Chromosome-Level Genome Assembly of *Herpetospermum pedunculosum* (Cucurbitaceae)

Yixi Yang[1,2], Bowen Zhang[1,2], Ying Bao[1,2], Peng Huang[3], Jian Li[1,4], Rui Li[1,2,*], and Qi Zhao[1,2,*]

[1]Engineering Research Center of Sichuan-Tibet Traditional Medicinal Plant, Chengdu University, China
[2]School of Food and Biological Engineering, Chengdu University, China
[3]Tibet Rhodiola Pharmaceutical Holding Company, Lhasa, China
[4]School of Basic Medical Sciences, Chengdu University, China

*Corresponding authors: E-mails: lirui@cdu.edu.cn; zhaoqi@cdu.edu.cn.

Accepted: 11 January 2023

Abstract

This study presents a chromosome-level reference genome assembly of a traditional Tibetan medicinal plant, *Herpetospermum pedunculosum* belonging to the Cucurbitaceae family. Following a combined PacBio high-fidelity sequencing and Hi-C analysis, a final *H. pedunculosum* genome assembly, 804.11 Mb in length was obtained, 90.45% of which was anchored into ten pseudochromosomes with a contig N50 of 24.39 Mb. In addition, 579.55 Mb repetitive sequences and 23,924 high-confidence protein-coding genes were annotated. Phylogenetic analysis revealed that *H. pedunculosum* was sister to a clade formed by cucumber, zucchini, and wax gourd. Further whole-genome duplication analysis revealed no recent polyploidization event in the *H. pedunculosum* genome. The high-quality *H. pedunculosum* genome presented here will be highly useful in investigating the molecular mechanisms underlying the biosynthesis of its active compounds and adaptation strategies to the extreme environment. It will also provide great insights into comparative genomic studies of Cucurbitaceae and flowering plants.

Key words: *Herpetospermum pedunculosum*, genome assembly, PacBio high-fidelity sequencing, genome evolution.

Significance

For the first time, we report the reference-level genome assembly of the traditional Tibetan medicinal plant, *Herpetospermum pedunculosum*. This genome provides a valuable genomic resource for investigating the gene function and genetic breeding of *H. pedunculosum*. It is also beneficial for the comparative genomic studies of species in the Cucurbitaceae family and flowering plants.

Introduction

Cucurbitaceae, commonly known as the gourd family or cucurbits, is an ecologically important flowering plant family consisting of about 960 species distributed globally in tropical and subtropical areas. It consists of several important domesticated vegetables and fruits, including cucumber (*Cucumis sativus*), melon (*Cucumis melo*), watermelon (*Citrullus lanatus*), wax gourd (*Benincasa hispida*), bitter gourd (*Momordica charantia*), bottle gourd (*Lagenaria siceraria*), and zucchini (*Cucurbita pepo*) (Chomicki et al. 2020). In addition, some species in the Cucurbitaceae family have antioxidant, antidiabetic, antiinflammatory, and purgative properties; hence, they have been used for centuries for their medicinal value (Rolnik and Olas 2020). Cucurbits are characterized by shoot-derived tendrils with threadlike shapes and complex modifications during their development, making them an important model system for understanding the molecular regulation of tendril development (Sousa-Baena et al. 2018; Guo et al. 2020). With the rapid development of sequencing technologies,

genome sequencing of 18 species of Cucurbitaceae has been completed (Ma et al. 2022), including cucumber (Huang et al. 2009), bitter gourd (Matsumura et al. 2020), wax gourd (Xie et al. 2019), watermelon (Guo et al. 2013), and zucchini (Montero-Pau et al. 2018). These genomic resources are beneficial for studying the genome evolution, gene function, and molecular breeding of species in the Cucurbitaceae family.

Herpetospermum pedunculosum (Ser.) C. B. Clarke is an annual herbaceous plant in the family Cucurbitaceae that is mainly distributed in the high-altitude (2,300–3,500 m asl) areas of Southwest China, Nepal, and Northeast India. The dried ripe seeds of *H. pedunculosum* are utilized as a traditional Tibetan medicine to treat liver diseases, cholic disorders, and dyspepsia (Wei et al. 2020). Recent studies on *H. pedunculosum* have revealed its pharmacological and phytochemical properties and adaptation strategies to the extreme environment (Li et al. 2018; Ma et al. 2019; Wei et al. 2020, 2021). However, information on the genetic and genomic resources of *H. pedunculosum* is limited, which limits the deep understanding of its genetics and genomics characteristics.

Herein, we present a high-quality, chromosome-level reference genome sequence of *H. pedunculosum* by integrating PacBio high-fidelity (HiFi) sequencing and high-throughput chromosomal conformation capture (Hi-C) technology. A high-confidence set of protein-coding genes (PCGs) was generated, and the evolutionary history of *H. pedunculosum* was investigated. The high-quality genome assembly provides insights into the gene function and genetic breeding of *H. pedunculosum*, which is crucial for the future comparative genomic studies of Cucurbitaceae and flowering plants.

Results and Discussion

Genome Sequencing and Assembly

A total of 31.04 Gb Illumina short reads (supplementary table S1, Supplementary Material online) are generated for a 17-mer frequency distribution analysis. The genome size of *H. pedunculosum* is estimated to be 806.93 Mb with a heterozygosity of 1.20% and a repeat content of 74.92% (supplementary table S2, Supplementary Material online). In addition, 22.02 Gb PacBio HiFi reads were generated and de novo assembled into 239 contigs with a total length of 804.11 Mb, covering 99.65% of the estimated genome. These contigs are anchored into pseudochromosomes using a Hi-C scaffolding approach based on 82.66 Gb Hi-C sequencing data (supplementary table S1, Supplementary Material online). Ultimately, 727.23 Mb (90.45% genome) of genome sequences are successfully assigned to ten pseudochromosomes ranging from 56.95 to 88.34 Mb in length and an average gap of 6.1 per pseudochromosome (fig. 1*A* and supplementary table S3, Supplementary Material online). The final assembly contained 250 contigs and 189 scaffolds, with a contig and scaffold N50 of 24.39 and 71.40 Mb, respectively (table 1 and supplementary table S4, Supplementary Material online). The genome assembly had a guanine-cytosine (GC) content of 38.38%, with a higher GC content in central regions than in distal regions of the pseudochromosomes (fig. 1*B*). Besides, the genome had a benchmarking universal single-copy orthologs (BUSCO) completeness score of 97.65% (table 1 and supplementary table S5, Supplementary Material online) and an overall long terminal repeat (LTR) assembly index (LAI) score of 22.74 (supplementary fig. S1, Supplementary Material online), implying that the genome is of high quality and completeness.

Genome Annotation

A total of 579.55 Mb repetitive sequences representing 72.08% of the *H. pedunculosum* genome are identified (supplementary table S6, Supplementary Material online). LTR retrotransposons are the most abundant repeat sequences, consisting of 549.62 Mb (68.36%) of the *H. pedunculosum* genome, followed by tandem repeats (50.73 Mb), unclassified repeats (13.62 Mb), long interspersed nuclear elements (4.89 Mb), and DNA transposons (3.07 Mb) (supplementary table S7, Supplementary Material online). In addition, 23,924 high-confidence PCGs are annotated after masking all repetitive sequences (supplementary table S8, Supplementary Material online), with a BUSCO completeness score of 94.05% in protein code (supplementary table S5, Supplementary Material online). Among them, 97.90% (23,421 PCGs) were located on ten pseudochromosomes, with an overall gene density of 32.2 genes per Mb. These PCGs have an average transcript length of 4,079 bp, an average coding sequence length of 1,234 bp, 5.1 exons per gene, and an average intron length of 693 bp (table 1). Overall, 23,386 (97.75%) PCGs are functionally annotated using at least one of the publicly available databases, 14,071 (58.82%) PCGs are assigned to gene ontology (GO) terms, and 18,469 (77.20%) are mapped into known Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (supplementary table S9, Supplementary Material online). At the same time, 30,866 non-coding RNAs (ncRNAs), including 20,736 ribosomal RNAs (rRNAs), 8,022 transfer RNAs (tRNAs), 437 microRNAs (miRNAs), and 1,671 small nuclear RNAs (snRNAs) are annotated (supplementary table S10, Supplementary Material online).

Genome Evolution Analysis

A total of 3,369 single-copy orthogroups were identified in *H. pedunculosum* and other sequenced plant species,

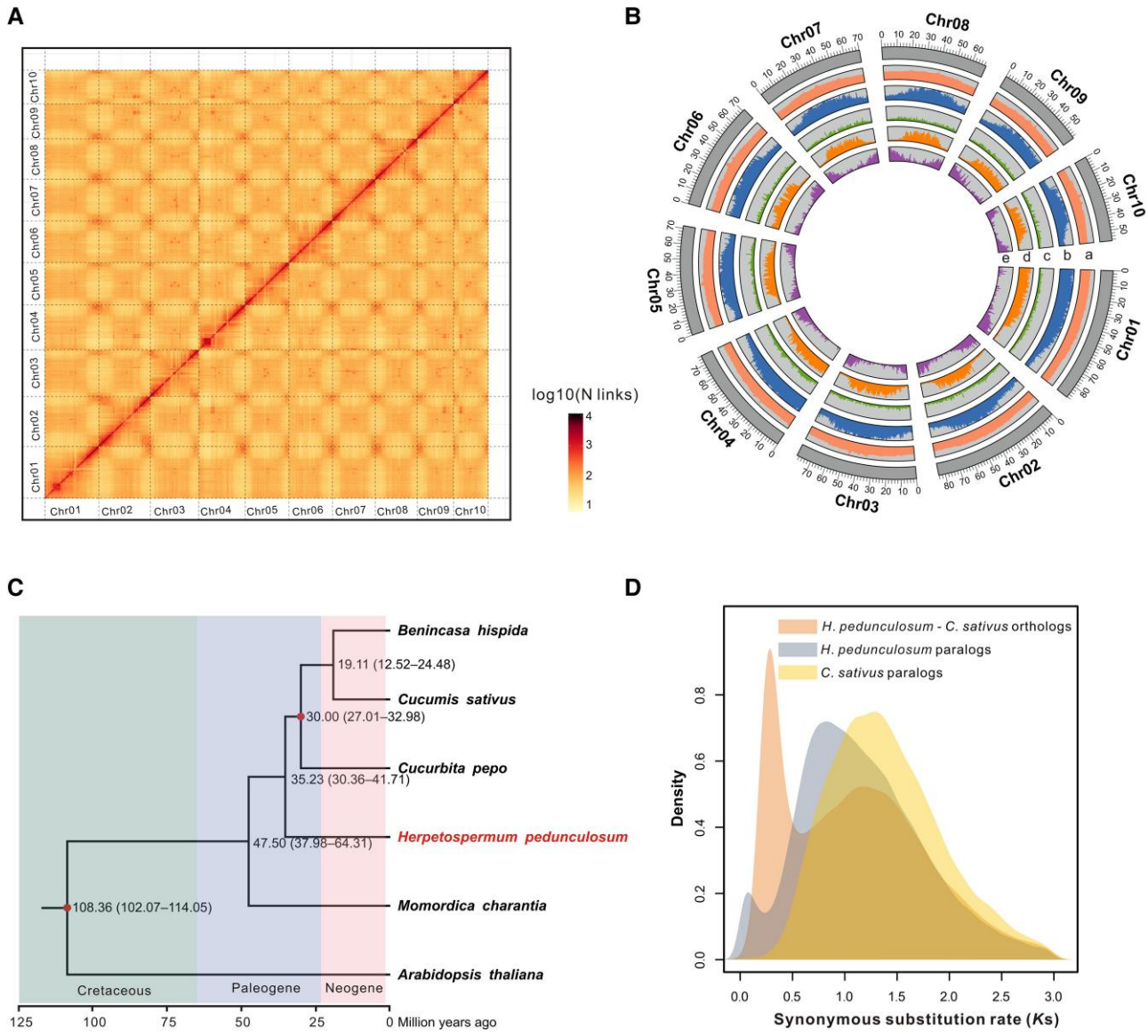

Fig. 1.—Features and evolutionary dynamics of the *H. pedunculosum* genome. (*A*) Heatmap showing Hi-C interactions among ten pseudochromosomes. (*B*) Genome features in non-overlapping windows of 500 kb across the *H. pedunculosum* genome. Tracks from outside to inside are as follows: (*a*) GC content, (*b*) repeat density, (*c*) LTR/*Copia* density, (*d*) LTR/*Gypsy* density, and (*e*) gene density. (*C*) A species tree on the basis of 3,369 single-copy orthogroups from six plant species. Numbers on nodes represent the species divergence times with 95% confidence intervals. Circles represent the fossil calibration points obtained from the TimeTree database. (*D*) K_s distributions for the whole paranome identified from the whole genome of *H. pedunculosum* and cucumber.

including four Cucurbitaceae species (cucumber, zucchini, wax gourd, and bitter gourd) and an outgroup species (*Arabidopsis thaliana*). A fully supported maximum likelihood phylogenetic tree is constructed based on these single-copy genes, where *H. pedunculosum* clustered with a clade formed by cucumber, zucchini, and wax gourd (CZW clade) (fig. 1*C*), which was consistent with the findings reported by Guo et al. (2020). The divergence time between *H. pedunculosum* and the CZW clade was estimated to be around 35 million years ago (Mya), which was slightly younger than the estimated divergence time (~42 Mya) obtained by Guo et al. (2020). Besides, the K_s distribution of orthologous gene pairs between *H. pedunculosum* and cucumber has a major peak at 0.28, which is younger than the two peaks identified from the paralogues gene pairs within *H. pedunculosum* (0.82) and cucumber (1.29) (fig. 1*D*), implying that no recent whole-genome duplication (WGD) event has occurred in the *H. pedunculosum* genome following its split from CZW clade, which is consistent with the results revealed by gene tree analysis in Guo et al. (2020).

Table 1
Statistics of the Assembly and Annotation of the *H. pedunculosum* Genome

Assembly	
Total length (Mb)	804.06
GC content (%)	38.38
Contig N50 (Mb)	24.39
Contig number	250
Scaffold N50 (Mb)	71.40
Scaffold number	189
Pseudochromosome number	10
BUSCO completeness score (%)	97.65
Annotation	
Total length of repeats (Mb)	579.55
Number of protein-coding genes	23,924
Mean transcript length (bp)	4,078.82
Mean coding sequence length (bp)	1,234.09
Mean exon length (bp)	241.78
Mean intron length (bp)	693.13
Average exons per gene	5.10
BUSCO completeness score (%)	94.05

Conclusions

This study presents a chromosome-level assembly of a highly heterozygous and repeat-rich *H. pedunculosum* genome. The final genome assembly is of high quality and completeness, with an average gap of 6.1 per pseudochromosome and a chromosome anchoring rate of 90.45%. Genome annotation of *H. pedunculosum* yields 23,924 high-confidence gene models and 30,866 ncRNAs. Besides, *H. pedunculosum* clustered with a clade formed by cucumber, zucchini, and wax gourd, with their split around 35 Mya. Still, there is no recent WGD event in the *H. pedunculosum* genome following its split from the CZW clade. The high-quality *H. pedunculosum* genome presented here will facilitate the genetic and genomic studies of *H. pedunculosum*, and comparative genomic studies of Cucurbitaceae and flowering plants.

Materials and Methods

Sample Preparation

Herpetospermum pedunculosum seeds were purchased from Tibet Pharmaceutical Planting Base and planted in a controlled laboratory environment at Chengdu University. Fresh leaves were collected from mature *H. pedunculosum* plants and frozen immediately in liquid nitrogen awaiting DNA extraction. In addition, young and mature seeds were also harvested from the same plants for RNA extraction.

DNA and RNA Sequencing

Total genomic DNA was extracted from the frozen leaves using the cetyltrimethylammonium bromide method (Doyle and Doyle 1987). Paired-end libraries with an insertion size of 350 bp were prepared and sequenced on an Illumina HisSeq 2500 platform. Next, 15–17 kb SMRTbell libraries were constructed and sequenced using the circular consensus sequencing (CCS) mode on the PacBio Sequel IIe platform. In addition, Hi-C libraries were constructed following the protocol described by Louwers et al. (2009) and sequenced on an Illumina HiSeq 2500 platform.

Subsequently, total RNA was extracted from the seeds, and residual DNA was removed using a DNA-free DNA removal kit (Thermo Fisher Scientific, USA). Next, the RNA-seq libraries were prepared using the TruSeq Stranded mRNA-Seq kit following the manufacturer's protocol and sequenced on the Illumina HiSeq 2500 platform.

Genome Size Estimation

The genome size of *H. pedunculosum* was estimated based on a 17-mer depth distribution of Illumina reads using Jellyfish v2.2.9 (Marçais and Kingsford 2011). First, the low-frequency *k*-mers were removed, and the genome size was calculated using the following formula: Genome size = *k*-mer number/homozygous peak depth.

Genome Assembly and Assessment

The HiFi reads were processed into a CCS analysis workflow using SMRT Link v8.0 (PacBio) and assembled into contigs using hifiasm v0.14 (Cheng et al. 2021). The resulting contigs were anchored into pseudochromosomes using the ALLHiC algorithm, following pruning, partition, rescue, optimization, and building steps (Zhang et al. 2019). Subsequently, the Hi-C contact maps were plotted using Juicebox v1.8.8 (Durand et al. 2016). Obvious scaffolding errors were manually adjusted according to the contact maps plotted. To assess the genome quality, the BUSCO completeness score of the *H. pedunculosum* genome was calculated using BUSCO v5.2.2 (Simão et al. 2015) based on the Embryophyta odb10 dataset. Finally, the LAI score of the genome was calculated using LTR_retriever v2.8 (Ou and Jiang 2018).

Identification of Repetitive Sequences

The repetitive sequences within the *H. pedunculosum* genome were identified using the homology-based and de novo methods. Tandem repeats were predicted using Tandem Repeats Finder (TRF) v4.09 (Benson 1999). For the homology-based prediction, Repbase database v22.11, RepeatProteinMask, and RepeatMasker v4.1.0 (Tarailo-Graovac and Chen 2009) were used to search known nucleotide and amino acid repeats in the genome. Subsequently, a de novo repeat library was built using LTR_FINDER v1.06 (Xu and Wang 2007), RepeatScout v1.0.5, and RepeatModeler v2.0.1 (Price et al. 2005).

Finally, all repeat libraries identified were combined into a non-redundant library, and a DNA-level repeat detection was performed using RepeatMasker v4.1.0.

Annotation of PCGs

Structural annotation of PCGs was performed using a combination of homology-based searches, de novo predictions, and transcriptome-based methods. First, protein sequences of cucumber, zucchini, wax gourd, bitter gourd, and *A. thaliana* were aligned to the *H. pedunculosum* genome using TBLASTN v2.2.26 (Camacho et al. 2009). Subsequently, gene structures were predicted using GeneWise v2.4.1 (Birney et al. 2004) based on homology alignments. Second, de novo gene model predictions were performed using Augustus v3.2.3 (Stanke et al. 2006), Geneid v1.4.4 (Blanco et al. 2007), GeneScan v1.0 (Burge and Karlin 1997), GlimmerHMM v3.04 (Majoros et al. 2004), and SNAP v2013.11.29 (Korf 2004). Third, a comprehensive transcriptome database was constructed by combining de novo and genome-guided RNA-seq assemblies, and potential protein-coding regions were predicted using PASA v2.2.0 (Haas et al. 2003) based on the transcript alignments to the assembly. Finally, all gene models predicted using the above three methods were integrated into a non-redundant reference gene set using EVM v1.1.1 (Haas et al. 2008).

Gene functions were assigned by aligning the protein sequences against the Swiss-Prot (Bairoch and Apweiler 2000) and NCBI non-redundant protein databases using DIAMOND v0.8.22 (Buchfink et al. 2015). Subsequently, the protein motifs and domains were identified using InterProScan v5.35 (Hunter et al. 2009). GO terms were assigned according to the corresponding InterPro entry. KEGG pathways in which the genes might be involved were annotated using the KEGG Automatic Annotation Server (Moriya et al. 2007).

Annotation of ncRNAs

Four ncRNAs, including tRNAs, miRNAs, rRNAs, and snRNAs, were annotated as previously outlined by Wang et al. (2022). The tRNAs were predicted using tRNAscan-SE v1.4, while snRNAs and miRNAs were predicted using the cmsearch program in Infernal v1.1.3 (Nawrocki and Eddy 2013) based on the Rfam database. The rRNAs were identified through homology searches against the publicly available rRNA sequences of *A. thaliana* and *Oryza sativa*.

Comparative Genomic Analysis

The species phylogeny and divergence times were estimated using previously described methods with minor modifications (Wang, Zhang et al. 2022). The protein sequences of *H. pedunculosum*, cucumber, zucchini, wax gourd, bitter gourd, and *A. thaliana* were clustered into families with OrthoFinder v2.3.11 (Emms and Kelly 2019) using the DIAMOND aligner and the Markov cluster algorithm. The protein sequences of single-copy orthogroups were aligned using MAFFT-LINSI v7.313 (Katoh et al. 2002) and concatenated into a supermatrix, followed by the extraction of the conserved sites using Gblocks v0.91b (Castresana 2000). Subsequently, a maximum likelihood tree was constructed using RAxML v8.2.11 (Stamatakis 2014) under the PROTGAMMAILGX model with 100 bootstraps. Next, the species divergence time was estimated using MCMCTREE in PAML v4.9e (Yang 2007). The zucchini–cucumber divergence (27–33 Mya) and *A. thaliana*–bitter gourd (102–114 Mya) were obtained from the TimeTree database and used as fossil calibration points.

In addition, all-against-all pairwise comparisons between *H. pedunculosum* and cucumber protein sequences were performed using the DIAMOND aligner tool. Next, ortholog and paralog gene pairs were identified using MCScanX v1.1 (Wang et al. 2012). Finally, the nonsynonymous substitution rate (K_a) and synonymous substitution rate (K_s) were calculated for each gene pair using the "add_ka_and_ks_to_collinearity.pl" script in MCScanX.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This study was supported by Tibet Autonomous Region Science and Technology Plan (high-tech social development) project (No. XZ202201ZY0031G) to Y.Y.

Data Availability

The *H. pedunculosum* assembly and all sequence data have been deposited at the NCBI under the BioProject PRJNA906014. The genome assembly and annotations are also available on FigShare at the link: https://doi.org/10.6084/m9.figshare.21626153.v2.

Literature Cited

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28(1): 45–48.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Birney E, Clamp M, Durbin R. 2004. Genewise and genomewise. Genome Res. 14(5):988–995.

Blanco E, Parra G, Guigó R. 2007. Using geneid to identify genes. Curr Protoc Bioinform. 18(4.3):1–28.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 268(1):78–94.
Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):1–9.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17(4): 540–552.
Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18(2):170–175.
Chomicki G, Schaefer H, Renner SS. 2020. Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. New Phytol. 226(5):1240–1255.
Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 19:11–15.
Durand NC, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3(1):95–98.
Emms DM, Kelly S. 2019. Orthofinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):1–14.
Guo J, et al. 2020. Phylotranscriptomics in Cucurbitaceae reveal multiple whole-genome duplications and key morphological and molecular innovations. Mol Plant. 13(8):1117–1133.
Guo S, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet. 45(1): 51–58.
Haas BJ, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31(19):5654–5666.
Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9(1):R7.
Huang S, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. Nat Genet. 41(12):1275–1281.
Hunter S, et al. 2009. Interpro: the integrative protein signature database. Nucleic Acids Res. 37(Database issue):D211–D215.
Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5(1): 1–9.
Li Q, et al. 2018. Natural medicines used in the traditional Tibetan medical system for the treatment of liver diseases. Front Pharmacol. 9:29.
Louwers M, Splinter E, Van Driel R, De Laat W, Stam M. 2009. Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C). Nat Protoc. 4(8):1216.
Ma L, et al. 2022. Cucurbitaceae genome evolution, gene function, and molecular breeding. Hortic Res. 9:1–24.
Ma Y, et al. 2019. Cytotoxic lignans from the stems of *Herpetospermum pedunculosum*. Phytochemistry 164:102–110.
Majoros WH, Pertea M, Salzberg SL. 2004. Tigrscan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16):2878–2879.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6): 764–770.
Matsumura H, et al. 2020. Long-read bitter gourd (*Momordica charantia*) genome and the genomic architecture of nonclassic domestication. Proc Natl Acad Sci U S A. 117(25): 14543–14551.
Montero-Pau J, et al. 2018. *De novo* assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the *Cucurbita* genus. Plant Biotechnol J. 16(6):1161–1171.
Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(suppl_2):W182–W185.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.
Ou S, Jiang N. 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 176(2):1410–1422.
Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. Bioinformatics 21(suppl_1):i351–i358.
Rolnik A, Olas B. 2020. Vegetables from the Cucurbitaceae family and their products: positive effect on human health. Nutrition 78:110788.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19): 3210–3212.
Sousa-Baena MS, Lohmann LG, Hernandes-Lopes J, Sinha NR. 2018. The molecular control of tendril development in angiosperms. New Phytol. 218(3):944–958.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9): 1312–1313.
Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(suppl_2):W435–W439.
Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 25(1):4–10.
Wang M, et al. 2022. Improved assembly and annotation of the sesame genome. DNA Res. 29(6):dsac041.
Wang M, Zhang L, Tong S, Jiang D, Fu Z. 2022. Chromosome-level genome assembly of a xerophytic plant, *Haloxylon ammodendron*. DNA Res. 29(2):dsac006.
Wang Y, et al. 2012. MCScanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40(7):e49.
Wei X, et al. 2020. Ethyl acetate extract of *Herpetospermum pedunculosum* alleviates α-naphthylisothiocyanate-induced cholestasis by activating the farnesoid x receptor and suppressing oxidative stress and inflammation in rats. Phytomedicine 76:153257.
Wei X, et al. 2021. Dehydrodiconiferyl alcohol, a lignan from *Herpetospermum pedunculosum*, alleviates cholestasis by activating pathways associated with the farnesoid X receptor. Phytomedicine 80:153378.
Xie D, et al. 2019. The wax gourd genomes offer insights into the genetic diversity and ancestral cucurbit karyotype. Nat Commun. 10(1):1–12.
Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 35(suppl_2): W265–W268.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 5(8):833–845.

Associate editor: Maud Tenaillon