving an average coverage depth of 14.80× (19.14 Gb per line).

Variant calling identified 1,163,869 high-confidence non-redundant SNPs and 542,709 InDels (Table S8), corresponding to a mutation density of 1.68 variants per kilobase (mutations/Kb). This mutation frequency is significantly lower than those reported in EMS-mutagenized soybean populations of Wm82 (6.7 mutations/Kb; Zhang et al., 2022) and Zhongpin 661 (11.8 mutations/Kb; Li et al., 2017), a difference likely attributable to generational selection effects. Whereas prior studies examined earlier generations (M_2 and M_4/M_5, respectively), our advanced M_7 population underwent rigorous homology-based selection, which not only reduced mutational burden but also enhanced genetic stability.

Genetic analysis revealed that 82.81% of SNPs and 85.16% of InDels were fixed in the homozygous state (Fig. 3A), with individual mutants harboring an average of 12,821 SNPs and 4820 InDels (Fig. 3B). Mutations were uniformly distributed across all 20 chromosomes (Fig. 3C), confirming genome-wide coverage of our mutant

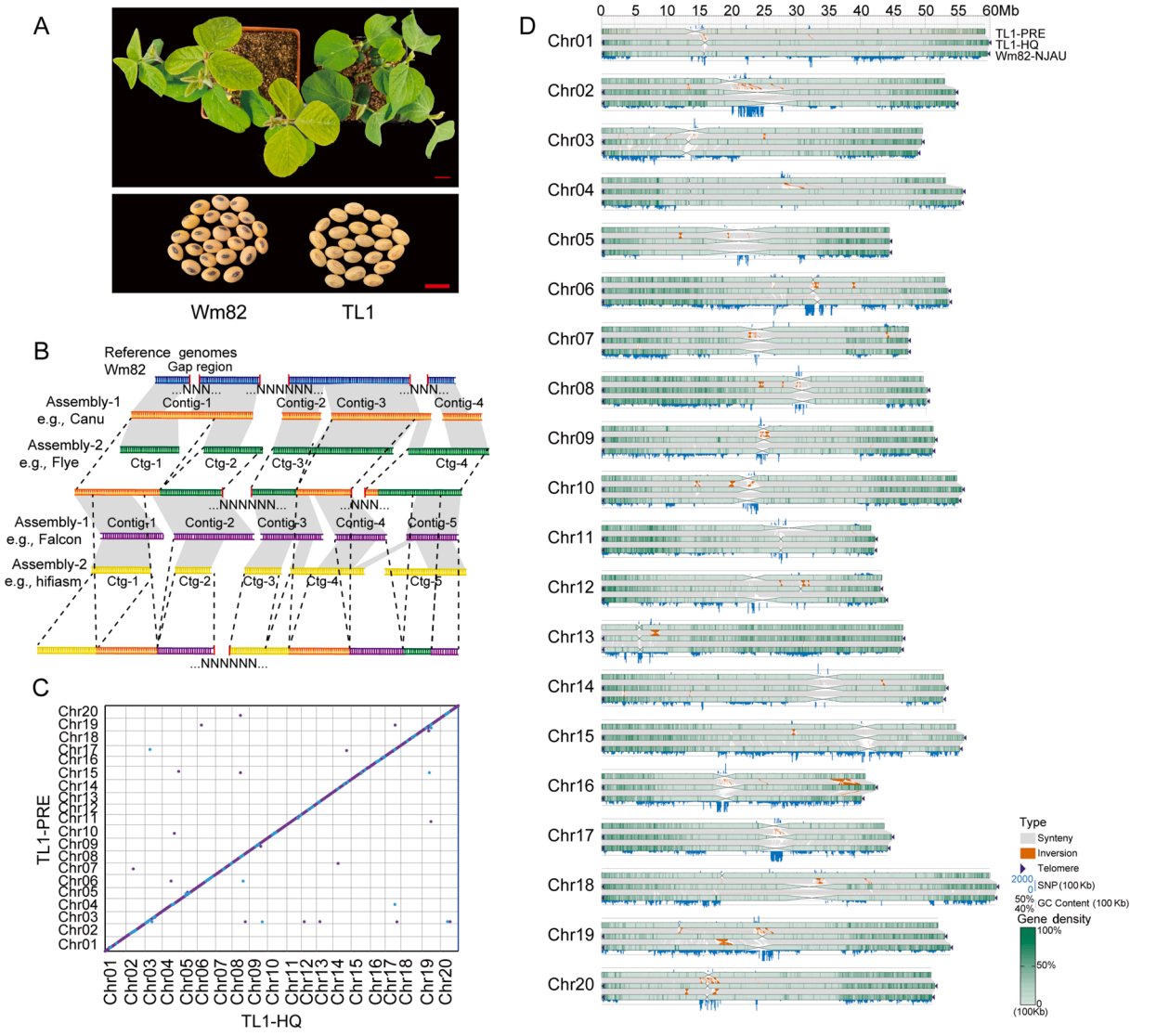

Fig. 1. Comparative genomic analysis of TL1 assemblies and the Wm82-NJAU reference genome. **A**: Phenotypic and seed morphology comparison between Wm82 and TL1 accessions. Top panel: representative images of 14-day-old seedlings (Wm82 vs. TL1) showing distinct growth phenotypes. Bottom panel: morphological differences in mature seeds (Wm82 vs. TL1), highlighting variations in seed size, shape, and coat color. **B**: Pipeline for the high-quality TL1-HQ genome assembly (modified from Huang et al., 2024). Top tier: initial assembly-1 and assembly-2 generated through GPM assisted scaffolding. Bottom tier: refined assembly-1 and assembly-2 produced by iterative GPM editing of all CCS reads within the assembly framework. **C**: Whole-genome synteny analysis between the previously published TL1-PRE (Jia et al., 2020) and the assembled TL1-HQ genomes. The dot plot displays homologous gene pairs across all 20 chromosomes (Chr1–Chr20), with purple points indicating homologous and continuous collinear regions and blue points highlighting major structural variations. Conserved syntenic blocks appear as diagonal patterns, where gene pairs maintain consistent genomic positions. **D**: Integrated visualization of genomic architecture across TL1-PRE, TL1-HQ, and Wm82-NJAU assemblies. Condensed central regions denote centromeric repeats. Black triangles indicate telomeric regions. Gray lines connect syntenic regions across genomes. Orange lines highlight inversion events. Blue density plots show SNP distribution (threshold line = 2000 SNPs per 100 Kb window). Scale bars, 2 cm (**A**, top panel); 1 cm (**A**, bottom panel). Wm82, Williams 82; TL1, Tianlong 1; GPM, Genome Puzzle Master; CCS, circular consensus sequencing; TL1-PRE, previously published Tianlong 1 genome; TL1-HQ, high-quality Tianlong 1 genome; Wm82-NJAU, telomere-to-telomere (T2T) assembly of soybean cultivar Wm82 from Wang et al. (2023b).

resource. Notably, the mutation spectrum exhibited a strong bias toward C:G>T:A transitions, accounting for 78.75% of all mutations (Fig. 3C), consistent with the canonical EMS mutational signature (Greene et al., 2003).

To validate variant calling accuracy, we performed Sanger sequencing on 22 exonic SNPs, including 19 G:C>A:T transitions and 3 non-canonical variants. This confirmed 95.5% overall accuracy, with only one false positive among non-G:C>A:T mutations (Tables S6 and S7). The high validation rate underscores the reliability of our dataset for downstream functional genomics applications.

Functional variants in coding and regulatory regions across mutant lines

Genome-wide analysis revealed extensive functional variation across the mutant population. Alignment to the TL1-HQ reference genome identified 414,544 SNPs (33.65%) and 338,018 InDels (56.33%) within gene regions (Table S8), affecting 91.89% of annotated transcripts (70,776/77,021). Functional annotation indicated that 44,697 SNPs (3.63%) resulted in missense mutations (amino acid substitutions), while 2209 (0.18%) introduced premature stop codons. Among InDels, 6083 (1.01%) led to amino acid changes, and 9132 (1.52%) caused frameshift mutations, which are expected to

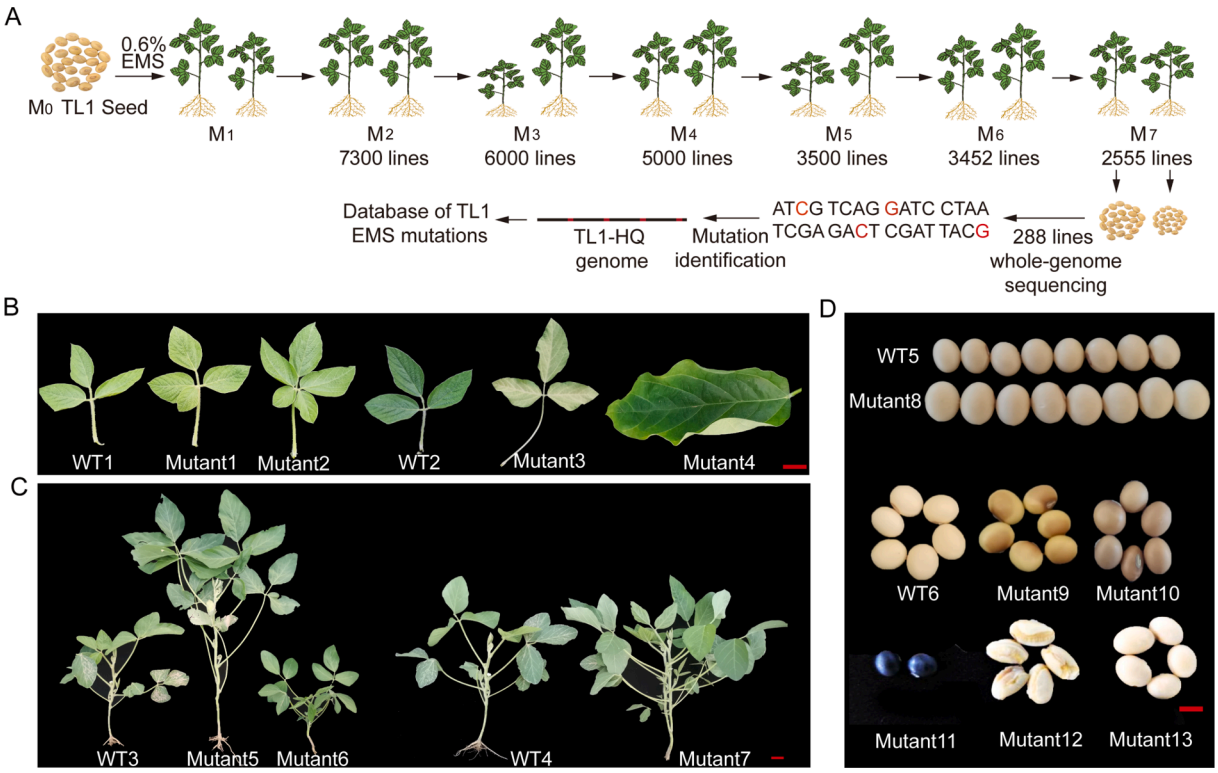

Fig. 2. Development and phenotypic characterization of the EMS-mutagenized TL1 population. **A**: Workflow for EMS mutagenesis and of the TL1 mutant library. For mutagenesis, ~5 kg of healthy M_0 seeds were treated with 0.6% EMS solution, and mutagenized populations were propagated to M_7 via SSD. For genomic analysis, 288 M_7 lines with WT control were whole-genome resequenced and the reads were aligned to the TL1-HQ reference genome for variant calling. Red bases highlight induced mutations. **B**–**D**: Phenotypic screening of M_4 generation mutants under field conditions. Leaf morphology variants: representative mutants showing altered leaf symmetry, color or size compared with WT (**B**). Plant architecture mutants: lines exhibiting aberrant growth patterns including height alterations (gigas phenotype or dwarfism) or branching defects (**C**). Seed phenotypic variations: mutants displaying significant alterations in seed size, color or coat texture relative to WT (**D**). Scale bars, 2 cm (**B** and **C**); 1 cm (**D**). TL1, Tianlong 1; SSD, single-seed descent; WT, wild-type; TL1-HQ, high-quality Tianlong 1 genome.

directly disrupt protein function (Tables S9–S11). Collectively, these mutations represent substantial functionally impactful variation within the population. Beyond coding regions, we detected 120,139 (9.75%) SNPs and 118,483 (19.75%) InDels in promoter regions (1 Kb upstream of transcription start sites; Figs. 3D and S5; Table S8), suggesting widespread potential for modulating transcriptional regulation and gene expression.

Further genomic analysis revealed an average of 521 genes carrying functionally impactful SNPs per line (range: 285–963; Fig. 3E), including nonsynonymous and premature stop codon mutations. Among these, genes with nonsynonymous variants predominated, averaging 510 per line. At the gene level, we detected an average of 1.81 functional SNPs per transcript (Fig. 3F). The high mutational density, spanning both coding and regulatory regions, ensures broad genomic coverage, making this population a valuable resource for functional genomics and gene discovery. Moreover, the abundance of protein-altering mutations, combined with extensive inter-line heterogeneity, enables precise genotype-phenotype association mapping and systematic investigation of gene function in soybean.

Protein-level analysis identified several high-frequency amino acid substitutions that significantly exceeded the background variation rate (average 0.32%), including alanine to threonine (A>T; 5.73%), alanine to valine (A>V; 5.58 %) and glutamic acid to lysine (E>K; 4.28%) (Fig. 3G; Table S9). Notably, the E>K substitution, resulting from G>A transitions, was among the most prevalent changes (Fig. 3G; Table S9). Given its charge-reversing effect (negative to positive), this substitution may exert substantial functional consequences on affected proteins, potentially altering enzymatic activity, protein-protein interactions, or subcellular localization.

Functional characterization of selected mutants

Seed coat color and leaf coloration represent important agronomic traits in soybean, serving as visible indicators of underlying physiological and metabolic processes. Seed coat coloration reflects the activity of the flavonoid biosynthesis pathway (Mach, 2017), while leaf coloration is closely associated with light signaling and senescence regulation (Fang et al., 2014). To explore the genetic basis of these traits, we selected two phenotypically distinct variants from our EMS-induced TL1 population for detailed genetic characterization: a seed coat color mutant *TM-29897* and a leaf color mutant *TM-30017*.

The *TM-29897* mutant displayed a distinctive color alteration specifically in the seed saddle region (Fig. 4A). Whole-genome sequencing revealed 426 transcripts carrying homozygous functional mutations in this line, including 228 transcripts with nonsynonymous substitutions (Table S12). Among these, we identified *GmAGO5* (*ARGONAUTE 5*) as a prime candidate, given its established role in regulating siRNA-mediated targeting of *Chalcone Synthase* (*CHS*) genes, which encode key enzymes in flavonoid biosynthesis and seed coat pigmentation (Mach, 2017). Targeted Sanger sequencing confirmed a causal SNP (Chr11:16471327, C>T) in *GmAGO5*, resulting in an aspartate-to-asparagine (D297N) substitution within the functionally critical PIWI domain (Fig. 4B and 4C). The perfect genotype-phenotype correlation establishes *GmAGO5* as an important regulator of soybean seed pigmentation and demonstrate the utility of EMS mutagenesis for dissecting complex agronomic traits.

Similarly, whole-genome sequencing of the *TM-30017* stay-green mutant identified 481 transcripts with homozygous mutations,

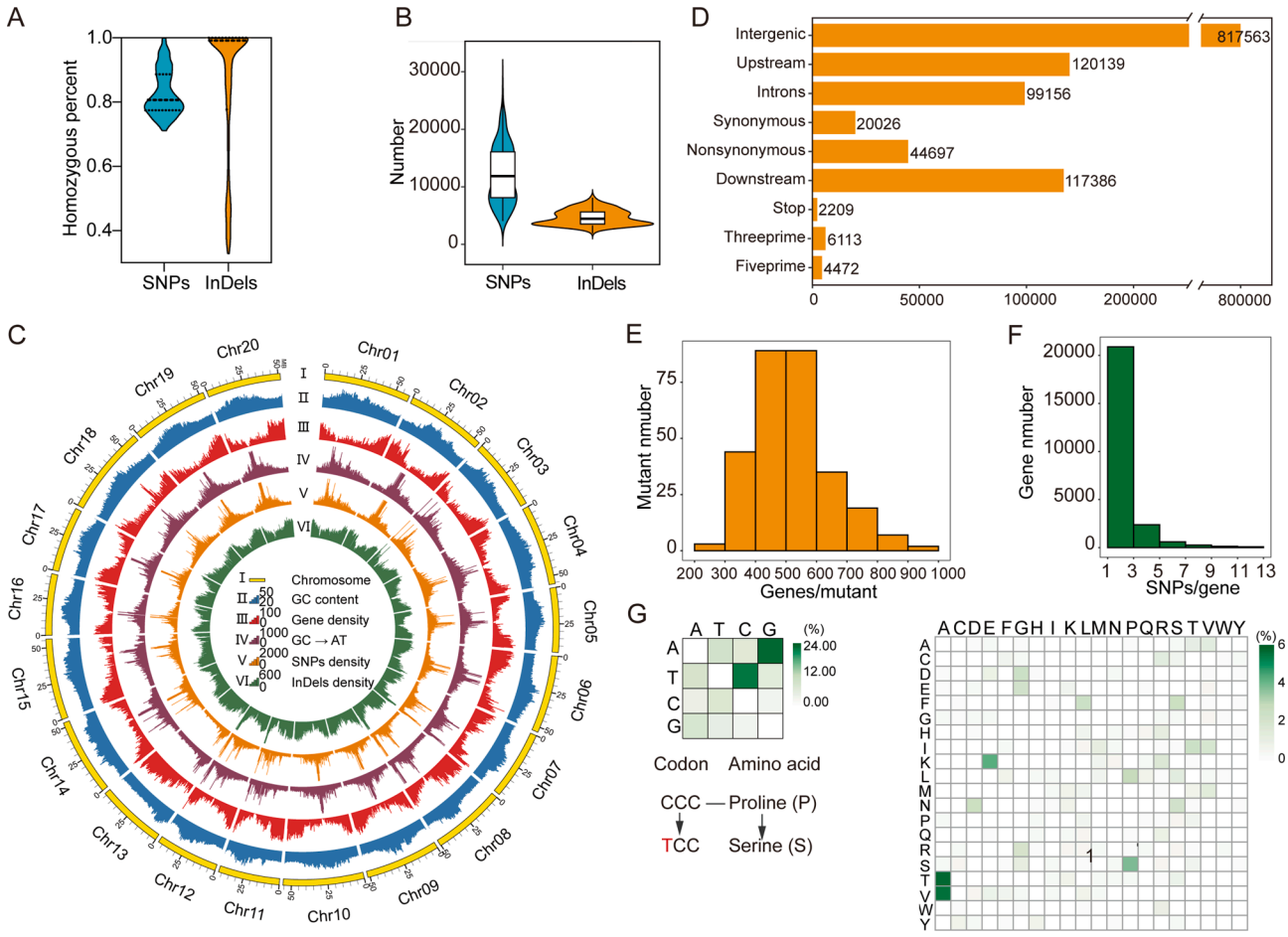

Fig. 3. Genome-wide characterization of EMS-induced mutations in the TL1 population. **A**: Violin plot displaying the percentage of homozygous SNPs and InDels across M_7 mutant lines, demonstrating the effectiveness of single-seed descent in fixation of mutations. **B**: Paired violin plots showing the density of total SNPs (left, median = 12,821 variants/line) and InDels (right, median = 4820 variants/line) across 288 mutant lines, revealing the spectrum of mutagenesis intensity. **C**: Genomic mutation landscape. Six integrated annotation tracks are: (I) Chromosomal positions (scale in Mb), (II) GC content variation (sliding window analysis), (III) Gene density profile (genes/Mb), (IV) GC>AT transition frequency (characteristic EMS signature), (V) SNP density (variants/Mb), and (VI) InDel distribution (indels/Mb). **D**: Histogram chart quantifying 44,697 non-synonymous SNPs and 120,139 regulatory SNPs in promoter regions, with additional categories including splice-site and UTR variants. **E**: Histogram of genes containing functional SNPs per line (mean = 521 genes, range: 285–963), demonstrating variable genetic impact across mutants. **F**: Histogram showing SNP frequency per gene. **G**: Molecular consequence analysis. Heatmap on the left: nucleotide substitution patterns (transition/transversion bias) with EMS-characteristic GC>AT changes highlighted in red and with green color intensity reflecting substitution frequency. Heatmap on the right: resultant amino acid substitutions, with color intensity reflecting substitution frequency. TL1, Tianlong 1.

including 343 nonsynonymous substitutions (Table S13). This mutant exhibited significantly delayed leaf senescence, retaining chlorophyll longer than wild-type plants under both normal and dark-induced conditions (Fig. 4D). We identified a key missense mutation (Chr11:1966372, G>A) in *GmSGR1* (*STAY-GREEN 1*), resulting in an alanine-to-threonine (A125T) substitution (Fig. 4E). As *GmSGR1* encodes a senescence-inducible chloroplast protein that regulates chlorophyll degradation (Fang et al., 2014), this mutation likely impairs normal protein function, leading to prolonged chlorophyll retention. Together with the *TM-29897* seed coat mutant, these results demonstrate the power of our EMS mutant library combined with whole-genome sequencing for identifying agronomically important alleles and validating gene-phenotype relationships, which are vital for dissecting complex agronomic traits in soybean.

Discussion

The extensive genomic rearrangements accumulated throughout soybean breeding history, including numerous mutations, insertions, and deletions, have posed significant challenges for functional genomics research and the genetic improvement of modern cultivars. To address these challenges, we focused on TL1, an elite soybean cultivar derived from a cross between Zhongdou 32 and Zhongdou 29. Widely cultivated in China's Yangtze River Basin due to its superior disease resistance and high yield potential, TL1 serves as an ideal model for studying complex agronomic traits (Guo et al., 2015; Jia et al., 2020). To overcome the lack of a high-quality reference genome for TL1, we performed de novo assembly of PacBio long-read sequencing data, generating an improved TL1-HQ reference genome. This assembly achieves 99.89% BUSCO completeness and resolves numerous structural variants compared to the previously published TL1 genome (Jia et al., 2020; Fig. 1), significantly enhancing mutation detection accuracy. Leveraging this genomic resource, we developed an EMS-induced mutant population (M_4 generation) exhibiting broad phenotypic variation, including alterations in seed size, leaf shape, and plant height (Fig. 2). Whole-genome re-sequencing of 288 mutants identified 1,163,869 high-confidence SNPs, including 44,697 nonsynonymous variants affecting protein-coding regions (Fig. 3; Table S8). This dataset provides a rich foundation for genotype-phenotype association studies. Together, these resources establish a powerful platform for advancing soybean functional genomics, enabling rapid identification of significant alleles linked to key agronomic traits, as demonstrated in our case studies (Fig. 4). Notably, beyond EMS-induced DNA sequence variation,

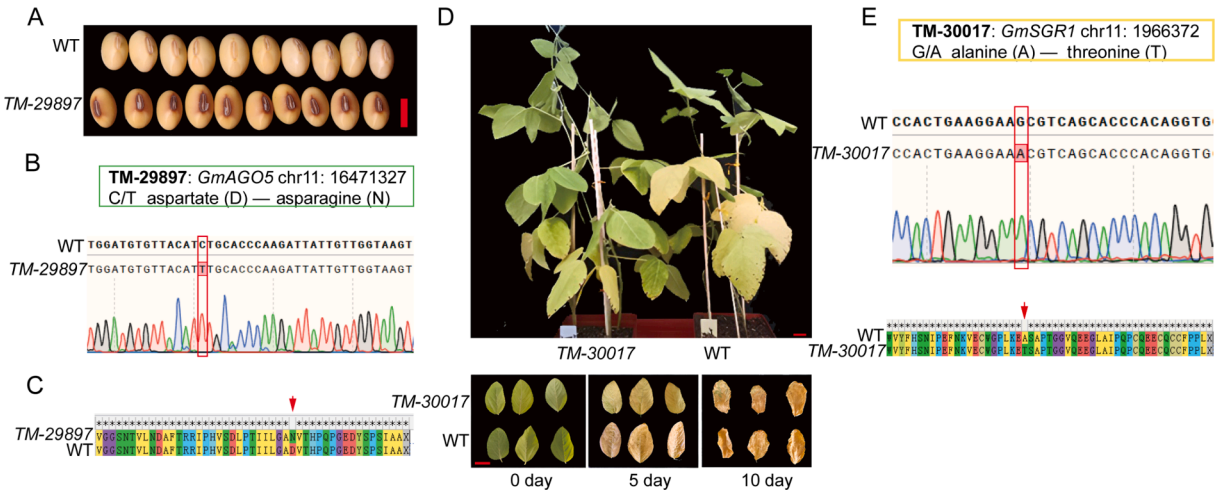

Fig. 4. Phenotypic characterization and causal mutation identification in selected TL1 mutants. **A**: Seed phenotype comparison between the *TM-29897* mutant and WT, showing distinct color variation in the saddle region. **B**: Molecular characterization of *GmAGO5* mutation. Top, schematic view of Chr11:16471327 (C/T) SNP in *GmAGO5*. Bottom, Sanger sequencing chromatograms (mutant vs. WT) with mutation site boxed in red. **C**: Predicted structural impact of the *GmAGO5* mutation, with an Aspartate (D)-to-Asparagine (N) substitution at the critical residue (red arrow). **D**: Dark-induced senescence assay comparing *TM-30017* mutant and WT leaves at 5-day intervals (0 days = treatment initiation). The mutant exhibits delayed chlorophyll degradation. **E**: Characterization of the *GmSGR1* mutation. Top: schematic view of the SNP (Chr11:1966372, G/A). Middle: Sanger sequencing validation (red boxes highlight the mutation site). Bottom: resulting Alanine (A)-to-Threonine (T) substitution in the GmSGR1 protein (red arrow indicates the altered residue). Scale bar, 1 cm (**A** and **D**). TL1, Tianlong 1; WT, wild-type.

chemical mutagenesis may concurrently trigger epigenetic modifications, such as changes in DNA methylation (Wang et al., 2023a). These epigenetic effects could be explored using whole-genome bisulfite sequencing (WGBS-seq) or assay for transposase-accessible chromatin sequencing (ATAC-seq), providing additional layers of regulatory insight.

From an applied perspective, the EMS-induced missense mutations in our collection offer exceptional opportunities for systematic screening of climate resilience traits, including drought and heat tolerance, as well as performance under nutrient-limited conditions (e.g., low nitrogen or phosphorus) (Zahra et al., 2022). These mutations serve as ideal targets for CRISPR-Cas9-mediated precision editing, while comparative analysis between natural and induced variants may further optimize gene-editing strategies. Additionally, this mutant library provides a unique resource for evaluating sgRNA off-target effects in genome editing experiments. Looking forward, this comprehensive mutant collection represents an ideal dataset for developing machine learning frameworks to predict genotype–phenotype–environment interactions. Establishing a global sharing platform for these resources could significantly accelerate international soybean breeding efforts through collaborative networks.

In conclusion, our integrated genomic and mutagenesis platform provides a powerful foundation for dissecting complex traits and accelerating functional gene discovery in soybean. This resource holds tremendous potential for advancing both conventional breeding and modern biotechnological approaches to soybean improvement, with possible applications extending to other legume crops.

Materials and methods

De novo assembly of the TL1-HQ genome

Genomic DNA was extracted from fresh young leaves of TL1 plants using the cetyltrimethylammonium bromide (CTAB) method (Allen et al., 2006). After quality control (1% agarose gel electrophoresis for degradation/contamination assessment and Qubit fluorometer quantification), high-molecular-weight DNA (50 μg) was used to prepare SMRTbell libraries with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) following established protocols (Pendleton et al., 2015). Libraries were sequenced on the PacBio Sequel II platform to generate 50× coverage of HiFi reads.

For genome assembly, 58.34 Gb of HiFi data were processed using 6 complementary assembly algorithms, including Canu v2.0 (Koren et al., 2017), FALCON toolkit v1.3.0, MECAT2 (Xiao et al., 2017), Flye 2.8.2-b1689 (Kolmogorov et al., 2020), hifiasm 0.10-r299 (Cheng et al., 2021), and Wtdbg2 v2.5 (Ruan and Li, 2020). The resulting assemblies were integrated and optimized using the Genome Puzzle Master (GPM) pipeline (Zhang et al., 2016). Subsequent polishing and error correction followed established protocols (Huang et al., 2024, Fig. 1B).

Telomeric regions were identified by screening the genome assembly for perfect matches to the conserved 7-bp plant telomeric repeat motifs (CCCTAAA) and their complementary sequences (TTTAGGG), following established protocols for soybean genome annotation (Schmutz et al., 2010). Similarly, centromeric regions were annotated based on known soybean centromeric repeats (Gill et al., 2009; Tek et al., 2010). Assembly quality was rigorously evaluated using (1) BUSCO v5 (Waterhouse et al., 2018) (*Fabales* ortholog set with 5366 genes) for completeness assessment, (2) GCI (Chen et al., 2024) for large-scale continuity evaluation, and (3) Clipping information for Revealing Assembly Quality (CRAQ) (Li et al., 2023) analysis for base-level accuracy verification.

Gene annotation

We annotated the TL1-HQ genome using an iterative pipeline combining *ab initio* prediction and homology-based evidence. Repetitive sequences were masked using RepeatMasker v4.0.7 with the Dfam_Consensus-20170127 and RepBase-20170127 databases. Gene prediction was then performed through three progressively refined cycles: initial untrained models from SNAP and Augustus v3.4.0 (Stanke et al., 2008) were integrated with Wm82 protein evidence via MAKER-P v3.01.03 (Campbell et al., 2014), then the resulting predictions were used to train improved species-specific models for a second-round annotation and a final optimization step was implemented to refine the model architecture,

followed by systematic integration of comprehensive genomic evidence to generate a high-confidence consensus gene set.

Comparative genomic analysis

To genome collinearity and structural variations were investigated through comparative analyses among the TL1-PRE, TL1-HQ, Wm82-NJAU and ZH13 genomes using multiple bioinformatics tools. Pairwise collinearity analysis was conducted using Mummer v4.0 (Marçais et al., 2018) with the 'show-diff' function to identify regions of insertion and/or deletion. The alignment results were visualized as a dot plot using the mummerplot utility. For large-scale synteny and structural variation comparisons, GenomeSyn (Zhou et al., 2022) and NGenomeSyn (He et al., 2023) were employed to enable high-resolution visualization of genome alignments and rearrangement events. Additionally, previously unknown genomic regions (>1 Kb) between the TL1-PRE and TL1-HQ genomes were identified using Syri (Goel et al., 2019), facilitating the detection of unaligned genomic segments and potential assembly improvements.

Ortholog identification and analysis

To establish high-confidence orthologous relationships among gene models in the TL1-HQ, Wm82.a4.v1 and Wm82.a6.v1 genomes, we performed a bidirectional BLAST search (BLASTN/BLASTP) using both transcript and protein sequences, with orthologs identified as reciprocal best hits meeting stringent criteria including syntenic alignment, consistent strand orientation and conservation of gene architecture. For cases with multiple homologous hits, only the top-scoring gene pair was retained to ensure unambiguous one-to-one orthology. The alignment parameters were optimized for specificity and sensitivity (-per_identity 60, -evalue 1e-6, -best_hit_score_edge 0.05, -best_hit_overhang 0.4, -max_target_seqs 2).

EMS mutagenesis and phenotypic characterization

The mutagenesis population was developed using the soybean cultivar TL1. In 2017, approximately 5 kg of healthy TL1 M_0 seeds were selected and subjected to chemical mutagenesis at Huazhong Agricultural University (Wuhan, Hubei Province, China). The seeds were immersed in 0.6% EMS solution and incubated in darkness at room temperature for 5 h. Following mutagenesis, seeds were thoroughly rinsed under running water for 30 min to remove residual EMS and then air-dried on filter paper. As shown in Fig. 2A, the mutagenized seeds were then planted in an isolated field with no history of soybean cultivation to prevent cross-contamination with wild-type plants. The experimental field was arranged in single-row plots (1 m length per row) with 8–10 seeds per row and 0.5 m inter-row spacing. The population was advanced using the single seed descent (SSD) method to establish subsequent generations. Comprehensive phenotypic evaluation was conducted across multiple generations, with particular focus on M_4 plants. Systematic observations were made during critical developmental stages (seedling establishment, branching, and seed maturation) and compared with wild-type controls.

DNA sequencing and analysis of mutant lines

A total of 288 mutant lines, along with the wild-type control, were subjected to whole-genome resequencing on the Illumina HiSeq 2000 platform. Basically, fresh young leaves were collected from seedlings of each genotype, flash-frozen in liquid nitrogen, and stored at −80°C until DNA extraction. Genomic DNA was isolated using the CTAB method. DNA integrity was assessed by electrophoresis on a 1% agarose gel, and only high-quality samples (without degradation or contamination) were processed further. DNA was randomly fragmented using a Covaris ultrasonicator, followed by library construction through end-repair, 3′ adenylation, adapter ligation, PCR amplification and size selection. Final libraries were quantified and quality-checked before sequencing.

For sequencing and data processing, paired-end sequencing (2 × 150 bp) was performed on the Illumina HiSeq 2000 platform, and raw sequencing data were processed using Trimmomatic (Bolger et al., 2014) to remove adapter sequences, duplicated reads, and low-quality reads (defined as those containing >10% N bases or >50% bases with Phred quality scores < 20). Clean reads were aligned to the TL1-HQ reference genome using BWA-MEM (Li and Durbin, 2009) with default parameters. SAM files were converted to sorted BAM files using SAMtools (Li et al., 2009), and sequencing depth was calculated using SAMtools 'depth'.

Mutation detection and variant analysis

To identify mutation sites in the mutant lines, sorted BAM files were processed using the Genome Analysis Toolkit (GATK v4.2.2) (McKenna et al., 2010) following best practices for variant detection. The TL1-HQ reference genome was indexed using BWA (Li and Durbin, 2009) prior to alignment. During preprocessing, PCR duplicates were removed from the BAM files using Sambamba (Tarasov et al., 2015) to reduce false-positive variant calls. Initial variant detection was performed using GATK's HaplotypeCaller to identify SNPs and InDels for each mutant relative to the wild-type (TL1-HQ) reference, with outputs stored in gVCF format. These individual gVCF files were then merged using CombineGVCFs and jointly genotyped with GenotypeGVCFs to generate a unified variant call set. For variant filtering, SNPs were subjected to hard-filtering criteria (QD < 2.0, MQ < 40.0, MQRankSum < −12.5, ReadPosRankSum < −8.0, FS > 60.0), while InDels were filtered using QD < 2.0, ReadPosRankSum < −20.0, and FS > 200.0. Finally, to isolate mutant-specific variants, wild-type background polymorphisms were excluded using BCFtools v1.15.1 (Danecek et al., 2021) isec module, retaining only mutations absent in the wild-type control.

Comprehensive variant annotation and classification

Comprehensive functional annotation and classification of all identified SNPs and InDels were performed using ANNOVAR (Wang et al., 2010) with the TL1-HQ genome as the reference. Variants were systematically categorized into exonic regions (including splice-site variants within ± 2 bp of exon-intron junctions), regulatory regions (5′ UTR, 3′ UTR, and upstream/downstream sequences within 1 Kb of transcription start/termination sites), and non-coding regions (intronic and intergenic sequences), following a hierarchical prioritization system: exonic (including splicing) > 5′/3′ UTR > intronic > upstream/downstream > intergenic. Exonic variants were further classified as synonymous (no amino acid change), nonsynonymous (missense), stop-gain (nonsense), frameshift (indels disrupting the reading frame) or non-frameshift (in-frame variants).

Genotypic verification and phenotypic characterization of soybean mutants

For genotypic verification of mutant lines, genomic DNA was extracted from wild-type (TL1) and mutant soybean lines using standard protocols. To confirm mutant genotypes, we amplified fragments of 24 target genes via PCR using gene-specific primers (Table S7), including *GmAGO5* and *GmSGR1*, which contain exon-located SNP variants. The resulting amplicons were Sanger-sequenced to identify genetic variations between wild-type and mutant lines.

For phenotypic assessment, wild-type and mutant plants were grown in a controlled greenhouse environment with the following conditions: 12-h light/dark photoperiod, constant temperature of 27°C and 50% relative humidity. The growth substrate was maintained at water saturation throughout the experiment. Seed color phenotypes were compared between *TM-29897* mutant and wild-type. Leaf color phenotypes were compared between *TM-30017* mutant and wild-type, with leaves at identical nodal positions assessed at 2 weeks post-sowing (day 0 of dark treatment) and subsequently evaluated at 5-day intervals until day 10 of dark treatment.

Data availability

Mutant seeds reported in this paper are available for non-profit research upon reasonable request. The TL1-HQ genome data and EMS mutations whole-genome variation data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center, Chinese Academy of Sciences (PRJCA040540) that are publicly accessible at https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA040540.

CRediT authorship contribution statement

Min Chen and **Jianwei Zhang**: Supervision, Conceptualization. **Min Chen**: Resources, Funding acquisition, Project administration, Investigation. **Yinghua Sheng**, **Yicheng Huang**, **Zilun Jin**: Data curation, Formal analysis, Methodology, Software, Visualization. **Xuyan Wang** and **Jingwen Zhang**: Validation. **Chenghui Liu**: Software, Visualization. **Yinghua Sheng**, **Jianwei Zhang**, **Min Chen**: Writing - Original draft, Writing - Review & Editing. **Chuang Ma** and **Zhipeng Zhou**: Writing - Review & Editing.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31970344) and Joint Funds of the Natural Science Foundation of Hainan Province, China (2021JJLH0065).

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgg.2025.08.006.

References

Allen, G.C., Flores-Vergara, M.A., Krasynanski, S., Kumar, S., Thompson, W.F., 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat. Protoc. 1, 2320–2325.

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

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al., 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164, 513–524.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods 18, 170–175.

Chen, Q., Yang, C., Zhang, G., Wu, D., 2024. GCI: a continuity inspector for complete genome assembly. Bioinformatics 40, btae633.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008.

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y., Li, B., et al., 2022. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Mol. Plant 15, 1268–1284.

Espina, M.J., Lovell, J.T., Jenkins, J., Shu, S., Sreedasyam, A., Jordan, B.D., Webber, J., Boston, L., Brůna, T., Talag, J., et al., 2024. Assembly, comparative analysis, and utilization of a single haplotype reference genome for soybean. Plant J. 120, 1221–1235.

Fang, C., Li, C., Li, W., Wang, Z., Zhou, Z., Shen, Y., Wu, M., Wu, Y., Li, G., Kong, L. A., et al., 2014. Concerted evolution of *D1* and *D2* to regulate chlorophyll degradation in soybean. Plant J. 77, 700–712.

Ge, F., Zheng, N., Zhang, L., Huang, W., Peng, D., Liu, S., 2018. Chemical mutagenesis and soybean mutants potential for identification of novel genes conferring resistance to soybean cyst nematode. J. Integr. Agric. 17, 2734–2744.

Gill, N., Findley, S., Walling, J.G., Hans, C., Ma, J., Doyle, J., Stacey, G., Jackson, S. A., 2009. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 151, 1167–1174.

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.

Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., Song, Q., Pantalone, V.R., Song, B.H., Yu, D., et al., 2022. *POWR1* is a domestication gene pleiotropically regulating seed quality and yield in soybean. Nat. Commun. 13, 3051.

Greene, E.A., Codomo, C.A., Taylor, N.E., Henikoff, J.G., Till, B.J., Reynolds, S.H., Enns, L.C., Burtner, C., Johnson, J.E., Odden, A.R., et al., 2003. Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. Genetics 164, 731–740.

Guo, G., Xu, K., Zhang, X., Zhu, J., Lu, M., Chen, F., Liu, L., Xi, Z., Bachmair, A., Chen, Q., et al., 2015. Extensive analysis of *GmFTL* and *GmCOL* expression in northern soybean cultivars in field conditions. PLoS ONE 10, e0136601.

He, W., Yang, J., Jing, Y., Xu, L., Yu, K., Fang, X., 2023. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. Bioinformatics 39, btad121.

Huang, Y., Koo, D.H., Mao, Y., Herman, E.M., Zhang, J., Schmidt, M.A., 2024. A complete reference genome for the soybean cv. Jack. Plant Commun. 5, 100765.

Jia, J., Ji, R., Li, Z., Yu, Y., Nakano, M., Long, Y., Feng, L., Qin, C., Lu, D., Zhan, J., et al., 2020. Soybean *DICER-LIKE2* regulates seed coat color via production of primary 22-nucleotide small interfering RNAs from long inverted repeats. Plant Cell 32, 3662–3673.

Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., et al., 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat. Methods 17, 1103–1110.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27, 722–736.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, K., Xu, P., Wang, J., Yi, X., Jiao, Y., 2023. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. Nat. Commun. 14, 6556.

Li, Z., Jiang, L., Ma, Y., Wei, Z., Hong, H., Liu, Z., Lei, J., Liu, Y., Guan, R., Guo, Y., et al., 2017. Development and utilization of a new chemically-induced soybean library with a high mutation density. J. Integr. Plant Biol. 59, 60–74.

Liang, S., Duan, Z., He, X., Yang, X., Yuan, Y., Liang, Q., Pan, Y., Zhou, G., Zhang, M., Liu, S., et al., 2024. Natural variation in *GmSW17* controls seed size in soybean. Nat. Commun. 15, 7417.

Mach, J., 2017. Saddle up, soybean seed pigments: Argonaute5 in spatially regulated silencing of *chalcone synthase* genes. Plant Cell 29, 604.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. 14, e1005944.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al., 2010. The Genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Pendleton, M., Sebra, R., Pang, A.W., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., et al., 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods 12, 780–786.

Ruan, J., Li, H., 2020. Fast and accurate long-read assembly with wtdbg2. Nat. Methods 17, 155–158.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al., 2010. Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183.

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.
Sun, X., Li, X., Lu, Y., Wang, S., Zhang, X., Zhang, K., Su, X., Liu, M., Feng, D., Luo, S., et al., 2022. Construction of a high-density mutant population of Chinese cabbage facilitates the genetic dissection of agronomic traits. Mol. Plant 15, 913−924.
Takagi, H., Tamiru, M., Abe, A., Yoshida, K., Uemura, A., Yaegashi, H., Obara, T., Oikawa, K., Utsushi, H., Kanzaki, E., et al., 2015. MutMap accelerates breeding of a salt-tolerant rice cultivar. Nat. Biotechnol. 33, 445−449.
Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., Prins, P., 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032−2034.
Tek, A.L., Kashihara, K., Murata, M., Nagaki, K., 2010. Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon. Chromosome Res. 18, 337−347.
Tian, Z., Nepomuceno, A.L., Song, Q., Stupar, R.M., Liu, B., Kong, F., Ma, J., Lee, S. H., Jackson, S.A., 2025. Soybean2035: a decadal vision for soybean functional genomics and breeding. Mol. Plant 18, 245−271.
Valliyodan, B., Cannon, S.B., Bayer, P.E., Shu, S., Brown, A.V., Ren, L., Jenkins, J., Chung, C.Y., Chan, T.F., Daum, C.G., et al., 2019. Construction and comparison of three reference-quality genome assemblies for soybean. Plant J. 100, 1066−1082.
Wang, D., Li, Y., Wang, H., Xu, Y., Yang, Y., Zhou, Y., Chen, Z., Zhou, Y., Gui, L., Guo, Y., et al., 2023a. Boosting wheat functional genomics via an indexed EMS mutant library of KN9204. Plant Commun. 4, 100593.
Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164.
Wang, L., Zhang, M., Li, M., Jiang, X., Jiao, W., Song, Q., 2023b. A telomere-to-telomere gap-free assembly of soybean genome. Mol. Plant 16, 1711−1714.
Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35, 543−548.
Xiao, C., Chen, Y., Xie, S., Chen, K., Wang, Y., Han, Y., Luo, F., Xie, Z., 2017. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. Nat. Methods 14, 1072−1074.
Yuan, X., Jiang, X., Zhang, M., Wang, L., Jiao, W., Chen, H., Mao, J., Ye, W., Song, Q., 2024. Integrative omics analysis elucidates the genetic basis underlying seed weight and oil content in soybean. Plant Cell 36, 2160−2175.
Zahra, S., Hussain, M., Zulfiqar, S., Ishfaq, S., Shaheen, T., Akhtar, M., Mehboob ur, R., 2022. EMS-based mutants are useful for enhancing drought tolerance in spring wheat. Cereal Res. Commun. 50, 767−778.
Zhang, C., Xie, L., Yu, H., Wang, J., Chen, Q., Wang, H., 2023. The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. Mol. Plant 16, 1715−1718.
Zhang, J., Kudrna, D., Mu, T., Li, W., Copetti, D., Yu, Y., Goicoechea, J.L., Lei, Y., Wing, R.A., 2016. Genome puzzle master (GPM): an integrated pipeline for building and editing pseudomolecules from fragmented sequences. Bioinformatics 32, 3058−3064.
Zhang, M., Zhang, X., Jiang, X., Qiu, L., Jia, G., Wang, L., Ye, W., Song, Q., 2022. iSoybean: a database for the mutational fingerprints of soybean. Plant Biotechnol. J. 20, 1435−1437.
Zhang, Y., Liu, S., Liang, X., Zheng, J., Lu, X., Zhao, J., Li, H., Zhan, Y., Teng, W., Li, H., et al., 2025. *GmFER1*, a soybean ferritin, enhances tolerance to salt stress and root rot disease and improves soybean yield. Plant Biotechnol. J. 23, 3094−3112.
Zhong, X., Wang, J., Shi, X., Bai, M., Yuan, C., Cai, C., Wang, N., Zhu, X., Kuang, H., Wang, X., et al., 2024. Genetically optimizing soybean nodulation improves yield and protein content. Nat. Plants 10, 736−742.
Zhou, Y., Zhang, T., Wang, X., Wu, W., Xing, J., Li, Z., Qiao, X., Zhang, C., Wang, X., Wang, G., et al., 2023. A maize epimerase modulates cell wall synthesis and glycosylation during stomatal morphogenesis. Nat. Commun. 14, 4384.
Zhou, Z., Yu, Z., Huang, X., Liu, J., Guo, Y., Chen, L., Song, J., 2022. GenomeSyn: a bioinformatics tool for visualizing genome synteny and structural variations. J. Genet. Genomics 49, 1174−1176.