

Giant chromosomes of a tiny plant—the complete telomere-to-telomere genome assembly of the simple thalloid liverwort *Apopellia endiviifolia* (Jungermanniopsida, Marchantiophyta)

Joanna Szablińska-Piernik ^{1,*}, Paweł Sulima ², and Jakub Sawicki ^{1,*}

¹Department of Botany and Evolutionary Ecology, University of Warmia and Mazury in Olsztyn, Plac Łódzki 1, Olsztyn 10-719, Poland

²Department of Genetics, Plant Breeding and Bioresource Engineering, University of Warmia and Mazury in Olsztyn, Plac Łódzki 3, Olsztyn 10-724, Poland

*Correspondence address. Joanna Szablińska-Piernik, Department of Botany and Evolutionary Ecology, University of Warmia and Mazury in Olsztyn, Plac Łódzki 1, Olsztyn 10-719, Poland. E-mail: joanna.szablinska@uwm.edu.pl; Jakub Sawicki, Department of Botany and Evolutionary Ecology, University of Warmia and Mazury in Olsztyn, Plac Łódzki 1, Olsztyn 10-719, Poland. E-mail: jakub.sawicki@uwm.edu.pl

Abstract

Background The liverwort *Apopellia endiviifolia*, a dioicous, simple thalloid species, is notable for its cryptic diversity, habitat adaptability, and genomic innovation, and it represents a clade that is sister to all other Jungermanniopsida. These features make *A. endiviifolia* an essential model for exploring speciation mechanisms and the evolution of genome structures within liverworts.

Findings We present the genome assembly of a haploid *A. endiviifolia* isolate with a total size of 2,914,960,273 bp and an N50 of 468,157,909 bp, demonstrating high completeness (99.2% BUSCO) and a high consensus quality (quality value 47.6). The assembly consisted of 9 chromosomes, which included 18 telomeres and 9 centromeres (ranging from 1.9 to 5 Mbp in length). RNA sequencing-based annotation identified 34,615 genes, predominantly protein coding. The transposable elements comprised 12.16% long terminal repeat elements and 57 Helitrons. Among the retroelements, the *Copia* and *Gypsy* superfamilies comprised 8.94% and 2.95% of the genome, respectively. The *Ty3/Gypsy* superfamily was significantly enriched in centromeric regions. The average GC content ranged from 38.8% to 39.6%, with gene density varying between 5.52 and 9.78 genes per 500 kbp. Synteny analysis of related liverwort species has revealed complex chromosomal relationships, indicating extensive genome rearrangements among species.

Conclusions This study provides the first high-quality reference genome assembly of the haploid liverwort *A. endiviifolia*. Assembly and annotation offer valuable resources for investigating liverwort evolution, centromere biology, and genome expansion in simple thalloid liverworts.

Keywords: *Apopellia endiviifolia*, liverworts, telomere-to-telomere (T2T), genome assembly, centromere

Background

Liverworts (Marchantiophyta) represent one of the earliest terrestrial plants, with fossil evidence indicating their emergence in the Middle Ordovician period approximately 419 to 447 million years ago [1]. They occupy a pivotal position in plant evolutionary history as members of the monophyletic bryophyte lineage, which, along with vascular plants, form 2 distinct monophyletic groups within terrestrial plants. Within bryophytes, liverworts are sister to mosses, forming the Setaphyta clade, which in turn is sister to hornworts. This phylogenetic relationship is supported by recent genomic analyses that affirm the monophyly of bryophytes and clarify the evolutionary relationships among early land plants [2]. Liverworts, together with other bryophytes such as mosses and hornworts, display a range of morphological and physiological traits important for plant adaptation to terrestrial environments. However, it remains debated whether these traits are ancestral characteristics or the result of secondary reductions. Recent evidence tends to support the idea of secondary reductions [3]. This group is often characterized by structural simplicity, a predominantly haploid life cycle, and a slow rate of molecular evolution, which corresponds to gradual morphological di-

versification [4]. However, such generalizations primarily reflect the characteristics of the Haplomitriopsida and Marchantiopsida classes (complex thalloid liverworts) [5]. Significant variation exists within the broader liverwort lineage, particularly in Jungermanniopsida, which includes leafy and simple thalloid liverworts. This class comprises most (>80%) of the extant liverwort species and demonstrates a significantly higher degree of structural complexity and a notably accelerated rate of molecular evolution. This is evidenced by their diverse leafy morphologies, intricate branching patterns, and specialized reproductive structures, which facilitate their adaptation to a wide array of ecological niches [6–8]. Within Jungermanniopsida, the order Pelliiales consists of 2 families: *Noterocladaceae* and *Pelliaceae* [9]. Recent molecular and morphological studies have refined the taxonomy within *Pelliaceae*, distinguishing *Apopellia* as a separate genus that includes 3 species: *Apopellia apicola*, *Apopellia megaspora*, and *Apopellia endiviifolia* [10].

A. endiviifolia (NCBI:txid:304445) is a dioicous, simple thalloid liverwort characterized by its cuneate apical cell, the absence of a midrib in the thallus, a spherical capsule, and a robust seta. It is

Received: June 27, 2025. Revised: September 30, 2025. Accepted: November 27, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

widely distributed across the Northern Hemisphere and thrives in a diverse array of habitats, including aquatic environments such as springs and streambanks, as well as in dry conditions often associated with limestone substrates. Its capacity to grow on limestone rocks, in arid soils, in aquatic habitats (hydrophytes), and on decaying wood (epixyl) underscores its remarkable ecological versatility [11]. This species exemplifies cryptic speciation, with European populations diverging into 2 lineages: A (typical form) and B (water form), which are differentiated by molecular markers and microhabitat preferences [8, 10, 12, 13]. This pattern reflects broader taxonomic revisions within *Pellia* s.l., which have been split into *Apopellia* and *Pellia* s.s. through integrative approaches, highlighting its importance in studying speciation mechanisms [10]. The combination of cryptic diversity, habitat versatility, and genomic novelty renders *A. endiviifolia* a valuable model for exploring speciation mechanisms and the evolution of genomic structure in liverworts.

Bryophytes, including liverworts, mosses, and hornworts, are generally characterized by relatively small nuclear genomes compared to other plant groups. However, significant variation exists within these groups, particularly among liverworts. On average, hornwort genomes measure 244 Mbp (median 205 Mbp), moss genomes average around 504 Mbp (median 433 Mbp), and liverwort genomes tend to be larger, averaging 1,844 Mbp with a median of 751 Mbp [14]. Flow cytometry data further reveal that liverwort genome sizes vary widely, ranging from 206.2 Mbp in *Lejeunea cavifolia* to 20,006 Mbp in *Phyllohallia fuegiana* [15, 16]. Within liverworts, the *Pelliaceae* family is notable for its particularly large genomes: *Pellia borealis* has a genome size of 7,238.3 Mbp, *P. epiphylla* 3,719.2 Mbp, and *A. endiviifolia* 3,364.0 Mbp [15]. Despite efforts to determine the nuclear genome size of over 100 liverwort species, comprehensive genomic resources remain limited, with only a handful of liverwort genomes sequenced at the chromosomal level [17–21].

Although liverwort genome sizes are less variable than those of angiosperms [22], there is still substantial variation in genome size among liverwort species. This variation presents an intriguing area of research, especially given the limited understanding of the patterns and rates of structural changes within liverwort genomes [16]. Studies of genome evolution, particularly in complex thalloid liverworts, have revealed the absence of ancient whole-genome duplication events, minimal rates of gene duplication and chromosomal rearrangements, and rare occurrences of transposable element (TE) bursts [4]. Nevertheless, a comparative analysis of the nuclear genomes of the model liverwort *Marchantia polymorpha* (286.7 Mbp) and *Lunularia cruciata*, which is distinguished by a genome size nearly twice as large (565.6 Mbp), highlights the role of Ty3/Gypsy retrotransposon proliferation in genome size expansion [4]. Furthermore, recent advances in long-read sequencing technologies have enabled the assembly of entire genomes at the telomere-to-telomere scale, providing unprecedented insights into highly repetitive regions such as centromeres and telomeres [23]. These advances have shown that centromere structure is linked to chromosome length evolution. In particular, chromosomes with longer centromeres tend to have a higher proportion and longer stretches of *Copia* TEs within their centromeric regions. This enrichment of *Copia* elements may contribute to a positive association between centromere length and overall chromosome length, potentially influencing karyotype evolution [24]. In bryophytes, the application of near telomere-to-telomere genome assembly has facilitated the identification of centromere sequences in the moss species *Physcomitrium patens*. This has enabled the precise characterization of its centromeres,

revealing 26 monocentric chromosomes, each containing a single centromeric region enriched with RLC5 retrotransposons from the *Bryco* clade of the *Copia* superfamily [25]. Moreover, this study elucidated the evolutionary dynamics of centromeres in nonseed plants, highlighting their unique composition and recent evolution. It also provides a gap-free genomic framework for investigating its role in chromosome stability and segregation. In contrast, a study of *M. polymorpha* revealed that its centromeres consist of simple 162-bp satellite repeats and lack extensive pericentromeric heterochromatin and long terminal repeat (LTR) retrotransposon enrichment typical of flowering plants. Instead, these centromeres are flanked by a specific long interspersed nuclear element (LINE) transposon family [26].

To further advance our understanding of genome evolution and the pivotal role of repetitive elements in the centromere architecture and biology of liverworts, our study provides the first comprehensive, high-quality reference genome assembly of the haploid liverwort *A. endiviifolia* using primary Oxford Nanopore long-read sequencing and Pore-C technology. This high-quality genome assembly and annotation serves as a vital resource for further exploration of liverwort evolution and provides new perspectives on centromere biology and genome expansion mechanisms in simple thalloid liverworts.

Material and Methods

Sample collection and in vitro culture

A. endiviifolia plants were collected from the Nature Reserve of the Sources of the Łyna River (NE Poland; 54.6208°N, 21.2267°E). *In vitro* cultures of *A. endiviifolia* were established in the Plant Biotechnology Laboratory at the Department of Genetics, Plant Breeding and Bioresource Engineering, University of Warmia and Mazury in Olsztyn (Poland). *A. endiviifolia* fragments were washed for 15 minutes in tap water, surface-disinfected with 0.5% calcium hypochlorite and 0.05% TWEEN-20 for 10 minutes, and then triple-rinsed with sterile distilled water (5, 10, and 15 minutes). The sterilization method was highly effective, achieving a culture sterility of 95.65%. Sterile explants were cultured on solid ½ Gamborg's B5 medium (½ basal salts, organics, vitamins [27], 20 g/L sucrose, 8 g/L agar, pH 6.0) at 24°C under a light cycle of 16 hours light/8 hours dark. The upper segments of the sterilized plants served as secondary explants for micropropagation on the same medium, spaced 1–2 cm apart. Micropropagated plants were used in the subsequent experiments (Fig. 1).

A voucher specimen for the sequenced sample has been prepared and deposited in the Herbarium of the University of Warmia and Mazury in Olsztyn (OLS), with the accession number OLS-H2024P004.

DNA extraction

For genome assembly, genomic DNA was extracted from the aerial parts of the thallus using a modified ultra-long DNA extraction protocol [28]. Briefly, 500 mg of the material was ground into a fine powder in liquid nitrogen and incubated in 30 mL of homogenization buffer (HB) for 15 minutes on ice. The suspension was then filtered through a 40-µm cell strainer. Following centrifugation (3,000 × g) and 2 wash cycles with HB buffer, the pellet was dissolved in SDS lysis buffer containing 5 µL RNase A (20 µg/µL) and 75 µL proteinase K (20 µg/µL) and incubated at 50°C for 3 hours. Subsequently, the DNA was extracted using chilled phenol/chloroform/isoamyl alcohol. The DNA was precipitated with isopropanol, washed with 80% ethanol, and eluted in 50 µL of water. The fragment length distribution and DNA integrity were as-



Figure 1 *A. endiviifolia* plants located in the Nature Reserve of the Sources of the Łyna River (A), in vitro culture (B), and thalli morphology under in vitro conditions (C).

Table 1 Assembly statistics

Genome assembly statistics	Value
Genome size [bp]	2,914,960,273
Genome coverage [median]	47.6
Number of chromosomes	9
Number of telomeres	18
Number of centromeres	9
Contig N50 [bp]	76,579,043
Scaffold N50 [bp]	468,157,909
Number of gaps	6
Number of total/protein-coding genes	34,615/33,513
Quality value	47.6
LAI	20.06
BUSCO—genome (Eukaryota/Viridiplantae/Embryophyta) [%]	C:99.2/95.5/63.2 [S:93.0/91.1/60.0, D:6.2/4.4/3.2]

essed using TapeStation with the Genomic DNA ScreenTape Assay (Agilent), reaching a maximum intensity peak with a length of >60,000 bp and DIN 8.2, respectively. The concentration was determined using the Qubit fluorometer HS DNA assay kit and amounted to 65.2 ng/ μ L.

Pore-C procedure

The restriction enzyme Pore-C protocol for plant samples (RE-Pore-C; Oxford Nanopore Technologies) was employed to capture 3-dimensional DNA interactions within chromatin, with a few modifications as previously described by Krawczyk et al. [18]. Chromatin was preserved with formaldehyde, and the crosslinked plant material was cryogenically ground. The resulting suspension was filtered through a 40- μ m strainer, purified, and digested with the NlaIII (NEB) restriction enzyme for 18 hours at 37°C, followed by heat denaturation. Subsequently, a proximity ligation reaction was performed using 40,000 U of T4 DNA ligase for 6 hours at 16°C. This was followed by protein degradation and chromatin de-crosslinking with 100 μ L proteinase K (20 μ g/ μ L) and 3 μ L RNase A (100 μ g/ μ L) for 18 hours at 56°C, with additional rounds of proteinase K digestion, as per the HiPore-C v1 protocol [29]. DNA was extracted using chilled phenol/chloroform/isoamyl alcohol and EDTA, precipitated with NaCl, washed with ethanol, and eluted with TE buffer. DNA quality was assessed using TapeStation, which reached a peak length of 11,142 bp and DIN 5.8. The concentration was 43 ng/ μ L, as measured using a Qubit fluorometer.

Nanopore sequencing

Libraries were prepared for nanopore sequencing of native DNA and proximity-ligated DNA fragments using the Ligation Sequencing Kit V14 (SQK-LSK114, Oxford Nanopore Technologies) following the manufacturer's protocol. DNA sequencing data were generated using the Oxford Nanopore Technologies PromethION 2 platform on R10.4.1 flowcells (PRO-114M) and the MinKNOW sequencing software.

RNA sequencing

Total RNA for the short-read procedure was extracted from 4 samples of male *A. endiviifolia*: (i) land and (ii) water form, (iii) antheridia, and (iv) surrounding thallus. This extraction was conducted using the RNA Plant Mini Spin Kit (Qiagen) according to the manufacturer's protocol. Short-read RNA sequencing (RNA-seq) libraries were prepared using 2 distinct protocols. For RNA extracted from the water and land form of *A. endiviifolia*, libraries were constructed using the TruSeq Stranded Total RNA Library Prep Kit (Illumina) using the Ribo-Zero rRNA option. For antheridia and surrounding thallus, the amount of extracted RNA was insufficient for TruSeq, and libraries were prepared using the QI-Aseq FX Single Cell RNA Library Kit (Qiagen). All libraries were sequenced on an Illumina NovaSeq 6000 platform (Macrogen) in 2 $\times$ 150 paired-end mode. Raw reads were deposited in BioProject PRJNA1279829 and BioSample SAMN49506713.

Genome size estimation, assembly, and quality evaluation

Basecalling

Raw Nanopore signal data from both the standard and Pore-C libraries were basecalled using Dorado v0.9.1 (RRID:SCR_025883) [30]. The superior accuracy model SUP dna_r10.4.1_e8.2_400bps_sup@v5.0.0 was employed to generate high-fidelity basecalls in BAM format. In the case of high-molecular-weight sequencing reads, the DNA modification v3 models were used to detect all context 6mA, 4mC, and 5mC methylation at the single-base accuracy level.

A fastq file was used to count *k*-mer frequencies using *kmerfreq* v4.0 [31]. Subsequently, the genome size was estimated using the GCE v1.0.2 program [32] and *kmerfreq* files. The estimated genome size for *k*-mers from 14 to 21 fell within the 2,950- to 3,090-Mbp range.

Table 2 The identified telomeres and centromeres in the *A. endiviifolia* assembly

Chromosomes	Length [bp]	Telomeres		Centromeres	
		Number of repeats at the left end	Number of repeats at the right end	Start	End
Ch1	529,742,643	425	418	359,400,001	363,400,000
Ch2	473,504,363	352	422	232,800,001	237,800,000
Ch3	468,157,909	115 ^a	414	446,400,001	449,400,000
Ch4	317,755,521	408	402	239,340,762 ^b	241,950,761 ^b
Ch5	310,129,949	481	421	37,548,478 ^b	39,527,839 ^b
Ch6	244,886,949	419	176 ^a	242,160,000	244,159,999
Ch7	236,110,306	393	452	189,670,000	191,669,999
Ch8	234,121,349	428	403	202,730,000	204,729,999
Ch9	100,551,284	446	420	89,933,334	93,766,667

^aTelomeres with a noncanonical ACGCAGC/TGCGTTCG motif.

^bCentromere verification based only on quarTeT and Pore-C contact map.

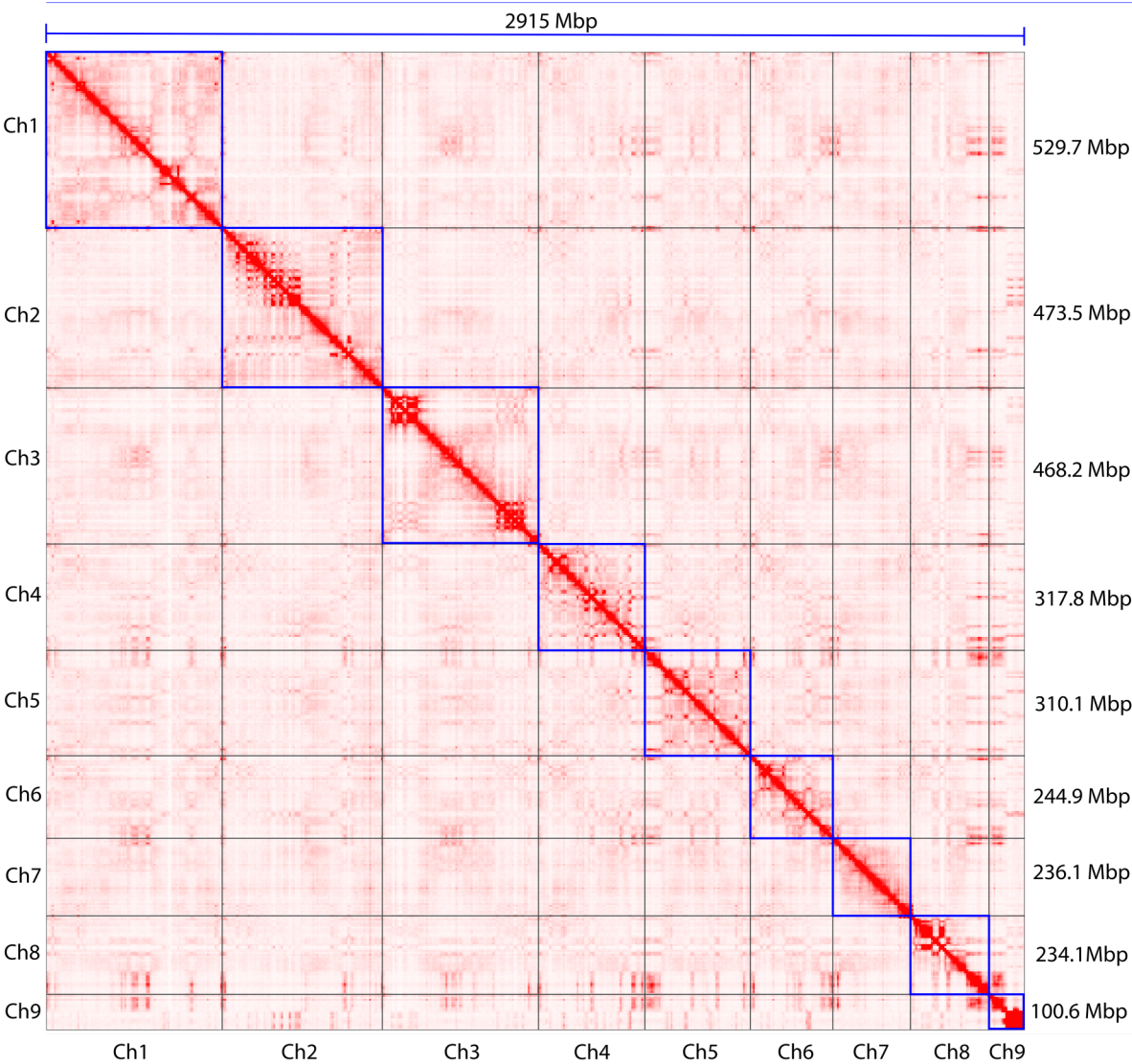


Figure 2 Pore-C interaction heatmap of the *A. endiviifolia* genome.

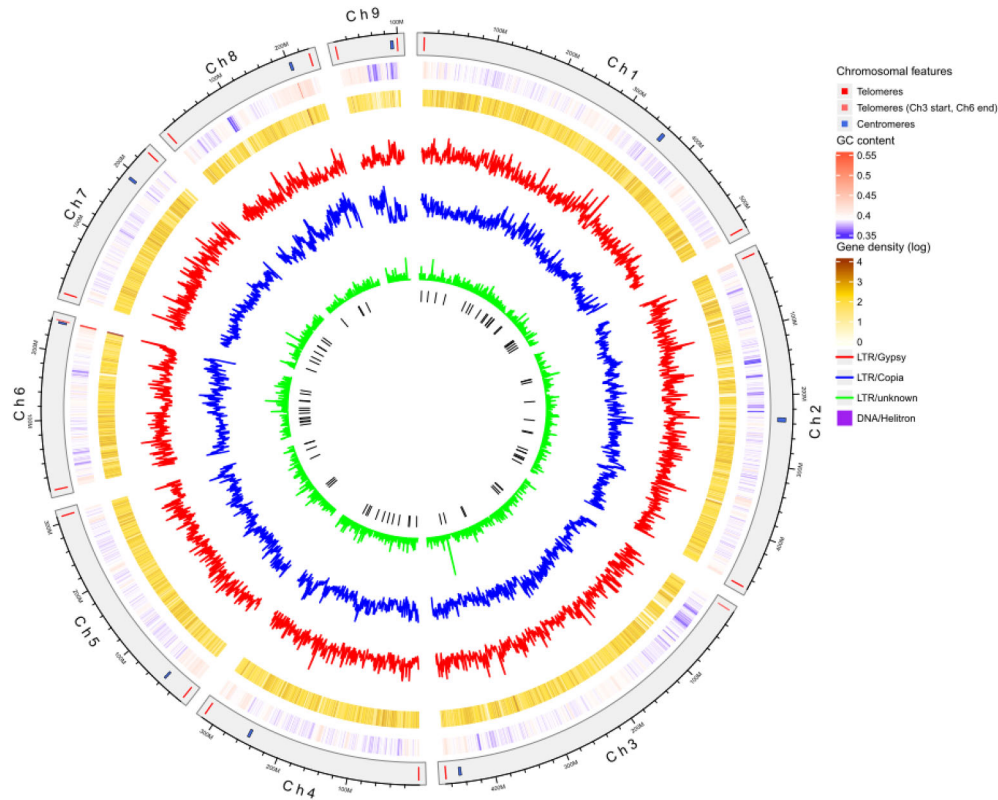


Figure 3 Circos plot showing detailed characterization of the 9 *A. endiviifolia* chromosomes. From outside to inside: length of chromosomes in Mbp with marked telomeric (red bars) and centromeric regions (blue bars), GC content, gene density, distribution of LTR/Gypsy transposons, distribution of LTR/Copia transposons, distribution of LTR/unknown transposons, and DNA/Helitron locations.

Initial contig assembly with Hifiasm

An initial *de novo* assembly was generated directly from the base-called Nanopore long reads using Hifiasm v0.25.0 (RRID:SCR_021069) [33], which supports assembly from high-quality Nanopore data using the `-ont` option. Simplex basecalled reads were used as input. Hifiasm was run with parameters suitable for Nanopore reads `-t64-l0`. This produced a set of high-quality initial contigs. To organize the initial contig assembly into chromosome-level structures, we utilized proximity ligation sequencing reads from the Pore-C library processed using the `epi2me-labs/wf-pore-c` pipeline v1.3.0 [34] to generate a BED file of pairwise chromatin contacts. This contact map served as the input for scaffolding with YaHS v1.2.2 (RRID:SCR_022965) [35] with 3 iterative rounds. Following automated scaffolding, the resulting assembly and the Pore-C contact map were loaded into Juicerbox v2.17 [36] for manual inspection and curation. Gap closing of the assembly was performed using SAMBA with default options, with the exception of the `-m` parameter (minimum matching length) set to 9,000, as recommended for large plant genomes [37].

The gap-closed assembly was polished using Dorado v0.9.1 aligner and polishing functionalities. The required input for polishing, a sorted BAM file containing alignments of basecalled reads with necessary metadata, was generated using a dorado aligner. This dorado aligner output BAM, along with the gap-closed assembly FASTA, was then processed using the dorado polish command to achieve the final polished assembly.

Quality check

To ensure high quality of the final genome assembly, we employed a multifaceted approach to evaluate the completeness, structural integrity, and assembly of repetitive regions. Gene content completeness was assessed using BUSCO v5.8.2 (RRID:SCR_015008) [38] with the `viridiplantae_odb10`, `eukaryota_odb10`, and `embryophyta_odb10` lineage datasets. The structural accuracy and continuity of the 9 assembled chromosomes were evaluated using Inspector v1.3.1 [39]. This was performed in a reference-free manner by mapping the long reads back to the assembly to identify potential misassemblies, structural variants, and other inconsistencies. Finally, to assess the completeness of the repetitive landscape, the LTR Assembly Index (LAI) was calculated using the EDTA v2.2.2 (Extensive *de novo* TE Annotator, RRID:SCR_022063) package [40].

Genome annotation

Structural and functional annotation of the genome was conducted using the NCBI EGAPx v0.3.2 pipeline [41] empirical evidence from all 4 RNA-seq libraries. Based on the provided taxonomy ID, the pipeline automatically selected appropriate protein sets for homology-based evidence, which were aligned to the assembly using miniprot v0.15 [42]. The RNA-seq reads were aligned using STAR v2.7.11 (RRID:SCR_004463) [43] to generate transcript-based evidence. The core of the annotation was performed using Gnomon [44], which first chained the protein and transcript alignments into putative gene models. Gnomon supplemented these with *ab initio* predictions derived from hidden markov model (HMM) models to identify genes lacking direct evidence. Finally,

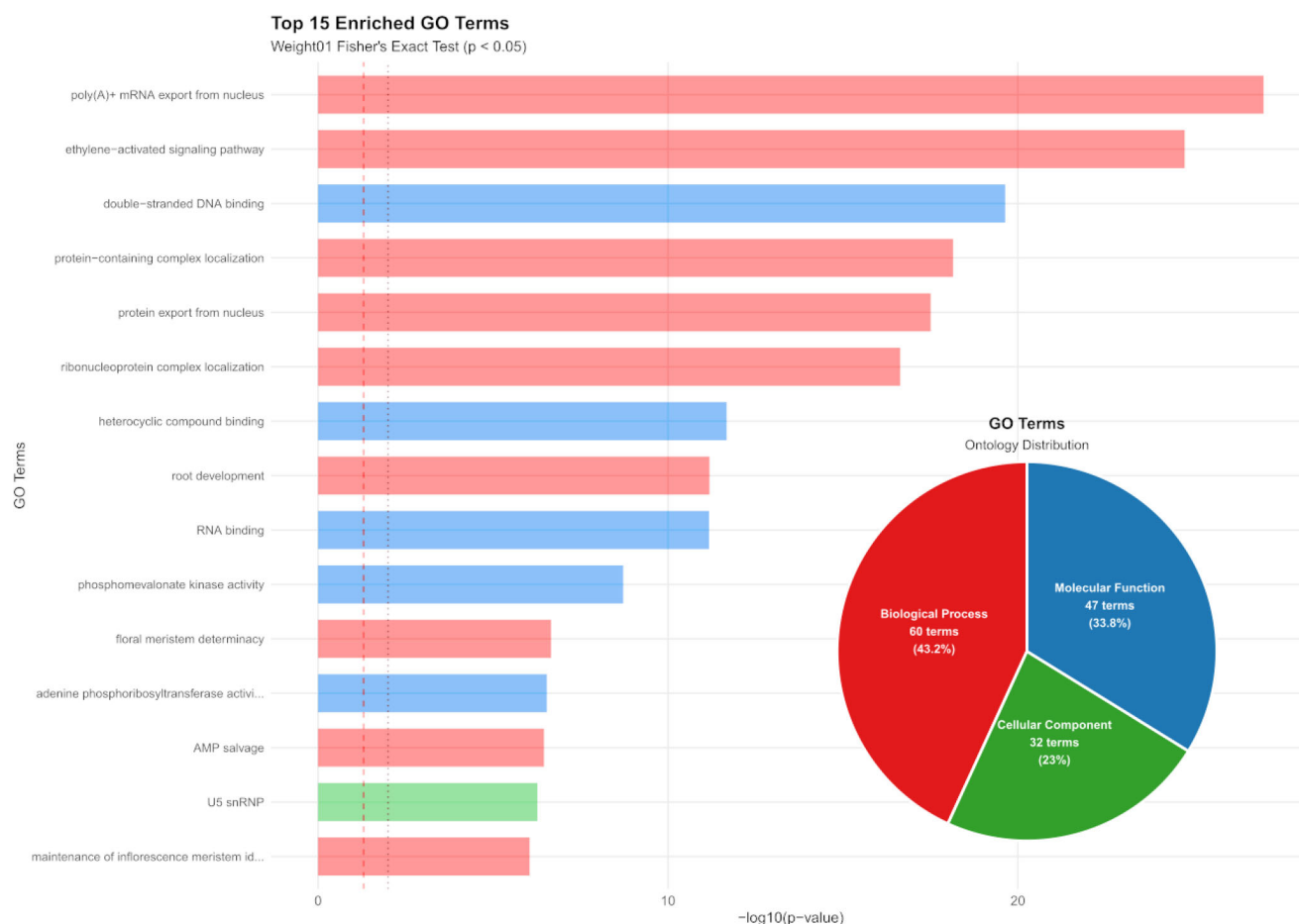


Figure 4 Distribution of all significant GO terms in the genome and the top 15 enriched GO terms across all ontologies in the genome.

functional information was added based on the model quality and orthology, and the complete annotation set was generated as a GFF3 file.

Repeatome annotation

To identify and characterize repetitive elements within the *A. endiviifolia* genome, comprehensive repeatome annotation was performed. Initially, repetitive sequences were identified using EDTA v2.2.2 (RRID:SCR_022063) [40]. Further refinement and classification of TE families were performed using TESorter [45], which leverages machine learning to accurately classify diverse TE types. A phylogenetic approach was employed for the specific annotation of LTR retrotransposons. Individual LTR sequences were first aligned using MAFFT (RRID:SCR_011811) [46] with the default settings. Subsequently, based on the obtained alignment and evolutionary relationships among the distinct LTR retrotransposon families, phylogenetic trees were constructed using IQ-TREE 2 (RRID:SCR_017254) with 1,000 bootstrap replicates [47].

Identification and characteristics of telomeres and centromeres

The telomeric sequences and centromeric regions within the *A. endiviifolia* genome assembly were identified using quartet (RRID:SCR_025258) [48]. The centromeres of each chromosome were further analyzed and confirmed using CentIER [49]. Both types of software were used with default settings.

A custom Python script (analyze_ltr_enrichment.py) [50] was developed and employed to identify LTR retrotransposon domain families that are preferentially associated with centromeric regions. Initially, all LTR element sequences were extracted from the whole-genome LTR GFF3 annotation file using reference genome assembly and subjected to 6-frame translation. The resulting protein sequences were scanned for conserved LTR protein domains using HMMER (hmmsearch) against the REXdb database. This process generated a comprehensive genome-wide map linking LTR elements (identified by their unique GFF IDs) to their constituent REXdb protein domain types (e.g., Athila, Bryco, and SIRE). Next, these genome-wide LTR domain annotations were used for enrichment analysis within the predefined centromeric regions. For each REXdb domain type, the script calculated the total occupied length and the total count of distinct LTR elements containing that domain separately for centromeric and noncentromeric portions of the genome. The noncentromeric portion was determined by subtracting the total length of the defined centromeric regions from the total genome size. To assess the statistical significance of domain enrichment in centromeres, Fisher's exact test (1-tailed, testing for enrichment) was performed for each domain type based on both its total length and element count in centromeric versus noncentromeric regions.

Syntenic analysis

The genomic sequences (FASTA) and gene annotation files (GFF3) for *M. polymorpha* (Tak-1/2 v7.1, GCA_037833965.1) and *A. endivi-*

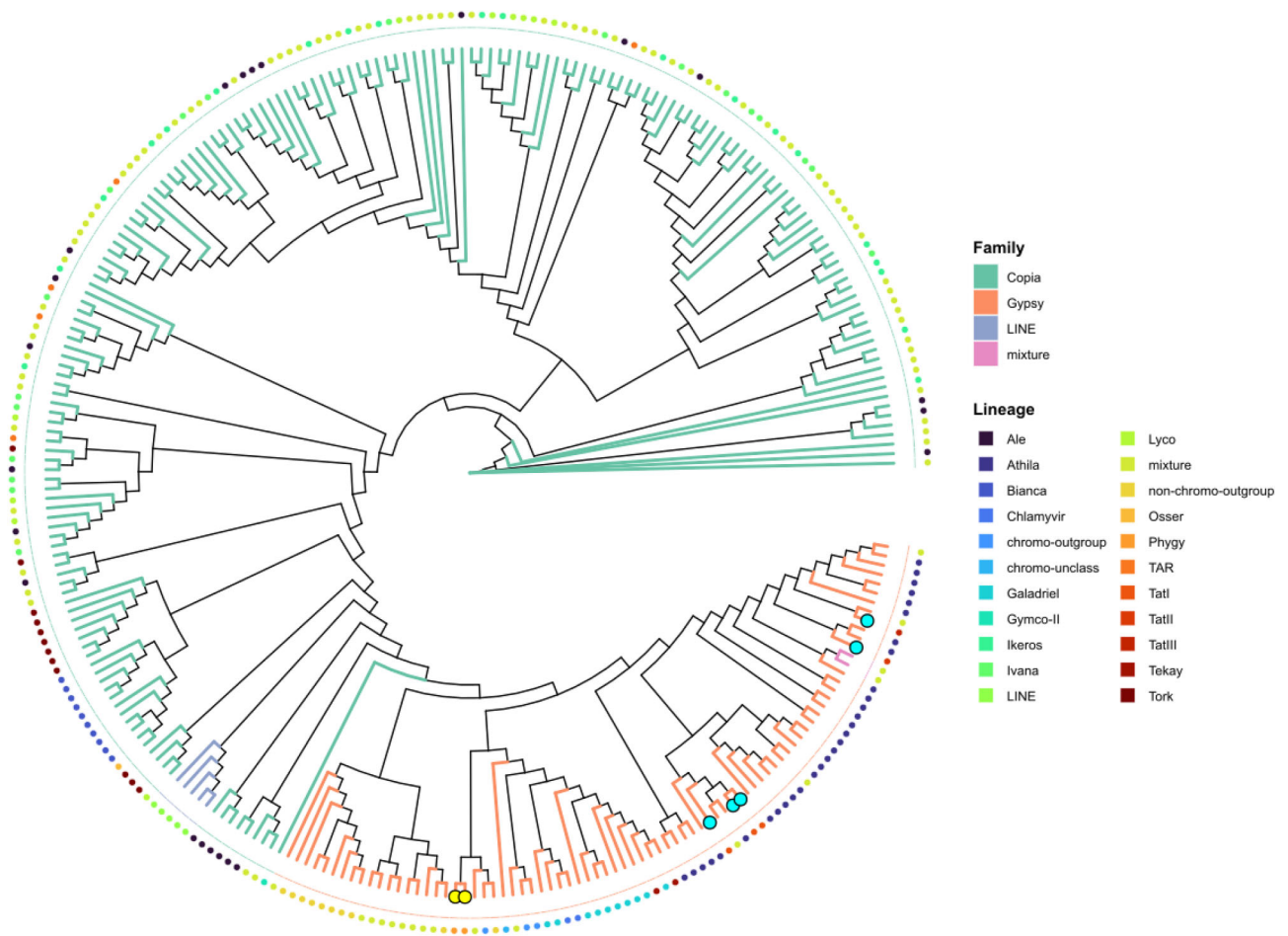


Figure 5 Maximum likelihood phylogenetic tree of LTR retrotransposon lineages identified in the analyzed genome. The tips highlighted in color represent LTR elements enriched in the centromeric regions.

ifolia (JBRAX000000000) were obtained from GenBank. The protein sequences for each species were extracted from their respective genomic FASTA and GFF3 files using gffread v0.12.7 (RRID:SCR_018965) [51]. For each gene, the longest protein isoform was retained for downstream analysis.

Gene coordinates were converted into BED6 format using the jvarkit.formats.gff bed (v1.5.4 from JCVI utilities [52]) from the GFF3 files. The gene identifiers in these BED files were subsequently cleaned to remove prefixes (e.g., “rna-”) and complex locus tag components to ensure compatibility across tools. Specifically, *Marchantia* gene IDs were processed in the format MPTK2_... and *A. endiviifolia* gene IDs were processed in egapxtmp_...-R... by using custom awk scripts. These cleaned BED files provided the chromosome, start, end, cleaned gene ID, dummy score, and strand information. Orthologous gene pairs between *M. polymorpha* and *A. endiviifolia* were identified using OrthoFinder v3.0.1b1 (RRID:SCR_017118) [53]. The extracted protein sequences from both species were used as inputs. OrthoFinder was run with DIAMOND v2.1.11 (RRID:SCR_016071) [54] as the sequence search tool for all-versus-all protein comparisons. Syntenic blocks between *M. polymorpha* and *A. endiviifolia* were identified using the scan action within the jvarkit.compara.synteny module (v1.5.4; JCVI utilities [52]). The cleaned BED files and pairwise ortholog files were used as inputs. The analysis was performed using relaxed parameters that are suitable for distantly related species. A minimum of 2 collinear gene pairs ($-\text{min_size}=2$) was required to define a

syntenic block, and a maximum gap of 100 noncollinear genes ($-\text{dist}=100$) was allowed within a block. The output anchor file, containing the identified syntenic gene pairs, was retained. The identified syntenic relationships were visualized as a dot plot using jvarkit.graphics.dotplot (v1.5.4, JCVI utilities [52]).

Synteny analysis of *A. endiviifolia* and the second available T2T genome assembly of Jungermanniopsida, *Herbertus hutchinsiae*, was performed using ntSynt v1.0.2 [55] since the gene annotation file was not available in GenBank (GCA_965112325.1). The ntSynt workflow begins by generating ordered minimizer sketches for each genome. These sketches were filtered to retain only the single-copy minimizers present in all assemblies, which were then used to construct an initial graph. After simplifying this graph, linear paths were identified to compute an initial set of syntenic blocks. The algorithm refines these by reanalyzing regions not covered by the initial blocks with a smaller window size and augmenting the graph. The pipeline was run with default settings and $-d$ 40 divergence value, and the resulting syntenic blocks were then visualized using the ntSynt-viz v1.0.0 pipeline [55]. The *Apopellia* genome was used as the reference, and the strands of chromosomes in the *Herbertus* genome were normalized relative to it. The final visualization of syntenic relationships was rendered as a ribbon plot using gggenomes v1.0.1 [56] R package.

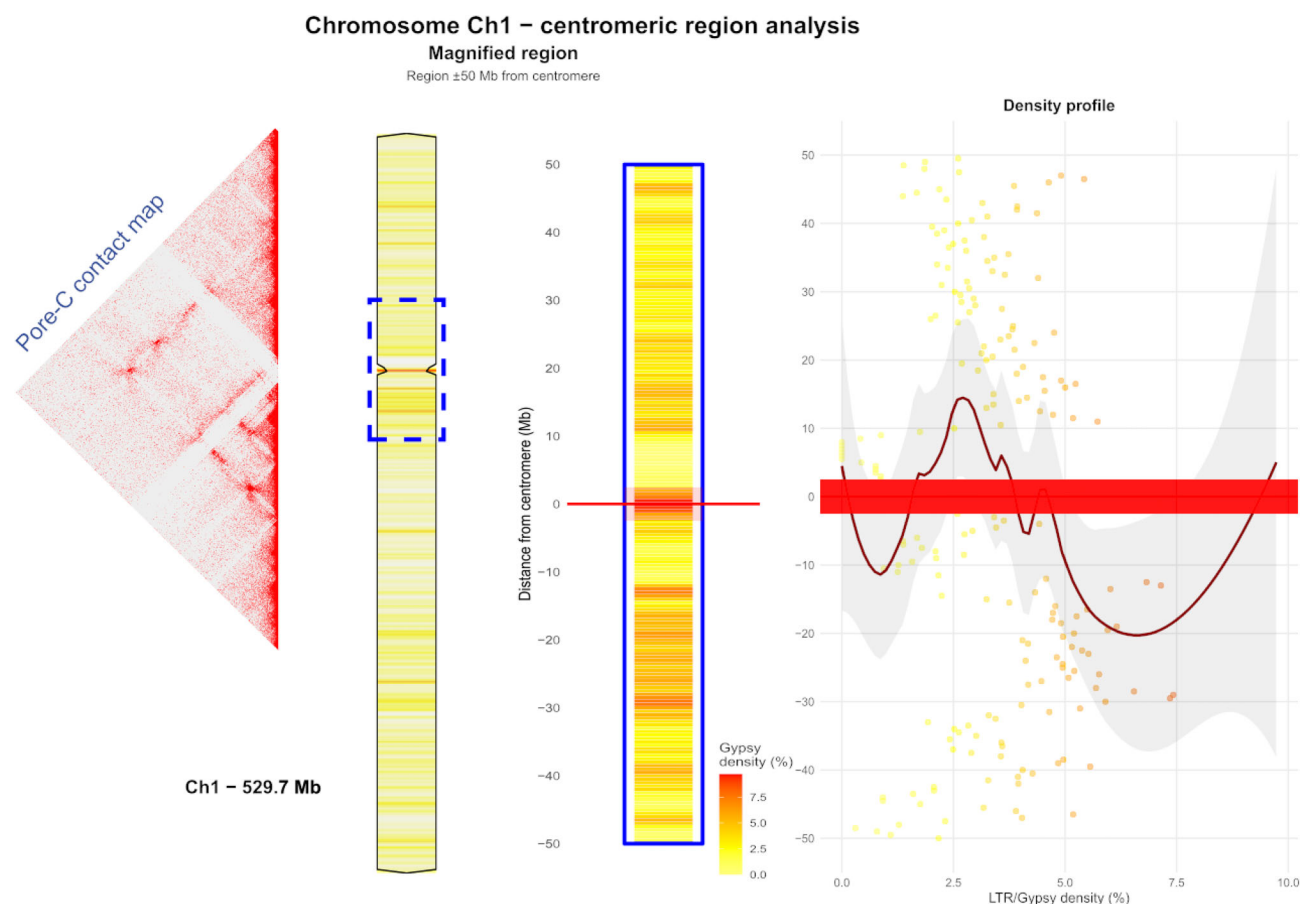


Figure 6 Detailed examination of the example centromere—analysis of *Gypsy* retrotransposon distribution and chromatin interactions across chromosome 1. In the density profile, each dot represents a single genomic window, plotting its LTR/*Gypsy* density (x-axis) against its chromosomal position (y-axis). The solid dark red line shows the overall trend in density, calculated using a locally estimated scatterplot smoothing (LOESS) regression, while the surrounding light gray shaded area indicates the 95% confidence interval for this trend. The thick red horizontal bar marks the position of the centromere across the center of the magnified plots.

Phylogenomic analysis

Chromosome-scale genome assemblies for 13 liverwort species ranging in size from 200 to 3 Gbp were obtained from the GenBank genome database in FASTA format. Single-copy orthologous genes were identified using BUSCO v5.4.7 (RRID:SCR_015008) [38] with the Embryophyta_odb10 database, which contains 1,614 conserved ortholog groups from 50 plant species. BUSCO was run in genome mode with 32 threads using the default parameters. Custom Python scripts [50] were used to extract the protein sequences of the single-copy orthologs shared across all 14 species. Only orthologs present in a single copy in all genomes were retained for downstream analysis, resulting in 502 shared single-copy orthologs. Individual protein sequences for each ortholog were aligned using MAFFT v7.505 (RRID:SCR_011811) [46] with the automatic algorithm selection option (`-auto`), which selects the optimal alignment strategy based on sequence characteristics. The alignments were subsequently trimmed to remove poorly aligned regions using trimAl v1.4 (RRID:SCR_017334) [57] with the `-automated1` option, which applies a heuristic selection of the optimal automated trimming method. Trimmed alignments of all 502 single-copy orthologs were concatenated into a supermatrix using custom Python scripts. Partition information was recorded for potential partitioned phylogenetic analyses, with each gene treated as a separate partition. Maxi-

mum likelihood phylogenetic analysis was performed using IQ-TREE v2.2.0 (RRID:SCR_017254) [47] with automatic model selection using ModelFinder [58] implemented in IQ-TREE (`-m MFP` option). The analysis included 1,000 ultrafast bootstrap replicates [59] to assess branch support ($\geq B$ 1000). The final phylogenetic tree was visualized using the R package *ape* v5.6 (RRID:SCR_017343) [60], *phangorn* v2.10 (RRID:SCR_017302) [61], and *phytools* (RRID:SCR_015502) [62]. All custom scripts used in this analysis are available in the Github repository [50].

Results

Complete reference genome assembly and annotation for *A. endiviifolia*

The integration of 68.5 Gbp of raw data, consisting of 7,751,459 ONT long reads (N50 = 31 kbp) with a quality score of Q20 of 90.1% and Q30 of 81.3%, along with 103 Gbp of proximity-ligated DNA fragments, producing 434,262,093 reads with a mean quality score of 19, facilitated the generation of a complete assembly of the *A. endiviifolia* reference genome. Raw reads were assembled using hifiasm, resulting in 1,860 contigs with a total size of 2,956 Mbp and an N50 length of 76.58 Mbp (Table 1). The initial contigs were then used as the backbone to build scaffold contigs for

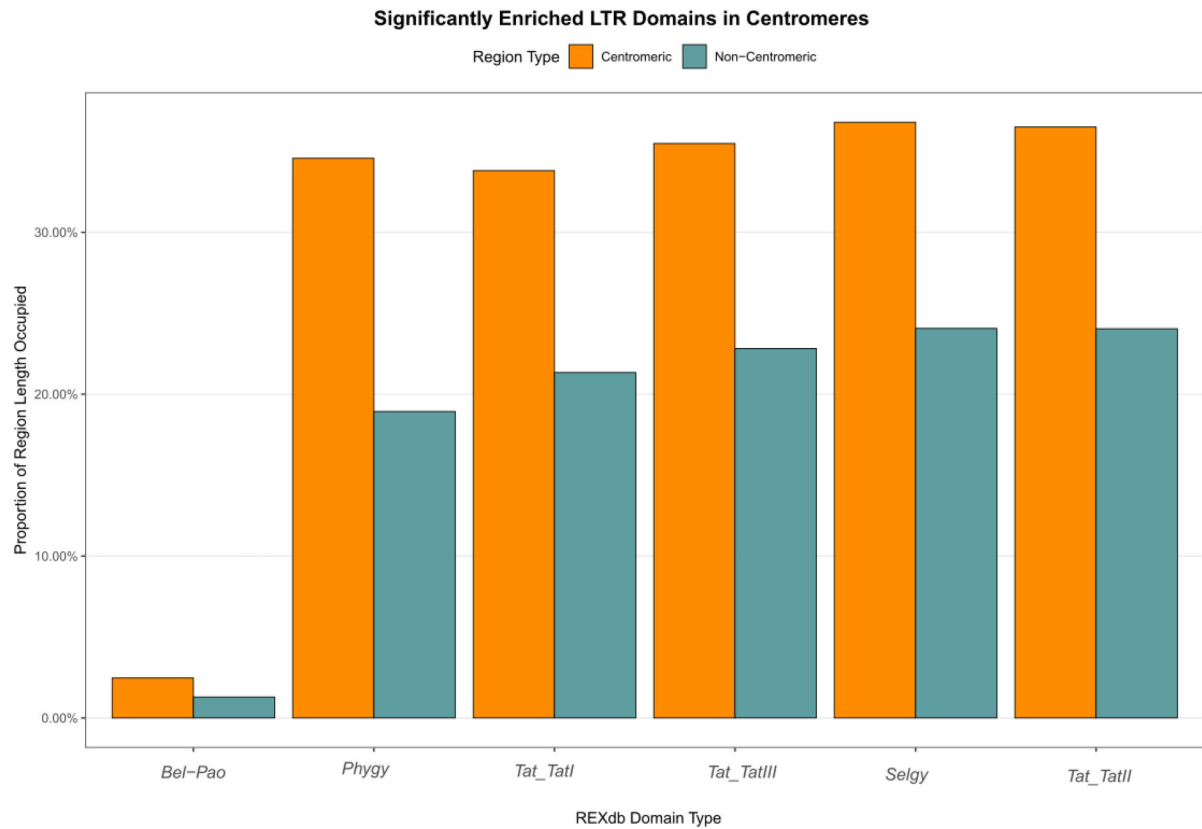


Figure 7 The proportion of regional length occupied by the top 6 significantly ($P < 0.05$) enriched LTR domains in centromeric regions compared with noncentromeric regions.

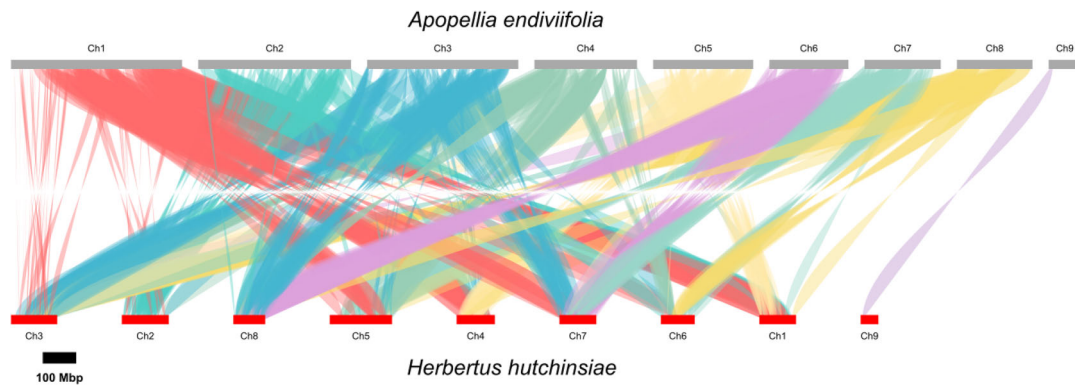


Figure 8 Synteny plot of the differences and similarities between the assembly of *A. endiviifolia* chromosomes and that of *H. hutchinsiae*. Colored lines connect collinear regions between the species' genomes.

the chromosomes, incorporating Pore-C data. After gap filling and polishing, the final assembly had a total size of 2,914,960,273 bp with an N50 of 468,157,909 bp (Table 1), comprising 9 chromosomes (with 6 gaps in intergenic regions) ranging from 100,551,284 to 529,742,643 bp in length (Table 2, Figs. 2 and 3).

Assembly accuracy and completeness were evaluated using multiple methodologies. The Pore-C interaction heatmap demonstrated a high degree of consistency across all chromosomes, thereby providing robust evidence of the precision of genome sequencing (Fig. 2). The BUSCO scores for Viridiplantae_odb10 and Eukaryota_odb10 were 99.2% and 95.5%, respectively, with low gene duplication levels of 6.2% and 4.4%, respectively. Furthermore, integrity assessments of the LTRs indicated an assembled

LAI of 20.06. The genome exhibited a consensus quality value (QV) of 47.6 (Table 1). Collectively, these findings underscore the high accuracy and reliability of the *A. endiviifolia* genome assembly. The assembled genome and gene annotation can be found in the NCBI assembly with submission number JBRAUX000000000.

In the process of annotating the genome of *A. endiviifolia*, short-read RNA sequencing was performed on the thallus of both aquatic (16.8 Gbp) and terrestrial forms (20.6 Gbp), as well as on the antheridia themselves (6 Gbp) and surrounding thallus (6.3 Gbp), generating a total of 49.7 Gbp reads. Genomic annotation identified 34,615 genes, including 33,513 protein-coding genes (Table 1). In the comprehensive analysis of the overall distribution of all significant Gene Ontology (GO) terms, the “Biological Pro-

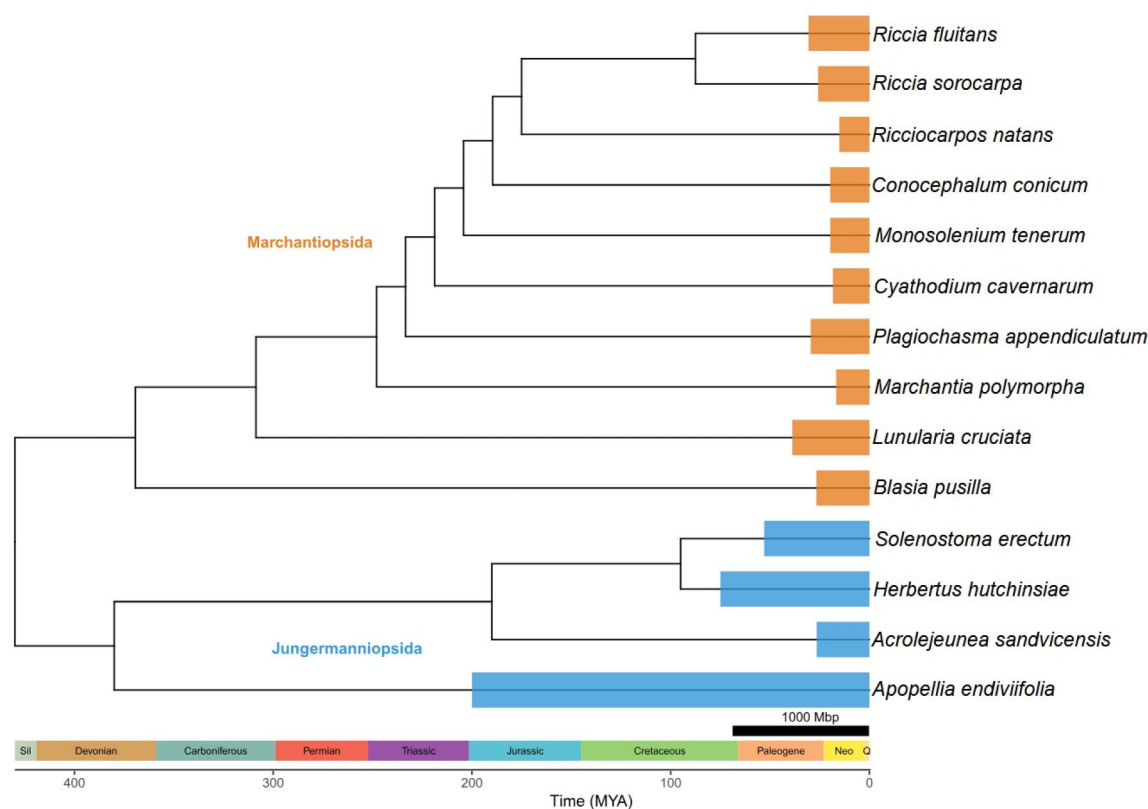


Figure 9 Phylogenetic relationships among liverworts determined by conserved single-copy orthologs. Size of orange and blue boxes is proportional to the genome size of the species. All nodes have the maximum statistical support.

cess” category contained the highest number of unique terms, significantly surpassing the other categories with 96 terms, accounting for 45.1% of the total. The “Molecular Function” category ranked second in abundance (66 terms, 31%), while the “Cellular Component” category had the fewest unique GO terms (51 terms, 23.9%) (Fig. 4). GO enrichment analysis revealed that 6 terms predominated among the top 15 significantly enriched GO terms across all ontologies: “poly(A)+ mRNA export from nucleus”, “ethylene-activated signaling pathway”, “double-stranded DNA binding”, “protein-containing complex localization”, “protein export from nucleus”, and “ribonucleoprotein complex localization” (Fig. 4). The remaining categories contain progressively fewer terms. The composition of the TEs included 12.16% LTRs elements and 57 Helitrons. Among the classified retroelements, the *Copia* and *Gypsy* superfamilies accounted for 8.94% and 2.95% of the assembly, respectively (Fig. 3). The phylogenetic tree illustrates the diversity and evolutionary relationships among LTR retrotransposon families identified in the analyzed genome, including major lineages such as *Athila*, *Phygy*, *Tekay*, and others (Fig. 5). The GC content and gene density were assessed in 500-kbp windows across each chromosome. The average GC content across all chromosomes ranged from 38.8% to 39.6%. The highest gene density was observed on chromosome 6, with a value of 9.78, whereas the lowest gene density was recorded on chromosome 9, with a value of 5.52 (Fig. 3).

Detection and characteristics of telomeres and centromeres

The completion and accuracy of genome sequencing have enabled the identification of telomeres and centromeres (Table 2,

Fig. 3). Examination of the telomeric regions by scanning chromosome ends for high-copy tandem repeats showed that both ends of the 7 *A. endiviifolia* chromosomes (except chromosomes 3 and 6) feature telomere repeat sequences (CCCTAAA/TTTAGGG) that align with telomeric structures typical of most plant species. For these 7 chromosomes, the number of repeats at the left end ranged from 352 to 481, whereas those at the right end ranged from 402 to 452 (Table 2). In chromosomes 3 and 6, the typical telomeric sequence was identifiable at one end, whereas the opposite end exhibited a substantial number of repeats of an alternative motif. Specifically, the left end of chromosome 3 contained 115 repeats of the ACGCAGC motif, whereas the right end of chromosome 6 contained 176 repeats of the TGCGTCG motif (Table 2).

The application of *quarTeT*, a tool that identifies centromeres by computationally detecting and mapping tandem repeats and associated retrotransposons in assembled genomes, in conjunction with *CentIER*, which identifies centromeres by clustering tandem repeats (Supplementary Table S1), mapping their abundance and distribution, and designating regions with extensive dense arrays as candidate centromeres, facilitated reliable identification of centromeric regions across all 9 chromosomes. However, to confirm the presence of centromeres in chromosomes 4 and 5, it was necessary to verify the results obtained from the *quarTeT* analysis by comparing them with the gaps observed in the Pore-C interaction heatmap (Supplementary Fig. S1). Furthermore, all sites identified as centromeres on the remaining 7 chromosomes were corroborated by the Pore-C interaction heatmap. Examination of these centromeric regions confirmed that LTR/*Gypsy* elements were highly enriched in the immediate vicinity of the cen-

trromere, and their abundance decreased as the distance from the centromere increased, as presented in detail for chromosome 1 (Fig. 6). Additionally, this analysis demonstrated a correlation between high LTR/Gypsy density and the formation of a specialized chromatin domain at the centromere. Moreover, LTR domains from the Ty3/Gypsy superfamily, particularly those belonging to the *Tat* (*TatI*, *TatII*, and *TatIII*), *Phygy*, and *Selgy* families, were significantly enriched in centromeric regions compared to non-centromeric chromosomal regions. Although LTR elements from the *Bel-Pao* family are infrequently present in centromeres, they nonetheless showed a statistically significant enrichment in these regions relative to the rest of the genome (Fig. 7). Finally, the validated centromeres exhibited length variation ranging from 1.9 to 5 Mbp (Table 2). Compared to *A. endiviifolia*, *M. polymorpha* centromeres are similarly structured with abundant retroelements and are typically centrally located, ranging from about 176 to 604 Kbp [63]. *P. patens* centromeres share this organization but tend to be smaller, varying from 64 to 221 Kbp, likely reflecting its more compact genome [25]. These differences highlight centromere size and composition diversity within bryophytes and demonstrate the value of long-read sequencing for resolving complex genomic regions.

Genome synteny analysis

Collinearity between the 9 chromosomes of *A. endiviifolia* and the leafy liverwort *H. hutchinsiae*, both belonging to the Jungermanniopsida class, revealed complex and divergent synteny patterns characterized by fragmentation into small syntenic blocks distributed across multiple chromosomes rather than forming extensive chromosome-scale fusions. A total of 429 syntenic blocks spanning approximately 4.8 Mb were detected, with a median size of about 5.2 kb and a mean size of 11.2 kb, indicating small, fragmented conserved regions. These blocks were evenly distributed in terms of strand orientation, with 201 blocks on the positive strand and 228 on the negative strand. Details of the detected syntenic blocks between *A. endiviifolia* and *H. hutchinsiae* chromosomes, including chromosome coordinates and strand orientation, are provided in Supplementary Tables S2 and S3. Structural variation analysis revealed numerous chromosomal breakpoints across chromosomes, predominantly in chromosome 1 (54 translocation and 18 inversion breakpoints), while chromosome 9 showed no breakpoints, consistent with its slightly more conserved relationship (Fig. 8, Supplementary Table S3). Furthermore, analysis of genomic conservation and structural variations between *A. endiviifolia* and the model liverwort *M. polymorpha* did not reveal significant collinearity (Supplementary Fig. S2).

Phylogenetic relationship analysis

The protein sequences of single-copy orthologs conserved across *A. endiviifolia* and 13 other liverwort species with chromosome-scale genome assemblies were selected for phylogenetic analysis to elucidate evolutionary relationships among these species. The resulting phylogenetic tree clearly separated Marchantiopsida and Jungermanniopsida into distinct evolutionary lineages. Notably, all internal nodes in the tree were supported by 100% bootstrap values, reflecting the maximal confidence in the overall tree topology. This strong support underscores the reliability of the inferred evolutionary relationships. Additionally, the observed variation in genome size, particularly in the large genomes of some Jungermanniopsida, may reflect lineage-specific differences in genome dynamics, such as variation in transposable el-

ement activity, gene duplication, and structural rearrangements (Fig. 9).

Reuse potential

The telomere-to-telomere reference genome assembly of the haploid liverwort *A. endiviifolia* is a high-quality genomic resource with a broad potential for reuse across multiple research fields. Generated using Oxford Nanopore long-read sequencing combined with Pore-C technology, this 2,914,960,273-bp assembly achieves chromosome-scale resolution with exceptional completeness (99.2% BUSCO) and accuracy (QV 47.6). It comprises 9 experimentally validated chromosomes, featuring 18 telomeres and 9 predicted centromeres enriched in LTR/Gypsy retrotransposons, and includes 34,615 annotated genes. The unusual presence of noncanonical telomeric-like repeats (ACGCAGC and TGCGTCG) on chromosomes 3 and 6 in *A. endiviifolia* suggests species-specific telomere evolution or chromosomal rearrangements, although their biological significance is unclear. Further comparative and functional studies are needed to understand their origin and role. These findings contribute to the genomic landscape of *A. endiviifolia* and support future research on plant telomere diversity and evolution. Notably, the centromeres have been predicted solely through *in silico* analyses, representing a plausible but not yet fully validated hypothesis.

This comprehensive dataset enables comparative genomics within liverworts and across land plants, facilitating studies of genome evolution, centromere biology, chromosome end structures, genome stability, and chromosome segregation. By providing a complete, well-annotated, and experimentally validated genome, this resource provides a robust foundation for future research in plant genomics, cytogenetics, and evolutionary biology, extending beyond the scope of the current study.

Additional Files

Supplementary Fig. S1. Detailed examination of the predicted centromere of chromosome 5—analysis of Gypsy retrotransposon distribution and chromatin interactions across chromosome 5. In the density profile, each dot represents a single genomic window, plotting its LTR/Gypsy density (x-axis) against its chromosomal position (y-axis). The solid dark red line shows the overall trend in density, calculated using a LOESS regression, while the surrounding light gray shaded area indicates the 95% confidence interval for this trend. The thick red horizontal bar marks the position of the centromere across the center of the magnified plots.

Supplementary Fig. S2. Synteny dot plot comparing the genomes of *M. polymorpha* and *A. endiviifolia*.

Supplementary Table S1. Tandem centromeric repeats of *in silico* validated centromeric regions in the *A. endiviifolia* assembly. Centromeres in chromosomes 1–3 and 6–9 were identified by the Centier program, and centromeres in chromosomes 4 and 5 were identified by the quarTeT program. All centromeres were validated using a Pore-C interaction heatmap.

Supplementary Table S2. Synteny block between *Herbertus hutchinsiae* and *Apopellia endiviifolia* detected by ntSynt software.

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; GO: Gene Ontology; HB: homogenization buffer; HMM: hidden markov model; LAI: LTR Assembly Index; LINE: long interspersed nuclear

element; LOESS: locally estimated scatterplot smoothing; LTR: long terminal repeat; QV: quality value; RNA-seq: RNA sequencing; TE: transposable element.

Acknowledgments

We thank Dr. Katarzyna Krawczyk for her assistance with the Pore-C procedure. We also extend our gratitude to the staff of the Department of Botany and Evolutionary Ecology at the University of Warmia and Mazury in Olsztyn for the pleasant working atmosphere.

Author Contributions

J.S.-P. and J.S. contributed to research design, analyzed the data, and wrote the original draft. J.S.-P. carried out laboratory analyses. J.S. carried out bioinformatic analyses, visualized data, and prepared figures. P.S. was responsible for *in vitro* cultures. J.S. obtained funding. All authors revised and approved the final version of the manuscript.

Funding

This study was financially supported by the National Science Center, Kraków, Poland (Grant No. 2020/39/B/NZ8/02504).

Data Availability

The genomic and transcriptomic sequence data have been deposited in NCBI BioProject PRJNA1279829. All additional supporting data are available in the GigaScience repository, GigaDB [64].

References

- Bechteler J, Peñaloza-Bojacá G, Bell D, et al. Comprehensive phylogenomic time tree of bryophytes reveals deep relationships and uncovers gene incongruences in the last 500 million years of diversification. *Am J Bot.* 2023;110:e16249 <https://doi.org/10.1002/AJB2.16249>.
- Donoghue PCJ, Harrison CJ, Paps J, et al. The evolutionary emergence of land plants. *Curr Biol.* 2021;31:R1281–98. <https://doi.org/10.1016/j.cub.2021.07.038>.
- Harris BJ, Clark JW, Schrepf D, et al. Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nat Ecol Evol.* 2022;6:1634–43. <https://doi.org/10.1038/S41559-022-01885-X>.
- Linde AM, Singh S, Bowman JL, et al. Genome evolution in plants: complex thalloid liverworts (Marchantiopsida). *Genome Biol Evol.* 2023;15:evad014. <https://doi.org/10.1093/GBE/evad014>.
- Villarreal A. JC, Crandall-Stotler BJ, Hart ML, et al. Divergence times and the evolution of morphological complexity in an early land plant lineage (Marchantiopsida) with a slow molecular rate. *New Phytol.* 2016;209:1734–46. <https://doi.org/10.1111/NPH.13716>.
- Renner MAM, Heslewood MM, Patzak SDF, et al. By how much do we underestimate species diversity of liverworts using morphological evidence? An example from Australasian Plagiochila (Plagiochilaceae: jungermanniopsida). *Mol Phylogenet Evol.* 2017;107:576–93. <https://doi.org/10.1016/j.ympev.2016.12.018>.
- Söderström L, Hagborg A, Von Konrat M, et al. World checklist of hornworts and liverworts. *PhytoKeys.* 2016;59:1–828. <https://doi.org/10.3897/PHYTOKEYS.59.6261>.
- Paukszto Ł, Górski P, Krawczyk K, et al. The organellar genomes of Pellidae (Marchantiophyta): the evidence of cryptic speciation, conflicting phylogenies and extraordinary reduction of mitogenomes in simple thalloid liverwort lineage. *Sci Rep.* 2023;13:8303. <https://doi.org/10.1038/S41598-023-35269-3>.
- Crandall-Stotler B, Stotler RE, Zhang L, et al. On the morphology, systematics and phylogeny of Notoclada (Notocladaceae, Marchantiophyta). *Nova Hedwigia.* 2010;91:421–50. <https://doi.org/10.1127/0029-5035/2010/0091-0421>.
- Schütz N, Quandt D, Nebel M. The position of the genus Apopellia stat. Nov. within the Pelliales (Marchantiophytina: jungermanniopsida). *Taxon.* 2016;65:221–34. <https://doi.org/10.12705/652.1>.
- Crandall-Stotler B, Stotler RE, Long DG. Phylogeny and classification of the Marchantiophyta. *Edinburgh J Bot.* 2009;66:155–98. <https://doi.org/10.1017/S0960428609005393>.
- Sawicki J, Krawczyk K, Ślipiko M, et al. Comparative analysis of apopellia endiviifolia plastomes reveals a strikingly high level of differentiation between its terrestrial and water form. *Diversity (Basel).* 2021;13:674. <https://doi.org/10.3390/d13120674>.
- Interpretation of electrophoretic patterns in population genetics of bryophytes: VI. Genetic variation and evolution of the liverwort genus Pellia with special reference to Central European territory on JSTOR. <https://www.jstor.org/stable/20149587?seq=1>. Accessed 25 June 2025.
- Pellicer J, Hidalgo O, Dodsworth S, et al. Genome size diversity and its impact on the evolution of land plants. *Genes.* 2018;9:88. <https://doi.org/10.3390/GENES9020088>.
- Temsch EM, Greilhuber J, Krisai R. Genome size in liverworts. *Preslia.* 2010;82:63–80. <https://www.preslia.cz/article/pdf?id=224>.
- Bainard JD, Forrest LL, Goffinet B, et al. Nuclear DNA content variation and evolution in liverworts. *Mol Phylogenet Evol.* 2013;68:619–27. <https://doi.org/10.1016/j.ympev.2013.04.008>.
- Singh S, Bowman JL. The monoicous secondarily aquatic liverwort Ricciocarpos natans as a model within the radiation of derived Marchantiopsida. *Front Plant Sci.* 2023; 14:1260596. <https://doi.org/10.3389/fpls.2023.1260596>.
- Krawczyk K, Szablińska-Piemiak J, Paukszto Ł, et al. Chromosome-scale telomere to telomere genome assembly of common crystalwort (Riccia sorocarpa Bisch.). *Sci Data.* 2025;12:1. <https://doi.org/10.1038/s41597-025-04373-6>.
- Levins J, Paukszto Ł, Krawczyk K, et al. Evolution of sexual systems and regressive evolution in Riccia. *New Phytol.* 2025;246:769–84. <https://doi.org/10.1111/NPH.20454>.
- Fu Y, Zhang X, Zhang T, et al. Evidence for evolution of a new sex chromosome within the haploid-dominant Marchantiales plant lineage. *J Integr Plant Biol.* 2025;67(6):1533–50. <https://doi.org/10.1111/JIPB.13867>.
- Bowman JL, Kohchi T, Yamato KT, et al. Insights into land plant evolution garnered from the Marchantia polymorpha genome. *Cell.* 2017;171:287–304.e15. <https://doi.org/10.1016/j.cell.2017.09.030>.
- Linde AM, Eklund DM, Cronberg N, et al. Rates and patterns of molecular evolution in bryophyte genomes, with focus on complex thalloid liverworts, Marchantiopsida. *Mol Phylogenet Evol.* 2021;165:107295. <https://doi.org/10.1016/j.ympev.2021.107295>.
- Miga KH. Centromere studies in the era of ‘telomere-to-telomere’ genomics. *Exp Cell Res.* 2020;394:112127. <https://doi.org/10.1016/j.yexcr.2020.112127>.

24. Peng D, Hong Z, Kan S, et al. The telomere-to-telomere (T2T) genome provides insights into the evolution of specialized centromere sequences in sandalwood. *Gigascience*. 2024;13:giae096. <https://doi.org/10.1093/gigascience/giae096>.
25. Bi G, Zhao S, Yao J, et al. Near telomere-to-telomere genome of the model plant *Physcomitrium patens*. *Nat Plants*. 2024;10:327–43. <https://doi.org/10.1038/s41477-023-01614-7>.
26. Montgomery SA, Tanizawa Y, Galik B, et al. Chromatin organization in early land plants reveals an ancestral association between H3K27me3, transposons, and constitutive heterochromatin. *Curr Biol*. 2020;30:573–88.e7. <https://doi.org/10.1016/j.CU.2019.12.015>.
27. Sawicki J, Krawczyk K, Kurzyński M, et al. Nanopore sequencing of organellar genomes revealed heteroplasmy in simple thalloid and leafy liverworts. *Acta Soc Bot Pol*. 2023;92:1–13. <https://doi.org/10.5586/ASBP/172516>.
28. Lu D, Liu C, Ji W, et al. Nanopore ultra-long sequencing and adaptive sampling spur plant complete telomere-to-telomere genome assembly. *Mol Plant*. 2024;17:1773–86. <https://doi.org/10.1016/j.MOLP.2024.10.008>.
29. Zhong JY, Niu L, Bin Lin Z, et al. High-throughput pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. *Nat Commun*. 2023;14:1. <https://doi.org/10.1038/s41467-023-36899-x>.
30. GitHub—nanoporetech/dorado: Oxford Nanopore's Basecaller. <https://github.com/nanoporetech/dorado> Accessed 10 September 2025.
31. Liu B, Shi Y, Yuan J, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv*. 2013. <https://arxiv.org/abs/1308.2012>.
32. Wang H, Liu B, Zhang Y, et al. Estimation of genome size using k-mer frequencies from corrected long reads. *arXiv*. 2020. <https://arxiv.org/abs/2003.11817>.
33. Cheng H, Concepcion GT, Feng X, et al. 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>.
34. GitHub—epi2me-labs/wf-pore-c. <https://github.com/epi2me-labs/wf-pore-c>. Accessed 10 September 2025.
35. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Periodical hi-C scaffolding tool. *Bioinformatics*. 2023;39:btac808. <https://doi.org/10.1093/BIOINFORMATICS/BTAC808>.
36. Durand NC, Robinson JT, Shamim MS, 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>.
37. GitHub—alekseyzimin/masurca. <https://github.com/alekseyzimin/masurca>. Accessed 10 September 2025.
38. Manni M, Berkeley MR, Seppey M, et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021;38:4647–54. <https://doi.org/10.1093/MOLBEV/MSAB199>.
39. Chen Y, Zhang Y, Wang AY, et al. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol*. 2021;22:312. <https://doi.org/10.1186/S13059-021-02527-4>.
40. Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019;20:275. <https://doi.org/10.1186/S13059-019-1905-Y>.
41. GitHub—ncbi/egapx: Eukaryotic Genome Annotation Pipeline—external caller scripts and documentation. <https://github.com/ncbi/egapx>. Accessed 10 September.
42. Li H. Protein-to-genome alignment with minimap. *Bioinformatics*. 2023;39:btad014. <https://doi.org/10.1093/BIOINFORMATICS/BTAD014>.
43. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21. <https://doi.org/10.1093/BIOINFORMATICS/BTS635>.
44. Gnomon—the NCBI eukaryotic gene prediction tool. https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/gnomon/. Accessed 10 September 2025.
45. Zhang RG, Li GY, Wang XL, et al. TESorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res*. 2022;9:uhac017. <https://doi.org/10.1093/HR/UHAC017>.
46. 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>.
47. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37:1530–34. <https://doi.org/10.1093/MOLBEV/MSAA015>.
48. Lin Y, Ye C, Li X, et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res*. 2023;10:uhad127. <https://doi.org/10.1093/HR/UHAD127>.
49. Xu D, Yang J, Wen H, et al. CentIER: accurate centromere identification for plant genomes. *Plant Commun*. 2024;5:101046. <https://doi.org/10.1016/j.XPLC.2024.101046>.
50. GitHub—kubek78/PhylogenomicsME: Phylogenomics Made Easy. <https://github.com/kubek78/PhylogenomicsME/tree/main>. Accessed 10 September 2025.
51. Pertea G, Pertea M. GFF utilities: gffRead and GffCompare. *F1000Res*. 2020;9:304. <https://doi.org/10.12688/F1000RESEARCH.23297.2>.

52. Tang H, Krishnakumar V, Zeng X, et al. JCVI: a versatile toolkit for comparative genomics analysis. *iMeta*. 2024;3:e211. <https://doi.org/10.1002/IMT2.211>.
53. 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>.
54. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021;18:366–68. <https://doi.org/10.1038/S41592-021-01101-X>.
55. Coombe L, Kazemi P, Wong J, et al. Multi-genome synteny detection using minimizer graph mappings. *Biorxiv*. 2024. <https://doi.org/10.1101/2024.02.07.579356>.
56. Hackl T, Ankenbrand M, van Adrichem B, et al. gggenomes: effective and versatile visualizations for comparative genomics. *arXiv*. 2024; arXiv:2411.13556. <https://doi.org/10.48550/arXiv.2411.13556>.
57. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25:1972–73. <https://doi.org/10.1093/BIOINFORMATICS/BTP348>.
58. Kalyaanamoorthy S, Minh BQ, Wong TKF, et al. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;419:456–9. <https://doi.org/10.1038/>.
59. Hoang DT, Chernomor O, Von Haeseler A, et al. UFBoot2: improving the Ultrafast bootstrap approximation. *Mol Biol Evol*. 2018;35:518–22. <https://doi.org/10.1093/MOLBEV/MSX281>.
60. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35:526–28. <https://doi.org/10.1093/BIOINFORMATICS/BTY633>.
61. Schliep KP. phangorn: phylogenetic analysis in R. *Bioinformatics*. 2011;27:592–93. <https://doi.org/10.1093/BIOINFORMATICS/BTQ706>.
62. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and Periodical things). *PeerJ*. 2024;12:e16505. <https://doi.org/10.7717/PEERJ.16505>.
63. Tanizawa Y, Mochizuki T, Yagura M, et al. MarpolBase: genome database for Marchantia polymorpha featuring high quality reference genome sequences. *Biorxiv*. 2025. <https://doi.org/10.1101/2025.03.30.646155>.
64. Szablinska-Piernik J, Sulima P, Sawicki J. Supporting data for “Giant Chromosomes of a Tiny Plant—The Complete Telomere-to-Telomere Genome Assembly of the Simple Thalloid Liverwort *Apopellia endiviifolia* (Jungermanniopsida, Marchantiophyta).” *GigaScience Database*. 2025. <https://doi.org/10.5524/102776>.