

Complete telomere-to-telomere genomes of cowpea reveal insights into centromere evolution in Phaseoleae

Chuanzheng Wei^{1,†}, Shichao Sun^{1,†}, Yinzi Wang^{1,2}, Li Liu³, Sofie Pearson⁴, Yanbo Wang¹, Tashi Dorjee¹, Emma Mace⁴, David Jordan⁴, Yan Yang^{5,*} and Yongfu Tao^{1,*} 

¹State Key Laboratory for Tropical Crop Breeding, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China

²School of Life Sciences, Henan University, Kaifeng, Henan 475004, China

³National Key Lab of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China

⁴Queensland Alliance for Agriculture and Food Innovation (QAAFI), The University of Queensland, Hermitage Research Facility, Warwick 4370 QLD, Australia

⁵Tropical Crops Genetic Resources Institute, National Key Laboratory for Tropical Crop Breeding, Chinese Academy of Tropical Agricultural Sciences, Haikou/Sanya, Hainan 571101/572024, China

*Corresponding authors. E-mail: taoyongfu@caas.cn; yzqi@126.com

[†]These authors contributed equally

Abstract

Cowpea (*Vigna unguiculata*) is a versatile legume crop providing a critical source of grain, vegetable and forage globally. Cultivated cowpea is classified into two main subspecies, subsp. *sesquipedalis* for fresh-pod vegetable and subsp. *unguiculata* for grain production. Here, we present two complete telomere-to-telomere (T2T) assemblies for the grain-type inbred lines HJD and vegetable-type FC6 through integrating PacBio HiFi reads, Oxford Nanopore ultralong reads, and Hi-C data. The T2T genomes demonstrated improved contiguity, completeness, and accuracy compared to existing genomes, revealing clear telomeric and centromeric features. Comparative analysis of the T2T genomes highlighted inversions underlying subspecies divergence in cowpea. Evolutionary analysis uncovered contraction of gene families related to symbiosis in HJD, consistent with its reduced root nodules compared to FC6. Distribution and composition of tandem repeat arrays and transposable elements in centromeric regions were largely conserved in cowpea, but displayed pronounced variation among Phaseoleae. Furthermore, frequent shifts of centromeric locations coincided with inversions found in Phaseoleae. Overall, this study provides a set of fundamental resources for cowpea improvement and enhances our understanding of cowpea subspecies divergence and genome evolution in Phaseoleae.

Introduction

Cowpea (*Vigna unguiculata*) is an annual herbaceous species in the Fabaceae family, native to the semiarid tropics of Africa [1, 2]. It is widely cultivated across semiarid tropical regions of Africa, Asia, and Latin America, due to its well-known adaptation to infertile soils and low-rainfall conditions [3, 4]. Cultivated cowpea has two major subspecies, *unguiculata* and *sesquipedalis* [5]. The *unguiculata* subspecies is the second most grown legume in Africa for seed production, providing a critical source of protein for the most vulnerable people threatened by malnutrition [6, 7]. The *sesquipedalis* subspecies, also known as the yardlong bean, is predominantly cultivated in East Asia for fresh pod consumption [8, 9]. In addition, cowpea is also a popular forage crop due to its high protein content, digestibility, and fast-growing character [10]. Therefore, genetic improvement of cowpea is critical for global food security.

Obtaining high-quality reference genomes is critical for exploring genomic variation to improve crop productivity [11]. The

first reference genome of cowpea, IT97K-499-35, was assembled using Pacific Biosciences (PacBio) continuous long reads, covering 519 Mb DNA sequence [12]. Despite recent improvements, this reference genome is still highly fragmented with 46.0 Mb unanchored sequence and dozens of unfilled gaps [13]. Recent genome sequencing efforts have resulted in chromosome-scale assemblies of additional cowpea genomes [14, 15]. However, these incomplete genomes still hinder a holistic view of genomic variation and further functional investigations. Complete genomes also present a unique opportunity to systematically characterize highly repetitive regions, such as centromeres and telomeres [16–18].

The Phaseoleae tribe, including numerous widely grown grain and forage species, represents an important lineage in the Papilionoideae subfamily of Fabaceae [19]. Phaseoleae is dominated by three major genera: *Phaseolus*, *Vigna*, and *Glycine* [20]. These closely related genera share strong macro-synteny with large conserved genomic blocks identified among their genomes [13, 21–23]. Centromeres are crucial chromosomal functional

Received: 18 August 2025. Accepted: 21 December 2025. Published: 29 December 2025. Corrected and Typeset: 1 April 2026

© The Author(s) 2025. Published by Oxford University Press on behalf of the Nanjing Agricultural University.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

domains that guarantee the faithful replication and precise separation of chromosomes during mitosis and meiosis [24]. Centromere repositioning events have been reported among *Vigna angularis*, *V. unguiculata*, and *Phaseolus vulgaris* through sequence comparison and oligonucleotide-FISH chromosomal painting [25, 26]. However, centromere evolution and variation in Phaseoleae are still largely elusive due to the lack of high-quality references covering the highly repetitive regions. Nevertheless, the recent completion of telomere-to-telomere (T2T) genome assemblies for major Phaseoleae species has laid a solid foundation for in-depth investigation of centromere evolution [27–30].

In this study, we aim to (i) employ a combination of PacBio HiFi, Oxford Nanopore sequencing technologies to assemble complete T2T genomes of cultivated cowpea; (ii) use multiple measures to demonstrate the superior quality of our T2T genomes and their potential in enhancing genetic discovery; (iii) evaluate conservation and variation of repetitive elements and centromeres in Phaseoleae using the T2T genomes. This study will provide a set of fundamental genomic resources for cowpea improvement and shed new lights on centromere evolution in Phaseoleae.

Results

Telomere-to-telomere genome assembly and annotation of two cowpea lines

To assemble the genomes of HJD and FC6, we generated 37.33 Gb ($\sim 73\times$ coverage) of PacBio HiFi reads, 61.01 Gb ($\sim 119\times$ coverage) of Oxford Nanopore ultralong reads and 32.12 Gb Hi-C reads for HJD. Additionally, we retrieved 28.59 Gb ($\sim 57\times$ coverage) of HiFi reads, 67.27 Gb ($\sim 133\times$ coverage) of Oxford nanopore technologies (ONT) reads and 24.10 Gb Hi-C reads of FC6 from a previous study [15] (Fig. 1A, Table S1). K-mer analysis ($k = 21$) based on HiFi reads estimated the genome size of HJD at ~ 481.93 Mb with a heterozygosity rate of 0.06%, while the FC6 genome was similarly sized at 474.17 Mb with a higher heterozygosity rate of 0.18% (Fig. S1, Table S2).

The preliminary assemblies were generated using Hifiasm by integrating PacBio HiFi, Oxford Nanopore ultralong, and Hi-C reads. The draft assembly of HJD was 525.19 Mb in 15 contigs, and the FC6 genome amounted to 517.59 Mb in 17 contigs (Table S3). After removing organellar sequences, we employed HapHiC software's quickview mode to orient and cluster contigs before manually refining the assignments in Juicerbox to anchor the assemblies onto 11 chromosomes. The resultant HJD assembly contained six unassembled telomeric regions and four unresolved gaps on chromosome 2, 3, 6, and 10 with an average size of 10.9 Kb, while the FC6 assembly had three missing telomeric regions and six gaps on chromosome 4, 6, 9, 10, and 11 with an average size of 7.7 Kb. To fill the scaffold gaps, we used preliminary contigs generated by verkko and NextDenovo to close two gaps in HJD and three in FC6, and then closed remaining gaps using ONT ultralong reads (Table S4). We appended all unassembled telomeric sequences following Liu *et al.* [31], yielding two gapless genome assemblies (Fig. 1B). The final HJD assembly reached 523.32 Mb with a contig N50 of 46.48 Mb, while FC6 totaled 517.39 Mb with a contig N50 of 46.29 Mb (Table 1).

Multiple strategies were employed to assess the accuracy and completeness of the T2T assemblies. Chromosome-level

Hi-C heatmaps revealed high interaction intensity within each chromosome and no anomalous contact patterns, validating the proper ordering and orientation of all pseudomolecules (Fig. 1C and D). HiFi and ONT reads were remapped to the final assemblies, yielding mapping rates $>99.50\%$ for all datasets except for ONT reads in FC6 with a mapping rate of 98.94% (Table S5). The uniform genome-wide coverage across all types of reads also supported the high quality of our assembly (Fig. S2). Assembly completeness of the two genomes was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO), yielding scores of 98.40% for HJD and 98.50% for FC6 (Fig. S3). The LTR Assembly Index was used to evaluate completeness of long terminal repeat (LTR) sequences, scoring 13.56 and 13.55 for HJD and FC6, respectively. Using the k-mer-based method implemented in Merqury, quality values of >50.55 were calculated for both assemblies, corresponding to base accuracy of $>99.99\%$ (Fig. S4, Table S6). Overall, these assessments demonstrated the excellent quality of our T2T assemblies.

After repeat masking, we predicted 27 357 genes in HJD and 27 670 genes in FC6 combining *ab initio* gene prediction with homology and RNA-seq evidence. Both genomes contain $\sim 78\%$ multiexonic genes, with an average of 6.49 and 6.47 exons per multiexonic gene in HJD and FC6, respectively, consistent with patterns observed in other legumes. Gene models also show balanced strand distribution in both assemblies (Table S7). A total of 23 176 allelic gene pairs were found between the two genomes, representing $\sim 84\%$ of both gene sets. Predicted proteins were matched to public databases for 26 984 HJD genes (98.64%) and 27 290 FC6 genes (98.63%) (Table S8).

T2T genomes improve genetic analysis in cowpea

The complete sequence of our T2T genomes allows identification of telomeric and centromeric regions in cowpea. We found a total of 22 telomeric regions on the 11 chromosomes in each genome, spanning 535–24 326 bp in length (Table S9). In plants, centromeric regions are generally composed of arrays of tandemly repeated monomers [32]. We searched for the cowpea-specific satellite repeat CEN455 to localize centromeres in our T2T assemblies. Intact CEN455 arrays were identified on seven of the 11 chromosomes, while partial CEN455 arrays were found on the remaining chromosomes (Table S10). We analyzed the density of tandem repeat using quarTeT and manually merged the most repeat-dense regions with CEN455 hits to define centromere boundaries. For the remaining four chromosomes, centromeric regions were inferred based on repeat density profiles in combination with the distribution of CEN455 fragments. The size of centromere varied markedly among different chromosomes in both genomes (Table S11). However, these centromeric regions are highly conserved between HJD and FC6.

Compared to the reference genome IT97K-499-35, our two T2T assemblies anchored ~ 50 Mb of new sequence into chromosomes with improved contiguity, completeness, and accuracy (Fig. 2A and B). Around 85% of these new sequences are repetitive DNA with 30 and 23 genes annotated in HJD and FC6, respectively. Comparative synteny assessment of the T2T assemblies against IT97K-499-35 uncovered significant sequence omissions in the latter, particularly around centromeric regions (Fig. 2C). Further

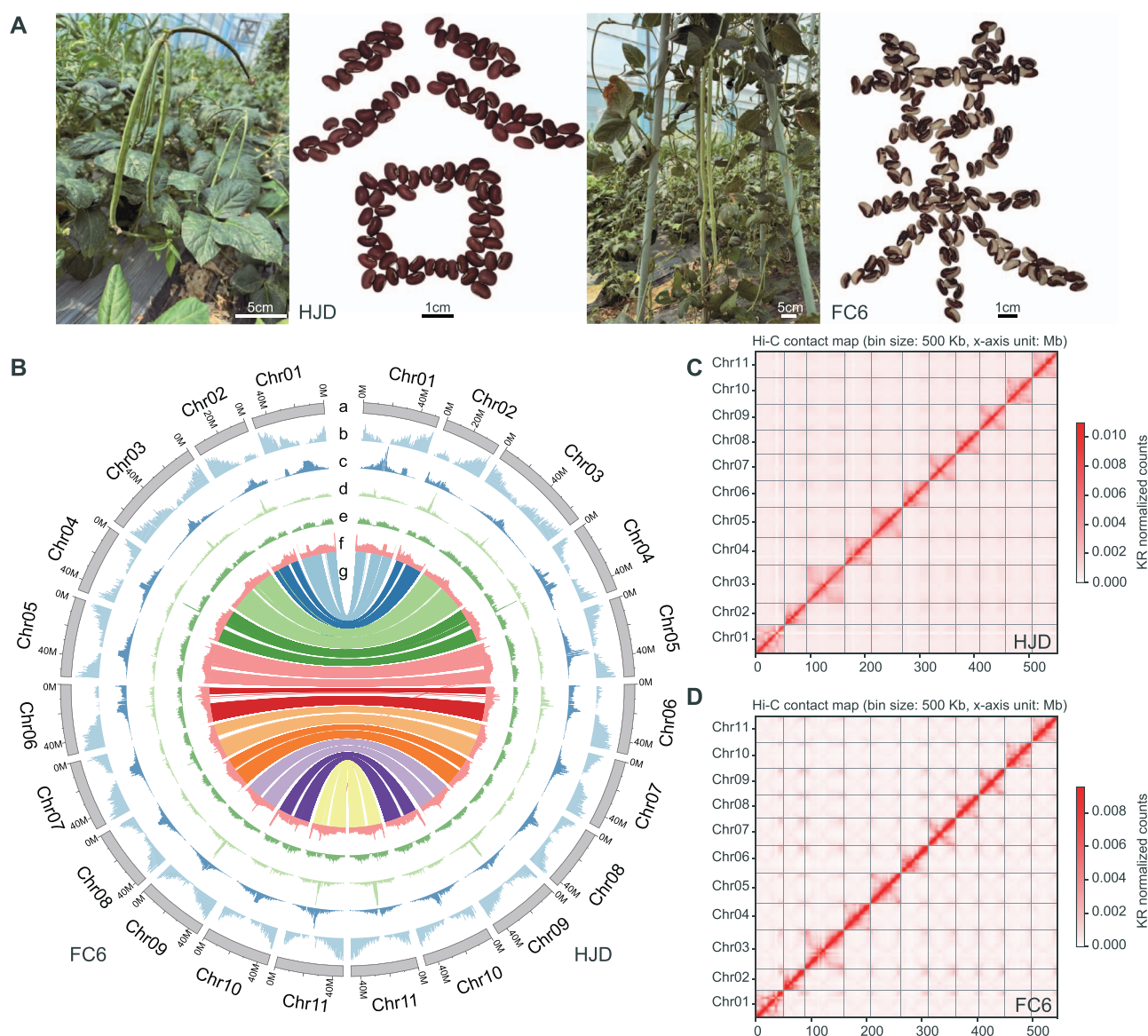


Figure 1 Assembly and genomic features of grain-type cowpea HJD and vegetable-type cowpea FC6. (A) Field-grown pods and seeds of HJD and FC6. (B) Circos plots of the HJD and FC6 assemblies showing: (a) chromosome layout; (b) gene density; (c) LTR/Gypsy element density; (d) LTR/Copia element density; (e) DNA transposon density; (f) GC content; and (g) syntenic relationships among homologous genes. All tracks were computed using 500-kb windows. (C) Hi-C interaction maps at 500-kb bin size. Interaction frequencies are indicated by a color gradient from the plot periphery toward the diagonal, reflecting low- to high-contact intensity.

comparisons of IT97K-499-35 with the T2T assemblies identified large inversions on chromosomes 3, 4, 5, 7, and 10 in IT97K-499-35, likely due to previous assembly errors (Table S12).

To assess the potential applications of our T2T genomes in genetic analysis, we firstly aligned a recently published cowpea genetic linkage map [12] to the HJD T2T assembly. Obvious marker intervals around centromeric regions were found on nearly every chromosome (Fig. 2D, Figs S5 and S6). The high-precision physical coordinates of our T2T assembly enabled accurate conversion of genetic distances to physical distances, and therefore, a refined assessment of recombination rate distribution across the genome. Secondly, resequencing reads of 270 cowpea accessions (45 *V. unguiculata* and 225 *Vigna sesquipedalis*) from a previous study were mapped to the T2T genomes to assess improvement in short-reads mapping [33]. The T2T assemblies reduced the

number of unmapped reads by 33% compared to IT97K-499-35, with higher proportion of the genomes covered by these short reads (Fig. S7, Tables S13 and S14). The mismatch rate also declined in the T2T genomes (Table S15). Thirdly, we identified 201 and 140 previously unannotated genes in the centromeric regions of HJD and FC6, respectively. Gene ontology (GO) functional annotation revealed that these genes were significantly enriched in multiple biological processes, including cellular respiration, immune response-activating cell surface receptor signaling, and positive regulation of defense response to insects (Fig. S8, Table S16). Transcriptome data from six FC6 tissues showed that 21 of the 140 annotated centromeric genes exhibited detectable expression in at least one tissue, suggesting that centromeric regions contain expressed genes with potential relevance to cowpea improvement.

Table 1 Cowpea genome assembly statistics

	HJD	FC6	IT97K-499-35
Size of assembly (bp)	548 737 058	542 521 633	519 435 864
Contig N50 (bp)	48 738 631	48 539 341	10 911 736
Number of contigs	11	11	754
Scaffold N50 (bp)	48 738 631	48 539 341	41 684 185
Number of scaffolds	11	11	686
Number of telomeres	22	22	5
Number of gaps	0	0	68
BUSCO of assembled genome (%)	98.4	98.5	98.3
Protein-coding genes	27 357	27 670	31 948
BUSCO of protein-coding genes (%)	98.4	98.4	98.6
Repeat ratio in genome (%)	51.80	51.26	48.35

T2T genomes identified inversions influencing subspecies divergence in cowpea

Chromosomal inversions can play a critical role in racial divergence by markedly suppressing recombination and reducing genetic exchange [34]. Comparing the two cowpea T2T genomes identified 54 inversion events (Fig. S9). Nearly, half (24/54) of these inversions showed pairwise fixation statistic (F_{ST}) values between the cowpea subspecies higher than the genome-wide average, suggesting their potential roles in driving subspecies differentiation (Table S17). In particular, the largest inversion spanning 9–13 Mb on Chr01 was unequivocally supported by Hi-C data (Fig. 3A). F_{ST} within this inversion reached 0.74, significantly higher than surrounding regions (Fig. 3B). Within each subspecies, nucleotide diversity (π) values inside the inversion were comparable to adjacent regions (Fig. 3C and D), indicating that the inversion region remains polymorphic. Principal component analysis (PCA) based on single nucleotide polymorphism (SNPs) within the inversion showed the two subspecies were clearly divided along the first principal component, which explained ~89% of the variance (Fig. 3E). This evidence suggests that this inversion might play a critical role in the divergence of the two cowpea subspecies. GO enrichment analysis of genes within the inversion revealed significant overrepresentation of genes related to gibberellin biosynthesis and toxin metabolism pathways (Table S18).

We further genotyped the inversion in the 270 cowpea accessions [33]. In the inversion region, *V. sesquipedalis* accessions showed pronounced coverage voids and reduced overall read depth (Fig. 3F), while *V. unguiculata* accessions retained consistently high coverage (Fig. 3G). Based on the different coverage patterns, the inversion was genotyped across the cowpea population (Fig. 3H). We found the HJD allele accounted for ~69% of *V. unguiculata* accessions compared to 24% of *V. sesquipedalis* (Fig. 3I). This marked divergence of allele frequency further supported the role of the inversion in subspecies differentiation in cowpea.

Cowpea T2T genomes facilitated evolutionary analysis

Using the T2T genomes, we inferred the phylogenetic placement and divergence times of cowpea alongside 20 other plant species

(Table S19). As expected, the phylogeny constructed from single-copy gene families was congruent with previously published trees. Molecular dating indicated that HJD and FC6 represented the most recently diverged lineages relative to IT97K-499-35. The split of cowpea from glycine was dated at ~22.62 Mya (Figs 4A and S10), preceding soybean's most recent whole-genome duplication (~13 Mya) [35, 36].

A comparative analysis of four Phaseoleae species revealed 13 701 orthologous gene families conserved in all species and 3601 gene families specific to cowpea (Fig. 4B). Synteny analysis among Phaseoleae species revealed that regions present as a single block in cowpea generally corresponded to two blocks in soybean, but only one block in mung bean and common bean (Fig. 4C). Within cowpea, a total of 194 syntenic blocks comprising 8618 homologous genes (29.12% of the total) were identified, reflecting ancient polyploidy (Fig. S11, Table S20). Legumes trace their origin to a shared tetraploid ancestor, which underwent two polyploidization events at ~130 and 58 Mya, with soybean subsequently experiencing an additional tetraploidization ~13 Mya [35, 37]. The K_s distribution of soybean paralogs exhibited peaks at 0.59 and 0.12, corresponding to whole-genome duplication events dated to ~58 and ~13 Mya. The only K_s peak in cowpea was at ~0.7, indicated no subsequent whole-genome duplication (WGD) events since ~58 Mya (Figs 4D and S10).

Further gene family evolution analysis identified expansions of 162 and 121 families in HJD and FC6, respectively, while 974 and 308 families exhibited contractions (Table S20). The expanded gene families in HJD and FC6 were enriched for different functions, potentially related to their distinct characters (Fig. S12). Intriguingly, GO terms associated with symbiosis and immune response were overrepresented in contracted gene families in HJD (Fig. 4E, Table S21). Our rhizobial inoculation assays revealed that HJD formed only ~50% of the root nodules produced by FC6 (Fig. 4F and G). The reduced root nodules in HJD could be related to the contraction of genes related to symbiosis and immune response.

Tandem repeat arrays are highly diverse among Phaseoleae species

Tandem repeat arrays (TRAs) are highly repetitive elements that are difficult to resolve in genome assemblies. The cowpea T2T

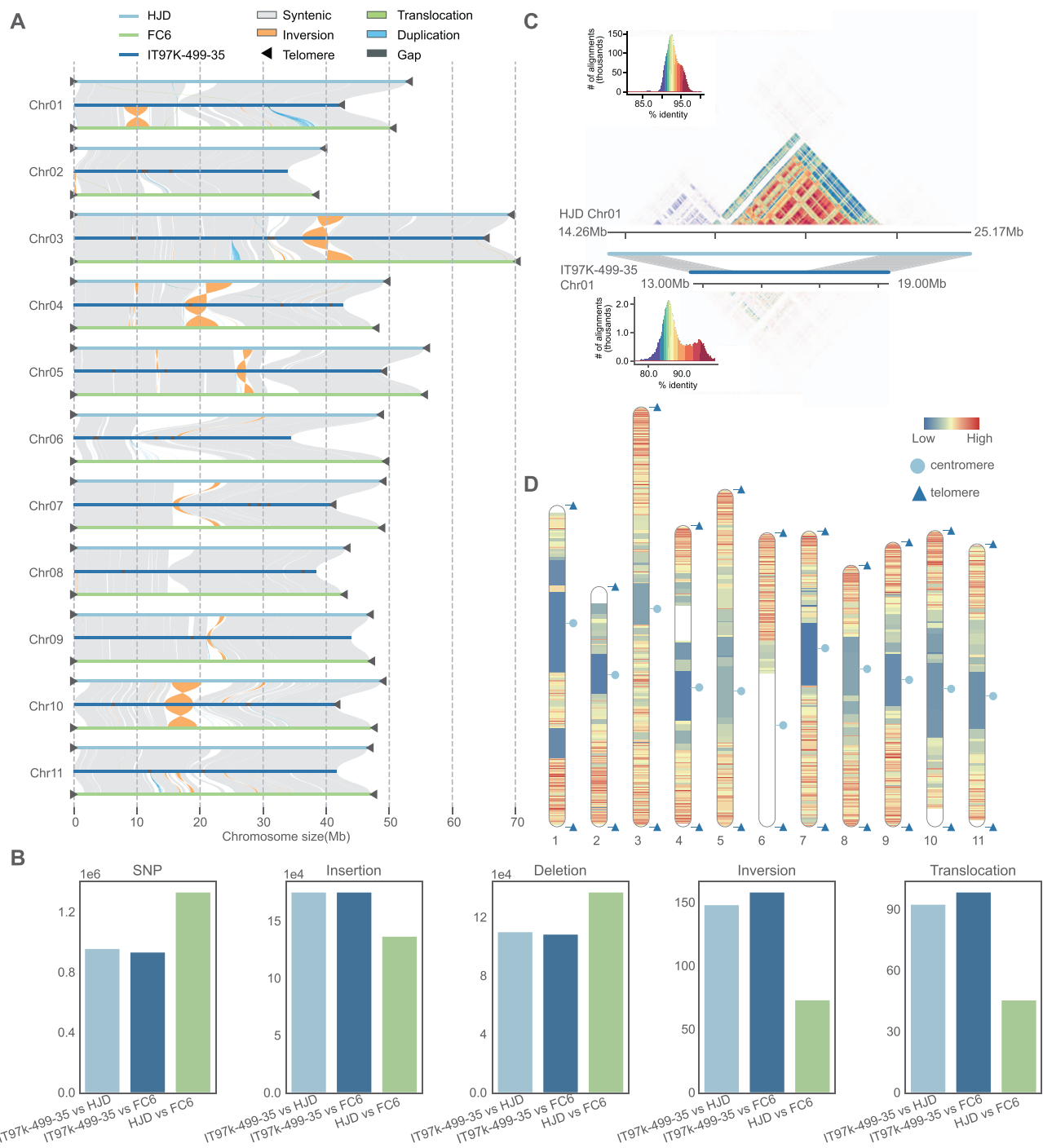


Figure 2 Comparison of cowpea T2T assemblies with IT97K-499-35 reference. (A) Syntenic overview of HJD (light blue), FC6 (dark blue) and IT97K-499-35 (cyan). Gray regions denote synteny, orange indicates inversions, green indicates translocations, teal indicates duplications, black indicates gaps, and black triangles mark telomeres. (B) Bar charts showing counts of SNPs, insertions, deletions, inversions, and translocations for each genome pair (IT97K-499-35 vs HJD, IT97K-499-35 vs FC6, HJD vs FC6). (C) Assembly improvement at the Chr01 centromere in HJD relative to IT97K-499-35. Gray blocks show synteny flanking the centromere; colored contact map displays similarity of tandem arrays. (D) Genome-wide recombination rate landscape across each chromosome of the HJD T2T assembly. Rates were calculated as genetic distance (cM) per physical distance (Mb) using linkage-map markers.

genomes offered an opportunity to compare the distribution and composition of TRAs across Phaseoleae (Table S22). We found almost all TRAs overlapped with centromeric regions in *Glycine max* (Fig. S13). This concentration of TRAs around the centromeres was not observed in other Phaseoleae. In particular, eight out of 11 TRAs in *P. vulgaris* resided in noncentromeric regions. These

TRAs exhibited extensive interspecific length variation, ranging from 12.89 Mb in HJD to just 0.59 Mb in *P. vulgaris* (Figs S14–S17). The major component of TRAs, monomeric repeat units, showed distinct length distribution profiles across the examined species. *Glycine max* had the shortest TRA monomers, predominantly clustering at ~100 bp, followed by *V. radiata* and cowpea peaking

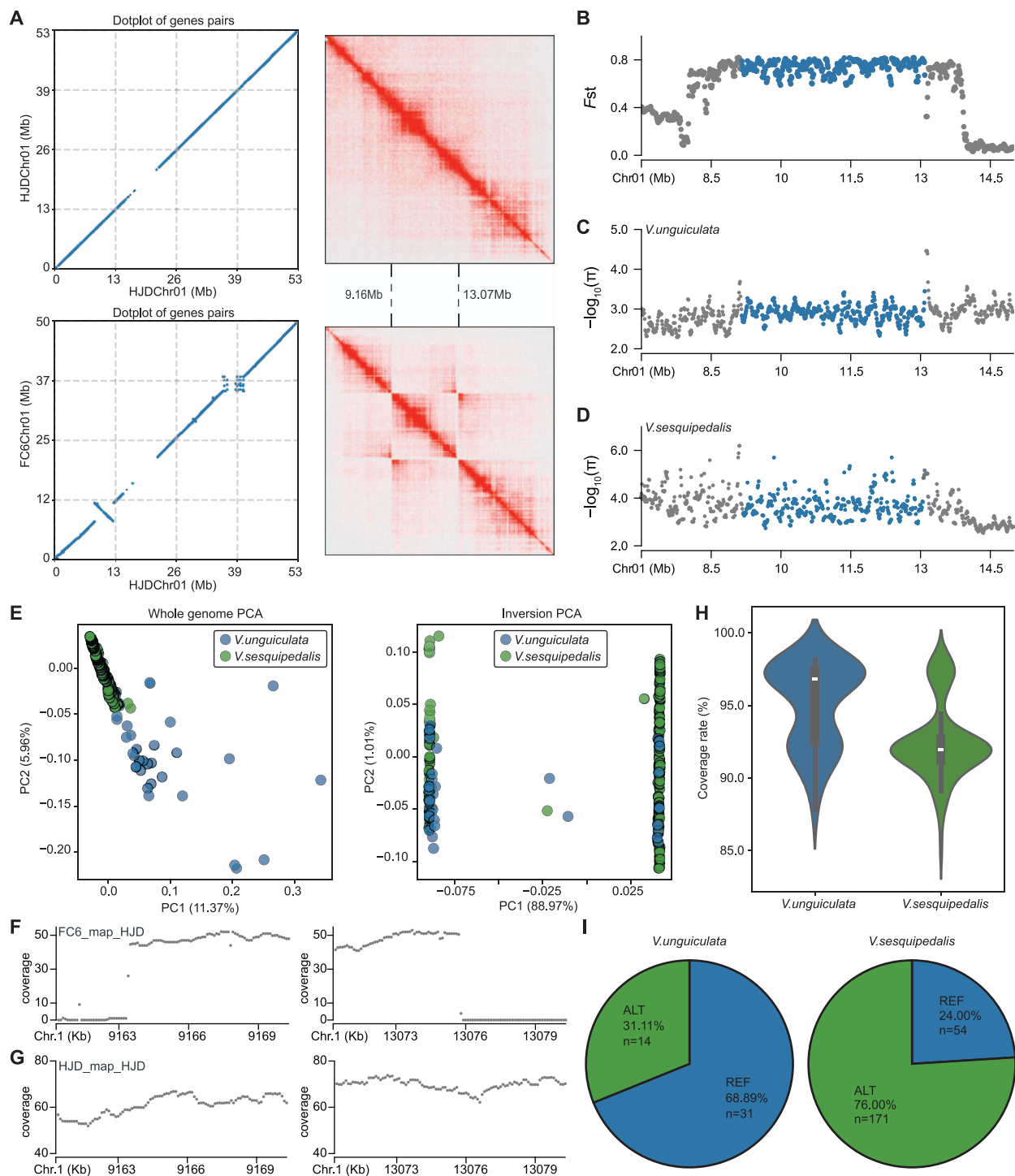


Figure 3 Inversion landscape on Chr01 and population genotyping. (A) Dot plot of syntenic gene pairs on Chr01 between HJD and FC6, alongside Hi-C contact matrices for the inverted region. Hi-C data from FC6 mapped to the HJD assembly reveal a butterfly pattern around the inversion. (B) Sliding window F_{ST} distribution across the inversion and flanking regions, with the inversion highlighted in blue. (C) π distribution within the inversion and adjacent regions in *V. unguiculata*, inversion marked in blue. (D) π distribution for the same regions in *V. sesquipedalis*, inversion marked in blue. (E) Principal component analysis of genome-wide SNPs vs inversion-region SNPs. (F) Alignment gaps generated when mapping FC6 long reads to the HJD reference. (G) Continuous coverage profile of HJD long reads mapped to the HJD assembly. (H) Short-read coverage of population samples over the inversion region, using HJD as reference. (I) Distribution of reference and inverted genotypes within the population.

at ~175 and 455 bp, respectively (Fig. 5A and B). *Phaseolus vulgaris* showed three peaks of monomer lengths at ~100, ~270, and ~530 bp (Fig. 5A).

CEN455 is the dominant monomeric repeat unit in cowpea. A total of 40 566 and 38 720 CEN455 repeat units were identified in HJD and FC6, respectively (Figs 5C and S13). An evolutionary

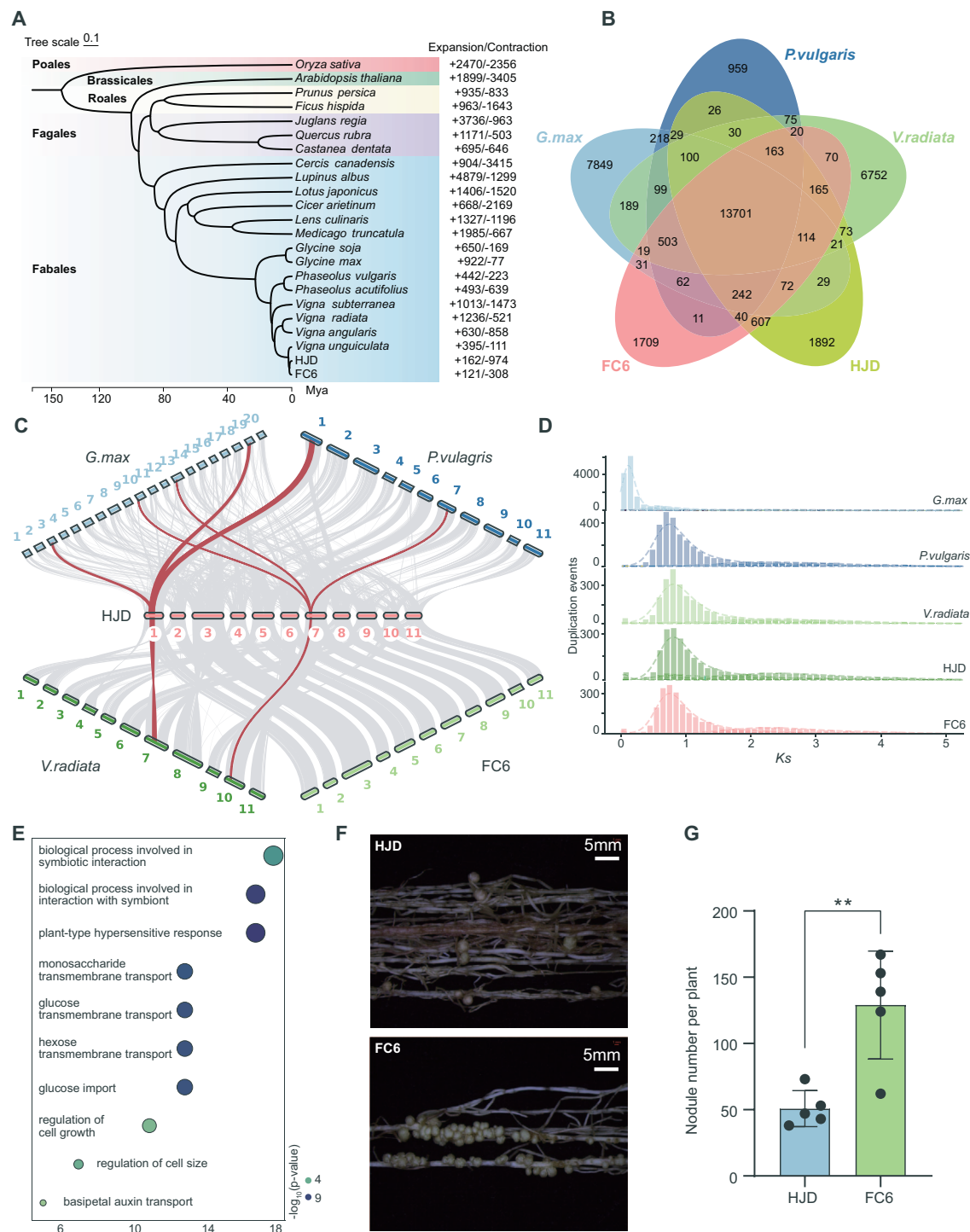


Figure 4 Comparative genomics and evolutionary analyses of cowpea. (A) Phylogenetic tree reconstructed from single-copy orthologs. Colored blocks (top to bottom) indicate Poales, Brassicales, Rosales, Fagales, and Fabales. The timescale at the base shows divergence times, and numbers on the right indicate counts of contracted and expanded gene families. (B) Numbers of orthologous gene groups among Phaseoleae species (*G. max*, *P. vulgaris*, *V. radiata* and *V. unguiculata*). (C) Intergenomic syntenic blocks among Phaseoleae species. Gray ribbons connect genome-wide syntenic blocks. Red ribbons highlight representative regions where a single block in cowpea corresponds to two blocks in soybean but only one block in mung bean and common bean. (D) Synonymous substitution (K_s) distributions for paralogous gene pairs in each species, where the y-axis reflects anchor-pair counts per K_s bin. (E) GO enrichment of contracted gene families in HJD. (F) Nodule phenotype of HJD and FC6 30 days after rhizobial inoculation. (G) Nodule counts of HJD and FC6 at 30 days (mean $\pm$ SD, $n = 5$). Statistical significance between two groups was assessed by two-sided Student's t -test; ** indicates $P < 0.01$.

tree showed that CEN455 repeats segregate mainly according to their chromosomal provenance rather than subspecies (Fig. 5D). CEN455 repeat units showed elevated sequence similarity within

a genome (Fig. 5E). The satellite repeat CEN440 displayed a similar pattern to CEN455 (Fig. S18). In contrast, CEN315 repeats were detected almost exclusively in FC6 (Fig. 5F).

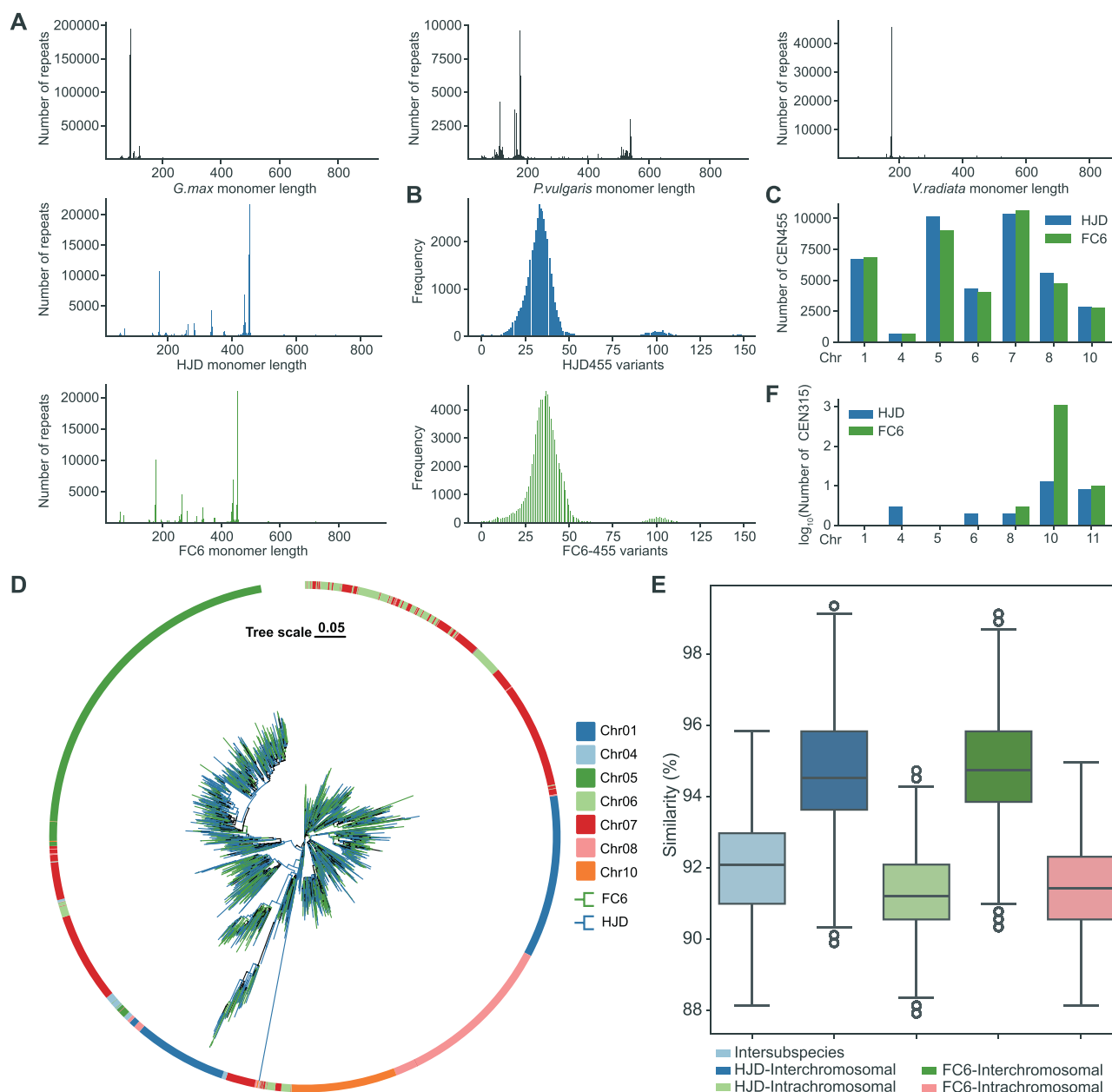


Figure 5 Characteristics of Phaseoleae TRAs. (A) Length distribution of monomeric tandem repeats identified by TRASH in each genome. (B) Distribution of variation counts in CEN455 compared with the consensus monomer. (C) Copy number of CEN455 on each chromosome in HJD and FC6. (D) Copy number of CEN315 on each chromosome in HJD and FC6. (E) Phylogenetic tree of sampled CEN455 monomers. Outer ring colors denote chromosomes; inner branch colors distinguish HJD and FC6. (F) Boxplot comparison of CEN455 monomer similarity across subspecies, within chromosomes, and among chromosomes.

Sequence similarity was observed among different classes of TRAs within species, such as Gm103 and Gm123 in *G. max*, Pv111 and Pv166 in *P. vulgaris*, and CEN440 and CEN455 in cowpea. Additionally, interspecies homology was detected among repeats including Gm335, Vr215, Pv539, and CEN338. Phylogenetic inference indicates these satellites might have derived from a shared ancestral repeat and undergone lineage-specific expansion following speciation (Fig. S19). Our results underscore substantial interspecific variability of TRAs within Phaseoleae.

Comparison of transposable elements reveals phaseoleae divergence

To investigate the impact of transposable element (TE) dynamics on genome evolution in Phaseoleae, we quantified TE content and composition. TEs comprised roughly 50% of each genome in all four species, despite their differences in genome size. LTR retrotransposons were the dominant components of TE, accounting for 31%–40% of the genome sequence (Fig. 6A, Table S23). TE

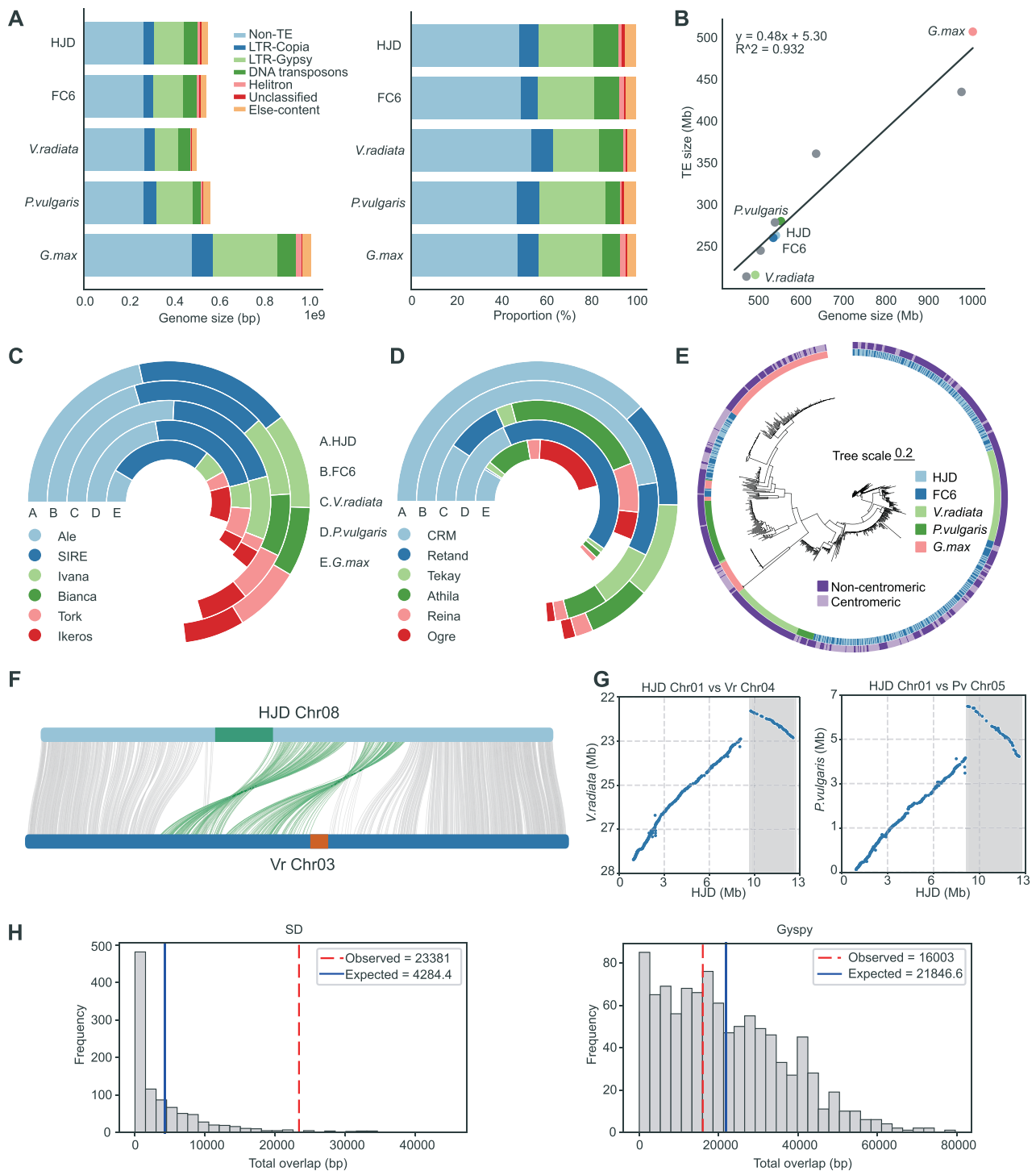


Figure 6 Analysis of repetitive sequences in Phaseoleae. (A) Genome composition of each species and proportions of repeat classes. (B) Relationship between transposable element content and genome size. (C) Circular distribution of Copia superfamily members (Panels A–E represent five genomic components; colors indicate proportional lineage composition) (D) Circular distribution of Gypsy superfamily members. (E) Phylogenetic tree of full-length CRM elements. (F) Synteny between cowpea (HJD Chr08) and mung bean (Vr Chr03). Each horizontal bar represents a chromosome; centromeric regions are shown as colored segments distinct from the chromosome backbone. Gray ribbons connect genome-wide syntenic gene pairs, while green ribbons highlight collinear gene blocks located adjacent to the centromeric regions. (G) Dot plots of inversions at shared breakpoints in cowpea vs mung bean and cowpea vs common bean; blue dots represent orthologous gene pairs, gray shading indicates aligned regions. (H) SD and Gypsy occupancy at inversion breakpoints estimated from 1000 random resampling iterations.

accumulation accounted for nearly half of genome size variation in these species (Fig. 6B, Table S24). Interestingly, the proportion of specific TE families differed markedly among species with

comparable TE content. Bianca-type Copia was nearly exclusively found in cowpea. Within Gypsy elements, CRM was the most abundant subfamily in cowpea, whereas Athila and Retand

predominated in mung bean and common bean, respectively (Fig. 6C and D, Table S25).

We further examined centromeric sequence composition, which showed that TEs occupied 24.88%–98.78% of centromeric regions. Gypsy elements were the predominant TE class in these regions, accounting for 69.29%–88.38% of centromeric TEs. The CRM subfamily represented the major component in cowpea centromeres, comprising 61.42% and 52.53% of centromere sequence in HJD and FC6, respectively (Table S26). A phylogeny based on CRM elements showed that CRM monomers from different species cluster together within multiple strongly supported lineages, suggesting recurrent incorporation of CRM into centromeric arrays (Fig. 6E). In *V. unguiculata*, CRM repeats were tightly interspersed across both centromeric and noncentromeric clades, reflecting recent bursts of element amplification in cowpea. In contrast, soybean CRM elements formed well-resolved lineages, consistent with an ancient expansion event in *G. max* (Fig. 6E). These patterns indicate that CRM elements have retained ancestral diversity while undergoing lineage-specific expansions throughout Phaseoleae evolution.

Frequent centromere renewal and repositioning in Phaseoleae

Comparison of centromere positions in Phaseoleae found frequent centromere renewal and repositioning. For example, the centromere of cowpea Chr08 did not align with the syntenic mung bean Chr03 (Fig. 6F). Changes in arm ratio of corresponding chromosomes in Phaseoleae species further illustrated centromere repositioning. The arm ratios for the syntenic group consisting of cowpea Chr07, mung bean Chr10, and common bean Chr07 were 1.64, 2.01, and 2.57, respectively. Remarkably, centromere repositionings coincided with nearby inversion events in all 11 syntenic chromosome pairs between cowpea and mung bean and in 10 out of 11 pairs between cowpea and common bean (Figs S20 and S21). This suggests that centromere repositioning is likely driven by inversion events.

Multiple centromere inversions shared nearly identical breakpoints, such as these breakpoints on Chr01 (~13 Mb), Chr03 (~21 Mb), and Chr06 (~1 Mb) of HJD (Fig. 6G, Fig. S22). Repetitive elements are a major force driving inversion formation. Therefore, we further assessed whether TEs and segmental duplications (SDs) were associated with inversion breakpoints (Table S27). SDs were found to be highly enriched at breakpoints of these inversions, indicating they play a role in inversion formation around centromere and centromere repositioning (Figs 6H and S23).

Discussion

High-quality reference genomes are fundamental resources for genomic investigation and genetic improvement of plants [38, 39]. This study presented the first two complete T2T genome assemblies of cowpea, resolving previously intractable regions, including centromeres and telomeres. These T2T genomes showed much higher contiguity, completeness, and accuracy compared to previous cowpea genomes. The improved mapping statistics in analysis of resequencing data demonstrated that the superior quality of our T2T genomes can enhance genetic

discovery. Comparative analysis of our T2T genomes found that inversions could play an important role in cowpea subspecies divergence. The holistic view of centromeres and repetitive elements from the T2T genomes uncovered frequent centromere renewal and repositioning in Phaseoleae, which often coincided with inversions. In summary, the findings of this study provided a set of foundational resources for cowpea improvement and enhanced our understanding of cowpea diversification and genome evolution in Phaseoleae.

T2T genomes provide new opportunities to generate highly accurate and complete references to explore the full genomic landscape in a species [40]. Recent advances in sequencing technology has enabled assembly of T2T genomes for multiple species, including rice [41], maize [42], sorghum [43], soybean [27], common bean [29], and mung bean [30]. These genomes uncovered genomic sequence in highly complex and previously intractable genomic regions and have proved their value in facilitating investigation of genomic pattern and gene mining. Likewise, the T2T genomes in this study captured full genome sequence, including centromere regions in cowpea. The identification of hundreds of new protein-coding genes in centromeres of our cowpea T2T genomes with GO enrichment in immunity and stress responses demonstrated the functional importance of these regions. Access to the T2T genomes provides a solid foundation for full-scale exploration of genomic variation to improve cowpea productivity.

The T2T assemblies have enabled the detailed examination of centromeres in cowpea and comparison among Phaseoleae species. We found composition and position of centromeres are largely conserved within cowpea. In contrast, Phaseoleae species displayed frequent repositioning of centromere and distinct centromere composition. Rapid centromere evolution is well documented in previous studies [44]. Despite their conserved function, centromeres show pronounced variation in terms of sequence components and physical position in *Gossypium hirsutum* [45] and *Fragaria x ananassa* [46]. Centromere-specific proliferation of TEs can create new binding sites for centromere proteins and disrupt old ones, thereby causing positional shifts of centromeres [47, 48]. Our results supported the role of TEs as key drivers of sequence variation and movement of centromeres in Phaseoleae. Similarly, centromere relocation promoted by pervasive TE insertions was reported in Salicaceae species [49]. In addition, the enrichment of SDs at the inversion breakpoints around centromere suggests that duplication-mediated inversions could also play a role in centromere shifts. Overall, our examination of centromeres in Phaseoleae underscores that centromeres are hotspots of genomic innovation.

Comparison of our cowpea T2T assemblies indicated that inversions likely underlie the divergence between the two cultivated subspecies. In particular, the ~4-Mb inversion on chromosome 1 showed clear allele frequency difference between *V. unguiculata* and *V. sesquipedalis*. Chromosomal inversions can prevent gene exchange by suppressing recombination in the region, therefore locking together co-adapted allele combinations [34]. This mechanism is well recognized as a driver of speciation and racial divergence. Large chromosomal inversions related to population divergence have been reported in *Sesbania bispinosa* [50] and Lake Malawi cichlid [51]. Nevertheless, it is still unclear what traits are affected by the large inversion in cowpea. GO enrichment analysis showed genes in the inversion region were primarily enriched

for gibberellin biosynthesis and toxin metabolism (Table S18), which are likely related to the different plant architecture and growing environments of the two subspecies. Future functional investigation of these genes will clarify how this inversion contributed to cowpea diversification.

Evidence from several plant systems indicates a strong association between chromosomal inversions and centromere movement. Pericentric inversions linked to shifts in centromere position have been reported in *Triticum monococcum* [52] and *Oryza* [53]. A similar pattern across Phaseoleae was observed in this study, where inversions and centromere repositioning frequently co-occur. Segmental duplications often cluster at inversion breakpoints in mammals [54] and other vertebrates [55], suggesting a potential role in driving inversion formation. Consistent with this, we found enrichment of SDs around inversion boundaries in cowpea. Together, these observations support a model in which SD-rich regions promote inversion formation through nonallelic homologous recombination, and the resulting pericentric inversions disrupt the structural and epigenetic stability of the ancestral centromere, thereby facilitating centromere renewal.

In summary, we generated complete T2T genome assemblies for the grain-type cowpea HJD and the vegetable-type cowpea FC6. The superior quality of the T2T genomes and their potential in improving genetic discovery were demonstrated. The T2T genomes enabled identification of inversions influencing subspecies divergence in cowpea and enhanced our understanding of evolutionary history of Phaseoleae. Analysis of repetitive sequence highlighted frequent centromeres changes with extensive content variation among Phaseoleae species. The T2T genomes in our study provide accurate and complete references for cowpea genetic improvement.

Materials and methods

Plant materials and sequencing

The cowpea cultivar HJD selected for genome sequencing was cultivated under greenhouse conditions for 30 days. Genomic DNA was extracted from young leaves using an improved CTAB method [56], and DNA integrity was verified by electrophoresis on a 0.75% agarose gel. High-quality genomic DNA was then prepared with the Pacific Biosciences SMRTbell Express Template Prep Kit 2.0 following the manufacturer's standard protocol. Each SMRTbell library was generated using the Pacific Biosciences SMRTbell Express Template Preparation Kit 2.0. The constructed libraries underwent size selection for ~15-kb fragments using BluePip-pinTM, and polymerase binding to the SMRTbell templates was performed using the DNA Polymerase Binding Kit. Subsequently, SMRT cell libraries were constructed and sequenced using the Pacific Biosciences Sequel II sequencing platform. For Oxford Nanopore ultralong sequencing, high-molecular-weight genomic DNA fragments were enriched using the Short Read Eliminator XL following Oxford Nanopore's ultralong sequencing guidelines. Libraries were then constructed using the Oxford Nanopore SQK-LSK110 Kit as per the standard protocol and sequenced on the PromethION platform. Hi-C libraries were constructed following previously established procedures [57], and sequencing data were generated using the Illumina HiSeq X Ten platform.

Genome assembly and quality evaluation

Genome sizes and heterozygosity for the HJD and FC6 genomes were estimated by Jellyfish v2.2.10 [58] in combination with GenomeScope2 [59]. Initial genome assemblies were generated using hifiasm v0.20.0 [60] in hybrid mode, integrating HiFi reads, ONT reads, and Hi-C sequencing data. Hi-C data facilitated scaffold anchoring and removal of low-quality short contigs. Due to the extremely low heterozygosity of both accessions, the primary assembly was selected as the sole input for scaffolding. Following read mapping onto the initial assembly using BWA-mem v0.7.17 [61], scaffolds were quickly ordered and oriented using the quick view mode of HapHic v1.0.6 [62]. Further refinement of genome assemblies was performed using interaction signals visualized in Juicebox, generating scaffolds with identified gaps. Concurrently, two additional initial assemblies were independently generated using verkko v1.4.1 [63] and NextDenovo v2.5.2 [64], which were subsequently employed for genome gap filling. Genome gaps were further filled using Winnowmap v2.03 [65]. For each gap bridged perfectly by both HiFi and ONT long reads, the longest and highest quality alignment was chosen to replace the missing sequence and close the gap. In total, four gaps in HJD and six in FC6 were successfully resolved (Table S28). Missing telomeric sequences were supplemented according to the procedure described by Liu *et al.* [31]. Briefly, hifiasm was used to assemble two read sets separately: (i) reads not mapped to the genome and (ii) HiFi reads containing telomeric repeats (TTAGGG). The assembled telomeric contigs were then aligned to the chromosome scaffolds with minimap2 v2.28, leveraging sequence overlaps, to generate a final genome assembly that incorporates complete telomeric sequences. After polishing with NextPolish v1.4.1 (<https://github.com/Nextomics/NextPolish>), the assembled HJD and FC6 genomes were compared against the previously published IT97K-499-35 reference genome to ascertain chromosome numbering and orientation. Ultimately, gap-free genome assemblies were achieved, and genomic interaction heatmaps were visualized using the plot mode of HapHic v1.0.6 [62].

Genome completeness was assessed using BUSCO v5.5.0 [66] based on the Fabales single-copy ortholog dataset (fabales_odb10). Genome assembly accuracy was evaluated by mapping whole-genome sequencing reads onto the HJD and FC6 assemblies using minimap2 v2.28 [67]. Alignment rates and coverage were calculated with pandepth v2.26 [68], and assembly quality values (QVs) were estimated using the k-mer-based Merqury software [69]. Genome continuity was assessed by calculating contig N50 values, while genome assembly quality was further evaluated using the LTR Assembly Index (LAI).

Whole-genome resequencing (WGS) reads from the same accessions were aligned to each assembly using BWA-MEM v0.7.17 with default parameters. Mapping rate and coverage rate were calculated from the resulting BAM files using pandepth v2.26. Assembly accuracy was evaluated with an in-house Python script (https://github.com/ChuanzhengWei/cowpea_T2T/blob/main/03.population/calculate_error_rates_fixed.py), which extracts the NM tag from each aligned read to quantify mismatches and small indels. The total number of mismatched bases divided by the total aligned query bases was used to compute the base-level error rate. Structural consistency was assessed by calculating the proportion of improperly paired reads (i.e. discordant orientation or insert size). Lower mismatch and improper-pair rates indicate

higher assembly accuracy. The same procedure was applied to all three assemblies (IT97K-499-35, HJD, and FC6), ensuring that the results are directly comparable.

Repetitive element annotation

HiTE v3.2.0 [70] was utilized for high-accuracy identification of full-length TEs, employing a rapid and precise dynamic boundary adjustment strategy with parameters: -plant 1 -recover 1 -annotate 1. HiTE integrates the strengths of both *de novo* and signature-based approaches, fully exploiting TE repeat characteristics, conserved motifs, and structural features to achieve precise detection and generate a high-quality repeat library. TE classification at the superfamily level was performed using TESorter v1.4.7 [71], with LTR-RTs further subclassified into distinct lineages. SDs in each assembled genome were identified using BISER v1.3 [72] with default parameters, employing the soft-masked versions of the respective genomes as inputs. We defined SD as repetitive segments longer than 1 kb exhibiting $\geq 90\%$ sequence identity.

Genome annotation

Following soft-masking of repetitive sequences, protein-coding genes were comprehensively predicted and annotated through an integrative approach, combining *ab initio* gene prediction, homology-based searches, and RNA-seq assembly-based predictions. Protein sequences from *Lotus japonicus*, *Medicago truncatula*, *G. max*, *P. vulgaris*, and *Vigna radiata* were employed as the protein family database. RNA-seq reads were mapped onto the reference genome using HISAT2 v2.2.1 [73] with default parameters. All data were integrated into the fully automated BRAKER3 [74] pipeline to train gene prediction models and accurately annotate genes via GeneMark-ETP and AUGUSTUS. Transcripts were assembled *de novo* using StringTie v2.2.1 [75], and the PASA v2.5.3 [76] software tool was utilized to predict and refine gene structures from these assembled transcripts through two iterative rounds, resulting in the final gene annotation models.

The completeness of the final protein dataset was assessed using BUSCO v5.0.0 in protein mode with the fabales_odb10 dataset. Structural features of gene models were further evaluated using gFACs v1.1.2 [77]. Functional annotation of protein-coding genes was initially conducted via Diamond Blastp (v2.1.9, -evalue $1e-5$) [78] searches against UniProt, NR, GO, and KEGG databases. Additionally, conserved protein sequences, functional motifs, and structural domains were characterized by employing InterProScan and Hmmscan searches against the InterPro and Pfam databases.

Identification of centromeres and telomeres

Telomeric regions were identified using seqtk v1.4 (<https://github.com/lh3/seqtk>) by searching for canonical 5'-CCCTAAA and 3'-TTAGGG repeats with the parameter seqtk telo -m CCCTAAA. Centromeric regions were defined based on the previously identified 455-bp centromere-specific satellite sequence in *V. unguiculata* by fluorescence *in situ* hybridization (FISH). We first used Blastn v2.14.0+ to search for CEN455 repeats across the genome, and candidate regions were selected using a $\geq 90\%$ sequence identity threshold. For chromosomes lacking

a complete CEN455 array, centromere positions were inferred using a combination of computational prediction and repeat-based validation. The CentroMiner module in the quarTeT toolkit [79] was used to automatically predict putative centromeric regions based on sequence composition, repeat density, and other structural features. These predictions were then cross-validated using (i) tandem repeat annotation and structural hierarchy (TRASH) v1.2 [80], which identifies and quantifies centromere-associated tandem repeats, and (ii) StainedGlass v0.6 [81], which visualizes local self-homology patterns to delineate tandem repeat arrays. The approximate centromeric boundaries were defined based on consensus results obtained from these multiple analytical approaches and were subsequently refined through manual inspection. The centromere locations in soybean were sourced from previous research [27].

Hi-C data analysis for inversion

Hi-C reads from HJD and FC6 were each aligned to the HJD reference genome using BWA v0.7.17 [61]. Optical duplicates, unmapped reads, and singleton mates were filtered out to produce the final alignment files with the command: `bwa mem -5SP -t 48 HJD.fasta reads1 reads2 | samblaster | samtools view - -@ 48 -S -h -b -F 3340`. Juicer was then used to generate .assembly and .hic files for visualization. Juicebox heatmaps of the inversion regions revealed the characteristic 'butterfly' interaction pattern indicative of inversion events.

Variant calling and population genetic analyses

A total of 270 genomic resequencing datasets (NCBI accession number PRJNA890023) were retrieved and employed for population genetic analyses. Raw sequencing reads were filtered with fastp v0.23.4 and aligned to the HJD genome assembly using BWA v0.7.17, followed by sorting and indexing of alignment files with SAMtools v1.21. SNP calling was performed using GATK v4.6.2 [82], and variants were filtered to retain high-confidence SNPs according to the criteria: $QD < 2.0$ || $QUAL < 30.0$ || $SOR > 3.0$ || $FS > 60.0$ || $MQ < 40.0$ || $MQRankSum \leq 12.5$ || $ReadPosRankSum \leq 8.0$. A total of 7 685 972 SNPs were retained for subsequent analyses. For PCA, SNPs were further filtered using VCFtools v0.1.16 ($MAF \geq 0.03$ and missing data rate ≤ 0.1), and PCA was conducted with PLINK v1.90b7 [83]. π and F_{ST} were calculated and analyzed using VCFtools v0.1.16 [83] with a window size of 50 000 bp. Accessions were grouped into the two subspecies (*V. unguiculata* and *V. sesquipedalis*) defined in a previous study to evaluate genomic differentiation between the two cultivated forms [33].

Genotyping of inversions within populations

HiFi sequencing reads from HJD and FC6 were independently aligned to the HJD genome using minimap2 v2.28. The genomic coverage within inverted regions was notably higher in the HJD_map_HJD combination compared to FC6_map_HJD. Alignment of Illumina resequencing reads revealed markedly distinct patterns of genomic coverage within inversion regions between *V. unguiculata* and *V. sesquipedalis*. Regions exhibiting

genomic coverage >94% were classified as the reference genotype (no inversion), whereas lower coverage indicated the presence of inversions. PCA was performed using 19 754 SNPs from inversion regions, and inversion genotypes were assigned based on observed clusters. When discrepancies between the two genotyping approaches occurred, inversion genotypes were manually resolved using IGV v2.16.2 [84] by inspecting read coverage patterns near inversion breakpoints.

Collinearity analysis of different genomes

HJD and FC6 genome assemblies were aligned against the reference genome IT97K-499-35 (retrieved from <https://phytozome-next.jgi.doe.gov/>) using minimap2 v2.28 [67]. Syntenic regions and structural variations between genomes were identified using Syri v1.6.5 [85]. Structural variant statistics (number and length) were calculated with custom scripts, and genome alignments were visualized using plotsr v1.1.1 [86].

Comparative genomics analysis

Single-copy orthologous protein sequences were identified with OrthoFinder v2.5.5 [87], aligned independently with MUSCLE v5.1, and then concatenated to form a superalignment matrix. A phylogenetic tree encompassing 21 species was constructed using RAxML with the maximum likelihood method, supported by 100 bootstrap replicates. Divergence time estimation based on the phylogenetic tree was performed using the MCMCTree program in PAML [88] with the following parameters: burn-in period = 10 000, number of samples = 100 000, and sampling frequency = 2. Furthermore, gene family expansion and contraction analyses were carried out using CAFÉ [89], and functional annotation of these gene families was conducted using clusterProfiler [90].

Gene-based synteny analyses were performed using JCVI v1.5.2 [91] to identify orthologous and paralogous genes among genomes. Synonymous substitution rates (K_s) of homologous gene pairs, commonly used to infer WGD events, were calculated and visualized using WGD v2.0.38 [92].

Rhizobial inoculation experiment

Seeds of HJD and FC6 were surface-sterilized using chlorine gas for 12–14 h before being planted in pots and cultivated in a growth chamber under a 16-h light/8-h dark cycle at temperatures of 27°C (day) and 22°C (night). Seedlings were inoculated with *Bradyrhizobium japonicum* USDA110 at Day 7 postgermination. *Bradyrhizobium japonicum* was cultured in liquid YMA medium at 28°C prior to inoculation. Each seedling was inoculated with 5 ml of bacterial suspension ($OD_{600} = 0.1$ prepared in distilled water), and the number of root nodules was recorded after 30 days.

Tandem repeat annotation and analyses

The TRASH pipeline v1.2 [80] was employed to identify tandem repeat arrays across all genomes. Genome-wide internal sequence similarities were computed and visualized using StainedGlass v0.6 [81] using default parameters. Monomer diversity was evaluated based on variant sites within repeat monomers. Using CEN455 as an example, all identified monomers were initially aligned using MUSCLE v5.1 to generate a consensus sequence, and a custom

script was employed to determine the number of variant bases at corresponding positions relative to the consensus. Similarity was defined based on the ‘identity’ value obtained from Blastn alignments. To investigate the evolutionary origins of monomers, proportional random sampling by species and chromosome was performed to obtain ~1800 monomers. Maximum-likelihood phylogenetic trees were subsequently constructed using FastTree v2.2 (<https://github.com/morgannprice/fasttree>), and tree structures were visualized via the chiPlot website [93].

Identification of chromosomal inversions

Inversions were identified from JCVI-based gene synteny analysis using NGenomeSyn v1.42 [94] visualization, with evident funnel-shaped patterns considered inversion regions. Breakpoints were defined as 40-kb regions flanking the genes located at inversion boundaries in the HJD genome assembly. Enrichment analyses of repeats at inversion breakpoints were performed through permutation tests involving four categories of repetitive elements: LTR/Copia, LTR/Gypsy, DNA transposons, and segmental duplications. Specifically, centromeric and telomeric regions were masked genome-wide, and an in-house script was employed to conduct 1000 permutation tests. Repeat lengths in equivalent genomic intervals provided expected values, with enrichment determined by the frequency and magnitude of observed values exceeding these expectations. In practice, we performed 1000 permutations and computed the P -value as $P = (N - k) / N$, where k is the number of permutations in which the observed repeat occupancy exceeded its expected value.

Acknowledgments

This work was financially supported by State Key Laboratory for Tropical Crop Breeding and the start-up package from Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences.

Author contributions

Y.F.T. and Y.Y. conceived the project. C.Z.W., S.C.S., Y.Z.W., and S.P. performed bioinformatic analyses. Y.Z.W., Y.B.W., and T.D. conducted plant cultivation and experiments. C.Z.W. and Y.F.T. wrote the manuscript. E.M. and D.J. supervised the work. All authors contributed to the article and approved the submitted version.

Data availability

Raw sequencing reads and assembly results for HJD have been deposited under project PRJCA044301 at the National Genomics Data Center. Protein-coding gene annotation files and additional assembly resources have been uploaded to the online repository Figshare (URL: https://figshare.com/articles/dataset/cowpea_genome_project/29878346). The scripts used for the analysis are available in a GitHub repository at https://github.com/ChuanzhengWei/cowpea_T2T.

Conflicts of interest statement

The authors declared that they have no conflict of interest in relation to this work.

Supplementary material

Supplementary material is available at *Horticulture Research* online.

References

- Ehlers JD, Hall AE. Cowpea (*Vigna unguiculata* L. Walp.). *Field Crop Res.* 1997;53:187–204
- Chen H, Chen H, Hu L. *et al.* Genetic diversity and a population structure analysis of accessions in the Chinese cowpea [*Vigna unguiculata* (L.) Walp.] germplasm collection. *Crop J.* 2017;5:363–72
- Chipeta MM, Kafwambira J, Yohane E. Cowpea genetic diversity, population structure and genome-wide association studies in Malawi: insights for breeding programs. *Front Plant Sci.* 2025;15:1461631
- Boukar O, Fatokun CA, Huynh B-L. *et al.* Genomic tools in cowpea breeding programs: status and perspectives. *Front Plant Sci.* 2016;7:757
- Pasquet RS. Allozyme diversity of cultivated cowpea *Vigna unguiculata* (L.) Walp. *Theor Appl Genet.* 2000;101:211–9
- Kouakou A-G, Ogundapo A, Smale M. *et al.* IITA's genebank, cowpea diversity on farms, and farmers' welfare in Nigeria. *CABI Agric Biosci.* 2022;3:14
- Fang J, Chao C-CT, Roberts PA. *et al.* Genetic diversity of cowpea [*Vigna unguiculata* (L.) Walp.] in four West African and USA breeding programs as determined by AFLP analysis. *Genet Resour Crop Evol.* 2007;54:1197–209
- Xu P, Wu X, Wang B. *et al.* Development and polymorphism of *Vigna unguiculata* ssp. *unguiculata* microsatellite markers used for phylogenetic analysis in asparagus bean (*Vigna unguiculata* ssp. *sesquipedialis* (L.) Verdc.). *Mol Breed.* 2010;25:675–84
- Jayathilake C, Visvanathan R, Deen A. *et al.* Cowpea: an overview on its nutritional facts and health benefits. *J Sci Food Agric.* 2018;98:4793–806
- Affrifah NS, Phillips RD, Saalia FK. Cowpeas: nutritional profile, processing methods and products—a review. *Legume Sci.* 2022;4:e131
- Della Coletta R, Qiu Y, Ou S. *et al.* How the pan-genome is changing crop genomics and improvement. *Genome Biol.* 2021;22:3
- Muñoz-Amatriaín M, Mirebrahim H, Xu P. *et al.* Genome resources for climate-resilient cowpea, an essential crop for food security. *Plant J.* 2017;89:1042–54
- Lonardi S, Muñoz-Amatriaín M, Liang Q. *et al.* The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant J.* 2019;98:767–82
- Xia Q, Pan L, Zhang R. *et al.* The genome assembly of asparagus bean, *Vigna unguiculata* ssp. *sesquipedialis*. *Sci Data.* 2019;6:124
- Yang Y, Wu Z, Wu Z. *et al.* A near-complete assembly of asparagus bean provides insights into anthocyanin accumulation in pods. *Plant Biotechnol J.* 2023;21:2473–89
- Wang X, Zhou P, Hu X. *et al.* T2T genome, pan-genome analysis, and heat stress response genes in *Rhododendron* species. *iMeta.* 2025;4:e70010
- Gao W, Wang S, Jiang T. *et al.* Chromosome-scale and haplotype-resolved genome assembly of *Populus trichocarpa*. *Hortic Res.* 2025;12:uhaf012
- Li Q, Qiao X, Li L. *et al.* Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. *Plant Commun.* 2024;5:101000
- Kajita T, Ohashi H, Tateishi Y. *et al.* rbcL and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. *Syst Bot.* 2001;26:515–36
- Lackey JA. Chromosome numbers in the Phaseoleae (Fabaceae: Faboideae) and their relation to taxonomy. *Am J Bot.* 1980;67:595–602
- Vasconcelos EV, de Andrade Fonsêca AF, Pedrosa-Harand A. *et al.* Intra- and interchromosomal rearrangements between cowpea [*Vigna unguiculata* (L.) Walp.] and common bean (*Phaseolus vulgaris* L.) revealed by BAC-FISH. *Chromosom Res.* 2015;23:253–66
- de Oliveira Bustamante F, do Nascimento TH, Montenegro C. *et al.* Oligo-FISH barcode in beans: a new chromosome identification system. *Theor Appl Genet.* 2021;134:3675–86
- Oliveira AR, Martins L, Bustamante F. *et al.* Breaks of macrosynteny and collinearity among moth bean (*Vigna aconitifolia*), cowpea (*V. unguiculata*), and common bean (*Phaseolus vulgaris*). *Chromosom Res.* 2020;28:293–306
- Cleveland DW, Mao Y, Sullivan KF. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell.* 2003;112:407–21
- do Vale Martins L, de Oliveira Bustamante F, da Silva Oliveira AR. *et al.* BAC- and oligo-FISH mapping reveals chromosome evolution among *Vigna angularis*, *V. unguiculata*, and *Phaseolus vulgaris*. *Chromosoma.* 2021;130:133–47
- Montenegro C, do Vale Martins L, Bustamante F. *et al.* Comparative cytogenomics reveals genome reshuffling and centromere repositioning in the legume tribe Phaseoleae. *Chromosom Res.* 2022;30:477–92
- Wang L, Zhang M, Li M. *et al.* A telomere-to-telomere gap-free assembly of soybean genome. *Mol Plant.* 2023;16:1711–4
- Wang Y, Hao X, Chen C. *et al.* Telomere-to-telomere genome of common bean (*Phaseolus vulgaris* L., YP4). *GigaScience.* 2025;14:giaf001
- Zhao B, Zhang H, Zhao Q. *et al.* Gap-free genome assembly and metabolomics analysis of common bean provide insights into genomic characteristics and metabolic determinants of seed coat pigmentation. *J Genet Genomics.* 2025;52:852–5
- Jia K-H, Li G, Wang L. *et al.* Telomere-to-telomere, gap-free genome of mung bean (*Vigna radiata*) provides insights into domestication under structural variation. *Hortic Res.* 2025;12:uhaf337
- Liu J, Li Q, Hu Y. *et al.* The complete telomere-to-telomere sequence of a mouse genome. *Science.* 2024;386:1141–6
- Naish M, Henderson IR. The structure, function, and evolution of plant centromeres. *Genome Res.* 2024;34:161–78
- Pan L, Liu M, Kang Y. *et al.* Comprehensive genomic analyses of *Vigna unguiculata* provide insights into population differentiation and the genetic basis of key agricultural traits. *Plant Biotechnol J.* 2023;21:1426–39
- Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol Evol.* 2018;33:427–40

35. Wang J, Sun P, Li Y. *et al.* Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* 2017;174:284–300
36. Schmutz J, Cannon SB, Schlueter J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature.* 2010;463:178–83
37. Young ND, Debellé F, Oldroyd GED. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature.* 2011;480:520–4
38. Liu Y-Y, Yu C-L, Liu Y-J. *et al.* An autotetraploid genome of *Corydalis shearerii* provides insight into the evolution and benzyloisoquinoline alkaloids diversity of *Corydalis*. *Genomics Communications.* 2025;2:e002
39. Yang Y, Du W, Li Y. *et al.* Recent advances and challenges in de novo genome assembly. *Genomics Communications.* 2025;2:e104
40. Bilgrav Saether K, Eisfeldt J, Bengtsson JD. *et al.* Leveraging the T2T assembly to resolve rare and pathogenic inversions in reference genome gaps. *Genome Res.* 2024;34:1785–97
41. Shang L, He W, Wang T. *et al.* A complete assembly of the rice Nipponbare reference genome. *Mol Plant.* 2023;16:1232–6
42. Chen J, Wang Z, Tan K. *et al.* A complete telomere-to-telomere assembly of the maize genome. *Nat Genet.* 2023;55:1221–31
43. Wei C, Gao L, Xiao R. *et al.* Complete telomere-to-telomere assemblies of two sorghum genomes to guide biological discovery. *iMeta.* 2024;3:e193
44. Kumon T, Lampson MA. Evolution of eukaryotic centromeres by drive and suppression of selfish genetic elements. *Semin Cell Dev Biol.* 2022;128:51–60
45. Hu G, Wang Z, Tian Z. *et al.* A telomere-to-telomere genome assembly of cotton provides insights into centromere evolution and short-season adaptation. *Nat Genet.* 2025;57:1031–43
46. Jin X, Du H, Chen M. *et al.* A fully phased octoploid strawberry genome reveals the evolutionary dynamism of centromeric satellites. *Genome Biol.* 2025;26:17
47. Courret C, Hemmer LW, Wei X. *et al.* Turnover of retroelements and satellite DNA drives centromere reorganization over short evolutionary timescales in *Drosophila*. *PLoS Biol.* 2024;22:e3002911
48. Zhao J, Xie Y, Kong C. *et al.* Centromere repositioning and shifts in wheat evolution. *Plant Commun.* 2023;4:100556
49. Wang Y, Zhao L, Wang D. *et al.* Four near-complete genome assemblies reveal the landscape and evolution of centromeres in Salicaceae. *Genome Biol.* 2025;26:111
50. Huang G, Wang X, Liu C. *et al.* Genomic variation underpins genetic divergence and differing salt resilience in *Sesbania bispinosa*. *Adv Sci.* 2025;12:e02600
51. Blumer LM, Burskaia V, Artiushin I. *et al.* Introgression dynamics of sex-linked chromosomal inversions shape the Malawi cichlid radiation. *Science.* 2025;388:eadr9961
52. Ahmed HI, Heuberger M, Schoen A. *et al.* Einkorn genomics sheds light on history of the oldest domesticated wheat. *Nature.* 2023;620:830–8
53. Zhang T, Bao Y, You H. *et al.* T2T assembly of *Oryza australiensis* reveals centromere repositioning mechanisms and identifies the EE genome as direct donor of the extinct DD lineage. *Research Square Preprint.* 2025. <https://doi.org/10.21203/rs.3.rs-7517070/v1>
54. Porubsky D, Sanders AD, Höps W. *et al.* Recurrent inversion toggling and great ape genome evolution. *Nat Genet.* 2020;52:849–58
55. Francis BA, Gozashti L, Costello K. *et al.* Complete genome assemblies of two mouse subspecies reveal structural diversity of telomeres and centromeres. *Nat Genet.* 2025;57:2852–62
56. Porebski S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol Biol Report.* 1997;15:8–15
57. Rao SSP, Huntley MH, Durand NC. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 2014;159:1665–80
58. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 2011;27:764–70
59. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 2020;11:1432
60. Cheng H, Jarvis ED, Fedrigo O. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol.* 2022;40:1332–5
61. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.* 2013;arXiv:1303.3997
62. Zeng X, Yi Z, Zhang X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat Plants.* 2024;10:1184–200
63. Rautiainen M, Nurk S, Walenz BP. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol.* 2023;41:1474–82
64. Hu J, Wang Z, Sun Z. *et al.* NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* 2024;25:107
65. Jain C, Rhie A, Hansen NF. *et al.* Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Meth.* 2022;19:705–10
66. Manni M, Berkeley MR, Seppey M. *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 2021;38:4647–54
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–100
68. Yu H, Shi C, He W. *et al.* PanDepth, an ultrafast and efficient genomic tool for coverage calculation. *Brief Bioinform.* 2024;25:bbae197
69. Rhie A, Walenz BP, Koren S. *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 2020;21:245
70. Hu K, Ni P, Xu M. *et al.* HiTE: a fast and accurate dynamic boundary adjustment approach for full-length transposable element detection and annotation. *Nat Commun.* 2024;15:5573
71. Zhang R-G, Li G-Y, Wang X-L. *et al.* TESorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res.* 2022;9:uhac017
72. Išerić H, Alkan C, Hach F. *et al.* Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol Biol.* 2022;17:4

73. Kim D, Paggi JM, Park C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37:907–15
74. Gabriel L, Brůna T, Hoff KJ. *et al.* BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* 2024;34:769–77
75. Pertea M, Pertea GM, Antonescu CM. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 2015;33:290–5
76. Haas BJ, Delcher AL, Mount SM. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 2003;31:5654–66
77. Lan L, Hu H, Jia Y. *et al.* Tips for improving genome annotation quality. *Genomics Communications.* 2025;2:e005
78. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Meth.* 2015;12:59–60
79. Lin Y, Ye C, Li X. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res.* 2023;10:uhad127
80. Wlodzimierz P, Hong M, Henderson IR. TRASH: tandem repeat annotation and structural hierarchy. *Bioinformatics.* 2023;39:btad308
81. Vollger MR, Kerpedjiev P, Phillippy AM. *et al.* Stained-Glass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics.* 2022;38:2049–51
82. Van der Auwera GA, Carneiro MO, Hartl C. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics.* 2013;43:11.10.1–11.10.33
83. Chang CC, Chow CC, Tellier LCAM. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* 2015;4:s13742-015-0047-8
84. Robinson JT, Thorvaldsdóttir H, Winckler W. *et al.* Integrative genomics viewer. *Nat Biotechnol.* 2011;29:24–6
85. Goel M, Sun H, Jiao W-B. *et al.* SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 2019;20:277
86. Goel M, Schneeberger K. Plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics.* 2022;38:2922–6
87. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20:238
88. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24:1586–91
89. Mendes FK, Vanderpool D, Fulton B. *et al.* CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics.* 2021;36:5516–8
90. Wu T, Hu E, Xu S. *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation.* 2021;2:100141
91. Tang H, Krishnakumar V, Zeng X. *et al.* JCVI: a versatile toolkit for comparative genomics analysis. *iMeta.* 2024;3:e211
92. Chen H, Van Zwaenepoel A, de Peer Y. Wgd v2: a suite of tools to uncover and date ancient polyploidy and whole-genome duplication. *Bioinformatics.* 2024;40:btad272
93. Xie J, Chen Y, Cai G. *et al.* Tree visualization by one table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Res.* 2023;51:W587–92
94. He W, Yang J, Jing Y. *et al.* NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. *Bioinformatics.* 2023;39:btad121