



OPEN

DATA DESCRIPTOR

Chromosome-scale telomere to telomere genome assembly of common crystalwort (*Riccia sorocarpa* Bisch.)

Katarzyna Krawczyk^{1,3}✉, Joanna Szablińska-Piernik^{1,3}✉, Łukasz Pauksto¹, Mateusz Maździarz¹, Paweł Sulima², Jerzy Andrzej Przyborowski², Monika Szczecińska¹ & Jakub Sawicki¹

Riccia sorocarpa Bisch., commonly known as common crystalwort, is a plant belonging to the Marchantiales order with a cosmopolitan distribution among a wide range of habitats: fields, gardens, waste ground, on paths, cliff tops, and thin soil over rocks or by water bodies. However, research into the genetic aspects of this species is limited. In this study, the chromosome-scale telomere-to-telomere genome of *R. sorocarpa* was assembled exclusively by Oxford Nanopore long-read sequencing and Pore-C technology. A high-quality chromosomal-scale assembly was obtained with a final genome size of 376.690 Mbp, contig N50 of 49.132 Mbp and 97.02% of the assembled contigs associated with the eight chromosomes. Genome assembly completeness was confirmed by BUSCO analysis accounting 91.8%. Among 27,626 total genes, 23,562 (85.29%) were functionally annotated. Moreover, collinearity of Marchantiales was analyzed as well as gene family evolution and DNA methylation profile. The availability of this genome, which is the second telomere-to-telomere liverwort assembly, opens up new avenues for in-depth analysis of *R. sorocarpa* genetic diversity and genomic characteristics.

Background & Summary

Liverworts, members of the division Marchantiophyta, are non-vascular plants representing an early-diverging lineage in the evolution of land plants. Fossil records indicate the emergence of liverworts over 480 million years ago, suggesting their critical involvement in the initial colonization of land by plants¹. Their gametophyte-dominant life cycle, where the haploid stage is the conspicuous plant body, offers valuable insights into the evolution of alternation of generations within the plant lineage².

Liverworts present two primary morphologies: thallose, characterized by flattened, lobed structures, and leafy, exhibiting small leaf-like structures along a stem. The presence of specialized oil bodies within their cells is a unique feature of liverworts. These oil bodies contain a variety of terpenoids and aromatic compounds, which are hypothesized to serve protective functions against herbivory^{3,4}. Liverworts display both sexual and asexual reproductive strategies, facilitating dispersal and colonization of new habitats⁵. Liverworts occupy a wide range of habitats globally, with distribution contingent upon moisture availability⁶. They are commonly found in humid environments such as tropical and temperate forests, as well as specialized microhabitats like moist rock surfaces.

Riccia L. represents the most widespread and diverse genus within the complex thalloid liverworts (Class Marchantiopsida), comprising approximately 240 species⁷. These liverworts exhibit a range of growth habits, forming rosettes or extensive mats on exposed soil, or existing as semi-aquatic floating forms. Several *Riccia* species play a key role within delicate biological soil crusts, dominating as the primary bryophyte of arable fields. They also thrive in proximity to semi-permanent water bodies, on compacted dirt surfaces, and even

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. ³These authors contributed equally: Katarzyna Krawczyk, Joanna Szablińska-Piernik. ✉e-mail: katarzyna.krawczyk@uwm.edu.pl; joanna.szablinska@uwm.edu.pl

Genome assembly statistics	
Genome coverage [median]	163
Estimated genome size [Mbp]	320-390
Number of chromosomes	8
Chromosomes N50 [bp]	49,131,805
Longest chromosome [bp]	55,960,836
Number of contigs	3,218
Number of scaffolds	951
Scaffolds N50 [bp]	790,541
Scaffold NG50 [bp]	839,990
Longest scaffold [bp]	3,307,131
Genome size [bp]	376,689,626
Contig N50 [bp]	258,577
Longest contig [bp]	1,529,652
GC content [%]	41.5
Completeness BUSCO [%]	91.8
Single-copy BUSCO [%]	85.1
Duplicated BUSCO [%]	6.70

Table 1. Assembly statistics.

within urban environments, colonizing bare patches in lawns or cracks in pavement⁸. *Riccia* produces large, thick-walled spores with intricate ornamentation, and lack both elaters and oil bodies⁹.

Riccia sorocarpa Bisch. is a species that's both easy to identify and incredibly widespread. Its glaucous green color, turning brown with age, is a distinctive trait, but the most defining characteristic is thickness of thallus, which is 1-2 times wider than high. The margins of the thallus are well-defined, rising at an acute angle and lined with tiny, colorless and hyaline scales. common crystalwort has a cosmopolitan distribution, found across much of the globe, including Europe, Asia, Africa, Australia, North America, and South America¹⁰. This remarkable adaptability allows it to thrive in fields, gardens, waste ground, on paths, cliff tops, and even in thin soil over rocks or by water bodies. It tolerates both acidic and basic soil conditions growing often together with other *Riccia* species like *R. glauca* and *R. subbifurca*. Like other liverworts, it can reproduce both sexually through spore production or asexually by fragmentation. Interestingly, in contrast to model liverwort *Marchantia polymorpha*, *R. sorocarpa* is monoecious, meaning both male and female reproductive structures are found on the same plant. The preliminary phylogenetic studies resolved *R. sorocarpa* as monophyletic, forming common clade with species like *R. nigrella*, *R. macrocarpa*, *R. crinita* and *R. atromarginata*⁸ and later study resolved *R. sorocarpa* as sister to *R. crinita* and *subcrinita*¹¹. In comparison to other plant lineages, the nuclear genomes of liverworts are poorly explored, with only one, model species *Marchantia polymorpha* assembled to chromosome scale¹². The genomes of other species of complex thalloid liverworts were published as pseudochromosome or draft assemblies^{13,14}.

Nanopore sequencing stands apart in DNA sequencing with its remarkable capability to produce exceptionally long reads, detect modifications directly on the DNA molecule, and provide data in real-time. Long reads, often reaching hundreds of thousands of bases, offer an unparalleled advantage in tackling repetitive and complex regions of the genome. These long reads revolutionize genome assembly, enabling scientists to overcome the previous limitations of fragmented short-read approaches and unveil the true complexity of genomic architecture along with DNA modifications. Continuous improvements in nanopore sequencing technology have led to the capability of assembling and annotating chromosome-scale genomes using exclusively nanopore sequencing data¹⁵.

Liverwort genomic resources are still underrepresented, especially in comparison to mosses and vascular plants. Up to the date only two species were sequenced and assembled at chromosome scale: model species *M. polymorpha*¹² and *Ricciocarpos natans*¹⁶, while few other species were assembled at scaffold scale including *M. inflexa*, *M. palacea* and *Lunularia cruciata*^{13,17}.

To expand knowledge in the field of liverworts genomics, a high-quality chromosomal-scale telomere-to-telomere assembly of *R. sorocarpa* was generated applying exclusively Oxford Nanopore long-read sequencing and Pore-C technology. The assembled genome had a size of 376.690 Mbp with a contig N50 of 49.132 Mbp and GC content 41.53% (Table 1), and 97.02% of the assembled bases associated with the eight chromosomes. This genome will be a valuable resource as a genetic basis for further studies, including adaptive evolution, phylogenomics, population genomics and many others.

Methods

Material collection and cultivation. Small, young fragments of *Riccia sorocarpa* plants were collected from an arable field (Fig. 1) near Butryny (NE Poland, 53.603 N, 20.595E). Each fragment was treated as a possible separate plant. *In vitro* cultures of *R. sorocarpa* were carried out in the Department of Genetics, Plant Breeding and Bioresource Engineering of the University of Warmia and Mazury in Olsztyn (Poland). The *R. sorocarpa* fragments were washed in running tap water for 15 minutes. Thereafter, they were surface-disinfected for 10 min.

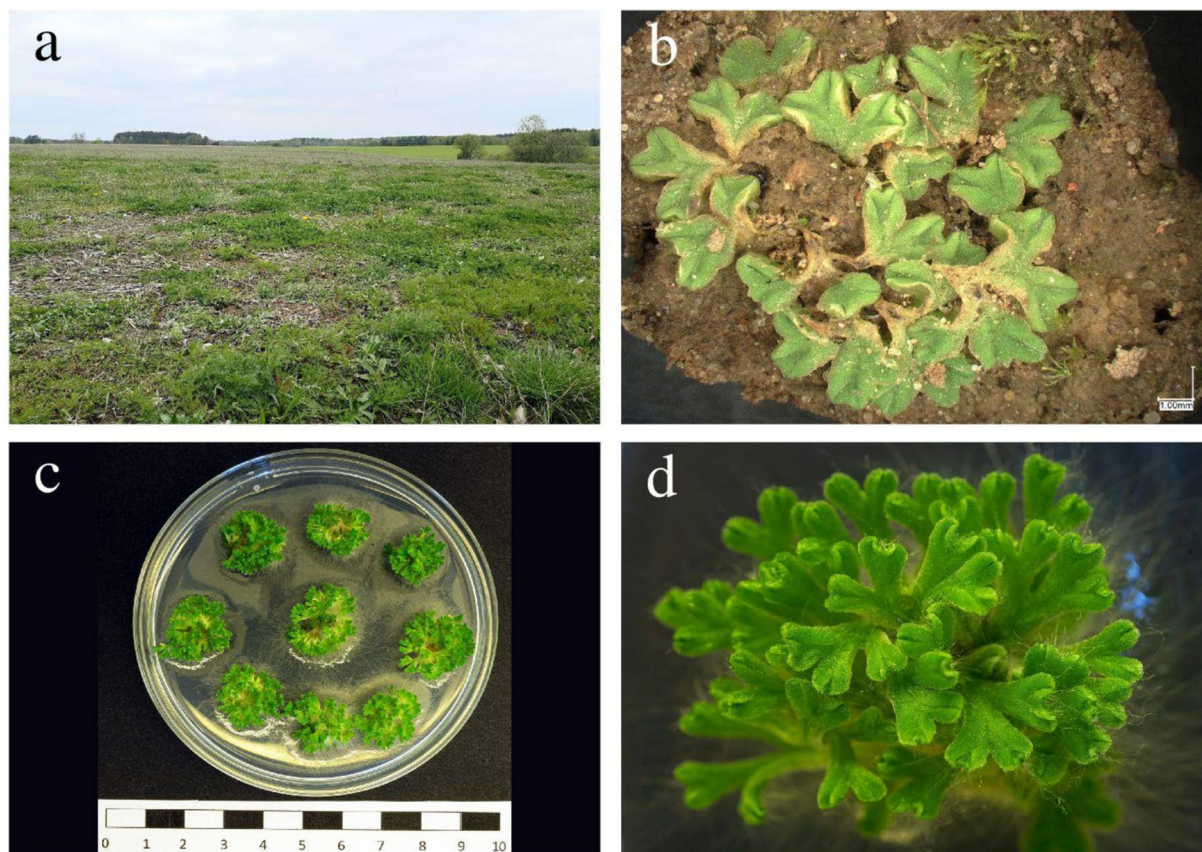


Fig. 1 Arable field near Butryny village (NE Poland) - the collection site (a). Thalli morphology on *Riccia sorocarpa*. (b) *In vitro* culture (c) and thalli morphology under *in vitro* conditions (d).

using 0.5% calcium hypochlorite (Chempur, Poland) with addition of 0.05% TWEEN-20 (POCH, Poland). After sterilization the explants were rinsed three times with sterile distilled water for 5, 10 and 15 min. Each *R. sorocarpa* explant was cultured separately on Petri dish with the solid half strength Gamborg's B5 medium ($\frac{1}{2}$ GB5)¹⁸ including $\frac{1}{2}$ basal salts (Sigma Aldrich Co., USA); organics and vitamins¹⁹ with 20 g/l sucrose, 8 g/l agar (A&A Biotechnology, Poland) and pH 6.0. Plants were grown in climate chambers at 24 °C under long-day conditions with 16 h and 8 h dark photoperiod. After first passage, the upper fragments of sterile plants of *R. sorocarpa* were used as secondary explants for micropropagation and were placed on the $\frac{1}{2}$ GB5 medium in the form of small clumps separated by approx. 1–2 cm. Micropropagated plants were used in the next parts of this study.

Long-read DNA sequencing and genome assembly. Before cryo-grinding the plants were kept in the dark for three days to reduce the abundance of secondary metabolites. DNA isolation was performed using DNeasy® Plant Mini Kit (Qiagen) according to the manufacturer's protocol. Quality and quantity control of the sample was carried out under Oxford Nanopore Technologies (ONT) recommendations. Sample concentration was checked using Qubit™ fluorometer and Qubit™ dsDNA HS Assay Kit and amounted 20 ng/μl, absorbance ratio $A_{260/230}$ and $A_{260/280}$ was checked with Carry spectrophotometer (Agilent) accounting 2.0 and 1.77, respectively. A total of 900 ng of native DNA was used for library preparation with Ligation sequencing kit V14 (LSK-SQK114) and NEBNext® Companion Module for ONT Ligation Sequencing (New England Biolabs). The sequencing experiment was run on the R10.4.1 flow cell (FLO-PRO114M) and the PromethION 2 Solo sequencing device.

The nanopore sequencing produced 96.206 Gbp of raw data composed of 40,443,226 long reads. Raw reads were basecalled using standalone Dorado 0.5.1 software (<https://github.com/nanoporetech/dorado>, ONT) with a super accuracy model (SUP) and stereo-duplex mode. After removing the low-quality reads by Porechop (v0.2.4) (<https://github.com/rrwick/Porechop>), clean data was evaluated by Whokaryote (v1.1.2)²⁰ and Centrifuge (v1.0.4)²¹ software to filter out potential bacterial DNA contamination. Among the 20,079,689 long reads, counting 64.795 Gbp with 3,929 N50 read length, 98.4% indicated the quality of sequence $Q > 10$ and 80.6% $Q > 15$ (Fig. 2). A fastq file was used to count K-mer frequencies using the *kmerfreq* program. Subsequently, the genome size was estimated using the *GCE* program and *kmerfreq* files²². The estimated genome size for k-mers from 14 to 21 fell within the 320–390 Mbp range. The preliminary assembly was performed by Flye (v2.9-b1774)²³ software with 3 rounds of polishing and -genome_size estimation equal to 400 m and -nano-hq parameters. As a result, the draft genome of *R. sorocarpa* was composed of 3,218 contigs with a total sequence length equal to 376,689,626 bp and a contig N50 = 258,577 (Table 1).

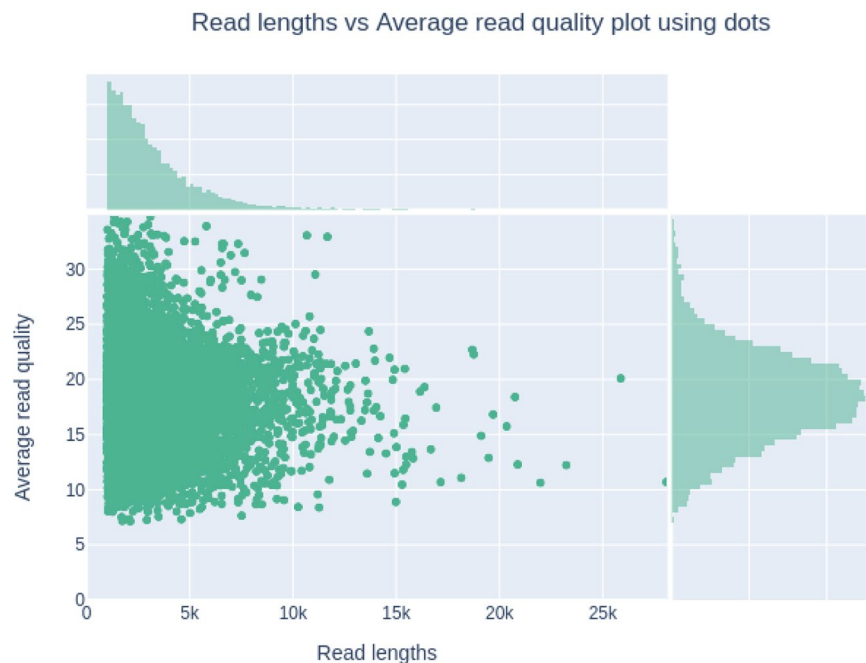


Fig. 2 Distribution of ONT DNA read lengths, considering the average read quality.

Pore-C procedure and chromosome-scale assembly. Approximately 2 g of fresh weight of the aerial parts of the thallus of *R. sorocarpa* were used for the Pore-C method. The restriction enzyme Pore-C for plant samples protocol (RE-Pore-C, ONT) was applied to capture three-dimensional interactions of DNA within chromatin (crosslinking of chromatin). Briefly, in the first step ‘in-nucleus’ chromatin was captured using formaldehyde to preserve nuclear structures within each chromosome. Then, the crosslinked plant material was cryo-grounded into a fine powder and proceeded to nuclei isolation. The suspension obtained after incubation in the nuclei isolation buffer was passed through a cell strainer (40 μ m). The filtrate was purified with a washing buffer and proceeded by chromatin digestion with NlaIII (NEB) restriction enzyme (18 h, 37 °C), which was then deactivated by heat denaturation.

In the next step, a proximity ligation reaction with 40,000 U of T4 DNA ligase (6 h, 16 °C) was performed, followed by protein degradation leading to chromatin de-crosslinking during which the sample was incubated with proteinase K (100 μ l) and RNaseA (100 ng/ μ l) for 18 h (56 °C). The reaction efficiency was increased by two additional rounds of proteinase K digestion (protocol HiPore-C v1)²⁴ performed using 10 μ l of Proteinase K (20 mg/ml) for each hour of reaction.

Finally, the extraction of DNA with chilled phenol:chloroform:isoamyl alcohol and EDTA was performed. The DNA was then precipitated with NaCl, washed with 80 and 70% ethanol, and eluted in 100 μ l TE buffer. Fragment length distribution and DNA integrity were analyzed with TapeStation using genomic DNA ScreenTape Assay (Agilent) reaching maximum intensity peak with length 29,362 bp and DIN 7.1, respectively. The concentration (10.2 ng/ μ l) was determined using the Qubit fluorometer HS DNA assay kit.

Two separate libraries were prepared for nanopore sequencing of proximity ligated DNA fragments with Ligation sequencing DNA V14 kit (SQK-LSK-114) according to the manufacturer's protocol. The only difference was time of end repair (NEBNext Companion Module for ONT Ligation Sequencing) which was extended to 20 min at 20 °C and 20 min at 65 °C. The libraries were sequenced on PromethION P2 Solo using R10.4.1 flow cells (FLO-PRO114M), generating 3,858,602 reads consisting of 12.058 Gb of raw data with a mean quality of 13.7. The trimmed Pore-C sequences were mapped to the draft assembly genome using the Pore-C-Snakemake pipeline (<https://github.com/nanoporetech/Pore-C-Snakemake>). The Pore-C contacts were generated using the 3D-DNA v.180419 procedure²⁵. After all, combining the draft genome and Pore-C contact data was used to create scaffolds by YaHS (v1.2) software²⁶. The procedure of Pore-C generation and scaffolding were iterated in the two loops and next the manual curation was done based on a heatmap of pairwise contacts using JuiceBox (v1.11.08)²⁷ (Fig. 3). The JuiceBox manual scaffolding result was reviewed using run-asm-pipeline-post-review.sh command implemented in 3D-DNA. Finally, the chromosome sequences were polished to reduce N bases into the *R. sorocarpa* draft genome using TGS-GapCloser software (v1.2.1)²⁸. The final genome assembly resulted in eight chromosomes what corresponded with literature data²⁹. It yielded a total length of 376,689,626 bp with the largest chromosome length of 55,960,836 bp. The 97.02% contigs were anchored to all chromosomes with an N50 length of 49.132 Mbp. After re-evaluating the primary assembly with ntLink software³⁰ 951 scaffolds from the contigs and trimmed Nanopore reads were generated, achieving an NG50 length of 839,990 bp. Evaluating the assembly genome by Inspector software³¹ the genome had a high consensus quality (QV) equal to 35.23 and reads mapping rate equal to 96.98%. To identify the basic repeated telomeric unit, the telomeric-identifier v.0.2.41 (<https://github.com/tolkkit/telomeric-identifier>) was used with default parameters (1% of the length of the chromosome either side). 2031 AAAACT motifs were detected as telomeric repeat units.

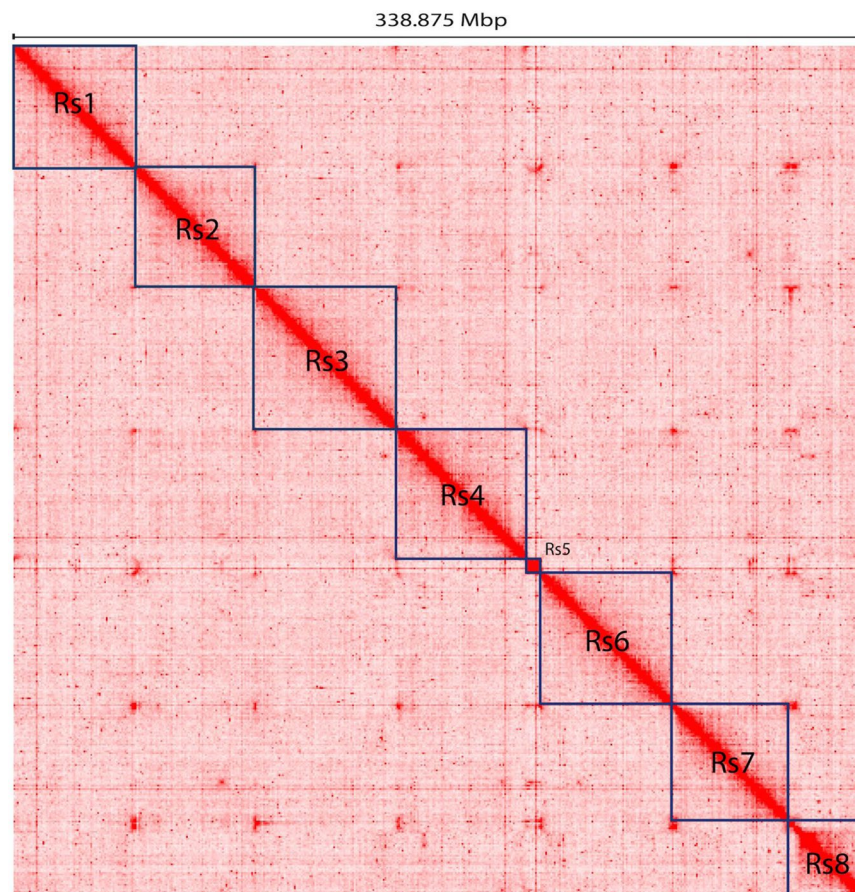


Fig. 3 Pore-C interaction heatmap for the *R. sorocarpa* genome.

Repetitive sequence and genome annotation. The chromosome-scale genome was scanned by EDTA³² and RepeatMasker (<https://github.com/rmhubble/RepeatMasker?tab=readme-ov-file#repeatmasker>) tools and repetitive sequences accounted for 42.17% of the genome assembly. In total, the prediction of repetitive elements distributed 61.62 Mbp (18.18%) transposon DNA, 76.10 Mbp (22.46%) long terminal repeats (LTRs) and 5.19 Mbp (1.53%) unclassified and miniature inverted-repeat transposable elements (MITEs) (Fig. 4).

The GC content and gene density were calculated in 500,000 nt windows for each chromosome using the GC.content function from the ape (v5.8) (<https://github.com/emmanuelparadis/ape>) package in the R environment. The highest average GC content was observed on chromosome 5 reaching 43.5%. The highest gene density (69) was observed on chromosome 8 between 26,500,000–27,000,000 bp (Fig. 5).

The cDNA long-read library was prepared for annotation of protein-coding genes. Briefly, total RNA was extracted from cryo-grinded thallus using RNeasy® Plant Mini Kit (Qiagen). The quantity of extracted RNA was measured with the Qubit™ RNA HS Assay Kit. Next approximately 150 ng of total RNA was reversely transcribed to cDNA and amplified to prepare the cDNA library according to the protocol provided by ONT for cDNA PCR Sequencing Kit (SQK-PCS109). Adapter-ligated cDNA sequences were sequenced on the MinION flow cell (vR9.4.1, FC106D) with MinION Mk1C sequencing device.

The cDNA long reads were assembled *de novo* by RNA-Bloom (v2.0.1)³³ to reconstruct 26,755 transcripts of *R. sorocarpa*. Next, the assembly transcriptome was implemented in the PASA pipeline (v2.5.3)³⁴ to model complete and partial protein-coding genes. Simultaneously, *ab initio* predictions were carried out with the assistance of BRAKER (v2.0.4)³⁵, GeneMark-ES/ET (v4.72)³⁶ and AUGUSTUS (v3.3.3)³⁷. Based on protein sequences of *Riccia fluitans* (PRJNA1158334), 30,615 potential gene annotations were predicted, respectively. Subsequently, a transcriptome-based set and two proteome-based predictions were integrated using EvidenceModeler (v2.1.0)³⁸ with weight bias parameters and finally 27,626 protein-coding genes were obtained.

Next, unique transcripts were translated into protein sequences and functionally annotated against multiple protein databases using blastp (v2.12.0)³⁹ and diamond (v2.0.15)⁴⁰. The best matches of protein alignments were assigned to seven functional databases as follows: 10,456 had homologous with SwissProt⁴¹, 13,711 with protein families database (Pfam)⁴², 16,284 Kyoto Encyclopedia of Genes and Genomes (KEGG)⁴³, 17,028 EggNOG-mapper⁴⁴, 18,177 Translation of EMBL (TrEMBL)⁴¹, 18,219 Non-Redundant Protein Sequence Database (<http://www.ncbi.nlm.nih.gov/protein>) and 21,030 Kofam (https://github.com/takaram/kofam_scan) (Fig. 5).

Collinearity of Marchantiales. Genome synteny between *R. sorocarpa*, *R. natans* and *M. polymorpha* were constructed by translating the longest isoforms of single-copy genes, utilizing GFF3 and BED format using JCVI

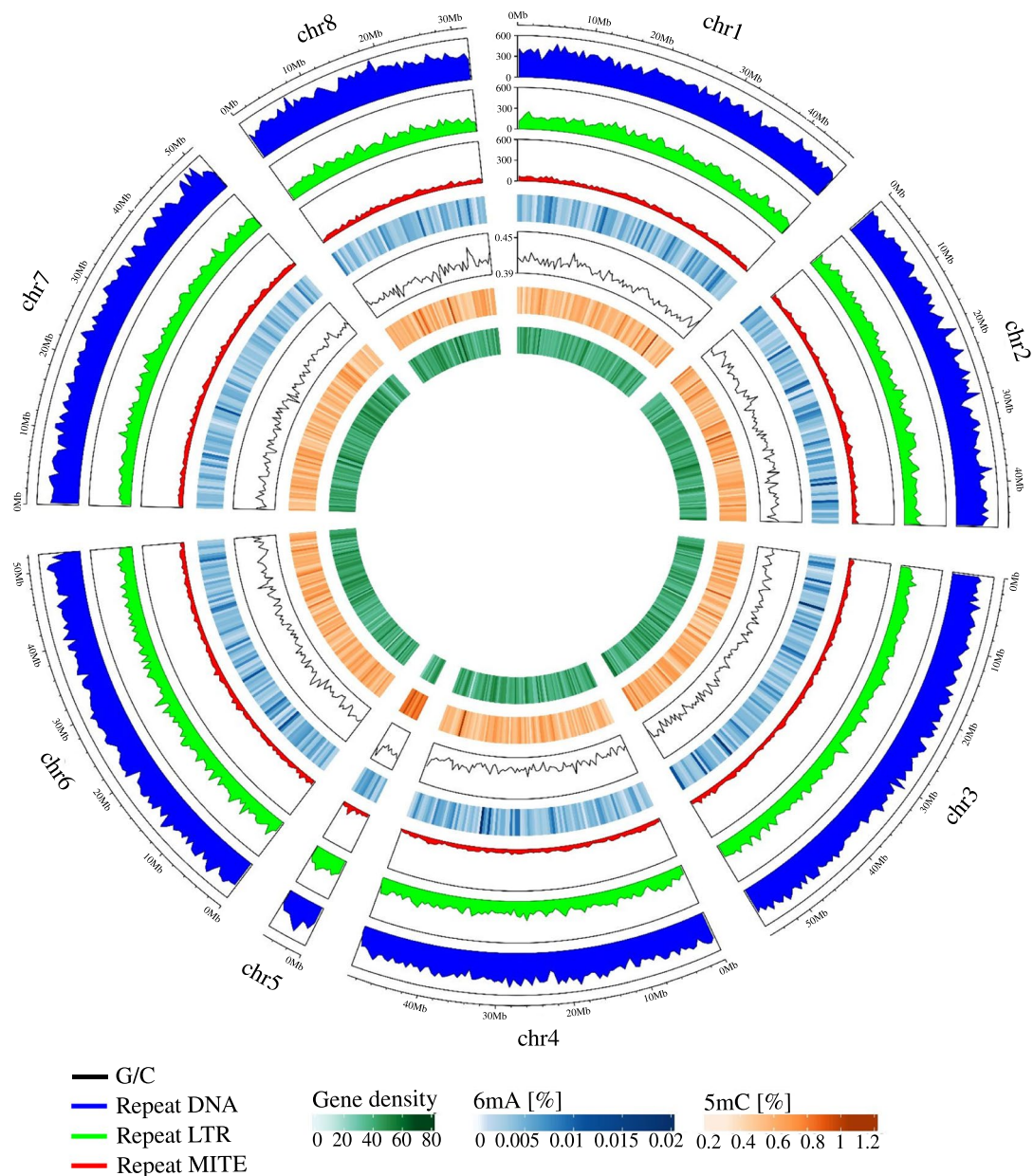


Fig. 4 Circos plot of eight *R. sorocarpa* chromosomes. The basic unit of information on the circus is 500,000 nucleotides. The first track (blue) shows the distribution of transposon DNA, the second track (green) shows the distribution of LTR transposons, the third track (red) shows the distribution of MITEs transposons, the fourth track reflects the blue heatmap of % 6mA to all adenines in one window, the fifth track shows the G/C ratio, the sixth track shows the red heatmap of % 5mC to all cytosines in the window, and the last track shows the gene density heatmap in the window.

software (v1.0.0)⁴⁵. The collinearity between eight chromosomes of each Marchantiales revealed that chromosomes 1, 2 and 4 had highly conserved relationships between each species. Other chromosomes showed more complex and divergent synteny patterns. Interestingly, *R. sorocarpa* indicated a deep reduction of chromosome 5 (Fig. 6).

Gene family evolution analysis. The trimmed alignment of single-copy proteins, along with the Newick format tree, was utilized to estimate the divergence times for the nodes in the phylogenetic tree using the *MegaX* program (v10.1.7)⁴⁶ was used to estimate the divergence time for nodes in the phylogenetic tree. The species divergence time-points were calibrated by fossil data extracted from Bechteler *et al.*⁴⁷, considering the divergence within *Syntrichia* genus [25 million years ago (mya)], *Sphagnum* genus (15 mya), between *S. ruralis* and *C. purpureus* (160 mya), *S. fallax* and *P. patens* (403 mya), *M. polymorpha* and *R. natans* (248 mya), and *M. polymorpha* and *P. patens* (455 mya). The estimated potential divergence time within *Marchantia* genus

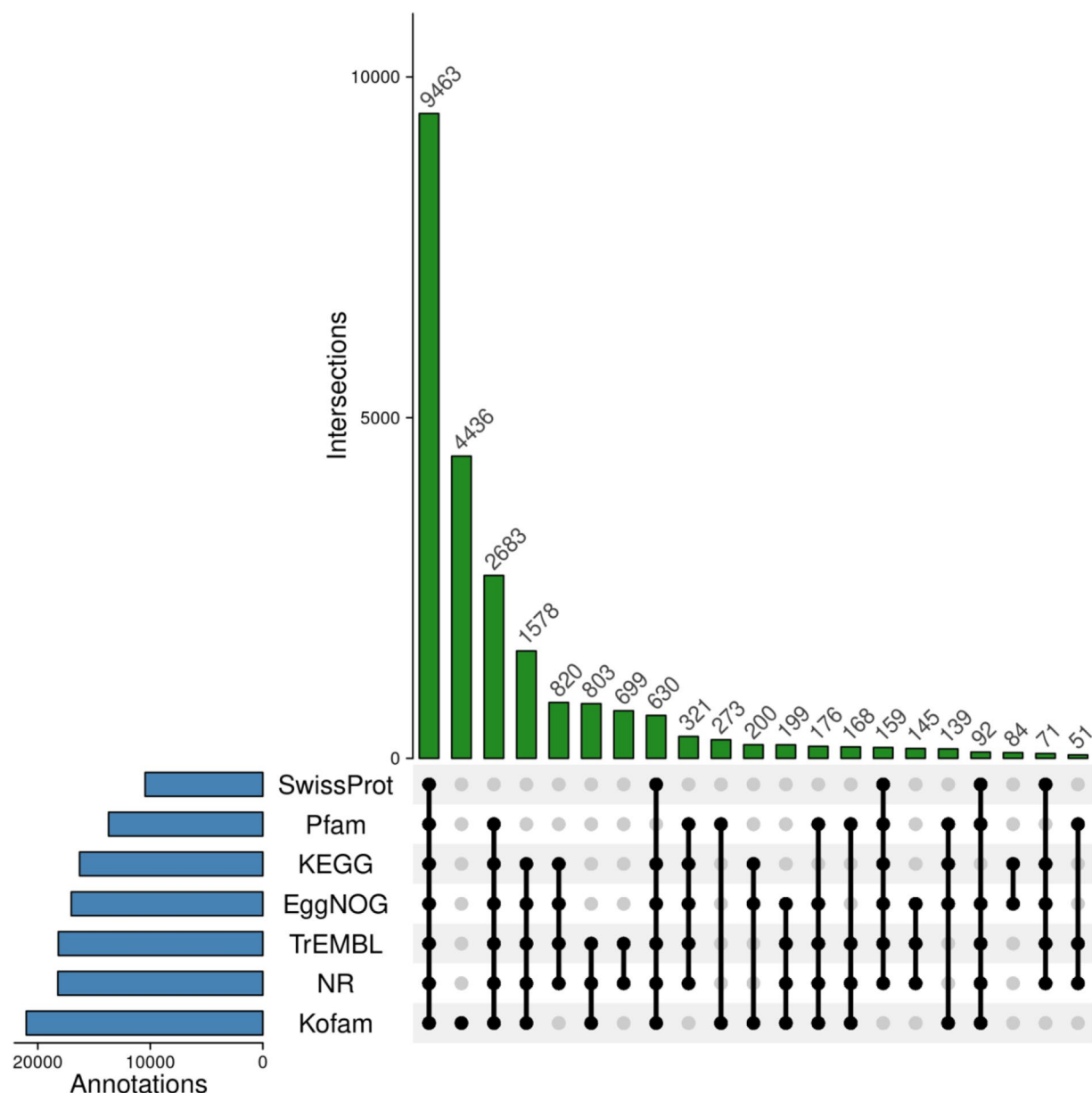


Fig. 5 Upset plot comparing functional gene annotations across seven databases: KEGG, Swiss-Prot, EggNOG, NR, TrEMBL, Kofam, and Pfam.

was established at 53.82 mya and between *R. natans* and *R. sorocarpa* at 174.8 mya. The dataset of 13 proteomes was evaluated by *Orthofinder* software (v2.5.4)⁴⁸. The longest representatives of each gene loci were extracted by *AGAT* software (v1.4.0)⁴⁹. To reconstruct phylogenetic orthology inference of *R. sorocarpa* (27,626 proteins) and other fully annotated 12 Bryophytes species, the following translation of single-copy genes was used: 24,742 for *Anthoceros agrestis*⁵⁰, 30,425 for *Ceratodon purpureus*⁵¹, 23,428 for *Lunularia cruciate*⁵², 18,566 for *Marchantia palaceae*⁵³, 17,944 for *Marchantia polymorpha* <https://marchantia.info/>, 32,926 for *Physcomitrium patens*⁵⁴, 17,728 for *Ricciocarpos natans*⁵², 25,100 for *Sphagnum fallax*⁵⁵, 25,227 for *Sphagnum magellanicum*⁵⁶, 16,544 for *Syntrichia caninervis*⁵², 21,169 for *Syntrichia ruralis*⁵², and 27,467 for *Takakia lepidozoides*⁵⁷.

Based on 18,261 single-copy orthogenes the gene family expansion and contraction analysis was done using the *CAFE* program (v5)⁵⁸ (Fig. 7). Following the divergence of the *Riccia* genus from its common ancestor with *R. natans*, 960 gene families expanded while 798 gene families contracted. According to single-copy genes present in the majority of 13 species, the multi-sequence alignment (MSA) was performed. Next, the *trimAl* software (v1.4)⁵⁹ cut the ambiguous positions within MSA and the phylogenetic tree was constructed using *IQ-TREE* 2 software (v2.2.0.3)⁶⁰ (Fig. 7).

DNA methylation calling. A previously generated BAM file was aligned using *Dorado* (v0.6.1) (<https://github.com/nanoporetech/dorado>) with our genome as a reference. It was then sorted and indexed employing

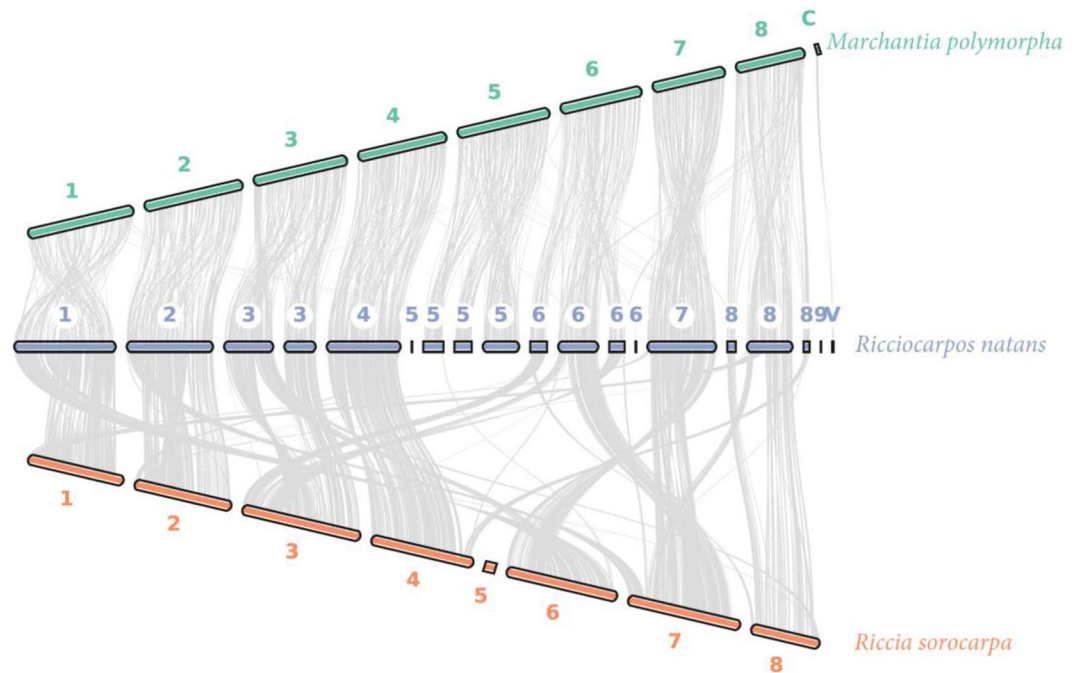


Fig. 6 Synteny plot shows the differences and similarities between our assembly of the *Riccia sorocarpa* (orange) chromosomes and the assembly of *Ricciocarpus natans* (blue) and *Marchantia polymorpha* (green) by other authors. Grey line thickness reflects collinearity between species genomes.

Samtools (v1.12)⁶¹. The next step was to extract DNA methylation information leveraging modkit (v0.2.7.) (<https://github.com/nanoporetech/modkit/releases>). Methylations were qualified for further analysis when `N_valid_cov > 20` and `fraction_modified = 100`. Next, methylations were divided into 5-methylcytosine (5mC) and 6-methyladenine (6mA). A total of 404,506 methylated cytosines and 4,071 methylated adenines were identified (Fig. 4). Their percentage occurrence was calculated within a 500,000 window across the genome in the R environment. Terminal chromosome regions were treated as a single window when there were not enough nucleotides to fill the 500,000-base window. The highest percentage of methylated cytosines to cytosines within a window was observed on chr4 between 49000000-49131805 (1.09%), chr1 46500000-46876028 (0.99%), and chr5 45000000-5374985 (0.98%). The highest percentage of methylated adenines was observed on chr3 18500000-19000000 (0.021%), chr4 40500000-41000000 (0.02%), and chr3 49000000-49500000 (0.018%).

Data Records

The Nanopore sequencing data used for genome assembly were deposited in the NCBI Sequence Read Archive database under BioProject [PRJNA1130527](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1130527)⁶². The DNA long reads Nanopore sequencing data was deposited under SRR29669162 accession number. The Pore-C data was stored under accession numbers SRR29695115 and SRR29695116. The transcriptomic sequencing data were deposited in the SRA at NCBI accession SRR29695421⁶³. The assembled genome and gene annotation can be found on NCBI assembly with the accession number [JBJQOH000000000](https://www.ncbi.nlm.nih.gov/assembly/JBJQOH000000000)⁶⁴.

Technical Validation

The Q10 and Q15 of Nanopore sequencing data was greater than 96% and 68%, respectively. The Q10 and Q15 of Pore-C sequencing data were greater than 95% and 52%, respectively. Genome completeness was confirmed by BUSCO (Benchmarking Universal Single-Copy Orthologs, v5.4.2)⁶⁵ analysis, which identified 85.1% single-copy, 6.7% duplicated genes, 1.6% fragmented and 6.6% missing in the eight longest scaffolds. To examine assembly integrity, the Nanopore trimmed sequences were checked using Minimap2 aligner (v2.17-r941)⁶⁶ and Geneious Prime software (v2023.0.1, <https://www.geneious.com>). The 97.02% of chromosome regions had high completeness after contigs remapping procedure and described by median equal to 163. The Pore-C trimmed reads were also aligned with a high map rate of 84.56% to the manually curated *R. sorocarpa* genome. The evidence supports the reliability of the data, validating its use in subsequent analytical procedures.

Usage Notes

The availability of this genome, the second near telomere-to-telomere assembly of a liverwort, provides a valuable resource for in-depth analysis of *Riccia sorocarpa*'s genetic diversity and will support further studies in adaptive evolution, phylogenomics, population genomics, and more. However, the low integrity of the raw DNA used for long-read sequencing has contributed to shorter ONT reads and reduced contig lengths. Future improvements to the genome assembly could be achieved with higher-quality DNA samples. The genome's QV is reported to be above 35, which is considered moderate. Nevertheless, the application of the Pore-C method

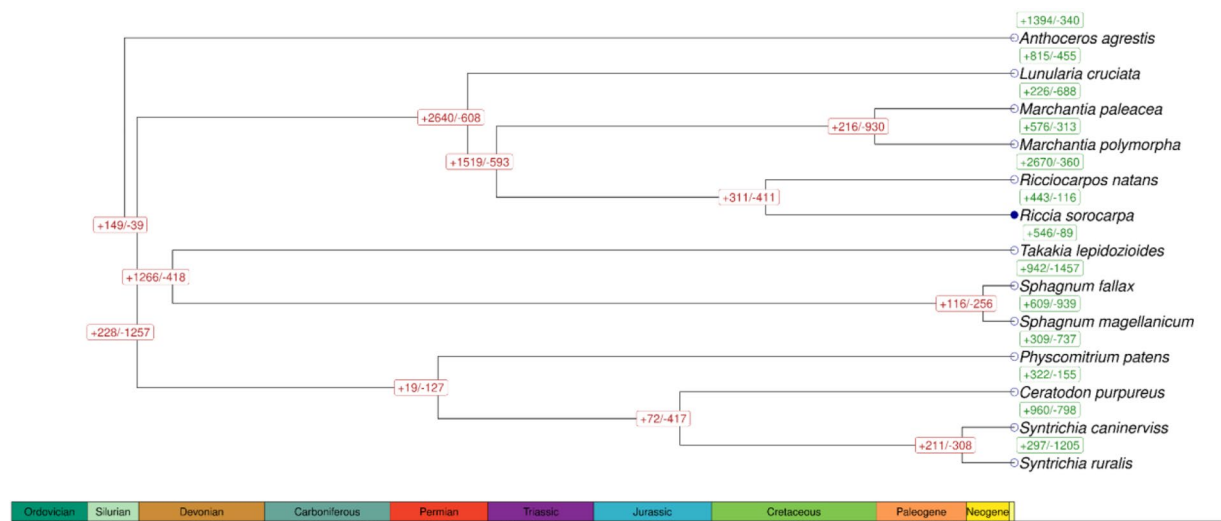


Fig. 7 Divergence time tree of *R. sorocarpa* and 12 other species. The species divergence time was estimated using MegaX. Green and red numbers above each species' name and divergence points reveal expansions (+) and contractions (−) of gene families.

enables a satisfactory assembly of the genome. The process of organizing the genome at the chromosomal scale involved multiple iterations that integrated data from both long-read ONT sequencing and Pore-C. This comprehensive approach facilitated the accurate assembly of the *Riccia sorocarpa* genome.

Code availability

The guidelines for all bioinformatics tools employed in this study were adhered strictly. The software and code utilized are publicly available, with specific versions and parameters detailed in the Methods section. Our data compilation process did not involve any custom code.

Received: 25 July 2024; Accepted: 1 January 2025;

Published online: 15 January 2025

References

- Harris, B. J. *et al.* Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nat Ecol Evol* **6**, 1634–1643 (2022).
- Graham, L. E., Cook, M. E. & Busse, J. S. *The Origin of Plants: Body Plan Changes Contributing to a Major Evolutionary Radiation*. PNAS vol. 97 www.pnas.org (2000).
- Kanazawa, T. *et al.* The liverwort oil body is formed by redirection of the secretory pathway. *Nature Communications* **11**, 1–11 (2020). 2020 11:1.
- Romani, F. *et al.* Oil Body Formation in *Marchantia polymorpha* Is Controlled by MpC1HDZ and Serves as a Defense against Arthropod Herbivores. *Curr Biol* **30**, 2815–2828.e8 (2020).
- Asakawa, Y. Chemosystematics of the Hepaticae. *Phytochemistry* **65**, 623–669 (2004).
- Kutnar, L., Kermavnar, J. & Sabovljević, M. S. Bryophyte diversity, composition and functional traits in relation to bedrock and tree species composition in close-to-nature managed forests. *Eur J For Res* **142**, 865–882 (2023).
- Söderström, L. *et al.* World checklist of hornworts and liverworts. *PhytoKeys* **59**, 1–828 (2016). 59: 1-828.
- Cargill, D. C., Neal, W. C., Sharma, I. & Gueidan, C. A preliminary molecular phylogeny of the genus *Riccia* L. (Ricciaceae) in Australia. *Aust Syst Bot* **29**, 197–217 (2016).
- Codd, L. E. W., Dyer, R. A., Rycroft, H. B., de Winter, B. & Africa, S. *Flora of Southern Africa: The Republic of South Africa, Basutoland, Swaziland and South West Africa. Flora of Southern Africa: the Republic of South Africa, Basutoland, Swaziland and South West Africa*. <https://doi.org/10.5962/bhl.title.119879> (Govt. Printer, Pretoria, 1999).
- Smith, A. E. The liverworts of Britain and Ireland. (1989).
- Xiang, Y. L. *et al.* Morphological and molecular evidence confirms a new species, *Riccia subcrinita* YouL.Xiang & R.L.Zhu) and *Riccia junghuhniana* Nees & Lindenb. (Ricciaceae, Marchantiophyta) new to China. *Phytotaxa* **531**, 41–53 (2022).
- Bowman, J. L. *et al.* Insights into Land Plant Evolution Garnered from the *Marchantia polymorpha* Genome. *Cell* **171**, 287–304.e15 (2017).
- Linde, A. M. *et al.* Genome Evolution in Plants: Complex Thaloid Liverworts (Marchantiopsida). *Genome Biol Evol* **15** (2023).
- Singh, S., Davies, K. M., Chagné, D. & Bowman, J. L. The fate of sex chromosomes during the evolution of monoicy from dioicy in liverworts. *Current Biology* **33**, 3597–3609.e3 (2023).
- Ong, C. T. *et al.* Chromosome-Scale Genome Assembly of the Sheep-Biting Louse *Bovicola ovis* Using Nanopore Sequencing Data and Pore-C Analysis. *Int J Mol Sci* **25**, 7824 (2024).
- Singh, S. & Bowman, J. L. The monoicous secondarily aquatic liverwort *Ricciocarpos natans* as a model within the radiation of derived Marchantiopsida. *Front Plant Sci* **14**, 1260596 (2023).
- Marks, R. A., Smith, J. J., Cronk, Q., Grassa, C. J. & McLetchie, D. N. Genome of the tropical plant *Marchantia inflexa*: implications for sex chromosome evolution and dehydration tolerance. *Sci Rep* **9** (2019).
- Gamborg, O. L., Miller, R. A. & Ojima, K. Nutrient requirements of suspension cultures of soybean root cells. *Exp Cell Res* **50**, 151–158 (1968).
- Sawicki, J. *et al.* Nanopore sequencing of organellar genomes revealed heteroplasmy in simple thalloid and leafy liverworts. *Acta Societatis Botanicorum Poloniae* **92** (2023).

20. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb Genom* **8**, 000823 (2022).
21. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26**, 1721–1729 (2016).
22. Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects (2013).
23. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of Long Error-Prone Reads Using Repeat Graphs. <https://doi.org/10.1101/247148>.
24. Zhong, J. Y. *et al.* High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. *Nature Communications* **14**, 1–18 (2023). 2023 14:1.
25. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* (1979) **356**, 92–95 (2017).
26. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39** (2023).
27. Dudchenko, O. *et al.* The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. *bioRxiv* <https://doi.org/10.1101/254797> (2018).
28. Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, 1–11 (2020).
29. Chromosome Numbers in Certain Ricciaceae on JSTOR. <https://www.jstor.org/stable/86681>.
30. Coombe, L. *et al.* LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* **22**, 1–13 (2021).
31. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol* **22**, 1–21 (2021).
32. Ou, S. *et al.* Differences in activity and stability drive transposable element variation in tropical and temperate maize. *bioRxiv* <https://doi.org/10.1101/2022.10.09.511471> (2022).
33. Nip, K. M. *et al.* Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nature Communications* **14**, 1–14 (2023). 2023 14:1.
34. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654 (2003).
35. Humann, J. L., Lee, T., Ficklin, S. & Main, D. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. *Methods Mol Biol* **1962**, 29–51 (2019).
36. Borodovsky, M. & Lomsadze, A. Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis... [et al.] CHAPTER*, Unit (2011).
37. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439 (2006).
38. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, 1–22 (2008).
39. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
40. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366–368 (2021). 2021 18:4.
41. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48 (2000).
42. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412 (2021).
43. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457–D462 (2016).
44. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825–5829 (2021).
45. Tang, H. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* e211, <https://doi.org/10.1002/IMT2.211> (2024).
46. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549 (2018).
47. Bechteler, J. *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* **110**, e16249 (2023).
48. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 1–14 (2019).
49. Dainat, J. NBISweden/AGAT: AGAT-v1.4.0. <https://doi.org/10.5281/ZENODO.11106497> (2024).
50. Download | Hornworts | UZH <https://www.hornworts.uzh.ch/en/download.html>.
51. Cpupureus_GG1 v1.1: Phytozome https://phytozome-next.jgi.doe.gov/info/CpupureusGG1_v1_1.
52. CoGe: Comparative Genomics. <https://genomevolution.org/coge/>.
53. Marchantia paleacea genome assembly ASM1418076v2 - NCBI - NLM. https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_014180765.2/.
54. Ppatens v3.3: Phytozome https://phytozome-next.jgi.doe.gov/info/Ppatens_v3_3.
55. Sfallax v1.1: Phytozome. https://phytozome-next.jgi.doe.gov/info/Sfallax_v1_1.
56. https://phytozome-next.jgi.doe.gov/info/Smagellanicum_v1_1 S.magellanicum v1.1: Phytozome.
57. Takakia lepidiozoides project | Home: Superman. <https://www.takakia.com/>.
58. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
59. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
60. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).
61. Daneczek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).
62. ID 1130527 - BioProject - NCBI. <https://identifiers.org/ncbi/bioproject:PRJNA1130527> (2024).
63. SRP517367 - SRA - NCBI. <https://identifiers.org/ncbi/insdc.sra:SRP517367>.
64. Pauksztó, L. Riccia sorocarpa isolate LP-2024, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBQOH000000000>.
65. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
66. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

Acknowledgements

This research was financially supported by the National Science Center, Poland [Grant No. 2020/39/B/NZ8/02504 (resources, methods establishment) and Grant No. 2023/07/X/NZ8/01263 (sequencing consumables)] and the statutory funding of Faculty of Agriculture and Forestry, Department of Genetics, Plant Breeding and Bioresource Engineering, University of Warmia and Mazury in Olsztyn (Grant No. 30.610.007-110 - *in vitro* cultures).

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.K. or J.S.-P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025