Received: 30 September 2024 | Revised: 4 March 2025 | Accepted: 18 March 2025

DOI: 10.1002/bod2.70001

ORIGINAL ARTICLE

Telomere-to-telomere reference genome reveals subgenome divergence and large structural variants in *Capsella bursa-pastoris* (Brassicaceae)

Xin Liu[1,2,3,4] | **Lu Liu**[1,2] | **Hai-Fei Yan**[1,2] | **Xue-Jun Ge**[1,2,5] | **Hui-Run Huang**[1,2]

[1]Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China

[2]South China National Botanical Garden, Guangzhou, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Stress Biology, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

[5]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China

Correspondence
Hui-Run Huang.
Email: huirun.huang@scbg.ac.cn

Funding information
Basic and Applied Basic Research Foundation of Guangdong Province, Grant/Award Number: 2021A1515012487; National Natural Science Foundation of China, Grant/Award Number: 31870353

Abstract

Capsella is a model genus for studying polyploid speciation and mating system transitions. Here, we report a subgenome-resolved telomere-to-telomere reference genome of *Capsella bursa-pastoris*, the only polyploid species in the genus, generated by incorporating PacBio HiFi reads, Nanopore ultra-long reads, and Hi-C data. The assembled genome, named Cbp_T2T, consisted of 16 chromosomes, with 8 in each of the two subgenomes. The genome size of the final assembly was 330.91 Mb. Multiple evaluation methods confirmed the contiguity (contig N50: 20.28 Mb), completeness (Benchmarking Universal Single - Copy Orthologue: 98.70%), and accuracy (QV: 35.76) of the genome. Repetitive sequences accounted for 37.04% of the genome size. In total, 27,626 and 28,096 protein-coding genes were annotated in the two subgenomes, respectively. Large structural variants were observed between the two subgenomes. Additionally, two large inversions on chromosomes 6 and 7, spanning approximately 1.93 and 2.33 Mb, respectively, were detected by comparing Cbp_T2T to a previously published chromosome-scale *C. bursa-pastoris* assembly. This high-quality reference genome will facilitate evolutionary biological and functional genomic studies in *Capsella* species.

KEYWORDS
Capsella bursa-pastoris, inversion, structural variant, subgenome divergence, telomere-to-telomere

Highlights

- Cbp_T2T is the first telomere-to-telomere reference genome in the genus *Capsella*, with multiple evaluations confirming its high quality.
- Significant divergences were observed between the subgenomes of Cbp_T2T.
- Genome comparison revealed two large inversions within *C. bursa-pastoris*.

Xin Liu and Lu Liu have contributed equally to the work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). Biological Diversity published by John Wiley & Sons Australia, Ltd on behalf of South China Botanical Garden (SCBG), Chinese Academy of Sciences (CAS).

1 | INTRODUCTION

The genus *Capsella* is considered an interesting model for plant evolutionary genetics because it presents two major trends in the evolution of flowering plants: the process of polyploidization and the transition from outbreeding to selfing.[1] This genus comprises a primarily self-compatible tetraploid *Capsella bursa-pastoris* as well as three diploids: the self-incompatible outcrosser *Capsella grandiflora* and two self-compatible species, *Capsella rubella*, and *Capsella orientalis*. These species differ greatly in their geographical distribution. *C. grandiflora* is limited to northwestern Greece and Albania, whereas *C. rubella* has a broader Mediterranean/central European distribution, and *C. orientalis* is found from Eastern Europe to Central Asia; in contrast, *C. bursa-pastoris* has spread or been introduced from Europe to all continents except Antarctica.[2–4] *Capsella* species are highly amenable to molecular and genomic analysis, particularly due to their close relatedness to *Arabidopsis thaliana*,[5,6] the best-understood plant model system. Therefore, this genus provides an ideal model for addressing evolutionary questions such as transitions in mating systems, polyploid speciation, subgenome evolution, and local adaptation to diverse habitats. The use of *Capsella* genomes would undoubtedly facilitate such studies. However, compared with *Arabidopsis*, high-quality reference genomes of *Capsella* species are still under-resolved.

C. bursa-pastoris is the only polyploid species in the genus *Capsella*. It is widely accepted that the species originated from the hybridization, *ca*. 100,000 to 300,000 years ago, of the two ancestral lineages: *C. orientalis* and *C. grandiflora*/*C. rubella*.[4,7] Within *C. bursa-pastoris*, morphological and life-history traits vary extensively.[8–11] In particular, there is a remarkable degree of ecotypic differentiation in flowering time among different local populations.[12–15] Thus, *C. bursa-pastoris* provides an outstanding opportunity to address important biological questions related to speciation, subgenome divergence, and local adaptation. Previously, two genome assemblies of *C. bursa-pastoris* have been released. Kasianov et al.[16] reported a de novo genome assembly of *C. bursa-pastoris* using short reads from Illumina HiSeq and MiSeq platforms, resulting in a genome size of approximately 268.70 Mb across 8362 scaffolds, of which 229.97 Mb was separated into subgenomes with a scaffold N50 of 744.15 kb. Most recently, Penin et al.[17] presented a subgenome-resolved chromosome-scale assembly with a total size of approximately 252.98 Mb and a scaffold N50 of 16.56 Mb, using a combination of PacBio high-fidelity (HiFi) reads, Hi-C data, and genetic mapping. These two genome assemblies have provided essential resources for advancing research on the evolution of the genus *Capsella*. However, given the 1C nuclear DNA content estimated at 391.20 Mb using flow cytometry,[18] some complex genomic regions, particularly repetitive sequences and structural variants, remain under-resolved in both assemblies.

In this study, we developed a telomere-to-telomere and nearly complete reference genome (Cbp_T2T) of the *C. bursa-pastoris* accession ZZH289 consisting of two subgenomes (CbpA and CbpB) by integrating PacBio HiFi reads, Oxford Nanopore Technology (ONT) ultra-long reads, and Hi-C reads. Multiple assessment methods were applied to evaluate the quality of this subgenome-resolved reference genome. The two primary assemblies harbored an initial contig N50 > 11 Mb, reaching approximately 300 × of the previously reported *C. bursa-pastoris* genome[16] (hereafter Cbp_version1) and around 10 × of the recently released one[17] (hereafter Cbp_version2), indicating a large increase in continuity. The genome size of the final assembly was estimated to be 330.91 Mb, much larger than the previous two versions (Cbp_version1: 229.97 Mb and Cbp_version2: 252.98 Mb). The utilization of multiple sequencing data combined with advanced assembly algorithms has emerged as a promising approach for addressing complex regions in the genome of this allotetraploid species, thus resulting in a complete reference genome with an assembly size closer to the expected size (322.06 Mb). Significant structural variations were identified between the subgenomes of Cbp_T2T, with centromeric and pericentromeric regions displaying lower nucleotide sequence collinearity compared to the chromosome arms. In addition, two large inversions were identified when comparing our Cbp_T2T genome assembly with Cbp_version2. The first inversion is located on chromosome 7 in the subgenome related to *C. grandiflora*/*C. rubella* lineage. It spans approximately 2.33 Mb in Cbp_T2T and 2.62 Mb in Cbp_version2. The second, on chromosome 6 in the *C. orientalis*-related subgenome, covers about 1.93 Mb in Cbp_T2T and 2.06 Mb in Cbp_version2. Overall, the characterization of the subgenome-resolved T2T genome highlighted previously unknown features of the *C. bursa-pastoris* genome. This high-quality genome will contribute to a better understanding of *Capsella* evolution, polyploidization, and adaptation.

2 | MATERIALS AND METHODS

2.1 | Sample collection and sequencing

The seeds of the *C. bursa-pastoris* accession ZZH289 were sampled from Zhengzhou, Henan, China (N34°45′, E113°40′). Seeds were stratified on moist filter paper for 1 week at 4°C and subsequently transferred to soil in pots, which were subsequently placed in a growth chamber at 23°C under 16 h light/8 h dark cycles. Tissue samples, including fresh leaves, stems, and roots, were harvested. Then, the tissues were immediately frozen in liquid nitrogen and preserved at −80°C for further DNA/RNA extraction.

For PacBio HiFi sequencing, high-molecular-weight genomic DNA was prepared from leaf samples using the Cetyltrimethylammonium Bromide method. DNA integrity was checked using gel electrophoresis. The concentration and purity of the extracted DNA were evaluated using a Qubit fluorometer (Thermo Fisher Scientific). A total of 5 μg of DNA was used for library construction. A standard SMRTbell library was constructed for sequencing with the SMRTbell Express Template Prep Kit 2.0 according to the manufacturer's protocols (Pacific Biosciences, CA, USA). The library was subsequently sequenced on a PacBio Sequel II system at Biozeron Biotechnology (Shanghai, China).

For ONT sequencing, ultra-long DNA was extracted using the Sodium Dodecyl Sulfate method without a purification step to maintain the length of the DNA. Approximately 8–10 μg of DNA was size-selected (>50 kb) with the SageHLS High Molecular Weight library system (Sage Science, USA) and used for library construction with the Oxford Nanopore SQK-LSK109 Kit following the manufacturers' instructions. The library was subsequently

sequenced on the Nanopore PromethION platform at Grandomics Biosciences (Beijing, China).

For Hi-C sequencing, the genomic DNA was cross-linked with formaldehyde and extracted for Hi-C library preparation. The library was paired-end sequenced on an MGI sequencing platform (Shenzhen, China). For DNAseq, paired-end 150-bp reads were generated on the MGI DNBSEQ-T7 platform. For RNAseq, total RNA was isolated from all tissue samples using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs, MA, USA). Paired-end 150-bp reads were also generated on the DNBSEQ-T7 platform. The above sequencing procedure was carried out at Grandomics Biosciences (Beijing, China).

2.2 | De novo haplotype-resolved genome assembly

Based on the obtained highly accurate HiFi reads, the K-mer distribution was analyzed using jellyfish[19] with jellyfish count -C -m 21 -s 100,000,000 and jellyfish histo -h 1,000,000. The results were subsequently imported to GenomeScope v2.0[20] with K-mer length = 21 and Ploidy = 4. Because *C. bursa-pastoris* is a predominantly selfing tetraploid species with disomic inheritance,[2,21] two homeologous genomes are expected. The assembly workflow is shown in Supporting Information S1: Figure S1. First, Hi-C reads were filtered using fastp v0.23.2[22] with default parameters. Then, hifiasm v0.16.1-r375[23] was used to generate the primary phased draft genome based on HiFi, ONT and Hi-C reads with -s 0.3. Genome assembly statistics were performed using QUality ASsement Tool[24] with default parameters. The cumulative lengths of the two primary assemblies were 255.32 and 209.93 Mb, with N50 values of 11.41 and 20.12 Mb, respectively. These assembler outputs were further polished using NextPolish v1.4.1[25] with HiFi and Illumina reads, followed by the reassignment of allelic contigs with the pipeline Purge Haplotigs.[26] Subsequently, ragtag v2.1.0[27] with default parameters was used to sort, orientate, and cluster the contigs guided by the reference genome of *C. rubella*.[28] Moreover, the two haploid assemblies were combined, and the contigs were anchored into 16 pseudochromosomes using Juicer v1.6[29] and 3D-DNA v180922[30] in turn. Then, based on the assembly file obtained from the ragtag and the hic file from Juicer and 3D-DNA, Juicebox v2.20.00[31] was introduced for visualizing Hi-C data, chromosome pair assignment, and manual correction to obtain a high-quality reference genome. For gap filling, ONT and HiFi reads were mapped to the genome using minimap2 v2.24-r1122[32]; then, we examined the breakpoints with the Integrative Genomics Viewer (IGV) tool[33] and manually filled the gaps based on the alignment results. To assign the chromosomes to subgenomes, we used sequences of the parental species *C. rubella* (SRA: SRR8394204) and *C. orientalis* (SRA: SRR8904471). We first sampled sequences by the same number of reads, then aligned the reads to the 16 chromosomes and calculated the coverage on 100-kb windows using Mosdepth.[34]

2.3 | Genome annotation

Telomere regions were identified with A Telomere Identification toolKit v0.2.1 (https://github.com/tolkit/telomeric-identifier) using seven-base repeats (CCCATTT at the 5′ end and TTTAGGG at the 3′ end) as queries. Short tandem repeats are highly enriched in centromeric regions of plants. By identifying these unique features, centromeric regions can be located. Using Tandem Repeats Finder v4.09,[35] the approximate locations of the centromeres in the Cbp_T2T assembly were identified. We used the CentroMiner module in quarTeT v1.1.7[36] to autoannotate centromeres. The EDTA v2.2.0 pipeline[37] was used to identify and group repeat elements in the Cbp_T2T assembly with default parameters. For gene annotation, we used funannotate v1.8.7 (https://github.com/nextgenusfs/funannotate) pipeline with RNAseq based, homology based, and de novo based strategies. Protein-coding genes were assigned to functional descriptions using eggNOG-mapper v2[38] based on the eggNOG v5.0 database.[39]

2.4 | Quality assessment

Multiple methods were used to evaluate the quality of the reference genome. The HiFi reads were mapped to the assembly using minimap2, whereas the Illumina reads were aligned with Burrows - Wheeler Aligner v0.7.17-r1188.[40] The mapping rate was calculated using BamTools v2.5.1.[41] Benchmarking Universal Single - Copy Orthologue (BUSCO) v5.5.0[42] was used for the evaluation of assembly completeness of the haplotype-resolved genome with the "embryophyta_odb10" database. In addition, the consensus quality value (QV) of the genome was assessed using Merqury v1.3[43] with meryl $k = 19$ count.

2.5 | Whole genome alignment and syntenic analysis

Nucmer v4.0.0rc1[44] was used to determine the syntenic relationship between CbpA and CbpB with default parameters. The delta-filter was then applied using the parameters "-i 95 -l 5000." Structural variants between the two subgenomes were identified using SyRI v1.6.346[45] with default settings. Similarly, Cbp_T2T was aligned to Cbp_version2 (National Center for Biotechnology Information (NCBI) BioProject: PRJNA352734) using Nucmer, and structural variants were identified using SyRI. *C. bursa-pastoris* is an allotetraploid species, with one subgenome more closely related to the *C. grandiflora/C. rubella* lineage and the other more closely related to *C. orientalis*. The comparison was performed separately for each subgenome to capture their specific divergences. To validate large structural variants (>1 Mb), PacBio HiFi reads were aligned to the assembly with minimap2, and the alignments around structural variant breakpoints were manually inspected with the IGV tool.

3 | RESULTS AND DISCUSSION

3.1 | Assembly of highly contiguous and phased reference genome

We successfully generated a high-quality *C. bursa-pastoris* genome assembly by integrating PacBio HiFi, ONT, and Hi-C sequencing technologies. The assembly process began with the generation of

22.81 Gb of PacBio HiFi data, yielding an N50 of 14.86 kb and covering approximately 69 × of the estimated genome size (Supporting Information S1: Table S1). In addition, 79.88 Gb of ONT reads were obtained, with an N50 of 86.41 kb, providing ~241 × coverage of the genome (Supporting Information S1: Table S1). To further enhance the assembly, 86.44 Gb of Hi-C data (~261 × coverage) and 49.10 Gb of DNAseq reads (~148 × coverage) were obtained (Supporting Information S1: Table S1). To support gene annotation, we generated RNAseq data from three different tissue libraries, resulting in 8.16, 8.58, and 9.76 Gb of reads, respectively (Supporting Information S1: Table S1).

The heterozygosity rate was estimated to be 4.19% (Supporting Information S1: Figure S2), with the aabb heterozygous type predominating at 4.14%, consistent with *C. bursa-pastoris* being a selfing tetraploid species with disomic inheritance.[2,21] Given its tetraploid nature, the genome size was expected to be twice the predicted haploid length (161.03 Mb), resulting in an estimated genome size of 322.06 Mb. This estimate is about 18% smaller than the 1C nuclear DNA content (391.20 Mb) obtained from flow cytometry.[18] Following Hi-C scaffolding and subgenome assignment, we assembled a highly contiguous, nearly complete, and phased reference genome named Cbp_T2T. The genome encompasses 16 chromosomes (8 for each subgenome) and spans a total length of 330.91 Mb, which is significantly longer than the estimated size of 322.06 Mb based on K-mer distribution, with only 19 gaps remaining (Figure 1, Table 1). The sizes of the subgenomes were determined to be 163.68 and 167.23 Mb.

Chromosomal coverage plots revealed that most chromosomes originated from one parental species, with the exception of chromosome 2, where homeologous exchange occurred between the two subgenomes (Supporting Information S1: Figure S3). This homeologous exchange was also observed in the previously published chromosome-scale assembly, Cbp_version2,[17] confirming that it is not an assembly artifact. For chromosome 2, subgenome assignment was based on the average coverage of *C. rubella* and *C. orientalis* accessions. We designated the subgenomes as CbpA and CbpB. CbpA, with pseudochromosome lengths ranging from 17,452,651 to 25,540,302 bp, is more closely related to the *C. grandiflora*/*C. rubella* lineage, whereas CbpB, with pseudochromosome lengths ranging from 17,411,671 to 26,347,669 bp, is more related to *C. orientalis* (Table 2). Notably, the contiguity of Cbp_T2T was significantly improved compared to earlier versions (Cbp_version1 and Cbp_version2). The scaffold N50 increased from 744.15 kb to 16.56 Mb in previous assemblies to 20.28 Mb in Cbp_T2T (Table 3), with six chromosomes each represented by a single contig without gaps (Table 2). Importantly, the cumulative genome size in Cbp_T2T is closer to the estimated genome size than in earlier assemblies (Cbp_version1: 229.97 Mb and Cbp_version2: 252.98 Mb), suggesting that the new assembly better captures complex genomic regions (Table 3). This improved assembly offers a more accurate and comprehensive representation of the *C. bursa-pastoris* genome, facilitating future studies on its evolution and genetics.

3.2 | Genome annotation and quality assessment

We identified a total of 28 telomeres and 12 telomere-to-telomere (T2T) chromosomes in the *C. bursa-pastoris* genome assembly Cbp_T2T. Of these, seven chromosomes are from subgenome CbpA and five from subgenome CbpB (Supporting Information S1: Table S2). The number of telomere repeats varied between the two subgenomes, with copy numbers ranging from 235 to 651 in CbpA and 351 to 651 in CbpB. The longest telomere sequence was located at the right end of chromosomes CbpA3 and CbpB5, whereas the shortest was found at the right end of CbpA1. Additionally, all centromeric regions were successfully captured, with sizes ranging from 1,959,688 to 5,320,005 bp in CbpA and from 1,307,211 to 7,075,834 bp in CbpB (Figure 1A, Supporting Information S1: Table S3). Resolving centromeres in plant genomes is a significant challenge, as these regions are often highly repetitive. The successful identification of the complete centromeric region reflects the high contiguity and quality of the Cbp_T2T assembly.

Using the Cbp_T2T assembly for de novo repeat library construction, we identified 122.57 Mb of repetitive regions, representing 37.04% of the genome. This proportion is slightly higher than the repeat content estimated in closely related diploid *Capsella* species (*C. rubella*, 30%; *C. grandiflora*, 27%; and *C. orientalis*, 25%)[46] but lower than that of many plant genomes.[47–51] Among the detected repeats, the largest fraction (21.87%) could not be classified into a known repeat family and was categorized as *repeat_region* (unknown repeat class), suggesting the presence of lineage-specific or poorly characterized repetitive sequences in *C. bursa-pastoris* (Table 4). Long terminal repeat retrotransposons (LTR-RTs) comprised 7.69% of the genome, whereas helitrons accounted for 4.95%, making them the most abundant repeat classes. In total, 55,722 high-confidence protein-coding genes were predicted in the Cbp_T2T assembly, slightly more than the 53,502 genes predicted in Cbp_version1, but fewer than the 63,776 identified in Cbp_version2. Of these, 27,626 were assigned to CbpA and 28,096 to CbpB. The observed differences in gene numbers across these assemblies may be attributed to improvements in assembly quality, gene prediction methods, and differing criteria for gene annotation. The total length of the predicted genes was 120.89 Mb, with an average gene length of 2170 bp. Functional annotation using eggNOG-mapper successfully assigned functions to 49,873 (89.50%) of the predicted genes, indicating a high level of gene annotation completeness. This high percentage of functionally annotated genes supports the accuracy of the gene predictions and highlights the biological relevance of the assembly.

The quality and completeness of the Cbp_T2T assembly were further evaluated using multiple approaches. Hi-C data, visualized with JuiceBox, demonstrated high consistency across all chromosomes, confirming the accuracy of chromosome ordering and orientation (Figure 1B). The high levels of sequence alignment to the assembly, with 96.37% of HiFi reads, 95.78% of DNAseq reads, and 99.58% of RNAseq reads successfully mapped to Cbp_T2T, further validate the robustness of the assembly. Additionally, the assembly captured 98.70% of the conserved core eukaryotic gene set, with 96.41% duplicated BUSCOs (Supporting Information S1: Table S4). The high proportion of duplicated BUSCOs reflects the disomic inheritance of *C. bursa-pastoris*.[2,21] In terms of accuracy, the consensus QV of the genome was estimated at 35.76, corresponding to genome accuracy greater than 99.97% (Supporting Information S1: Table S5). This extremely high accuracy underscores the reliability of the

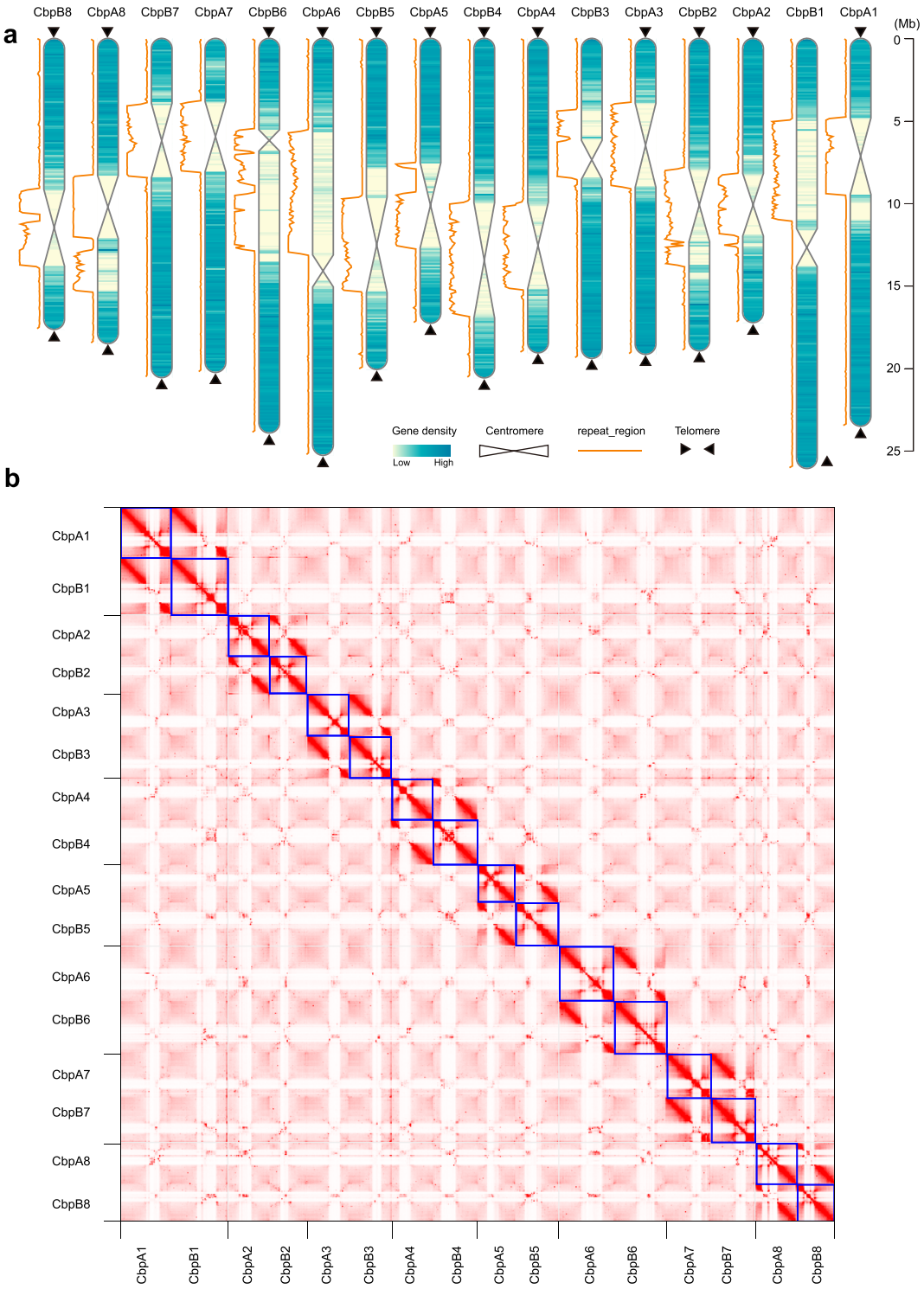


FIGURE 1 Overview of the *Capsella bursa-pastoris* genome assembly Cbp_T2T. (A) Eight chromosome pairs were defined. Gene density, repeat density, centromeres, and telomeres were characterized. (B) the Hi-C heatmap of Cbp_T2T. Each blue box represents a chromosome.

assembly for downstream analyses, such as comparative genomics, functional genomics, and evolutionary studies. Overall, the Cbp_T2T genome represents a significant improvement over previous assemblies of *C. bursa-pastoris*. The increase in contiguity, especially the recovery of telomere-to-telomere chromosomes and centromeres, ensures that the assembly provides a more comprehensive view of the genome structure. This level of detail is crucial for understanding genome evolution, particularly in polyploid species such as *C. bursa-pastoris*.

3.3 | Whole genome comparison and syntenic analysis

In this study, we first conducted a detailed comparison between the two subgenomes, CbpA and CbpB. A total of 191 translocations were identified, covering a cumulative size of 4.88 Mb (~2.98%) in CbpA and 5.13 Mb (~3.07%) in CbpB (Supporting Information S1: Table S6). Additionally, 48 inversions were detected, spanning 10.08 Mb (~6.16%) in CbpA and 8.74 Mb (~5.23%) in CbpB. Duplications were

prevalent, with 1186 events identified in CbpA, covering 31.67 Mb (~19.35%), and 982 duplications in CbpB, covering 21.65 Mb (~12.95%). Beyond these structural variants, we observed significant sequence polymorphism between the subgenomes, with 1,195,962 SNPs, 157,478 insertions, and 138,863 deletions. The syntenic analysis revealed that while the overall collinearity between the two subgenomes was maintained in chromosome arms, centromeric and pericentromeric regions showed lower levels of nucleotide sequence collinearity (Figure 2). Two large inversions were characterized in the arms of chromosomes 6 and 7 (Figure 2). On chromosome 6, the inversion spanned approximately 1.98 Mb in CbpA and 2.04 Mb in CbpB, whereas on chromosome 7, it measured around 2.20 Mb in CbpA and 2.19 Mb in CbpB. This level of structural complexity was not observed in Cbp_version2, where the two subgenomes were largely collinear, with no massive deletions, insertions, or rearrangements.[17] Additionally, several large inversions (>300 kb) were identified, including two in the centromeres of chromosomes 2 and 4 (Supporting Information S1: Figure S4). Although Hi-C data could not validate them due to low signals in these regions, the read alignments around the inversion breakpoints supported their existence (Supporting Information S1: Figure S5). Besides, the QV scores of chromosomes 2 and 4 all exceed 35 (Supporting Information S1: Table S5), and except for chromosome 4 in CbpA, the other three chromosomal assemblies are each represented by a single gapless contig (Table 2), confirming both the accuracy and contiguity of these chromosomes. Overall, this suggests that long-read sequencing is effective in resolving complex centromeric regions, producing a highly contiguous *C. bursa-pastoris* reference genome. Moreover, in addition to the substantial variations found between CbpA and CbpB, the significant differences in repeat content between CbpA and CbpB (Supporting Information S1: Table S7) further support the disomic

TABLE 1 Assembly quality metrics estimated using QUAST.

Assembly	Cbp_T2T	CbpA	CbpB
Number of contigs	16	8	8
Largest contig (bp)	26,347,669	25,540,302	26,347,669
Total length (bp)	330,911,955	163,683,258	167,228,697
GC (%)	35.58	35.56	35.61
N50	20,280,264	19,358,296	20,783,564
N90	17,452,651	17,452,651	17,411,671
L50	8	4	4
L90	15	8	8
N's per 100 kb	2.87	3.05	2.69

Abbreviations: Cbp, *Capsella bursa-pastoris*; GC, Guanine - Cytosine; T2T, telomere-to-telomere; QUAST, QUality ASsement Tool.

TABLE 2 The lengths of the pseudochromosomes of the *Capsella bursa-pastoris* genome.

Subgenome	CbpA		CbpB	
Chromosome	Length (bp)	Number of contigs	Length (bp)	Number of contigs
Chr1	23,761,493	2	26,347,669	3
Chr2	19,144,271	1	17,411,671	1
Chr3	19,358,296	2	19,603,588	5
Chr4	19,262,976	4	20,798,578	1
Chr5	17,452,651	2	20,280,264	2
Chr6	25,540,302	4	24,172,982	1
Chr7	20,463,744	1	20,783,564	1
Chr8	18,699,525	2	17,830,381	3

Abbreviation: Cbp, *Capsella bursa-pastoris*.

TABLE 3 Comparison of *Capsella bursa-pastoris* genome assemblies.

Assembly	Cbp_T2T	Cbp_version2[17]	Cbp_version1[16]
Assembly data	PacBio HiFi reads	PacBio HiFi reads	Illumina HiSeq reads
	ONT ultra-long reads	Hi-C reads	MiSeq reads
	Hi-C reads	Genetic map	Genetic map
Number of chromosomes	16	16	-
Total length (Mb)	330.91	252.98	229.97
Initial contig N50 (kb)	11,941	1251	38
Scaffold N50	20,280,264	16,555,871	744,145
BUSCOs	98.70%	98.10%	-
Number of protein-coding genes	55,722	63,776	53,502

Abbreviations: BUSCOs, Benchmarking Universal Single - Copy Orthologue; Cbp, *Capsella bursa-pastoris*; HiFi, high-fidelity; ONT, Oxford Nanopore Technology; and T2T, telomere-to-telomere.

inheritance of this allotetraploid species and suggest distinct evolutionary paths for the ancestral genomes.

To analyze the divergence between Cbp_T2T and the previously published chromosome-scale genome Cbp_version2,[17] we also conducted a comparative analysis. Cbp_T2T represents a genome assembly from a Chinese accession, whereas Cbp_version2 is from a Russian accession, with both originating from distinct genetic clusters.[52] We identified 149 translocations in Cbp_T2T, totaling 1.21 Mb (~0.37%), and 1.22 Mb (~0.48%) in Cbp_version2, along with 44 inversions spanning 7.44 Mb (~2.25%) in Cbp_T2T and 8.19 Mb (~3.24%) in Cbp_version2 (Supporting Information S1: Table S8). In addition, 2,029,446 SNPs, 354,850 insertions, and 250,129 deletions were identified. Regarding duplications, 228 events (totaling 1.82 Mb, ~0.55%) in Cbp_T2T and 151 events (1.03 Mb, ~0.41%) in Cbp_version2 were identified (Supporting Information S1: Table S8). Additionally, 738 unaligned regions were identified in Cbp_T2T, with a cumulative size of 117.34 Mb (~35.46%), compared to 22.17 Mb (~8.76%) in Cbp_version2. Notably, while the proportions of SNPs, insertions, and deletions were similar in the comparisons of Cbp_T2T versus Cbp_version2 and CbpA versus CbpB, structural variations were more abundant between CbpA and CbpB (Supporting Information S1: Tables S6 and S8), and their repeat content also showed significant divergence (Supporting Information S1: Table S7). These data indicate that Cbp_T2T and Cbp_version2 are less divergent than the two subgenomes, CbpA and CbpB. This is consistent with previous findings that *C. bursa-pastoris* retained a strong signature of parental legacy,[7] indicating that the divergence between the ancestral species likely accounts for the greater divergence observed between the subgenomes. However, we also cannot rule out the possibility that the smaller divergence between Cbp_T2T and Cbp_version2 may be due to differences in the completeness of the two genome assemblies, as Cbp_T2T is approximately 77.93 Mb longer than Cbp_version2. The lower levels of

TABLE 4 Summary of repetitive regions of the *Capsella bursa-pastoris* genome.

Class		Count	bpMasked	%Masked
LTR	Copia	5960	7,776,114	2.35%
	Gypsy	25,479	10,520,750	3.18%
	Unknown	10,833	7,142,930	2.16%
TIR	CACTA	3274	1,517,716	0.46%
	Mutator	14,450	5,270,928	1.59%
	PIF_Harbinger	2323	719,538	0.22%
	Tc1_Mariner	493	233,183	0.07%
	hAT	878	316,902	0.10%
non-LTR	LINE_element	360	234,331	0.07%
	Unknown	116	83,648	0.03%
non-TIR	Helitron	37,925	16,377,142	4.95%
repeat_region		447,125	72,381,250	21.87%
Total		549,216	122,574,432	37.04%

Abbreviations: CACTA, CACTA transposable element; LINE, long interspersed nuclear element; LTR, long terminal repeat; PIF, points - inguadona - foldback; TIR, terminal inverted repeat.

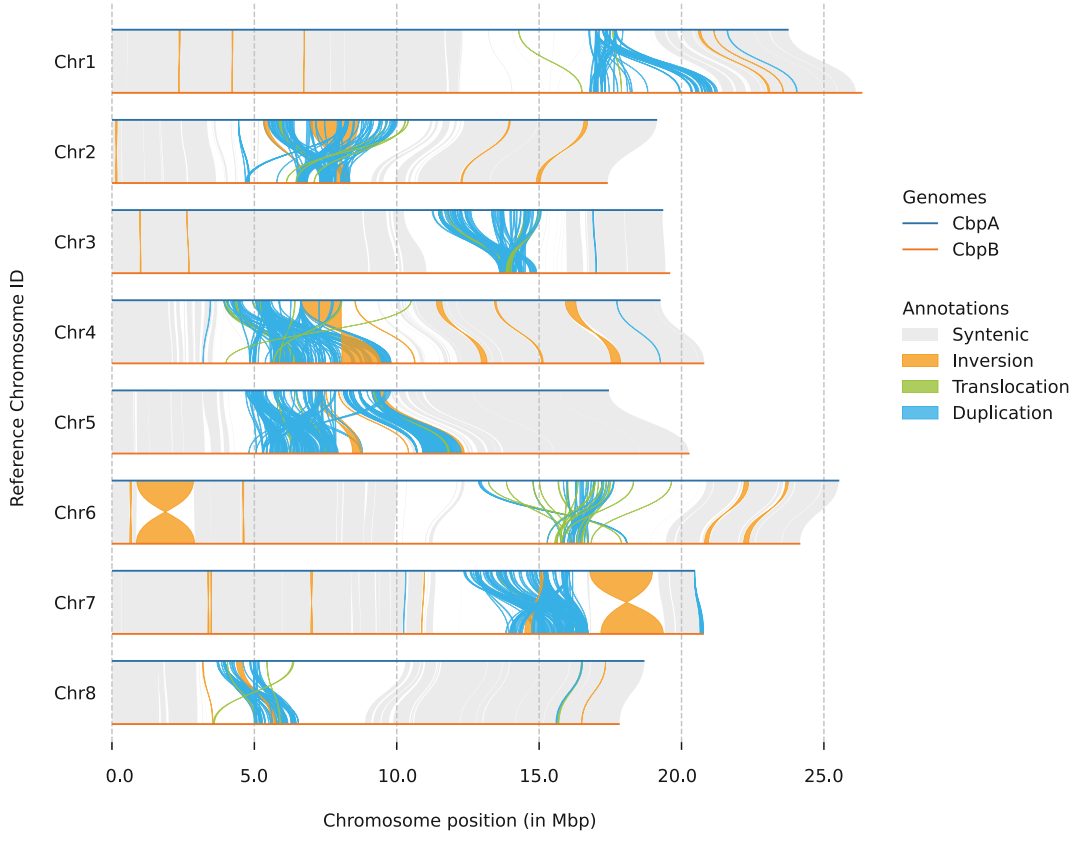


FIGURE 2 Syntenic analysis of the two subgenomes of Cbp_T2T. Syntenic, inversion, translocation, and duplication blocks were marked with corresponding colors. Frequent structural variants appear in centromeric and pericentromeric regions between the two subgenomes.

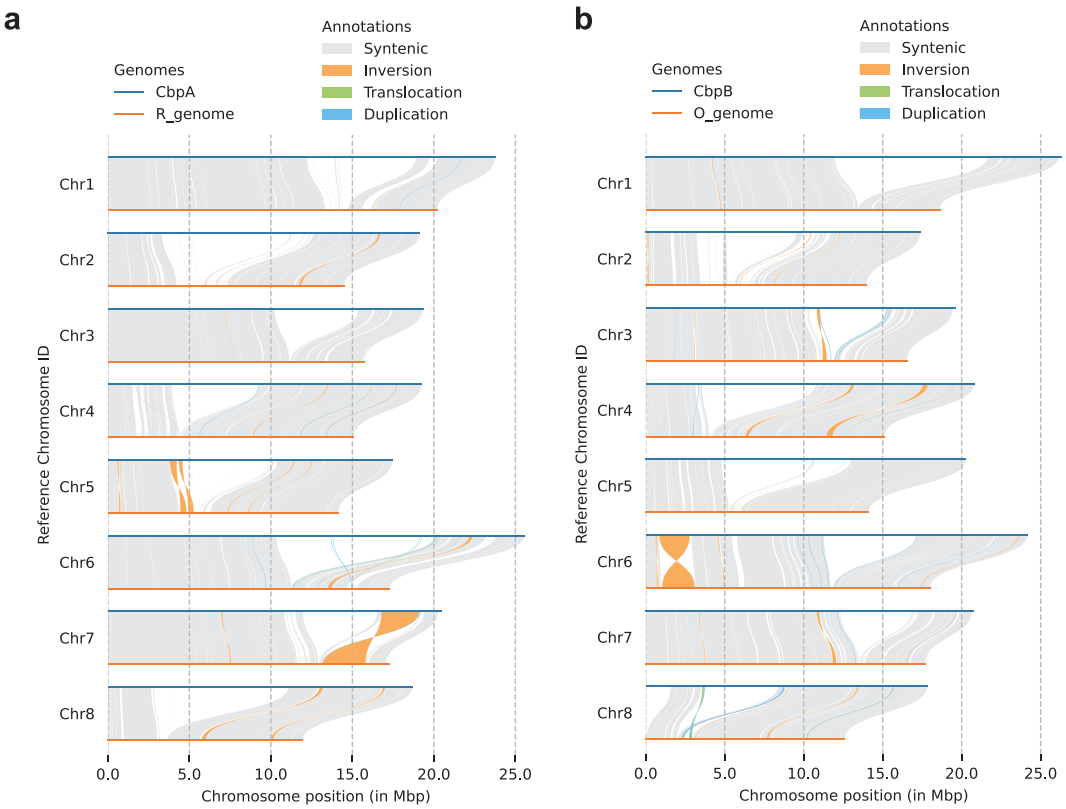


FIGURE 3 Syntenic relationship between Cbp_T2T and Cbp_version2. Syntenic, inversion, translocation, and duplication blocks were marked with corresponding colors. The left and right panels indicate the analyses on Cg_subgenome (A) and Co_subgenome (B).

duplications and the much higher proportion of unaligned regions in Cbp_T2T could be due to the large unresolved centromeric regions in Cbp_version2.

The studied species, *C. bursa-pastoris*, is an allotetraploid with one subgenome more closely related to the *C. grandiflora*/*C. rubella* lineage and the other to *C. orientalis* (hereafter referred to as Cg_subgenome and Co_subgenome). To capture their individual divergences, we compared the subgenomes separately. Overall, Cbp_T2T and Cbp_version2 are largely collinear (Figure 3); however, we identified several significant structural variants. In the Cg_subgenome, a large inversion on chromosome 7 was detected, spanning approximately 2.33 Mb in Cbp_T2T compared to about 2.62 Mb in the previously published assembly. In the Co_subgenome, a similar large inversion was observed on chromosome 6, with the inversion spanning 1.93 Mb in Cbp_T2T, versus nearly 2.06 Mb in Cbp_version2. These inversions were confirmed through manual inspection of HiFi read alignments at the breakpoints (Supporting Information S1: Figure S6), suggesting that structural variants have emerged among *C. bursa-pastoris* accessions from different genetic clusters. Notably, in chromosome 5 of Cbp_version2, a small fragment at the beginning of the Cg_subgenome (R genome) had no homeologous counterpart in the Co_subgenome (O genome), attributed to a deletion after the divergence of *C. grandiflora/rubella* and *C. orientalis*.[17] However, we found strong collinearity at the beginning of chromosome 5 between the two subgenomes of Cbp_T2T (Figure 2) and between Cbp_T2T and Cbp_version2 (Figure 3). Upon reviewing syntenic analyses, MUMmer showed good synteny at the start of chromosome 5 (Supporting Information S1: Figure S7A), but SyRI results, similar to Penin et al., showed no homeology (Supporting Information S1: Figure S7B). These results suggest that the discrepancy may be due to differences in analysis tools.

To summarize, comparative analysis revealed large inversions on chromosomes 6 and 7. In contrast to Cbp_version2, which displayed fewer structural complexities, Cbp_T2T uncovered substantial divergence, especially in centromeres. The improved resolution from long-read sequencing will facilitate evolutionary biological and functional genomic studies in *Capsella* species.

4 | CONCLUSION

In conclusion, we successfully generated a high-quality, highly contiguous, and phased reference genome assembly for *C. bursa-pastoris* (Cbp_T2T) using PacBio HiFi, ONT, and Hi-C sequencing data. This assembly represents a significant improvement over previously published versions, with enhanced contiguity, a more accurate representation of the genome, and fewer gaps. Notably, the Cbp_T2T assembly revealed substantial divergence between the two subgenomes, reflecting the distinct evolutionary histories of the parental species. Importantly, the comparison between Cbp_T2T and Cbp_version2 captured large structural variants, distinguishing accessions from different genetic clusters. For instance, we identified large inversions on chromosomes 6 and 7, spanning approximately 1.93 and 2.33 Mb, respectively, in Cbp_T2T, which differ from the corresponding regions in the Russian accession, Cbp_version2. These differences suggest that structural variations may have played a role in the genetic divergence among accessions, likely shaped by their distinct evolutionary histories and environmental adaptations.

The superior quality of the Cbp_T2T assembly, particularly its ability to resolve centromeric and telomeric regions, offers an unprecedented view into the genome structure of *C. bursa-pastoris*. The improved resolution of large inversions and other structural variations between accessions provides new insights into the genetic and evolutionary processes that underlie the diversity of this species. The Cbp_T2T assembly is an invaluable resource for future studies on polyploid evolution, genomic diversity, and functional genomics in the genus *Capsella*. Looking ahead, this genome will facilitate comparative genomic studies and provide critical insights into the evolution of complex plant genomes.

AUTHOR CONTRIBUTIONS

Hui-Run Huang and Xue-Jun Ge designed and supervised the research. Hai-Fei Yan collected the experimental materials. Lu Liu and Hai-Fei Yan grew the plants and collected the tissue samples. Xin Liu, Lu Liu, and Hui-Run Huang analyzed the data. Xin Liu and Hui-Run Huang wrote the manuscript. All authors have contributed to the manuscript revision and read and approved the final version.

ACKNOWLEDGMENTS

This work was financially supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515012487) and the National Natural Science Foundation of China (No. 31870353).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The raw sequencing reads and the genome assembly have been deposited in the National Genomics Data Center (NGDC) under BioProject ID PRJCA022749 and the NCBI under BioProject ID PRJNA1062754. The genome annotation files are available from the online open access repository Figshare database (https://doi.org/10.6084/m9.figshare.24974856.v3).

ETHICS STATEMENT

This study did not include any experiments involving human or animal subjects. All data utilized in the research were obtained in full compliance with relevant ethical standards and regulatory guidelines. No interventions or interactions with living organisms were conducted, and the study did not involve the collection or use of any personal, sensitive, or identifiable information. Ethical considerations were carefully followed to ensure that all data handling and analysis processes adhered to applicable regulations.

ORCID

Xin Liu https://orcid.org/0000-0003-4175-9358
Lu Liu https://orcid.org/0000-0001-7168-5017
Hai-Fei Yan https://orcid.org/0000-0002-5692-9391
Xue-Jun Ge https://orcid.org/0000-0002-5008-9475
Hui-Run Huang https://orcid.org/0000-0002-4656-5627

REFERENCES

1. Sicard, A., and M. Lenhard. 2018. "Capsella." *Current Biology* 28(17): R909–30. https://doi.org/10.1016/j.cub.2018.06.033.
2. Hurka, H., and B. Neuffer. 1997. "Evolutionary Processes in the Genus *Capsella* (Brassicaceae)." *Plant Systematics and Evolution* 206: 295–316.
3. Hurka, H., N. Friesen, D. A. German, A. Franzke, and B. Neuffer. 2012. "'Missing Link' Species *Capsella Orientalis* and *Capsella Thracica* Elucidate Evolution of Model Plant Genus *Capsella* (Brassicaceae)." *Molecular Ecology* 21(5): 1223–38. https://doi.org/10.1111/j.1365-294x.2012.05460.x.
4. Douglas, G. M., G. Gos, K. A. Steige, A. Salcedo, K. Holm, E. B. Josephs, R. Arunkumar, J. A. Ågren, K. M. Hazzouri, W. Wang, A. E. Platts, R. J. Williamson, B. Neuffer, M. Lascoux, T. Slotte, and S. I. Wright. 2015. "Hybrid Origins and the Earliest Stages of Diploidization in the Highly Successful Recent Polyploid *Capsella bursa-pastoris*." *Proceedings of the National Academy of Sciences* 112(9): 2806–11. https://doi.org/10.1073/pnas.1412277112.
5. German, D. A., N. Friesen, B. Neuffer, I. A. Al-Shehbaz, and H. Hurka. 2009. "Contribution to ITS Phylogeny of the Brassicaceae, With Special Reference to Some Asian Taxa." *Plant Systematics and Evolution* 283(1–2): 33–56. https://doi.org/10.1007/s00606-009-0213-5.
6. Franzke, A., M. A. Lysak, I. A. Al-Shehbaz, M. A. Koch, and K. Mummenhoff. 2011. "Cabbage Family Affairs: The Evolutionary History of Brassicaceae." *Trends in Plant Science* 16(2): 108–16. https://doi.org/10.1016/j.tplants.2010.11.005.
7. Kryvokhyzha, D., A. Salcedo, M. C. Eriksson, T. Duan, N. Tawari, J. Chen, M. Guerrina, J. M. Kreiner, T. V. Kent, U. Lagercrantz, J. R. Stinchcombe, S. Glémin, S. I. Wright, and M. Lascoux. 2019. "Parental Legacy, Demography, and Admixture Influenced the Evolution of the Two Subgenomes of the Tetraploid *Capsella bursa-pastoris* (Brassicaceae)." *PLoS Genetics* 15(2): e1007949. https://doi.org/10.1371/journal.pgen.1007949.
8. Neuffer, B., and H. Hurka. 1986. "Variation of Development Time Until Flowering in Natural Populations of *Capsella bursa-pastoris* (Cruciferae)." *Plant Systematics and Evolution* 152(3–4): 277–96. https://doi.org/10.1007/bf00989433.
9. Neuffer, B., and S. Bartelheim. 1989. "Gen-Ecology of *Capsella bursa-pastoris* From an Altitudinal Transect in the Alps." *Oecologia* 81(4): 521–7. https://doi.org/10.1007/bf00378963.
10. Neuffer, B. 1990. "Ecotype Differentiation in *Capsella*." *Vegetatio* 89(2): 165–71. https://doi.org/10.1007/bf00032168.
11. Ceplitis, A., Y. Su, and M. Lascoux. 2005. "Bayesian Inference of Evolutionary History From Chloroplast Microsatellites in the Cosmopolitan Weed *Capsella bursa-pastoris* (Brassicaceae)." *Molecular Ecology* 14: 4221–33. https://doi.org/10.1111/j.1365-294x.2005.02743.x.
12. Neuffer, B., and H. Hurka. 1999. "Colonization History and Introduction Dynamics of *Capsella bursa-pastoris* (Brassicaceae) in North America: Isozymes and Quantitative Traits." *Molecular Ecology* 8(10): 1667–81. https://doi.org/10.1046/j.1365-294x.1999.00752.x.
13. Neuffer, B., and R. Hoffrogge. 2000. "Ecotypic and Allozyme Variation of *Capsella bursa-pastoris* and *C. Rubella* (Brassicaceae) along Latitude and Altitude Gradients on the Iberian Peninsula." *Anales Del Jardín Botánico De Madrid* 57(2): 299–315. https://doi.org/10.3989/ajbm.1999.v57.i2.204.
14. Neuffer, B. 2011. "Native Range Variation in *Capsella bursa-pastoris* (Brassicaceae)." *Flora* 206(2): 107–19. https://doi.org/10.1016/j.flora.2010.03.001.
15. Huang, H.-R., P.-C. Yan, M. Lascoux, and X.-J. Ge. 2012. "Flowering Time and Transcriptome Variation in *Capsella bursa-pastoris* (Brassicaceae)." *New Phytologist* 194(3): 676–89. https://doi.org/10.1111/j.1469-8137.2012.04101.x.
16. Kasianov, A. S., A. V. Klepikova, I. V. Kulakovskiy, E. S. Gerasimov, A. V. Fedotova, E. G. Besedina, A. S. Kondrashov, M. D. Logacheva, and A. A. Penin. 2017. "High-quality Genome Assembly of *Capsella bursa-pastoris* Reveals Asymmetry of Regulatory Elements at Early Stages of Polyploid Genome Evolution." *Plant Journal* 91(2): 278–91. https://doi.org/10.1111/tpj.13563.
17. Penin, A. A., A. S. Kasianov, A. V. Klepikova, D. O. Omelchenko, M. S. Makarenko, and M. D. Logacheva. 2024. "Origin and Diversity of

Capsella bursa-pastoris From the Genomic Point of View." *BMC Biology* 22(1): 52. https://doi.org/10.1186/s12915-024-01832-1.
18. Lysak, M. A., M. A. Koch, J. M. Beaulieu, A. Meister, and I. J. Leitch. 2009. "The Dynamic Ups and Downs of Genome Size Evolution in Brassicaceae." *Molecular Biology and Evolution* 26(1): 85–98. https://doi.org/10.1093/molbev/msn223.
19. Marçais, G., and C. Kingsford. 2011. "A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers." *Bioinformatics* 27(6): 764–70. https://doi.org/10.1093/bioinformatics/btr011.
20. Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz. 2020. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11(1): 1432. https://doi.org/10.1038/s41467-020-14998-3.
21. Hurka, H., S. Freundner, A. H. D. Brown, and U. Plantholt. 1989. "Aspartate Aminotransferase Isozymes in the Genus *Capsella* (Brassicaceae): Subcellular Location, Gene Duplication, and Polymorphism." *Biochemical Genetics* 27(1–2): 77–90. https://doi.org/10.1007/bf00563019.
22. Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. "Fastp: An Ultra-Fast All-In-One FASTQ Preprocessor." *Bioinformatics* 34(17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.
23. Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. "Haplotype-Resolved De Novo Assembly Using Phased Assembly Graphs With Hifiasm." *Nature Methods* 18(2): 170–5. https://doi.org/10.1038/s41592-020-01056-5.
24. Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics* 29(8): 1072–5. https://doi.org/10.1093/bioinformatics/btt086.
25. Hu, J., J. Fan, Z. Sun, and S. Liu. 2020. "NextPolish: a Fast and Efficient Genome Polishing Tool for Long-Read Assembly." *Bioinformatics* 36(7): 2253–5. https://doi.org/10.1093/bioinformatics/btz891.
26. Roach, M. J., S. A. Schmidt, and A. R. Borneman. 2018. "Purge Haplotigs: Allelic Contig Reassignment for Third-Gen Diploid Genome Assemblies." *BMC Bioinformatics* 19(1): 460. https://doi.org/10.1186/s12859-018-2485-7.
27. Alonge, M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, F. J. Sedlazeck, Z. B. Lippman, and M. C. Schatz. 2019. "RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes." *Genome Biology* 20(1): 224. https://doi.org/10.1186/s13059-019-1829-6.
28. Slotte, T., K. M. Hazzouri, J. A. Ågren, D. Koenig, F. Maumus, Ya-L. Guo, K. Steige, A. E. Platts, J. S. Escobar, L. K. Newman, W. Wang, T. Mandáková, E. Vello, L. M. Smith, S. R. Henz, J. Steffen, S. Takuno, Y. Brandvain, G. Coop, P. Andolfatto, T. T. Hu, M. Blanchette, R. M. Clark, H. Quesneville, M. Nordborg, B. S. Gaut, M. A. Lysak, J. Jenkins, J. Grimwood, J. Chapman, S. Prochnik, S. Shu, D. Rokhsar, J. Schmutz, D. Weigel, and S. I. Wright. 2013. "The *Capsella rubella* Genome and the Genomic Consequences of Rapid Mating System Evolution." *Nature Genetics* 45(7): 831–5. https://doi.org/10.1038/ng.2669.
29. Durand, N. C., M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3(1): 95–8. https://doi.org/10.1016/j.cels.2016.07.002.
30. Dudchenko, O., S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden, and E. L. Aiden. 2017. "De Novo Assembly of the *Aedes aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science* 356(6333): 92–5. https://doi.org/10.1126/science.aal3327.
31. Durand, N. C., J. T. Robinson, M. S. Shamim, I. Machol, J. P. Mesirov, E. S. Lander, and E. L. Aiden. 2016. "Juicebox Provides a Visualization System for Hi-C Contact Maps With Unlimited Zoom." *Cell Systems* 3(1): 99–101. https://doi.org/10.1016/j.cels.2015.07.012.
32. Li, H. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34(18): 3094–100. https://doi.org/10.1093/bioinformatics/bty191.
33. Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov. 2013. "Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration." *Briefings in Bioinformatics* 14(2): 178–92. https://doi.org/10.1093/bib/bbs017.
34. Pedersen, B. S., and A. R. Quinlan. 2018. "Mosdepth: Quick Coverage Calculation for Genomes and Exomes." *Bioinformatics* 34(5): 867–8. https://doi.org/10.1093/bioinformatics/btx699.
35. Benson, G. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27(2): 573–80. https://doi.org/10.1093/nar/27.2.573.
36. Lin, Y., C. Ye, X. Li, Q. Chen, Y. Wu, F. Zhang, R. Pan, S. Zhang, S. Chen, X. Wang, S. Cao, Y. Wang, Y. Yue, Y. Liu, and J. Yue. 2023. "quarTeT: A Telomere-To-Telomere Toolkit for Gap-Free Genome Assembly and Centromeric Repeat Identification." *Horticultural Research* 10(8): uhad127. https://doi.org/10.1093/hr/uhad127.
37. Ou, S., W. Su, Y. Liao, K. Chougule, J. R. A. Agda, A. J. Hellinga, C. S. B. Lugo, T. A. Elliott, D. Ware, T. Peterson, N. Jiang, C. N. Hirsch, and M. B. Hufford. 2019. "Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline." *Genome Biology* 20(1): 275. https://doi.org/10.1186/s13059-019-1905-y.
38. Cantalapiedra, C. P., A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas. 2021. "eggNOG-Mapper V2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38(12): 5825–9. https://doi.org/10.1093/molbev/msab293.
39. Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, and P. Bork. 2019. "eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses." *Nucleic Acids Research* 47(D1): D309–14. https://doi.org/10.1093/nar/gky1085.
40. Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment With Burrows–Wheeler Transform." *Bioinformatics* 25(14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.
41. Barnett, D. W., E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth. 2011. "BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files." *Bioinformatics* 27(12): 1691–2. https://doi.org/10.1093/bioinformatics/btr174.
42. Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows Along With Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." *Molecular Biology and Evolution* 38(10): 4647–54. https://doi.org/10.1093/molbev/msab199.
43. Rhie, A., B. P. Walenz, S. Koren, and A. M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21(1): 245. https://doi.org/10.1186/s13059-020-02134-9.
44. Marçais, G., A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin. 2018. "MUMmer4: A Fast and Versatile Genome Alignment System." *PLoS Computational Biology* 14(1): e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
45. Goel, M., H. Sun, W.-B. Jiao, and K. Schneeberger. 2019. "SyRI: Finding Genomic Rearrangements and Local Sequence Differences From Whole-Genome Assemblies." *Genome Biology* 20(1): 277. https://doi.org/10.1186/s13059-019-1911-0.
46. Ågren, J., W. Wang, D. Koenig, B. Neuffer, D. Weigel, and S. I. Wright. 2014. "Mating System Shifts and Transposable Element Evolution in the Plant Genus *Capsella*." *BMC Genomics* 15(1): 602. https://doi.org/10.1186/1471-2164-15-602.
47. Shi, X., S. Cao, X. Wang, S. Huang, Y. Wang, Z. Liu, W. Liu, X. Leng, Y. Peng, N. Wang, Y. Wang, Z. Ma, X. Xu, F. Zhang, H. Xue, H. Zhong, Y. Wang, K. Zhang, A. Velt, K. Avia, D. Holtgräwe, J. Grimplet, J. T. Matus, D. Ware, X. Wu, H. Wang, C. Liu, Y. Fang, C. Rustenholz, Z. Cheng, H. Xiao, and Y. Zhou. 2023. "The Complete Reference Genome for Grapevine (*Vitis vinifera* L.) Genetics and Breeding." *Horticultural Research* 10(5): uhad061. https://doi.org/10.1093/hr/uhad061.
48. Huang, H.-R., X. Liu, R. Arshad, X. Wang, W.-M. Li, Y. Zhou, and X.-J. Ge. 2023. "Telomere-to-telomere Haplotype-Resolved Reference Genome Reveals Subgenome Divergence and Disease Resistance in Triploid Cavendish Banana." *Horticultural Research* 10(9): uhad153. https://doi.org/10.1093/hr/uhad153.

49. Chen, J., Z. Wang, K. Tan, W. Huang, J. Shi, T. Li, J. Hu, K. Wang, C. Wang, B. Xin, H. Zhao, W. Song, M. B. Hufford, J. C. Schnable, W. Jin, and J. Lai. 2023. "A Complete Telomere-To-Telomere Assembly of the Maize Genome." *Nature Genetics* 55(7): 1221–31. https://doi.org/10.1038/s41588-023-01419-6.
50. Wang, X., Z. Liu, F. Zhang, H. Xiao, S. Cao, H. Xue, W. Liu, Y. Su, Z. Liu, H. Zhong, F. Zhang, B. Ahmad, Q. Long, Y. Zhang, Y. Liu, Y. Gan, T. Hou, Z. Jin, X. Wu, G. Liu, Y. Wang, Y. Peng, and Y. Zhou. 2024. "Integrative Genomics Reveals the Polygenic Basis of Seedlessness in Grapevine." *Current Biology* 34(16): 3763–77. https://doi.org/10.1016/j.cub.2024.07.022.
51. Yang, X., Y. Su, S. Huang, Q. Hou, P. Wei, Y. Hao, J. Huang, H. Xiao, Z. Ma, X. Xu, X. Wang, S. Cao, X. Cao, M. Zhang, X. Wen, Y. Ma, Y. Peng, Y. Zhou, K. Cao, and G. Qiao. 2024. "Comparative Population Genomics Reveals Convergent and Divergent Selection in the Apricot-Peach-Plum-Mei Complex." *Horticultural Research* 11(6): uhae109. https://doi.org/10.1093/hr/uhae109.
52. Cornille, A., A. Salcedo, D. Kryvokhyzha, S. Glémin, K. Holm, S. I. Wright, and M. Lascoux. 2016. "Genomic Signature of Successful Colonization of Eurasia by the Allopolyploid Shepherd's Purse (*Capsella Bursa-pastoris*)." *Molecular Ecology* 25(2): 616–29. https://doi.org/10.1111/mec.13491.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu, Xin, Lu Liu, Hai-Fei Yan, Xue-Jun Ge, and Hui-Run Huang. 2025. "Telomere-To-Telomere Reference Genome Reveals Subgenome Divergence and Large Structural Variants in *Capsella Bursa-pastoris* (Brassicaceae)." *Biological Diversity* 2(1): 28–38. https://doi.org/10.1002/bod2.70001.