BMC Plant Biology

RESEARCH Open Access

A *de novo* assembled chromosome-level genome of *Ainsliaea gracilis* Franch. (Pertyeae, Asteraceae) and its comparative phylogenomics within Asteraceae

Xinyu Chen[1,2], Tianmeng Qu[2], Xinyi Zheng[2], Yanru Zhang[2], Liang Dou[3,4], Yizhen Shao[5], Mingcheng Wang[6*] and Zhixi Fu[1,2,7*]

Abstract

Background The genus *Ainsliaea* DC. is the largest group in the tribe Pertyeae (Pertyoideae, Asteraceae). *Ainsliaea gracilis* Franch., a medicinal plant endemic to China, is used as a traditional Chinese medicine to treat hemoptysis, phlegmon, and traumatic injuries. However, genomic research of molecular mechanisms of medicinal secondary metabolite biosynthesis for this species is limited by the absence of nuclear genome. To address this gap, we provided a high-quality chromosome-level genome of *A. gracilis* with comprehensive structural and functional annotations.

Results The genome assembly obtained by PacBio HiFi and Illumina Hi-C sequencing technologies was 1,161 Mb and anchored onto the 13 pseudochromosomes, achieving an anchor rate of 99.14%. The contig and scaffold N50 values were 60.20 Mb and 92.53 Mb, respectively. The BUSCO completeness score was 97.70%. The LTR Assembly Index (LAI) score was 22.75, surpassing the high-quality assembly threshold of 20. The genome annotations indicated 35,803 protein-coding genes (PCGs), 6,058 non-coding RNA genes (ncRNAs) and 72.20% repetitive elements. Gene family expansions and contractions were observed among 13 Asteraceae species, with contractions generally exceeding expansions. Phylogenomic analysis revealed that Pertyoideae was identified as the basal group of Asteraceae. Both *A. gracilis* and four Carduoideae species experienced a single, most recent whole genome duplication event. Molecular clock estimations pinpointed the divergence of the four subfamilies at 43.98 million years ago.

Conclusions This study fills a critical gap in the nuclear genome data of Pertyoideae (Asteraceae). It provides not only a high-quality genomic resource for developing molecular markers and exploring gene clusters related to medicinal secondary metabolite, but also proposes a practical framework for assembling other complex genomes within Pertyoideae. This work lays a solid foundation for the future phylogenomic studies of Pertyoideae and Asteraceae.

Keywords *Ainsliaea gracilis*, PacBio HiFi, Hi-C scaffolding, *De novo* assembly, Comparative genomics

*Correspondence:
Mingcheng Wang
wangmingcheng@cdu.edu.cn
Zhixi Fu
fuzx2017@sicnu.edu.cn

Full list of author information is available at the end of the article

Background

The genus *Ainsliaea* DC., the largest group in the tribe Pertyeae (Pertyeae, Asteraceae), comprises around 59 species of perennial herbs recorded [1–5]. The species of this genus are primarily distributed across Afghanistan, Himalayas, central, western, and eastern China, the Korean Peninsula, and Japan [2, 5–7]. In China, around 40 species (ca. 30 endemic) are found in the Yangtze River Basin and its southern provinces, particularly in the southwestern provinces (Yunnan, Sichuan, and Tibet) [8–13]. Currently, *Ainsliaea* is classified into two subgenera, including *A.* subg. *Ainsliaea* (58 species) and *A.* subg. *Diaspananthus* (*Ainsliaea uniflora* Sch.Bip., endemic to Japan) [5]. Within *A.* subg. *Ainsliaea*, four sections are recognized, namely *A.* sect. *Aggregatae* (24 species), *A.* sect. *Ainsliaea* (33 species), *A.* sect. *Alternae* (*Ainsliaea lancifolia* Franch., endemic to central and southwestern China), and *A.* sect. *Intermediae* (*Ainsliaea nana* Y.C.Tseng, endemic to Sichuan, China) [5]. Most species of the genus inhabit forest understory, significantly contributing to soil-water conservation and ecological sustainability [14]. Moreover, the genus includes several species of medicinal importance in traditional Chinese medicines [15, 16]. For example, *Ainsliaea gracilis* Franch., a species of sect. *Aggregatae*, has long been used for treating hemoptysis, phlegmon, and traumatic injuries [17]. This species typically grows in mountainous forests or streamside crevices at elevations of 400-1,600 m [10].

However, the phylogenetic position of *Ainsliaea* and its related genera has long been debated. Kim et al. [18] provided early evidence supporting the monophyly of *Ainsliaea* and placed it together with *Myripnois* Bunge and *Pertya* Sch.Bip. in the tribe Mutisieae (Mutisioideae), while Panero and Funk [4, 19] placed *Ainsliaea*, *Macroclinidium* Maxim. (endemic to Japan), *Myripnois*, and *Pertya* in the tribe Pertyeae (Pertyoideae). Subsequent studies using molecular or morphological data largely supported this treatment [1, 2, 6, 14, 20], although Gao et al. [10] and Hind [21] insisted on assigning these genera to Mutisieae. In addition, recent evidence even suggested merging *Myripnois* into *Pertya* [14, 22, 23]. Overall, despite advances in elucidating both intra- and intergeneric relationships within *Ainsliaea*, the circumscription and taxonomic status of Pertyeae remain unresolved. To date, most studies of *Ainsliaea* and its related genera have primarily relied on genome-skimming [4–6, 14, 18, 19, 23] or transcriptomes data [20, 24]. Broader taxon sampling and high-resolution nuclear genome data are therefore essential to further clarify the evolutionary history of these taxa. Currently, nuclear genome-based comparative and phylogenomic analyses have proven to be powerful tools for elucidating plant identification [25], evolutionary history [26–32], biosynthetic pathways [26, 32–35], and adaptive mechanisms [27, 28, 36, 37]. Assessing the current availability of nuclear genomes across Asteraceae provides critical context for such efforts.

Asteraceae is recognized as comrising 16 subfamilies, 45-50 tribes, and 176-190 subtribes to date [24, 38, 39]. Phylogenetic reconstruction of the family was nearly complete using transcriptomes and genome-skimming datasets from 706 species representing 16 subfamilies, 41 tribes, and 144 subtribes [24]. Nevertheless, nuclear genomic resources of Asteraceae have not been extensively reported. Recently, the rapid advancement of high-throughput sequencing technologies has greatly reduced the cost and increased the accessibility of nuclear genome data. *Helianthus annuus* L. [40, 41] was the first species in the family Asteraceae to have its nuclear genome sequenced. In 2011, Kane et al. [40] initiated the genome sequencing of *H. annuus* using next-generation whole-genome shotgun data. Although the final assembly covered only approximately 85% of predicted genome, it marked a pivotal milestone in the genomic study of Asteraceae. In 2017, Badouin et al. [41] reassembled the genome of *H. annuus* using third-generation sequencing data generated by the PacBio RS-II platform. They obtained a high-quality 3.0 Gb genome and successfully constructed 17 pseudochromosomes with a high anchor rate of 97%. These pioneering efforts laid the groundwork for assembling nuclear genome in Asteraceae, gradually leading to the sequencing of 35 species to date [42]. However, these sequenced species are limited to only three subfamilies [42], including Carduoideae [43–46], Cichorioideae [45, 47, 48], and Asteroideae [40, 41, 49–51]. These subfamilies also represent the three largest groups within Asteraceae [39].

Given that nuclear genomes have been sequenced for only three subfamilies, substantial gaps persist across the other 13. Particularly, nuclear genome data for Pertyoideae remain unavailable to date. Decoding the genome of the medicinal plant *A. gracilis* will provide essential resources for elucidating the biosynthesis pathways of medicinal secondary metabolites [52]. Ultimately, these resources will benefit not only the genus *Ainsliaea*, but also the subfamily Pertyoideae and the broader family Asteraceae. In this study, A high-quality nuclear genome of *A. gracilis* was sequenced and assembled using PacBio HiFi sequencing and Hi-C scaffolding. We successfully provided the first chromosome-level genome assembly and complete annotations of *A. gracilis* in the subfamily Pertyoideae. The objectives of this study are to: (i) fill the current gap in nuclear genomic resources for the subfamily Pertyoideae by sequencing *A. gracilis* and producing a high-quality chromosome-level genome assembly, (ii) establish the first functional genomic database for Pertyoideae through comprehensive annotation of repetitive elements, gene structures, gene functions, and ncRNAs,

and (iii) conduct the comparative and phylogenomic analyses within Asteraceae using 12 published genomes from three other subfamilies (Table S1).

Results

High-quality genome assembly of *A. gracilis*

After quality control of raw data (100.97 Gb), 96.70 Gb of clean data was obtained for the genome survey. The alignment rates of clean data to plant genomes were 8.64% for *Cynara cardunculus* var. *scolymus* (L.) Benth., 3.44% for *Lactuca sativa* L., and 3.29% for *Zabelia corymbosa* (Regel & Schmalh.) Makino, and others (Table S2). This indicated no significant contamination from other species. Based on the histogram of 17 k-mer frequency (Fig. S1), the genome size was estimated to be 1,248.94 Mb, with 0.40% heterozygosity and 71.80% repeat sequences. Two sequencing platforms, PacBio HiFi and Illumina Hi-C, were employed to sequence HiFi CCS reads and Hi-C paired-end reads, respectively. The CCS clean reads comprised 1,922,139 sequences, yielding 33.43 Gb with an estimated average coverage depth of 28×. The N50 length of reads was 17.38 kb. The distributions of read counts and base counts across read lengths were plotted (Fig. S2). The Illumina Hi-C clean data consisted of 698,879,178 paired-end reads, totaling 209.66 Gb with average coverage depth of 180×. The sequencing quality was high, with average Q20 = 96.375% and Q30 = 90.46%. The average GC content was evenly distributed at 39.98%.

The HiFi reads were firstly used to *de novo* assemble the genome of *A. gracilis*. It yielded a draft assembly consisting of 465 contigs with a total length of 1,172.93 Mb (Table S3), representing 93.91% of the estimated genome size. Contigs longer than 1 Mb totaled 675.00 Mb, indicating high assembly continuity. The contig N50 was 60.20 Mb (Table S3). The GC content was 38.81% (Table S3). However, some contigs exhibited extremely low coverage depth or an average GC content above 50% (Fig. S3), suggesting chimeric sequences. To obtain the final contig assembly, contigs with an average coverage depth of less than 5× and a GC content over 50% were excluded. The final draft assembly of the genome resulted in a total size of 1,160.95 Mb, comprising 262 contigs, with a contig N50 of 60.20 Mb and a GC content of 38.70% (Table S4).

Table 1 The detailed information of 13 pseudochromosomes in the genome of *A. gracilis*

Chromosome	Contig number	Chromosome length (bp)
Chr01	3	112,495,595
Chr02	1	107,077,711
Chr03	2	105,780,419
Chr04	1	105,435,098
Chr05	2	101,834,533
Chr06	1	92,531,016
Chr07	1	88,092,348
Chr08	2	87,632,738
Chr09	3	86,970,281
Chr10	2	73,476,384
Chr11	1	68,637,822
Chr12	2	67,670,192
Chr13	4	53,406,665
Total	25	1,151,040,802

Hi-C paired-end reads were used to assist the assembly, upgrading the draft assembly to the chromosome-level and producing the final genome assembly. Firstly, the Hi-C sequencing data were aligned to the draft assembly. After removing unmapped and multiple-mapped paired-end reads, a total of 421,802,108 valid read pairs (60.35%) were obtained for subsequent scaffolding. Assisted scaffolding assembly was performed based on the valid reads (Table S5). A total of 177 scaffolds were generated, with a combined length of 1,160,997,962 bp. The scaffold N50 was 92,53 Mb. The GC content distribution was normal (38.70%). Among them, 15 scaffolds exceeded 1 Mb in length. Thirteen scaffolds were longer than 5 Mb. These 13 largest scaffolds were considered as the 13 pseudochromosomes of the genome. After renaming the scaffolds in descending order of length, detailed information of 13 pseudochromosomes was provided in Table 1. A total of 25 contigs, with a combined length of 1,151.04 Mb, were anchored onto the 13 pseudochromosomes, achieving an anchor rate of 99.14%. The lengths of the pseudochromosomes ranged from 53.41 Mb to 112.50 Mb. Pseudochromosome 13 was composed of four contigs. The remaining 12 pseudochromosomes were assembled from three or fewer contigs. Finally, the Hi-C interaction heatmap (Fig. 1A) also revealed 13 distinct pseudochromosome groups. Intra-chromosomal interactions were much stronger than inter-chromosomal interactions. Interactions were stronger for pseudochromosomes at closer physical locations than at more distant physical locations.

The mapping rate of HiFi reads to the final genome assembly reached 98.86%, while RNA-Seq reads achieved a maximum mapping rate of 96.25%. The assembly was divided into 10 kb non-overlapping windows to calculate the GC content and average sequencing depth within each window. The highest GC content was 39%, and the sequencing depth peaked at 27× (Fig. S4). Moreover, no distinct scatter clusters were observed in the plot (Fig. S4), suggesting that the sequencing data was free from contamination by other species. BUSCO assessment of final assembly indicated a complete BUSCO score of 97.70%, with a single-copy BUSCO score of 91.90% (Table 2). The genome exhibited a LTR Assembly Index (LAI) score of 22.75, surpassing the high-quality assembly threshold of 20 (Fig. S5). The result of LAI further confirmed the high-quality chromosome-level assembly.

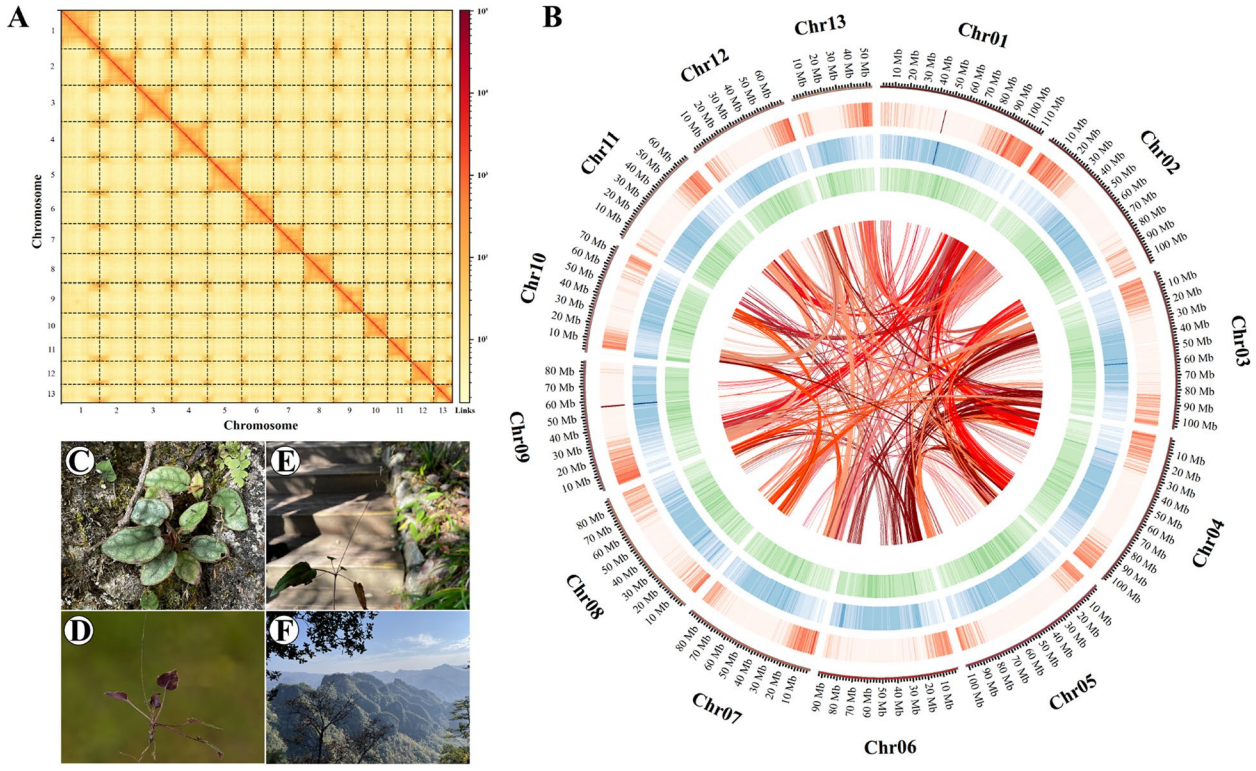

Fig. 1 Genomic and morphological characteristics of *Ainsliaea gracilis*. **A** Hi-C interaction heatmap of the genome. The x- and y-axes represent the assembled chromosomes. The colors ranging from yellow to red indicate increasing interaction strength between chromosomes **B** Circos plot arranged in a clockwise direction. The outer four tracks are organized according to the order of 13 pseudochromosomes. The outermost track displays 13 pseudochromosomes, with sizes labeled at 10 Mb intervals. The color intensity of the three middle tracks from the outer to the inner represents the density of genes, GC content, and LTR elements, respectively. The red lines in the center indicate syntenic blocks among 13 pseudochromosomes. **C** Adaxial surface of leaves **D** Abaxial surface of leaves **E** Habit **F** Habitat (Guangwu Mountain)

Table 2 The statistics of chromosome-level genome assembly of *A. gracilis*

Assembly features		Results
Genome survey	Predicted genome size (bp)	1,248,943,140
	Heterozygosity	0.40%
	Repetitive	71.80%
Assembly statistics	Assembled genome size (bp)	1,160,997,962
	Average GC content	38.70%
	Number of contigs	262
	N50 of contigs (bp)	60,202,890
	Number of scaffolds	177
	N50 of scaffolds (bp)	92,531,016
	Pseudochromosome number	13
	Anchor rate	99.14%
Assembly quality	Complete BUSCO (C)	1577 (97.7%)
	Single-copy BUSCO (S)	1483 (91.9%)
	Duplicated BUSCO (D)	94 (5.8%)
	Fragmented BUSCO (F)	17 (1.1%)
	Missing BUSCO (M)	20 (1.2%)
	Total BUSCO (C)	1614 (100%)
	LTR Assembly Index (LAI)	22.75
	HiFi reads mapping ratio	98.86%
	RNA-Seq mapping ratio	96.25%

These results all confirmed the high quality of the final genome assembly (Table 2). Finally, the genomic and morphological characteristics of *A. gracilis* were illustrated (Fig. 1), including a Hi-C interaction heatmap (Fig. 1A), a circos plot (Fig. 1B), and field photographs of the species (Fig. 1C-F).

High-resolution genome annotations of *A. gracilis*

A total of 872,155 repetitive sequences were identified, representing 838.20 Mb and accounting for 72.20% of the genome (Table 3). Long terminal repeat (LTR) retrotransposons were the most abundant class of repeats, with 452,816 elements detected. These accounted for a total of 593.44 Mb, representing 51.11% of the genome. *Gypsy* and *Copia* were the two primarily LTR families, accounting for 26.81% and 23.68% of the genome, respectively. DNA transposons, long interspersed nuclear elements (LINE), and short interspersed nuclear elements (SINE) comprised 4.74%, 1.61%, and 0.07% of the genome, respectively. For tandem repeats, the most abundant category was simple sequence repeats (SSR), with 8,386 elements totaling 4.26 Mb (0.37%), followed by satellite DNA, with 3,112 elements spanning 2.87 Mb (0.25%).

Table 3 The statistics of repetitive sequences categories in the genome of *A. gracilis*

Category	Type	Count	Length (bp)	Proportion of the genome (%)
DNA transposons	CMC-EnSpm	18,556	10,320,938	0.89
	hAT	13,927	6,630,172	0.57
	MULE-MuDR	14,244	12,383,245	1.07
	PIF-Harbinger	6,360	3,977,064	0.34
	Others	23,945	21,740,813	1.87
Retrotransposons	LINE	22,420	18,680,572	1.61
	SINE	2,932	801,303	0.07
	LTR	452,816	593,441,690	51.11
	Gypsy	259,927	311,225,150	26.81
	Copia	184,288	274,874,150	23.68
	Others	8,601	7,342,390	0.63
Tandem repeats	Satellite	3,112	2,874,681	0.25
	SSR	8,386	4,258,571	0.37
Other repeats	Low complexity	3	423	0.00004
	Unknown	305,454	210,695,808	18.15
Total	/	872,155	838,200,762*	72.20

"*": 4.10% of overlapping regions were excluded to avoid redundancy in the total repetitive sequence length estimation in the genome

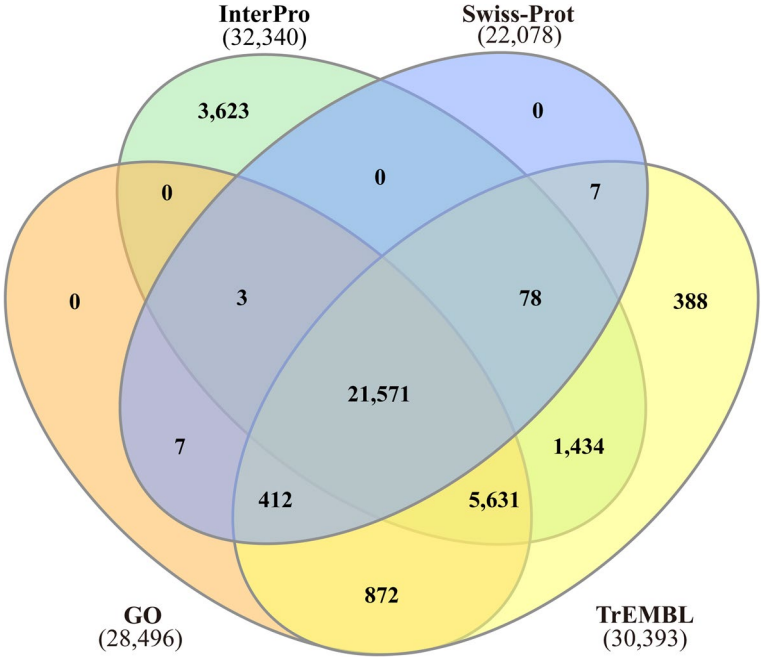

Fig. 2 Venn diagram of the functional annotation results for PCGs in the genome of *A. gracilis* based on four databases: InterPro, Gene Ontology (GO), Swiss-Prot, and TrEMBL. The numbers in parentheses indicate the total number of genes annotated by each database. The numbers in the overlapping areas represent genes annotated by multiple databases simultaneously. The numbers in the non-overlapping regions indicate genes uniquely annotated by a single database

The *de novo*, homology-based alignment, and RNA-Seq-based prediction methods were integrated to annotate the gene structure of *A. gracilis* (Table S6). A total of 35,803 protein-coding genes (PCGs) were identified in the genome, with an average gene length of 4,452.34 bp, an average exon length of 214.57 bp, and an average CDS length of 1,210.03 bp (Table S6). The average number of exons per gene is 5.64. Additionally, 166,102 introns were detected, with an average length of 698.60 bp. BUSCO assessment of predicted PCGs revealed a complete BUSCO score of 94.00%, with a single-copy BUSCO score of 87.2% (Table S7).

The CDS sequences obtained from structural annotation were aligned against four major protein databases: InterPro, Gene Ontology (GO), Swiss-Prot, and TrEMBL. The annotation results were then integrated and analyzed using a Venn diagram (Fig. 2) to obtain the functional annotation results. A total of 34,026 PCGs in the genome

of *A. gracilis* were successfully annotated, accounting for 95.04% of all genes (Fig. 2 and Table S8). Among these, InterPro annotated the highest number of genes, with 32,340 genes (90.33%). TrEMBL annotated 30,393 genes (84.89%), while GO annotated 28,496 genes (79.59%), and Swiss-Prot annotated 22,078 genes (61.67%) (Table S8). Notably, 21,571 genes were annotated by all four databases simultaneously (Fig. 2). Moreover, 388 genes were uniquely annotated by TrEMBL, and 3,623 genes were uniquely annotated by InterPro (Fig. 2).

The annotation of ncRNA (Table S9) identified a total of 2,317 transfer RNA (tRNA) genes, with an average length of 80.90 bp. Ribosomal RNA (rRNA) genes numbered 1,332, averaging 1,112.58 bp in length. A total of 120 microRNA (miRNA) genes were found, with an average length of 122.25 bp. There were also 926 small nuclear RNA (snRNA) genes, with an average length of 111.95 bp. In addition, 113 pseudogenes were annotated, with an average length of 77.86 bp and a total length of 8,798 bp.

Genome visualization and synteny analysis of *A. gracilis*

The genome of *A. gracilis* was visualized, and a circos plot was generated based on the analysis of internal genome synteny. From the outer to the inner tracks, the plot (Fig. 1B) presents 13 pseudochromosomes, gene density, GC content, LTR content, and internal chromosomal synteny in the genome of *A. gracilis*. Notably, genes in the genome were unevenly distributed, with most concentrated at the termini of each pseudochromosome. The contents of GC and LTR exhibited a relatively even distribution. A pronounced enrichment of interchromosomal syntenic regions was also observed within the genome (Fig. 1B). A total of 430 homologous genomic blocks were identified across 13 pseudochromosomes in the genome, comprising 10,866 collinear gene pairs. These accounted for 31.93% of the predicted genes within the genome (filtering with $p < 0.2$). These syntenic regions of *A. gracilis* were primarily distributed at the termini of each pseudochromosome.

Gene family evolution and phylogenomic divergence time in Asteraceae

The results of gene family clustering (Table S10) were obtained based on the nuclear protein sequences of *A. gracilis* and 13 other species. The data of orthogroups and genes obtained were also visualized (Fig. 3A and S6). A total of 16,671 gene families comprising 33,256 genes were identified in the genome of *A. gracilis* (Table S10). Among these gene families, 16,088 were shared homologous families, while 583 were species-specific. Additionally, 2,547 genes were unassigned and classified as orphan genes (Table S10). Species-specific evolutionary patterns were observed in the genome of *A.*

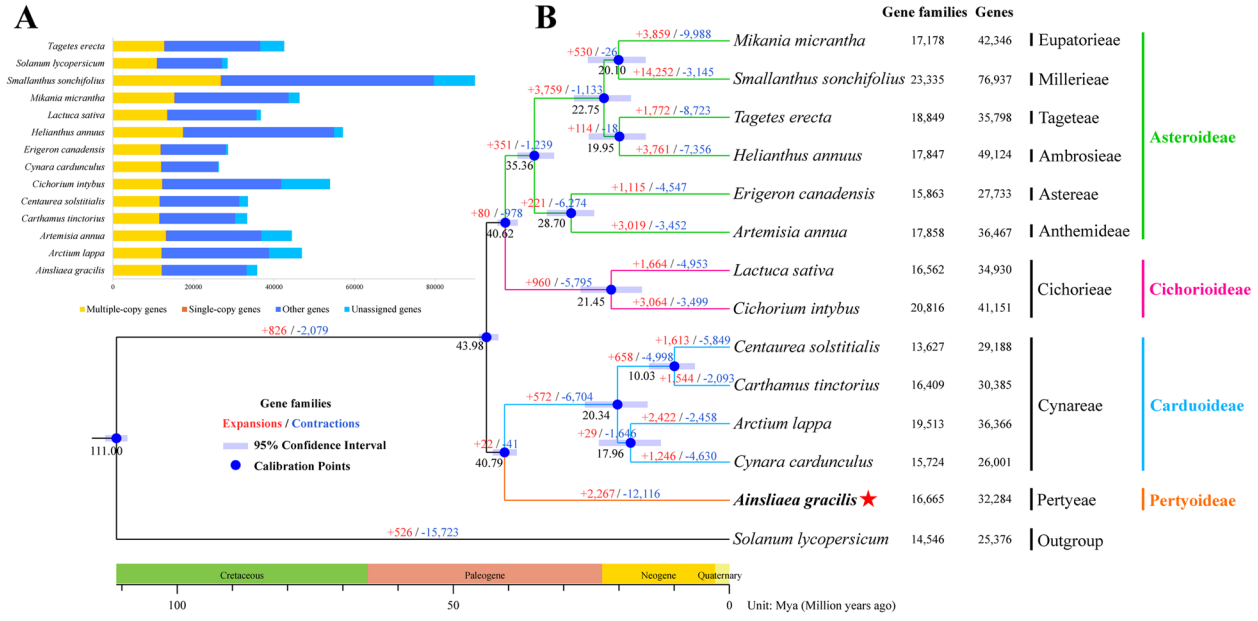

Fig. 3 Plot of gene family composition and a divergence time-calibrated maximum likelihood (ML) phylogenetic tree based on 98 single-copy orthologs from the genomes of *A. gracilis*, Asteraceae species, and one outgroup. **A** Bar chart of the number of single-copy, multiple-copy, other, and unassigned genes in orthogroups for each species **B** ML Phylogenetic tree with estimated divergence times (Mya) and the numbers of gene family expansions and contractions annotated along each branch or node. Blue bars represent the 95% confidential intervals for divergence time. Blue circles indicate calibration points assigned to common ancestors. The red star marks the newly sequenced species. The numbers on each branch denote the gene family expansion and contraction events. The numbers below each blue node indicate the estimated divergence times (Mya). The total numbers of gene families and genes for each species are shown next to the species names. The species of Asteraceae are color-coded to their respective tribes and subfamilies

gracilis. Neither phylogenetic affinity was detected in cross-species homologous comparisons, nor homologous sequence extension was found within the genome of this species. Among these 14 species, 98 single-copy and 10,909–26,786 multiple-copy orthologs were identified (Fig. 3A). A total of 6,438 conserved gene families were shared (Fig. S6). The number of species-specific gene families varied from 46 (*Cynara cardunculus* L.) to 2,775 (*Smallanthus sonchifolius* (Poeppig) H. Rob.) (Fig. S6).

Based on 98 single-copy orthologs identified from the genomes of 14 species, a phylogenetic tree (Fig. 3B and S7) was reconstructed using the Maximum Likelihood (ML) method, with *Solanum lycopersicum* L. (Solanaceae) as the outgroup. Most nodes showed strong bootstrap (BS) support (Fig. S7). In the resulting topology (Fig. 3B and S7), *A. gracilis* was resolved as a basal lineage within Asteraceae with high support (BS = 99). It formed a distinct clade separate from the other 12 Asteraceae species. Moreover, 13 Asteraceae species were clearly divided into two clades with strong support (BS = 100). Pertyoideae (*A. gracilis*) clustered together with Carduoideae (*Arctium lappa* L., *Carthamus tinctorius* L., *Centaurea solstitialis* L., and *Cynara cardunculus*) to form one clade (Fig. S7). Cichorioideae (*Cichorium intybus* L. and *L. sativa*) and Asteroideae (*Artemisia annua* L., *Erigeron canadensis*, *H. annuus*, etc.) formed another one (Fig. 3B and S7). As a result, the four subfamilies in Asteraceae could be well distinguished. To further explore the evolutionary history of gene families, gene family expansions and contractions at each evolutionary node of 14 species were quantitatively analyzed based on the phylogenetic topology. The number of gene family contraction events (12,116) was substantially higher than expansion events (2,267) in the genome of *A. gracilis* (Fig. 3B). This imbalance pattern was consistently observed across the most species, except for *Arctium lappa* and *Smallanthus sonchifolius*. In *Arctium lappa*, the numbers of expansion (2,422) and contraction (2,458) events were nearly equal. In *Smallanthus sonchifolius*, expansion events (14,252) were much higher than contraction events (3,145). Notably, a clade consisted of four species (*H. annuus*, *Mikania micrantha* Kunth, *Smallanthus sonchifolius*, and *Tagetes erecta* L.) in the subfamily Asteroideae exhibited a higher number of expansion events (3,759) than contraction events (1,113) (Fig. 3B).

The molecular clock hypothesis suggests that, across different phylogenetic lineages meeting the neutral evolution time threshold, nucleotide substitution rates (dN/dS) in homologous genes accumulate linearly over time. Based on the phylogenetic topology (Fig. S7), the divergence times among 14 species were further estimated (Fig. 3B). The results indicated that the common ancestor of Solanaceae and Asteraceae diverged as early as 111.00 Mya (Cretaceous). The two major clades within sampled Asteraceae separated 43.98 Mya (Paleogene). Subsequently, Pertyoideae and Carduoideae diverged 40.79 Mya (Paleogene) (Fig. 3B). It suggests that among the sampled 13 Asteraceae species, Pertyoideae (*A. gracilis*) was the earliest lineage to diverge, further forming the basal group of the family. Within Carduoideae, the most recent divergence event occurred 10.03 Mya (Neogene) between the common ancestor of *Carthamus tinctorius* and *Centaurea solstitialis*. The divergence between Asteroideae and Cichorioideae occurred 40.62 Mya (Fig. 3B). The latest divergence happened 19.95 Mya (Neogene) between the common ancestor of *H. annuus* and *T. erecta* within Asteroideae.

The history of selective pressure and whole genome duplication (WGD) events

Both self-alignment and interspecific alignment were performed for *A. gracilis* and five Carduoideae species. The alignment results were used to calculate Ka/Ks, Ks, and 4dTv values (Fig. 4). The results of Ka/Ks density distributions showed that most paralogs gene pairs within five species exhibited Ka/Ks values well below 1 (Fig. 4A). A single major peak of paralogs was observed at Ka/Ks values ranging from 0.11 (*Arctium lappa*) to 0.15 (*A. gracilis*) (Fig. 4A). This indicated that these genes of five species have predominantly undergone purifying selection during evolution. The encoded proteins have thus retained functional conservation. Similarly, it revealed that most orthologous gene pairs between *A. gracilis* and Carduoideae species displayed Ka/Ks values well below 1, with a primary peak ranging from 0.14 (*A. gracilis* vs *Centaurea solstitialis*) to 0.16 (*A. gracilis* vs *Arctium lappa*) (Fig. 4A). This suggested that most genes of *A. gracilis* have been subject to purifying selection since its divergence from Carduoideae species.

To further investigate the WGD events in the evolutionary history of *A. gracilis* and four Carduoideae species, Ks and 4dTv density distribution analyses were conducted (Fig. 4B-C). Based on the analysis of Ks and 4dTv distribution, a single, most recent WGD event might occur in the genome of *A. gracilis* (Fig. 4B-C). This lineage-specific event occurred relatively late in its evolutionary history after the origin of angiosperms. This is evidenced by a single dominant density peak in both the Ks distribution (Fig. 4B, Ks = 0.50) and the 4dTv distribution (Fig. 4C, 4dTv = 0.17). Statistical analysis revealed that 56.56% of paralogous gene pairs exhibit a strict 1:1 collinear depth, indicating that *A. gracilis* retains clear signatures of lineage-specific WGD event.

Similarly, each of four Carduoideae species may experience a single, most recent WGD event (Fig. 4B-C). However, these events occurred considerably earlier than that of *A. gracilis* and were relatively concentrated in time. Among them, the earliest two WGD events likely took

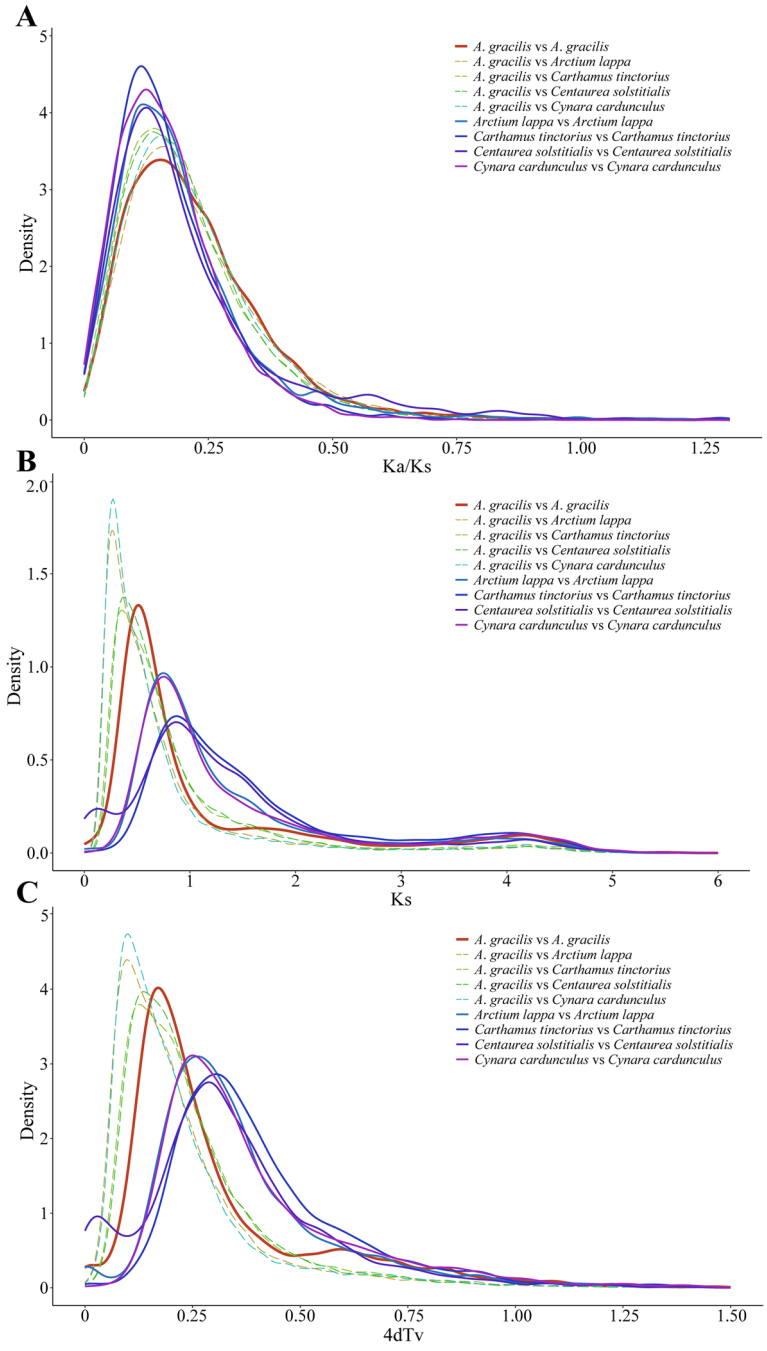

Fig. 4 Density distributions of substitution rates for orthologs and paralogs in the genomes of *A. gracilis* and four Carduoideae species. Solid curves indicate paralog distributions. Dashed curves denote ortholog distributions. **A** Distribution of Ka/Ks values **B** Distribution of Ks values **C** Distribution of 4dTv values

place in *Carthamus tinctorius* and *Centaurea solstitialis* (Fig. 4B-C). Moreover, the speciation of *A. gracilis* and four Carduoideae species occurred after the WGD event in *A. gracilis* (Fig. 4B-C).

Discussion

Genome survey, assembly, and annotations of *A. gracilis*

The final genome assembly accounted for 92.96% of the genome size estimated by 17 k-mer frequency analysis. Similar discrepancies have been frequently reported in the genomes of Asteraceae, as evidenced by large-scale evaluations of genome assemblies in public databases [53]. A substantial proportion of these assemblies exhibit

deviations around 10% from estimated genome sizes [53]. These discrepancies are largely attributed to the presence of highly repetitive, non-coding, and structural DNA elements [41, 53]. These sequences pose challenges for accurate resolution by current assembly algorithms. Notably, larger genomes tend to exhibit greater deviations due to the higher abundance of repetitive elements [41, 53]. A *de novo* genome assembly of *A. gracilis* was performed using HiFi CCS reads, generating a draft assembly of 1,160.95 Mb comprising 262 contigs, with a contig N50 of 60.20 Mb. The contig N50 is a key metric for evaluating genome assembly quality, reflecting the continuity of the assembly [54]. Generally, a higher N50 value indicates a greater number of longer contigs, suggesting better assembly quality and continuity.

Hi-C technology was used to cluster contigs and scaffolds by chromatin contact frequency. Given that adjacent genomic regions on the same chromosome interact more frequently than distant ones, Hi-C data enable precise ordering and orientation of sequences along pseudochromosomes [55]. By integrating Hi-C sequencing data, we successfully obtained the high-quality chromosome-level genome assembly of the medicinal plant *A. gracilis*, totaling 1,160,997,962 bp. This genome size is notably larger than those of other Asteraceae species, e.g., *Carthamus tinctorius* (1,032.34 Mb) [44], *Centaurea solstitialis* (756.58 Mb) [46], and *Erigeron canadensis* L. (421.64 Mb) [56]. This genome not only lays the groundwork for downstream annotations and analyses, but also acts as a practical pattern for assembling complex genomes of Pertyoideae relatives.

Genome annotation is a critical step in genomics research, involving the structural and functional interpretation of sequences. Repetitive sequences are closely associated with species evolution, genetics, and variation. They are widely distributed in eukaryotic genomes [57]. A total of 872,155 repetitive elements were identified, accounting for 72.20% of the genome. This proportion is not particularly higher than in some Asteraceae species, such as *Arctium lappa* (89.9%) [45], *Artemisia argyi* H.Lév. & Vaniot (81.03%) [58], and *Cichorium intybus* (77.11%) [45]. However, it remains greater than that of *Artemisia annua* (61.57%) [51] and *Senecio squalidus* (62%) [50]. The expansion of transposable elements (TE) in plant genomes could profoundly influence genome size and promote adaptation to diverse environments [59, 60]. Accordingly, most of the repeats in the genome were TE, including DNA transposons and retrotransposons. Particularly, LTR retrotransposons were dominant, accounted for 88.84% of TE. To ensure accurate and comprehensive protein-coding genes (PCGs) annotation, we integrated the methods of *de novo* prediction, homology-based alignment, and RNA-Seq data. As a result, 35,803 PCGs were identified, with 95.04% successfully annotated. These detailed gene structural and functional annotation data provide a valuable foundation for future functional genomics research of medicinal secondary metabolite in this species.

The genome of *A. gracilis* obtained in this study was the first chromosome-level high-quality reference genome in the subfamily Pertyoideae of Asteraceae. It provides a valuable reference for investigating the evolutionary history and patterns of plants within this subfamily. We identified the intragenomic synteny of *A. gracilis* and visualized large-scale collinear blocks, ultimately producing a comprehensive synteny map. Notably, these syntenic regions were predominantly enriched at the termini of pseudochromosomes. These telomere-proximal regions, where genes synteny are concentrated, may exhibit high transcriptional activity and serve as important hotspots for gene clustering and rapid evolution, facilitating species adaptation to environmental changes.

Phylogenomic insights into Asteraceae evolution

The phylogeny indicated that the relationships among the four subfamilies were consistent with the APG IV system [38] and latest classification framework of Asteraceae [39]. In the phylogenetic tree, Pertyoideae was identified as the basal group of Asteraceae, also in agreement with Chase et al. [38] and Susanna et al. [39]. Molecular clock estimation indicated that the divergence of four subfamilies occurred 43.98 million years ago, coinciding closely with the early accelerated uplift phase of the Qinghai-Tibet Plateau [61]. This suggests that the early evolutionary radiation of Asteraceae may have been driven by ancient geological events. Moreover, with increasing genome sequencing efforts, gene family losses and gains have been recognized as common phenomena [62–64]. We analyzed gene family expansions and contractions across these 14 species and found that all species experienced both gene family losses and gains, with losses generally outnumbering gains. The expansion of gene families is often accompanied by functional innovation and is associated with specific functional demands under environmental pressures. For example, the aphid's diet, characterized by high sugar and limited proteins and lipids (plant phloem sap), may have driven the expansion of cysteine protease genes of the family cathepsin B [65]. Similarly, during the domestication process of brewing yeast for beer production, artificial selection likely promoted the expansion of the flocculin gene family [63]. In contrast to the rapid expansion of gene families often driven by positive selection, the evolutionary pressures leading to gene family contraction remain poorly understood. Both neutral and adaptive hypotheses have been proposed, yet direct evidence for adaptive gene loss is limited. In most cases, the possibility that such contractions result from neutral mutations cannot be excluded

[62, 66]. Therefore, the expansion and contraction of gene families represent core driving forces in genome evolution. Studying these dynamics could not only elucidate the molecular mechanisms underlying organismal adaptation to environmental changes, but also offer critical insights into species diversity, metabolic innovation, and evolutionary history.

Whole genome duplication (WGD), also known as polyploidization, is the large-scale duplication of an entire genome, leading to a rapid expansion of both gene content and genome size. It provides a continuous source of genetic variation, enabling plants to better adapt to complex and dynamic environments. It is also widely regarded as one of the primary mechanisms driving plant evolution and biodiversity [67–69]. Many studies have demonstrated that WGD has played a pivotal role in the radiation and diversification of angiosperms [70, 71]. It is estimated that around 70% of flowering plants have undergone polyploidization throughout their evolutionary history. Polyploidization is a major and essential force in plant evolution. It has profoundly impacted across different timescales, from ancient to modern, and at various biological levels, from molecular to ecological. After genome polyploidization, reproductive isolation from the original species may develop. The duplicated genomic regions further accumulate various mutations over the course of evolution, ultimately resulting in the formation of new species. For instance, phylogenomic and phylotranscriptomic analyses have revealed that WGD events are pervasive throughout the evolutionary history of Asteraceae [20, 24, 42], including both ancient events during early Asteraceae evolution [20, 72–74] and more recent lineage-specific events [20, 41, 45, 49, 74]. Asteraceae experienced two successive ancient polyploidization events [20, 42]: the first one (WGD1) shared with Calyceraceae, and the second one (WGD2) shared by core Asteraceae species (Mutisioideae/Stifftioideae-Asteroideae) [20, 73, 75]. Specifically, *H. annuus* (sunflower) underwent an additional recent polyploidization event on top of two ancient events, ultimately shaping its current karyotype of 17 pseudochromosomes and contributing to a large genome size of 3.0 Gb [41].

During the evolutionary history of Asteraceae, extensive hybridization, polyploidization, and rapid radiation have greatly contributed to the remarkable species diversity within the family [20, 24]. Whole genome duplications have enhanced the adaptive capacity of Asteraceae species to various environmental pressures, potentially driving both rapid species proliferation and radiative diversification [74]. Moreover, the gene redundancy generated by polyploidization provides raw material for the evolution of novel gene functions through mutation and natural selection, thereby promoting evolutionary innovation [76]. Recent advances in plant genomics have provided important insights into the phylogenetic evolution and genome polyploidization processes of Asteraceae. However, compared to other plant families such as Poaceae and Fabaceae, nuclear genome sequencing efforts in Asteraceae are still at an early stage. High-quality chromosome-level genome assemblies are relatively limited. To better understand the genomic evolution of Asteraceae, broader sampling within unsequenced subfamilies and prioritizing genome sequencing of species at key phylogenetic positions are essential. Such genomic resources will offer valuable support for elucidating the phylogenetic relationships and evolutionary history of this diverse family.

Evolutionary implications of Pertyoideae

Based on the Angiosperm Phylogeny Group IV (APG IV) system [38] and the latest classification framework of Asteraceae [39], the species of the subfamily Carduoideae [43–46] are the closest sequenced relatives to *A. gracilis* (Pertyoideae). To identify WGD events in *A. gracilis*, we performed Ks and 4dTv analyses based on orthologs and paralogs between *A. gracilis* and four Carduoideae species. The Ks or 4dTv values of gene pairs within syntenic blocks can reflect both related species divergence events and WGD events throughout evolutionary history. Each polyploidization event leads to a surge of homologous genes, resulting in a corresponding density peak in the Ks or 4dTv distribution. The number and position of density peaks in the Ks or 4dTv distributions of paralogs indicate the number and timing of WGD events within a species. Meanwhile, the position of Ks or 4dTv peaks of orthologs in pairwise species comparisons indicates the speciation time between the two species. Generally, the larger Ks or 4dTv values corresponding to a density peak, the older the associated evolutionary event. WGD analyses revealed a prominent density peak at low Ks and 4dTv values in *A. gracilis* and Carduoideae species, indicating the possibility of a single, relatively recent WGD event. This pattern is consistent with *H. annuus* [41], *Arctium lappa* [45], and other Asteraceae species. Notably, this event is distinct from two ancient WGDs (WGD1 and WGD2) shared across Asteraceae and may represent a subfamily-specific polyploidization event (WGD6) [20] that may have promoted genomic plasticity and ecological adaptation in Pertyoideae without substantially increasing species richness (ca. 80 species) [20, 67]. By contrast, the four Carduoideae species experienced their most recent WGD events earlier in evolutionary history, underscoring the asynchronous timing of polyploidization across Asteraceae lineages.

Together with the reconstruction of ancestral morphological characters of Asteraceae by Zhang et al. [20], these results suggest that genome duplication and subsequent morphological innovation have jointly shaped

the evolutionary trajectory of the family. Following the divergence of Pertyoideae from the common ancestor of the three largest subfamilies, a transition from woody to herbaceous habit occurred within Asteraceae, postdating the core Asteraceae WGD but preceding the upshift in diversification rate observed in Asteroideae and Cichorioideae [20]. Floral diversification, including shifts in capitulum types from discoid to radiate-like, ligulate, and radiate, has been linked to duplicated MADS-box and *CYCLOIDEA2* (*CYC2*) genes since the origin of Asteraceae [77, 78]. These duplications likely contributed to the evolution of "flower-like" radiate head inflorescence, enhancing pollinator attraction and reproductive success [20]. Notably, smaller, early-diverging clades such as Pertyoideae have retained discoid capitula and woody traits, reflecting a more conservative evolutionary pathway. In this context, the lineage-specific WGD detected in Pertyoideae may have facilitated genomic flexibility and adaptation while preserving the ancestral woody habit (in most species), representing a distinct evolutionary strategy relative to the herbaceous lineages of core Asteraceae. Overall, these patterns illustrate multiple, lineage-specific strategies, ranging from ancestral trait retention in Pertyoideae to innovation-driven radiation in Asteroideae, Carduoideae, and Cichorioideae.

Conclusions

In this study, we sequenced, assembled, and annotated the nuclear genome of *Ainsliaea gracilis*. The first chromosome-level reference genome for this species was provided to successfully fill the gap in nuclear genomes within Pertyoideae (Asteraceae). This achievement was made possible by integrating PacBio HiFi and Illumina Hi-C sequencing technologies. Based on comparative and phylogenomic analyses, we offer new insights into the evolutionary history of *A. gracilis* and the broader Asteraceae family. The resulting high-quality genome could serve as a valuable genomic resource. It not only facilitates the development of molecular markers and the identification of gene clusters associated with medicinal secondary metabolite, but also proposes a practical framework for assembling other complex genomes within Pertyoideae. Overall, this work establishes a critical data foundation for the future phylogenomic studies of Pertyoideae and Asteraceae.

Methods

Plant sampling

In the summer of 2022, Prof. Dr. Zhixi Fu conducted a field investigation in Guangwu Mountain, located in Guangwushan Town, Nanjiang County, Bazhong City, Sichuan Province, China. Complete healthy flowering individuals of *A. gracilis* were collected and subsequently preserved as voucher specimens (no. DY159) in the Herbarium of Sichuan Normal University (SCNU!). Prof. Dr. Zhixi Fu formally identified the voucher specimens. Fresh, healthy leaves, stems, and flowers were preserved through both liquid nitrogen snap-freezing and silica gel desiccation methods. The field collection was legally authorized and ethically approved by the local government.

Genome sequencing

Genomic DNA was extracted from the leaves of *A. gracilis* using the modified CTAB method [79]. The extracted DNA was quantified and quality-controlled using NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, MA, USA), agarose gel electrophoresis, and Qubit Fluorometer (Thermo Fisher Scientific, MA, USA). Subsequently, total DNA was purified using AMPure PB beads (PacBio Biosciences 100–265–900, CA, USA). The High-Fidelity (HiFi) library of 15 kb was constructed using a SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA). The construction includes DNA shearing, AMPure PB Bead purification, ssDNA overhangs removing, damage repair, end repair, hairpin adapter ligation, and bead purification of the library. The circular consensus sequencing (CCS) was performed using a single 8 M SMAT Cell on the PacBio Sequel II platform (Pacific Biosciences, CA, USA) by Berry Genomics Co., Ltd., Beijing, China (https://www.berrygenomics.com/). To obtain clean HiFi data, the PacBio SMRT Link (https://www.pacb.com/smrt-link/) was used for the quality control of the raw polymerase reads.

To improve the assembly at the chromosomal level, a Hi-C library was prepared from the same individual following Rao et al. [80] and Xie et al. [81], including cross-linking cells with formaldehyde, digesting DNA with a suitable 4-cutter restriction enzyme (*DpnII*), filling ends and mark with biotin, ligating the resulting blunt-end fragments, purification and random shearing DNA into 300–500 bp fragments. After quality control test of the libraries using Qubit 2.0 Fluorometer, Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and q-PCR, 150 bp pair-end (PE) next-generation sequencing of the Hi-C library were performed on the Novaseq 6000 platform (Illumina, California, USA) by Berry Genomics Co., Ltd., Beijing, China (https://www.berrygenomics.com/). The clean data were obtained by filtering raw data containing more than three unidentified nucleotides (N), reads aligned to adaptor sequences, and reads with ≥ 20% of bases having a Phred quality score ≤ 5.

To assess genome quality and predict gene structure, the total RNA from fresh leaves, stems, and flowers collected from the same individual of *A. gracilis* were extracted for RNA sequencing (RNA-seq) using the RNeasy Plant Mini Kit (QIAGEN, NRW, Germany). The integrity, quality and concentration of RNA

were detected by NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, MA, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). cDNA libraries were constructed using the NEBNext Ultra II RNA Library Prep Kit for Illumina (New England Biolabs, MA, USA). The libraries were then sequenced on the Illumina NovaSeq™ X Plus platform (Illumina, California, USA) by Berry Genomics Co., Ltd., Beijing, China (https://www.berrygenomics.com/). Raw reads were filte red using Trimmomatic v.0.36 [82] with default parameters to obtain clean reads.

Genome survey

The libraries were constructed using the Agencourt AMPure XP-Medium Kit (Beckman Coulter Life Sciences, California, USA) and AxyPrep Mag PCR Clean-up Kit (Corning Life Sciences, California, USA). The size distribution of the libraries was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). The concentrations were quantified by real-time PCR. The libraries were then sequenced on the MGISEQ-2000 platform (MGI Tech, Shenzhen, China) by Beijing Grandomics Biotechnology Co., Ltd., China (https://www.grandomics.com/). To ensure the reliability of the reads, fastp v0.23.3 [83] was used with default parameters to filter low-quality, repetitive, poly-N, adapter, and plastid reads. A total of 100,000 clean reads were randomly selected and aligned to the NCBI Nucleotide (NT) Database (https://www.ncbi.nlm.nih.gov/nucleotide/) using blastn in BLAST + v.2.16.0 [84] to assess potential contamination from other species. The genome size of *A. gracilis* was estimated using the 17 k-mer frequency analysis based on clean sequencing data. The *k*-mer distribution was estimated using Jellyfish v.2.3.1 [85], further removing *k*-mer with a depth > 50,000 and a frequency > 250,000,000. The genome size was then calculated using the formula: Genome Size = $\sum$ (*k*-mer frequency × *k*-mer depth)/*k*-mer coverage depth at the major peak. Based on the histogram of 17 k-mer frequency, the heterozygosity ratio and repeat content of the genome were estimated using GenomeScope2 v.2.0.0 [86].

Genome assembly

The contig-level assembly was generated using Hifiasm v.0.15.2 [87] with default parameters based on CCS reads. Subsequently, the CCS reads were aligned to the draft assembly using minimap2 v.2.13 [88] with the parameters "-16G -ax map-pb -t 60" to assess assembly accuracy and filter out pseudo contigs. SAMtools v.1.17 [89] was used to evaluate the rate of alignment and coverage. A GC depth plot was generated by R script based on the alignment results. The RNA-Seq Illumina data were also aligned to the draft assembly using HISAT2 v.2.0.6 [90] with the parameters "-9 -dta-cufflinks" to evaluate the data recovery ratio. The clean Hi-C reads were mapped to the contig-level assembly using BWA v.0.7.17 [91], integrated within the Juicer v.1.6.2 [92], followed by manually filtering and adjustment. Hi-C-assisted genome scaffolding was performed using 3D-DNA v.180922 [93], based on uniquely mapped and valid PE reads. Juicebox v.1.9.8 [94] was then used to manually adjust the scaffolds and generate a visualized interaction map, resulting in the final chromosome-level assembly. To further validate the accuracy of the assembly, HiCExplorer v.3.3 [95] was employed to generate the plot of Hi-C interaction heatmap. Finally, LTR_retriever v.2.9.0 [96, 97] was applied to calculate the Long Terminal Repeat (LTR) Assembly Index (LAI) across the genome using a 3 Mb sliding window with a 300 kb step size. According to the LAI values, the assembled genomes could be divided into three quality levels: Gold (LAI ≥ 20), Reference (10 ≤ LAI < 20), and Draft (0 ≤ LAI < 10). Finally, the completeness and quality of draft and final assembly were assessed using BUSCO v.5.2.2 [98] with the parameters "-c 60-m geno –offline", against the Embryophyta odb10 database.

Genome annotations

Repeat sequences annotation

The combination of *de novo* and homology-based alignment method was used to predict repeat elements in the genome of *A. gracilis*. For *de novo* annotation, the software MITE-Hunter [99] was used to identify mini-inverted repeat transposable elements (MITE). LTR Retrotransposons (LTR-RT) were predicted based on structural features using LTR_Finder v.1.1 [100] and LTR_harvest v.1.5.9 [101]. The main parameters were set as follows: minimum LTR length 100 bp, maximum LTR length 7,000 bp, and minimum LTR similarity 90%. Finally, LTR_retriever v.2.9.0 [96, 97] was used to integrate the LTR prediction results and remove false positives. RepeatModeler v.2.0.4 [102] was used to *de novo* identify other repetitive sequences with repeat-masked genome. For homology-based annotation, RepeatMasker v.4.1.4 [103] was used to perform homology searches of repeats, based on the Viridiplantae repeat library from the Dfam (https://www.dfam.org/releases/current/families/FamDB/) [104] and Repbase (https://www.girinst.org/repbase/) [105] databases. Finally, RepeatMasker was used to integrate the above prediction results, producing the final repeats annotation of *A. gracilis*.

Gene structure annotation

A multimodal annotation strategy integrating *de novo*, homology-based alignment, and RNA-Seq-based prediction methods was adopted to systematically annotate the protein-coding genes (PCGs). For *de novo* method, AUGUSTUS v.3.5.0 [106] was employed to predict gene models of the genome of *A. gracilis*. For homology-based

method, protein sequences of *C. tinctorius* [44], *H. annuus* [40, 41], and *L. sativa* [47], were aligned to the genome of *A. gracilis* using tblastn in BLAST + v.2.16.0 [84]. The alignment results were subsequently processed with GeneWise v.2.4.1 [107] (https://www.ebi.ac.uk/jdispatcher/psa/genewise) to predict gene structures. For RNA-Seq-based method, Trinity v2.15.1 [108, 109] was used to *de novo* and reference-guided assemble transcripts based on RNA-Seq reads from flowers, stems, and leaves. The assembled transcripts were then analyzed with Program to Assemble Spliced Alignments (PASA pipeline) v2.5.2 [108, 110] to predict PCGs. Finally, EvidenceModeler (EVM) v.2.0.0 [111] was used to integrate the results from the three prediction methods into a non-redundant consensus gene set. The EVM results were further refined using PASA v2.5.2 [108, 110] by updating UTR annotations, validating alternative splicing models, and adjusting exons, ultimately generating the final genome structural annotation. The completeness and quality of PCGs were assessed using BUSCO v.5.2.2 [98] with the parameters "-c 60-m geno –offline", against the Embryophyta odb10 database.

Gene function annotation

The genome of *A. gracilis* was aligned against several protein databases to annotate the predicted PCGs. The PCGs were annotated using InterProScan v.5.61–93.0 [112] against the Pfam protein database (https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/) [113] to identify protein domains and motifs. DIAMOND v.2.1.6 [114] was used to perform homology searches of the PCGs against the GO (https://geneontology.org/docs/download-ontology/) [115], InterPro (https://ftp.ebi.ac.uk/pub/databases/interpro/current_release/) [116], Swiss-Prot (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz) [117], and TrEMBL (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.fasta.gz) [117] protein databases. The Gene Ontology (GO) IDs for each PCG were retrieved using Blast2GO v6.0 [118]. Finally, the functional annotation results were integrated by combining the outputs all the above methods. To visualize the functional annotation overlaps from the four protein databases, a Venn diagram was generated using InteractiVenn (https://www.interactivenn.net/index.html) [119].

Non-coding RNA (ncRNA) genes annotation

A combined strategy was employed to predict ncRNA genes. Specifically, tRNA genes were annotated using tRNAscan-SE v.2.0.12 [120]. rRNA genes was identified with RNAmmer v.1.2 [121]. miRNA, snRNA, and other ncRNA genes were predicted using Infernal v.1.1.5 [122] by searching against the Rfam 15.0 database (https://ftp.ebi.ac.uk/pub/databases/Rfam/CURRENT/) [123].

Intragenomic synteny visualization

The internal genome synteny of *A. gracilis* was analyzed using MCScanX v.1.0.0 [124]. The pseudochromosomes, gene density, GC content, LTR elements, and genome synteny were then visualized using Circos v.0.69.9 [125] to generate a genomic circos plot based on the annotation results of the assembly.

Comparative and phylogenomic analyses

Gene family clustering analysis

Given that only chromosome-level assemblies with complete annotations released are recognized as valid reference genomes, 18 genomes were qualified for this study. Among them, the subfamily Asteroideae contained the largest number of sequenced genomes (10), followed by Carduoideae (5) and Cichorioideae (3). However, selecting congeneric species and too many Asteroideae species was unnecessary and offered limited value for the analyses. Therefore, 12 representative genomes from different genera, each with relatively high-quality assemblies and annotations, were selected from these three subfamilies. In total, protein sequences from 14 species, including the above 12 representative Asteraceae species, *A. gracilis*, and one outgroup species (*Solanum lycopersicum*, Solanaceae), were used to construct a dataset (Table S1) for orthologous gene clustering analysis. GffRead v.0.12.7 [126] was used to filter redundant transcript sequences for each gene, retaining only the longest transcript per gene. Gene family clustering analysis was performed using OrthoFinder v.2.5.5 [127] with default parameters to identify single-copy, multi-copy, and species-specific orthologs.

Phylogenomic and divergence time analyses

A total of 98 single-copy orthologs identified from *A. gracilis*, 12 other Asteraceae species, and one outgroup (*Solanum lycopersicum*) (Table S1) were used for phylogenomic reconstruction. Protein sequences of each ortholog were first aligned using MAFFT v.7.526 [128] with the "–auto" strategy, and subsequently trimmed with trimAl v.1.5.0 [129] to remove gaps. The curated alignments were concatenated by species using SequenceMatrix v.1.10 [130] to generate a supermatrix for phylogenomic analysis. The phylogenetic tree was inferred using the Maximum Likelihood (ML) method in RAxML v.8.2.12 [131], with 1000 bootstrap replicates under the "JTT + I + G4 + F" substitution model, as predicted by ModelTest-NG v.0.1.7 [132]. The divergence time estimation among the 14 species was performed using the MCMCtree module in PAML v.4.10.6 [133] based on the resulting phylogenetic tree. The fossil points

obtained from Zhang et al. [20], Zhang et al. [24], and TimeTree database [134] (https://timetree.org/) to calibrate divergence time of these species. Finally, the phylogenetic tree with divergence time was visualized and formatted using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

Gene family expansion and contraction analysis

The expansion and contraction of gene families were analyzed using CAFE5 (Computational Analysis of gene Family Evolution) v.1.1 [135], based on clustering results after excluding families with more than 100 genes. The results were subsequently visualized on the phylogenetic tree.

Selective pressure and WGD analyses

The self and all-vs-all protein sequence alignments of *Arctium lappa*, *Carthamus tinctorius*, *Centaurea solstitialis*, *Cynara cardunculus*, and *A. gracilis* were performed using blastp in BLAST + v.2.16.0 [84] with default parameters. Based on these results, MCScanX v.1.0.0 [124] with default parameters was employed to identify inter- and intragenomic collinear regions among these five Asteraceae species. Orthologous and paralogous protein pairs were then extracted and aligned using MUSCLE v.5.3 [136]. The alignment results were then converted to CDS sequences with GffRead v.0.12.7 [126].

Selective pressure and WGD events were inferred by calculating Ka/Ks, Ks, and 4dTv values. Ka represents the nonsynonymous substitution rate, indicating the average number of nonsynonymous substitutions per nonsynonymous site. The Ks represents the synonymous substitution rate, indicating the average number of synonymous substitutions per synonymous site. 4dTv is the proportion of fourfold degenerate third-codon transversion sites and is commonly used to estimate synonymous substitution rates between homologous genes. The Ka/Ks ratio reflects selective pressure on PCGs: $Ka/Ks > 1$ suggests positive selection, $Ka/Ks = 1$ indicates neutral selection, $Ka/Ks < 1$ suggests negative or purifying selection. In the absence of WGD, the distributions of Ka/Ks, Ks, and 4dTv values among homologous genes typically follow exponential decay. In contrast, the occurrence of WGD events produce peaks resembling normal distributions in these values.

Ka/Ks and Ks values of orthologous and paralogous gene pairs among the five Asteraceae species were calculated from CDS alignments using KaKs_Calculator v.2.0 [137] under the YN model. The values of 4dTv were calculated under the HKY model using a Perl script (https://github.com/JinfengChen/Scripts/blob/master/FFgenome/03.evolution/distance_kaks_4dtv/bin/calculate_4DTV_correction.pl). The distributions of Ka/Ks, Ks, and 4dTv for orthologous and paralogous genes were plotted using a Perl script (https://github.com/JinfengChen/Scripts/blob/master/FFgenome/03.evolution/distance_kaks_4dtv/bin/draw_4dtv_kaks.pl). The Ka/Ks distribution was further analyzed to infer selective pressure. The Ks and 4dTv distributions were used to infer WGD events. To verify the WGD signature, the collinear depth of paralogous gene pairs in *A. gracilis* was assessed through statistical analysis of the collinearity results generated by MCScanX v1.0.0 [124].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-026-08155-z.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We deeply appreciate the editors and anonymous reviewers for their constructive criticism of the original manuscript.

Authors' contributions

MW and ZF conceived, designed, and funded this study. LD and ZF collected species samples. XC assembled the genome. MW annotated the genome. TQ, XZ, YZ, LD, and YS conducted the bioinformatic analysis. XC analyzed data and wrote the manuscript. XC, MW, and ZF revised the manuscript. All authors have read and approved the manuscript.

Funding

This study was financially supported by the National Natural Science Foundation of China (No. 32000158), the National Science & Technology Fundamental Resources Investigation Program of China (No. 2021XJKK0702), the Foundation of Sustainable Development Research Center of Resources and Environment of Western Sichuan, Sichuan Normal University (No. 2020CXZYHJZX03), and the Key Laboratory of Chemistry in Ethnic Medicinal Resources (Yunnan Minzu University), State Ethnic Affairs Commission & Ministry of Education (No. MZY2301).

Data availability

The datasets supporting the conclusions of this article are available in the National Genomics Data Center (https://ngdc.cncb.ac.cn/) under the BioProject accession number PRJCA038124, Figshare (https://doi.org/10.6084/m9.figshare.28714511.v2), and supplementary files including Figure S1-S7 and Table S1-S10.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Key Laboratory of Land Resources Evaluation and Monitoring in Southwest (Sichuan Normal University), Ministry of Education, Chengdu 610066, China
[2]College of Life Science, Sichuan Normal University, Chengdu 610101, China
[3]Sichuan University Museum, Sichuan University, Chengdu 610065, China
[4]The College of Life Sciences, Sichuan University, Chengdu 610065, China
[5]College of Life Science, Henan Agricultural University, Zhengzhou 450000, China

[6]Institute for Advanced Study, Chengdu University, Chengdu 610106, China
[7]Sustainable Development Research Center of Resources and Environment of Western Sichuan, Sichuan Normal University, Chengdu 610101, China

Received: 27 June 2025 / Accepted: 9 January 2026
Published online: 04 February 2026

References

1. Freire SE. Systematic revision and phylogeny of *Ainsliaea* DC. (Asteraceae, Mutisieae). Ann Mo Bot Gard. 2007;94:79–191. https://doi.org/10.3417/0026-6493(2007)94[79:SRAPOA]2.0.CO;2.
2. Freire SE. Pertyeae (Pertyoideae). In: Funk VA, Susanna A, Stuessy TF, Bayer RJ, editors. Systematics, Evolution, and Biogeography of Compositae. Vienna: International Association for Plant Taxonomy; 2009. p. 315–26.
3. Funk VA, Anderberg AA, Baldwin BG, Bayer RJ, Bonifacino JM, Breitwieser I, et al. Compositae metatrees: the next generation. In: Funk VA, Susanna A, Stuessy TF, Bayer RJ, editors., et al., Systematics, Evolution, and Biogeography of Compositae. Vienna: International Association for Plant Taxonomy; 2009. p. 747–77.
4. Panero JL, Funk VA. The value of sampling anomalous taxa in phylogenetic studies: major clades of the Asteraceae revealed. Mol Phylogenet Evol. 2008;47:757–82. https://doi.org/10.1016/j.ympev.2008.02.011.
5. Zhang CF, Tian J, Cheng YH, Peng S, Chen YS, Gao TG, et al. An updated phylogeny of *Ainsliaea* (Asteraceae: Pertyoideae) and its implications for classification and habit evolution. Taxon. 2024;73:1030–52. https://doi.org/10.1002/tax.13202.
6. Mitsui Y, Chen ST, Zhou ZK, Peng CI, Deng YF, Setoguchi H. Phylogeny and biogeography of the genus *Ainsliaea* (Asteraceae) in the Sino-Japanese region based on nuclear rDNA and plastid DNA sequence data. Ann Bot. 2008;101:111–24. https://doi.org/10.1093/aob/mcm267.
7. Mitsui Y, Setoguchi H. Demographic histories of adaptively diverged riparian and non-riparian species of *Ainsliaea* (Asteraceae) inferred from coalescent analyses using multiple nuclear loci. BMC Evol Biol. 2012;12:254. https://doi.org/10.1186/1471-2148-12-254.
8. Qian YY. A new species of *Ainsliaea* from Yunnan. J Trop Subtrop Bot. 2000;8:161–3. https://doi.org/10.3969/j.issn.1005-3395.2000.2.014.
9. Wang J, Ye YS, Ye HG. *Ainsliaea asaroides* sp. nov. (Asteraceae) from Guangdong, China. Nord J Bot. 2010;28:196–8. https://doi.org/10.1111/j.1756-1051.2009.00590.x.
10. Gao TG, Freire SE, Hind DJN. *Ainsliaea*. In: *Flora of China*, vol. 20–21. Edited by Wu ZY, Raven PH, Hong DY. Beijing & Saint Louis: Science Press & Missouri Botanical Garden; 2011: 15–27.
11. Zhang MJ, Yu WJ, Li HQ. *Ainsliaea simplicissima* (Asteraceae), a new species from southeast China and its phylogenetic position. Phytotaxa. 2019;424:243–52. https://doi.org/10.11646/phytotaxa.424.4.4.
12. Peng Y, Yang C, Luo Y. *Ainsliaea daheishanensis* (Asteraceae): a new species from China. PhytoKeys. 2020;138:233–9.
13. Zhang MJ, Su XX, An C, Li HQ, Zhang Z. *Ainsliaea polystachya* (Asteraceae), a new species from Fujian, China based on morphological and molecular evidence. Phytotaxa. 2021;497:277–84. https://doi.org/10.11646/phytotaxa.497.3.6.
14. Chen XY, Feng YF, Qu TM, Chen H, Liu XF, Pang L, et al. Complete chloroplast genomes of two *Ainsliaea* species and the phylogenetic analysis in the tribe Pertyeae. Front Genet. 2024;15:1408114. https://doi.org/10.3389/fgene.2024.1408114.
15. Fang X, Xu XK, Wang GW, Zeng RT, Tian XH, Shi ZR, et al. Guaianolide sesquiterpenoids from *Ainsliaea yunnanensis*. Phytochemistry. 2017;139:47–55. https://doi.org/10.1016/j.phytochem.2017.04.001.
16. Yuan WL, Dong XY, Huang ZR, Xiao SJ, Ye J, Tian XH, et al. Triterpenoids from *Ainsliaea latifolia* and their cyclooxyenase-2 (COX-2) inhibitory activities. Nat Prod Bioprospect. 2020;10:13–21. https://doi.org/10.1007/s13659-019-00228-x.
17. Zou YB, Zhi ZS. The species and distribution of medicinal plants of the genus *Ainsliaea* DC. in Guizhou Province. China J Chin Mater Med. 1998;23:389–92.
18. Kim HG, Loockerman DJ, Jansen RK. Systematic implications of *ndhF* sequence variation in the Mutisieae (Asteraceae). Syst Bot. 2002;27:598–609. https://doi.org/10.1043/0363-6445-27.3.598.
19. Panero JL, Funk VA. Toward a phylogenetic subfamilial classification for the Compositae (Asteraceae). In: *Proceedings of the Biological Society of Washington*. Washington: Washington Biological Society; 2002;115: 905–22.
20. Zhang C, Huang CH, Liu M, Hu Y, Panero JL, Luebert F, et al. Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. JIPB. 2021;63:1273–93. https://doi.org/10.1111/jipb.13078.
21. Hind D. Mutisieae. In: *Flowering Plants Eudicots: Asterales*. Edited by Kadereirt JW, Jeffrey C. Berlin & Heidelberg: Springer; 2007: 90–123. https://doi.org/10.1007/978-3-540-31051-8
22. Freire SE. Revision of the asian genus *Pertya* (Asteraceae, Pertyoideae). Syst Bot Monogr. 2017;101:1–90.
23. Fu ZX, Jiao BH, Nie B, Zhang GJ, Gao TG. A comprehensive generic-level phylogeny of the sunflower family: implications for the systematics of Chinese Asteraceae. J Syst Evol. 2016;54(4):416–37. https://doi.org/10.1111/jse.12216.
24. Zhang G, Yang J, Zhang C, Jiao B, Panero JL, Cai J, et al. Nuclear phylogenomics of Asteraceae with increased sampling provides new insights into convergent morphological and molecular evolution. Plant Communications. 2024;5:100851. https://doi.org/10.1016/j.xplc.2024.100851.
25. Wang M, Lin H, Lin H, Du P, Zhang S. From species to varieties: how modern sequencing technologies are shaping medicinal plant identification. Genes. 2025;16:16. https://doi.org/10.3390/genes16010016.
26. Ford SA, Ness RW, Kwon M, Ro D-K, Phillips MA. A chromosome level reference genome of Diviner's sage (*Salvia divinorum*) provides insight into salvinorin A biosynthesis. BMC Plant Biol. 2024;24:914. https://doi.org/10.1186/s12870-024-05633-0.
27. Lai Y, Ma J, Zhang X, Xuan X, Zhu F, Ding S, et al. High-quality chromosome-level genome assembly and multi-omics analysis of rosemary (*Salvia rosmarinus*) reveals new insights into the environmental and genome adaptation. Plant Biotechnol J. 2024;22:1833–47. https://doi.org/10.1111/pbi.14305.
28. Li WG, Li YY, Zheng CK, Li ZZ. Chromosome-level genome assembly of a cliff plant *Taihangia rupestris* var. *ciliata* provides insights into its adaptation and demographic history. BMC Plant Biol. 2024;24:596. https://doi.org/10.1186/s12870-024-05322-y.
29. Xue H, Zhao K, Zhao K, Han S, Chitikineni A, Zhang L, et al. A near complete genome of *Arachis monticola*, an allotetraploid wild peanut. Plant Biotechnol J. 2024;22:2110–2. https://doi.org/10.1111/pbi.14331.
30. Yan Z, Sang L, Ma Y, He Y, Sun J, Ma L, et al. A *de novo* assembled high-quality chromosome-scale *Trifolium pratense* genome and fine-scale phylogenetic analysis. BMC Plant Biol. 2022;22:332. https://doi.org/10.1186/s12870-022-03707-5.
31. Zhang H, Lian X, Gao F, Song C, Feng B, Zheng X, et al. A gap-free genome of pillar peach (*Prunus persica* L.) provides new insights into branch angle and double flower traits. Plant Biotechnol J. 2025;23:81–3. https://doi.org/10.1111/pbi.14480.
32. Zhang T, Zhou L, Pu Y, Tang Y, Liu J, Yang L, et al. A chromosome-level genome reveals genome evolution and molecular basis of anthraquinone biosynthesis in *Rheum palmatum*. BMC Plant Biol. 2024;24:261. https://doi.org/10.1186/s12870-024-04972-2.
33. Li Y, Wang F, Pei N, Li Q, Liu H, Yuan W, et al. The updated weeping forsythia genome reveals the genomic basis for the evolution and the forsythin and forsythoside A biosynthesis. Hortic Plant J. 2023;9:1149–61. https://doi.org/10.1016/j.hpj.2022.09.004.
34. Luo X, Li H, Wu Z, Yao W, Zhao P, Cao D, et al. The pomegranate (*Punica granatum* L.) draft genome dissects genetic divergence between soft- and hard-seeded cultivars. Plant Biotechnol J. 2020;18:955–68. https://doi.org/10.1111/pbi.13260.
35. Yuan Z, Fang Y, Zhang T, Fei Z, Han F, Liu C, et al. The pomegranate (*Punica granatum* L.) genome provides insights into fruit quality and ovule developmental biology. Plant Biotechnol J. 2018;16:1363–74. https://doi.org/10.1111/pbi.12875.
36. Yan X, Zhao L, Ren Y, Zhang N, Dong Z, Chen F. Identification of genetic loci and a candidate gene related to flag leaf traits in common wheat by genome-wide association study and linkage mapping. Mol Breed. 2020;40:58. https://doi.org/10.1007/s11032-020-01135-7.
37. Zhang J, Wang M, Guo Z, Guan Y, Guo Y, Yan X. Variation in ploidy level and genome size of *Cynodon dactylon* (L.) Pers. along a latitudinal gradient. Folia Geobot. 2019;54:267–78. https://doi.org/10.1007/s12224-019-09359-y.
38. Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 2016;181:1–20. https://doi.org/10.1111/boj.12385.

39. Susanna A, Baldwin BG, Bayer RJ, Bonifacino JM, Garcia-Jacas N, Keeley SC, et al. The classification of the Compositae: a tribute to Vicki Ann Funk (1947–2019). Taxon. 2020;69:807–14. https://doi.org/10.1002/tax.12235.
40. Kane NC, Gill N, King MG, Bowers JE, Berges H, Gouzy J, et al. Progress towards a reference genome for sunflower. Botany. 2011;89:429–37. https://doi.org/10.1139/b11-032.
41. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546:148–52. https://doi.org/10.1038/nature22380.
42. Feng T, McKibben M, Lovell J, Michelmore R, Rieseberg LH, Barker M, et al. Phylogenomic synteny analysis tracks conserved ancient polyploid-derived triplicated genomic blocks across Asteraceae genomes. bioRxiv. 2025:2025.01.08.631874. https://doi.org/10.1101/2025.01.08.631874
43. Acquadro A, Barchi L, Portis E, Mangino G, Valentino D, Mauromicale G, et al. Genome reconstruction in *Cynara cardunculus* taxa gains access to chromosome-scale DNA variation. Sci Rep. 2017;7:5617. https://doi.org/10.1038/s41598-017-05085-7.
44. Dong Y, Wang X, Ahmad N, Sun Y, Wang Y, Liu X, et al. The *Carthamus tinctorius* L. genome sequence provides insights into synthesis of unsaturated fatty acids. BMC Genomics. 2024;25:510. https://doi.org/10.1186/s12864-024-10405-z.
45. Fan W, Wang S, Wang H, Wang A, Jiang F, Liu H, et al. The genomes of chicory, endive, great burdock and yacon provide insights into Asteraceae palaeo-polyploidization history and plant inulin production. Mol Ecol Resour. 2022;22:3124–40. https://doi.org/10.1111/1755-0998.13675.
46. Reatini B, Pelosi JA, Cang FA, Jiang Q, McKibben MTW, Barker MS, et al. Chromosome-scale reference genome and RAD-based genetic map of yellow starthistle (*Centaurea solstitialis*) reveal putative structural variation and QTL associated with invader traits. Genome Biol Evol. 2024;16:evae243. https://doi.org/10.1093/gbe/evae243.
47. Wang K, Jin J, Wang J, Wang X, Sun J, Meng D, et al. The complete telomere-to-telomere genome assembly of lettuce. Plant Communications. 2024;5:101011. https://doi.org/10.1016/j.xplc.2024.101011.
48. Zhang B, Wang Z, Han X, Liu X, Wang Q, Zhang J, et al. The chromosome-scale assembly of endive (*Cichorium endivia*) genome provides insights into the sesquiterpenoid biosynthesis. Genomics. 2022;114:110400. https://doi.org/10.1016/j.ygeno.2022.110400.
49. Liu B, Yan J, Li W, Yin L, Li P, Yu H, et al. *Mikania micrantha* genome provides insights into the molecular mechanism of rapid growth. Nat Commun. 2020;11:340. https://doi.org/10.1038/s41467-019-13926-4.
50. Nevado B, Chapman MA, Brennan AC, Clark JW, Wong ELY, Batstone T, et al. Genomic changes and stabilization following homoploid hybrid speciation of the Oxford ragwort *Senecio squalidus*. Curr Biol. 2024;34:4412–23. https://doi.org/10.1016/j.cub.2024.08.009.
51. Shen Q, Zhang L, Liao Z, Wang S, Yan T, Shi P, et al. The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. Mol Plant. 2018;11:776–88. https://doi.org/10.1016/j.molp.2018.03.015.
52. Wang M, Zhang S, Li R, Zhao Q. Unraveling the specialized metabolic pathways in medicinal plant genomes: a review. Front Plant Sci. 2024;15:1459533. https://doi.org/10.3389/fpls.2024.1459533.
53. Hjelmen CE. Genome size and chromosome number are critical metrics for accurate genome assembly assessment in eukaryota. Genetics. 2024;227:iyae099. https://doi.org/10.1093/genetics/iyae099.
54. Wang P, Wang F. A proposed metric set for evaluation of genome assembly quality. Trends Genet. 2023;39:175–86. https://doi.org/10.1016/j.tig.2022.10.005.
55. Louwers M, Splinter E, van Driel R, de Laat W, Stam M. Studying physical chromatin interactions in plants using chromosome conformation capture (3C). Nat Protoc. 2009;4:1216–29. https://doi.org/10.1038/nprot.2009.113.
56. Laforest M, Martin SL, Bisaillon K, Soufiane B, Meloche S, Page E. A chromosome-scale draft sequence of the Canada fleabane genome. Pest Manag Sci. 2020;76:2158–69. https://doi.org/10.1002/ps.5753.
57. Zhang QJ, Li W, Li K, Nan H, Shi C, Zhang Y, et al. The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. Mol Plant. 2020;13:935–8. https://doi.org/10.1016/j.molp.2020.04.009.
58. Miao Y, Luo D, Zhao T, Du H, Liu Z, Xu Z, et al. Genome sequencing reveals chromosome fusion and extensive expansion of genes related to secondary metabolism in *Artemisia argyi*. Plant Biotechnol J. 2022;20:1902–15. https://doi.org/10.1111/pbi.13870.
59. Samoluk SS, Vaio M, Ortíz AM, Chalup LMI, Robledo G, Bertioli DJ, et al. Comparative repeatome analysis reveals new evidence on genome evolution in wild diploid *Arachis* (Fabaceae) species. Planta. 2022;256:50. https://doi.org/10.1007/s00425-022-03961-9.
60. Zhang QJ, Gao LZ. Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-genome *Oryza* species. G3 Genes|Genomes|Genetics. 2017;7:1875–85. https://doi.org/10.1534/g3.116.037572.
61. Ding L, Kapp P, Cai F, Garzione CN, Xiong Z, Wang H, et al. Timing and mechanisms of Tibetan Plateau uplift. Nat Rev Earth Environ. 2022;3:652–67. https://doi.org/10.1038/s43017-022-00318-4.
62. Demuth JP, Hahn MW. The life and death of gene families. Bioessays. 2009;31:29–39. https://doi.org/10.1002/bies.080085.
63. Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 2005;15:1153–60. https://doi.org/10.1101/gr.3567505.
64. Zhang J, Fu XX, Li RQ, Zhao X, Liu Y, Li MH, et al. The hornwort genome and early land plant evolution. Nat Plants. 2020;6:107–18. https://doi.org/10.1038/s41477-019-0588-4.
65. Rispe C, Kutsukake M, Doublet V, Hudaverdian S, Legeai F, Simon JC, et al. Large gene family expansion and variable selective pressures for cathepsin B in aphids. Mol Biol Evol. 2007;25:5–17. https://doi.org/10.1093/molbev/msm222.
66. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016;17:379–91. https://doi.org/10.1038/nrg.2016.39.
67. Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, et al. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant. 2018;11:414–28. https://doi.org/10.1016/j.molp.2018.01.002.
68. de Van Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18:411–24. https://doi.org/10.1038/nrg.2017.26.
69. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 2016;30:159–65. https://doi.org/10.1016/j.pbi.2016.03.015.
70. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature. 2011;473:97–100. https://doi.org/10.1038/nature09916.
71. Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, et al. The opium poppy genome and morphinan production. Science. 2018;362:343–7. https://doi.org/10.1126/science.aat4096.
72. Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, et al. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol. 2008;25:2445–3255. https://doi.org/10.1093/molbev/msn187.
73. Barker MS, Li Z, Kidder TI, Reardon CR, Lai Z, Oliveira LO, et al. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am J Bot. 2016;103:1203–11. https://doi.org/10.3732/ajb.1600113.
74. Huang CH, Zhang C, Liu M, Hu Y, Gao T, Qi J, et al. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. Mol Biol Evol. 2016;33:2820–35. https://doi.org/10.1093/molbev/msw157.
75. Kong X, Zhang Y, Wang Z, Bao S, Feng Y, Wang J, et al. Two-step model of paleohexaploidy, ancestral genome reshuffling and plasticity of heat shock response in Asteraceae. Hortic Res. 2023;10:uhad073. https://doi.org/10.1093/hr/uhad073.
76. Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison JS. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Ann Bot. 2017;120:183–94. https://doi.org/10.1093/aob/mcx079.
77. Chen J, Shen CZ, Guo YP, Rao GY. Patterning the Asteraceae capitulum: duplications and differential expression of the flower symmetry *CYC2*-like genes. Front Plant Sci. 2018;9:551. https://doi.org/10.3389/fpls.2018.00551.
78. Elomaa P, Zhao Y, Zhang T. Flower heads in Asteraceae—recruitment of conserved developmental regulators to control the flower-like inflorescence architecture. Hortic Res. 2018;5:36. https://doi.org/10.1038/s41438-018-0056-8.
79. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc. 2006;1:2320–5. https://doi.org/10.1038/nprot.2006.384.
80. Rao Suhas SP, Huntley Miriam H, Durand Neva C, Stamenova Elena K, Bochkov Ivan D, Robinson James T, et al. A 3D map of the human genome

at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
81. Xie T, Zheng JF, Liu S, Peng C, Zhou YM, Yang QY, et al. *De novo* plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. Mol Plant. 2015;8:489–92. https://doi.org/10.1016/j.molp.2014.12.015.
82. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
83. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2:e107. https://doi.org/10.1002/imt2.107.
84. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
85. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70. https://doi.org/10.1093/bioinformatics/btr011.
86. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11:1432. https://doi.org/10.1038/s41467-020-14998-3.
87. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5. https://doi.org/10.1038/s41592-020-01056-5.
88. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100. https://doi.org/10.1093/bioinformatics/bty191.
89. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:1–4. https://doi.org/10.1093/gigascience/giab008.
90. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15. https://doi.org/10.1038/s41587-019-0201-4.
91. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
92. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3:95–8. https://doi.org/10.1016/j.cels.2016.07.002.
93. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5. https://doi.org/10.1126/science.aal3327.
94. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101. https://doi.org/10.1016/j.cels.2015.07.012.
95. Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. Nucleic Acids Res. 2020;48:W177–84. https://doi.org/10.1093/nar/gkaa220.
96. Ou S, Jiang N. Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2017;176:1410–22. https://doi.org/10.1104/pp.17.01310.
97. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR assembly index (LAI). Nucleic Acids Res. 2018;46:e126. https://doi.org/10.1093/nar/gky730.
98. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 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. https://doi.org/10.1093/molbev/msab199.
99. Han Y, Wessler SR. Mite-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38:e199. https://doi.org/10.1093/nar/gkq862.
100. Xu Z, Wang H. Ltr_finder: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265-8. https://doi.org/10.1093/nar/gkm286.
101. Ellinghaus D, Kurtz S, Willhoeft U. *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18. https://doi.org/10.1186/1471-2105-9-18.
102. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7. https://doi.org/10.1073/pnas.1921046117.
103. Tarailo-Graovac M, Chen N. Using repeatmasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009;25:1–14. https://doi.org/10.1002/0471250953.bi0410s25.
104. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12:2. https://doi.org/10.1186/s13100-020-00230-y.
105. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11. https://doi.org/10.1186/s13100-015-0041-9.
106. Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinformatics. 2019;65:e57. https://doi.org/10.1002/cpbi.57.
107. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14:988–95. https://doi.org/10.1101/gr.1865504.
108. Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, et al. Comparative functional genomics of the fission yeasts. Science. 2011;332:930–6. https://doi.org/10.1126/science.1203357.
109. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. https://doi.org/10.1038/nbt.1883.
110. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66. https://doi.org/10.1093/nar/gkg770.
111. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9:R7. https://doi.org/10.1186/gb-2008-9-1-r7.
112. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. Interproscan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. https://doi.org/10.1093/bioinformatics/btu031.
113. Paysan Lafosse T, Andreeva A, Blum M, Chuguransky Sara R, Grego T, Pinto Beatriz L, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 2024;53:D523–34. https://doi.org/10.1093/nar/gkae997.
114. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8. https://doi.org/10.1038/s41592-021-01101-x.
115. Consortium TGO, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The gene ontology knowledgebase in 2023. Genetics. 2023;224:iyad031. https://doi.org/10.1093/genetics/iyad031.
116. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2020;49:D344–54. https://doi.org/10.1093/nar/gkaa977.
117. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28:45–8. https://doi.org/10.1093/nar/28.1.45.
118. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36:3420–35. https://doi.org/10.1093/nar/gkn176.
119. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics. 2015;16:169. https://doi.org/10.1186/s12859-015-0611-3.
120. Chan PP, Lin BY, Mak AJ, Lowe TM. TRNAscan-se 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49:9077–96. https://doi.org/10.1093/nar/gkab688.
121. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35:3100–8. https://doi.org/10.1093/nar/gkm160.
122. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5. https://doi.org/10.1093/bioinformatics/btt509.
123. Ontiveros-Palacios N, Cooke E, Nawrocki EP, Triebel S, Marz M, Rivas E, et al. Rfam 15: RNA families database in 2025. Nucleic Acids Res. 2024;53:D258–67. https://doi.org/10.1093/nar/gkae1023.
124. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49. https://doi.org/10.1093/nar/gkr1293.
125. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45. https://doi.org/10.1101/gr.092759.109.

126. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:304. https://doi.org/10.12688/f1000research.23297.2.
127. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238. https://doi.org/10.1186/s13059-019-1832-y.
128. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.
129. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3. https://doi.org/10.1093/bioinformatics/btp348.
130. Vaidya G, Lohman DJ, Meier R. Sequencematrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics. 2011;27:171–80. https://doi.org/10.1111/j.1096-0031.2010.00329.x.
131. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3. https://doi.org/10.1093/bioinformatics/btu033.
132. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. Modeltest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol. 2020;37:291–4. https://doi.org/10.1093/molbev/msz189.
133. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91. https://doi.org/10.1093/molbev/msm088.
134. Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, et al. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol. 2022;39:msac174. https://doi.org/10.1093/molbev/msac174.
135. Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2020;36:5516–8. https://doi.org/10.1093/bioinformatics/btaa1022.
136. Edgar RC. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun. 2022;13:6968. https://doi.org/10.1038/s41467-022-34630-w.
137. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. Kaks_calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010;8:77–80. https://doi.org/10.1016/s1672-0229(10)60008-3.

Publisher's Note