Title: Crossroads of assembling a moss genome: navigating contaminants and horizontal gene transfer in the moss *Physcomitrellopsis africana*

Vidya S. Vuruputoor[1], Andrew Starovoitov[1], Yuqing Cai[2], Yang Liu[2], Nasim Rahmatpour[1], Terry A. Hedderson[3], Nicholas Wilding[4], Jill L. Wegrzyn[1,5]*, Bernard Goffinet[1]*

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA, 06269

[2]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China; Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake

[3]Bolus Herbarium, Department of Biological Sciences, University of Cape Town, Private Bag, 7701, Rondebosch, South Africa

[4]Université de La Réunion, UMR PVBMT, BP 7151, chemin de l'IRAT, 97410 Saint-Pierre, La Réunion, France; Missouri Botanical Garden, P.O. Box 299, St. Louis, MO, 63166-0299, U.S.A.

[5]Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut, USA, 06269

ORCID numbers:

VV: 0000-0002-5836-7054

YC: 0009-0009-3715-2482

YL: 0000-0002-5942-839X

TH: 0000-0002-3537-6599

NW: 0000-0003-4029-5387

JW:0000-0001-5923-0888

BG: 0000-0002-2754-3895

* Corresponding authors: Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, 06269. Email: jill.wegrzyn@uconn.edu and bernard.goffinet@uconn.edu

Running head: *Physcomitrellopsis* genome

Journal: G3 (Genome Report)

Abstract

The first reference genome assembly of the moss *Physcomitrellopsis africana*, a rare narrow endemic terricolous species from southeastern coastal forests of South Africa, is presented here. Phylogenetically, *Physcomitrellopsis africana* bridges the major evo-devo moss models *Physcomitrium patens* and *Funaria hygrometrica*, which diverged from their common ancestor 60-80 million years ago. The *Physcomitrellopsis africana* genome was assembled with both long Nanopore reads (73x) and short Illumina reads (163x). The 440 Mb assembly comprises 2,396 contigs and 23,493 protein-coding genes (BUSCO: C:96.0%[D:13.9%]), including two unique genes of putative microbial origin absent in close relatives. While the informatic approaches to genome assembly are becoming more standardized, best practices for contamination detection are less defined. The long reads sequenced for the *Physcomitrellopsis africana* genome contained approximately 12% contamination originating from microbial sources. This study describes the informatic processes employed to distinguish contaminants from candidate horizontal gene transfer events. Following the assembly and annotation, examination of whole genome duplication, and patterns of gene family expansion and contraction, were conducted. The genome bears signatures of two whole genome duplications shared with *Physcomitrium patens* and *F. hygrometrica*. Comparative analyses of gene family evolution revealed contractions associated with the light harvesting regulatory network in *Physcomitrellopsis africana* in comparison to *Physcomitrium patens* and *F. hygrometrica*. This first high-quality African bryophyte genome provides insights into genome evolution and HGT in an understudied moss lineage.

Keywords: Funariaceae; reference genome; *Physcomitrella*; Bryophyta; HGT; contamination

Article Summary

The first draft genome of the rare moss, *Physcomitrellopsis africana*, endemic to South Africa's southeastern coastal forests, is presented here. The 440 Mb genome fills a critical phylogenetic gap, enabling the first comparative analysis of the genomes of three moss species that diverge over the last 60-80 million years. The analysis uncovered 23,493 genes and provides new insights into genome evolution and gene family expansion/contraction in mosses. Rigorous contaminant filtering also identified two genes uniquely acquired through horizontal gene transfer.

Introduction

Horizontal gene transfer (HGT) is the lateral movement of genetic material between divergent branches of the tree of life. This process is ubiquitous among bacteria, facilitating rapid adaptation through exchange of ecologically important genes (Aminov, 2011). While HGT is less common in eukaryotes compared to prokaryotes, it does play a role in shaping eukaryotic evolution with about 0.04-6.49% of eukaryotic genomes originating from HGT from microbes (Van Etten & Bhattacharya, 2020). During the evolution of land plants, increasing interactions with rhizosphere microbes, particularly bacteria and mycorrhizal fungi, have enabled occasional horizontal transfer of functional genes between these distantly related lineages (Martin et al., 2017). For example, the gene for Killer Protein 4 (KP4) was likely acquired by mosses through HGT from ascomycetes fungi (Guan et al., 2023). Similarly, the HET domain, common in fungal heterokaryon incompatibility genes and involved in self/non-self recognition, was identified as a truncated form in *Physcomitrium patens* (G. Sun et al., 2020). This gene is associated with moss-fungus interactions, possibly as a defense mechanism. Van Etten & Bhattacharya (2020) suggest that rather than merely anecdotal cases, HGT has been an important evolutionary force driving land plant adaptation to new habitats and stressors throughout their evolution.

Accurately detecting the taxonomic origin of genes and confirming horizontal gene transfer (HGT) events is challenging. Many reported cases of HGT in published genomes have proven to be artifacts resulting from contamination that went undetected. For example, the initial analysis of the genome of the tardigrade *Hypsibius dujardini* reported that 17% of its genes originated from HGT (Boothby et al., 2015), a number subsequently revised to only 0.2% of genes (Koutsovoulos et al., 2016). Similarly, claims of novel DNA modifications in mammals

also turned out to be erroneous due to bacterial contamination, highlighting the need for careful analyses and experimental validation (Douvlataniotis et al., 2020). A systematic screening of 43 arthropod genomes by François et al. (2020) revealed extensive bacterial contaminants, often outnumbering true horizontal gene transfer (HGT) events. For example, in the bumblebee *Bombus impatiens*, most contaminating genes were concentrated on just 30 contaminant scaffolds. Based on the size and number of contaminating sequences, it was concluded that the genome of the symbiont *Candidatus Schmidhempelia bombi* was co-assembled with its host. Strategies to confidently identify HGT, including taxonomic assignment of candidates via BLASTp/DIAMOND, phylogenetic analysis with candidate and donor proteins, synteny analysis of flanking genes, and quantitative PCR validation, have been presented to address these issues (Francois et al., 2020).

These strategies have been applied in recent land plants to identify HGT candidates For instance, the genomes of two fern species, *Azolla filiculoides* and *Salvinia cucullata*, were sequenced and screened for HGT by Li et al. (F.-W. Li et al., 2018)). A fern-specific insect resistance gene identified in *A. filiculoides* appears to have originated from bacteria via HGT based on phylogenetic analyses showing clustering of the fern gene with bacterial orthologs. The authors confirmed candidates by analyzing flanking genes and performing RT-PCR. A similar methodology was used in a large analysis of HGT across land plants, comparing candidates to "donor" and "recipient" databases (Ma et al., 2022).

Physcomitrellopsis africana inhabits transitional zones between grassland and forests in coastal habitats in South Africa. It is currently the sole species of the genus, which likely should also include several species of the paraphyletic genus *Entosthodon* (Wilding, 2015). This resource complements those available for *Funaria hygrometrica* and *Physcomitrium patens* (Kirbis et al., 2022; Rensing et al., 2008) and thereby constitutes a fundamental resource to further the reconstruction of the evolution of the genome of *Physcomitrium patens*, which is widely used in comparative genomics and land plant phylogenomics (Rensing et al., 2020). The first genome of an African bryophyte, a 440 Mb reference for the moss, *Physcomitrellopsis africana*, is presented here. Through rigorous contamination screening and verification, two unique candidate HGT events are described from this assembly. Characterizing validated HGT events in *Physcomitrellopsis africana* yields evolutionary insights while underscoring the need for stringent standards to support HGT conclusions from genomics data.

Materials and Methods

Sample collection and culturing

A population of *Physcomitrellopsis africana* was sampled in October 2010 from a coastal forest in the Eastern Cape Province of South Africa (Dwesa National Park, along trail to chalets behind campground, coordinates 32°18.222' S 28°49.666', at ± sea level). The voucher specimens collected by Goffinet (collection numbers 10326 and 10329, with T Hedderson and N Wilding) are deposited in the CONN herbarium under accession numbers CONN00235389 and CONN00235388, respectively. Specimen 10326 (culture and long read library DNA #5074) provided DNA for genome sequencing and assembly, while 10329 (RNA and Illumina DNA #5075) provided RNA and DNA for Illumina sequencing. Sterile cultures were first established on Knop medium using spores from a single operculate capsule. The gametophytes were harvested, ground and spread on a rich sandy loam soil in PlantCon tissue culture containers (MP Biomedicals, Solon, OH, USA), and maintained in a growth chamber under 16 h of daylight at about 24℃.

Genomic DNA and RNA extraction

Gametophytic tissue, stems and leaves, of *Physcomitrellopsis africana* was harvested from fresh soil cultures under a dissecting microscope and ground in liquid nitrogen. DNA was extracted by following a modified protocol by Young LA (2022). The quality of the DNA sample 5074 was assessed by quantitative PCR prior to sequencing, yielding a DIN score of 7.0 and concentration of 40.9 ng/µL. The DNA was then prepared for PromethION sequencing through a DNA repair step, generating blunt ends, and ligating sequencing adapters followed by priming of the flow cell, as described in the Oxford Nanopore Technologies Amplicons by Ligation (SQK-LSK109) protocol.

DNA for short read sequencing was extracted using the NucleoSpin Plant midi DNA extraction kit, following the manufacturer's protocol (Macherey-Nagel, Düren, Germany). DNA quality was evaluated using a Qubit® 3.0 Fluorometer (Thermo Fisher Scientific, USA). Total RNA was extracted from approximately 1 g of fresh gametophytic tissue using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA).

141 **Genome and transcriptome library preparation and sequencing**

142 The HMW DNA library was sequenced on the Oxford Nanopore PromethION long-read
143 sequencing instrument. The short-read DNA libraries were sequenced (150 bp PE) on two lanes
144 of a BGISEQ-500 sequencing platform (BGI-Shenzhen, China), library preparation followed the
145 methods used by Yu et al (2020).

146 Approximately 1 μg of RNA was used to generate two paired-end libraries with an insert
147 fragment size of 200–300 bp of the corresponding cDNA. RNA purification, reverse
148 transcription, library construction and sequencing were performed at WuXi NextCode (Shanghai,
149 China). The captured coding regions of the transcriptome from total RNA were prepared using
150 the TruSeq® RNA Exome Library Preparation Kit. The two RNA libraries were sequenced on
151 one lane of an Illumina HiSeq 2000 instrument (100 bp PE) at the WuXi NextCode (Shanghai,
152 China).

153 **Quality control of genomic and transcriptomic reads**

154 The genomic short reads were first assessed with FASTQC v.0.11.7 (Andrews, 2010). In
155 preparation for assembly with Haslr and Wengan, Sickle v.1.33 with a minimum quality score
156 threshold of 30 (-q) and a minimum length of 50 bp (-l) was employed for reads trimming.
157 Nanoplot v.1.21.0 was used to quantify and assess the quality of the Oxford Nanopore
158 PromethION long-read sequences. To detect potential contamination, the long reads were
159 aligned against the provided bacterial, human and viral databases with Centrifuge v.1.0.4-beta
160 (p+h+v; min-hit length was increased to 50 bp) (Kim et al., 2016). Reads aligning to the target
161 loci in the databases were removed. The quality of the transcriptomic short reads was assessed
162 with FASTQC v.0.11.7 (Andrews, 2010). The reads were trimmed with Sickle v1.33 (Joshi NA,
163 2011) with a minimum quality score threshold of 30 (-q) and a minimum length of 40 bp (-l).

164 **Genome size estimation**

165 A single lane of Illumina short-read genomic data from accession 5075 was employed to
166 estimate the genome size. The k-mer distribution was calculated using Jellyfish v2.2.6 (Marçais
167 & Kingsford, 2011) and size estimates were processed with GenomeScope v2.0 (Ranallo-
168 Benavidez et al., 2020; Fig. S1).

169 **Transcriptome assembly**

170 The transcriptome was independently assembled to provide protein-level evidence for the
171 structural annotation of the genome, using Trinity v2.6.6 (Grabherr et al., 2011), with a minimum
172 contig length of 300bp. Contigs with minimal read support, post assembly, were removed
173 (FPKM > 0.5) with RSEM v1.3.0 (B. Li & Dewey, 2011). Transdecoder v3.0.1 (Haas et al.,
174 2016) was used to translate the remaining contigs into Open Reading Frames (ORFs) and remove
175 sequences without a viable frame. To aid Transdecoder in the identification of ORFs, searches
176 against the Pfam database were performed with HMMER v3.1b2 (Z. Zhang & Wood, 2003).
177 Transdecoder annotated the putative transcripts as complete, partial, and internal. Those without
178 a defined start and stop codon (defined as internals) were removed (split_frames.py). The final
179 set of peptide sequences were functionally annotated with EnTAP v0.8.0 (Hart et al., 2020)
180 against NCBI's nr protein database and the UniProt/Swiss-prot reference databases. EnTAP was
181 run with contaminant filters that included bacteria, archaea, fungi, and insecta. Transcripts with
182 high confidence alignments to these organisms were removed (contam_removal.py).

183 **Genome assembly**

184 Hybrid genome assembly, integrating both long and short-read data, was conducted with
185 MaSuRCA v.4.0.3 (Zimin et al., 2013), Wengan v.0.2 (Di Genova et al., 2021), and Haslr
186 v.0.8a1 (Haghshenas et al., 2020). Additionally, a separate assembly using only long-reads as
187 input was conducted with Flye v.2.5 (with three polishing iterations) (Kolmogorov et al., 2019).

188 The Flye, Wengan, and Haslr assemblies were polished following long-read alignment
189 with Medaka v1.3.2 (github.com/nanoporetech/medaka). Post assembly with Wengan, the
190 assembly was filtered to remove scaffolds less than 3 Kb. To further improve the accuracy of the
191 assemblies, the hybrid assembly generated by MaSuRCA was polished with the Illumina short
192 reads using Pilon v.1.24 (Walker et al., 2014). Subsequently, the selected MaSuRCA assembly
193 was processed with Purge Haplotigs v.1.0 (Roach et al., 2018).

194 The Purge Haplotigs pipeline categorized the assembled scaffolds into four coverage
195 levels based on the distribution of mapped reads. This categorization enabled the identification
196 and removal of redundant sequences exhibiting low coverage, presumed to represent erroneous
197 duplicates given the haploid genome. Specific cutoff values of 0, 7, and 65 for the coverage
198 levels were selected to delineate scaffolds to be retained or discarded, based on the k-mer
199 coverage distribution peaks (Fig. S2). The term "allele" is used to describe these redundant

200 sequences for convenience, although technically inaccurate for a haploid genome. The coverage
201 analysis and purging allowed isolation of the primary genome sequence from duplications and
202 artifacts generated during assembly.

203 To evaluate the quality of the assemblies, QUAST v5.2.0 (Gurevich et al., 2013) and
204 BUSCO v4.1.2 (viridiplantae_odb10) (Manni et al., 2021) were employed. Each assembly was
205 also evaluated with Merqury v1.3 (Rhie et al., 2020).

206

207 **Structural and functional genome annotation**

208 *Repeat library construction and masking*

209 The repeat library for the final MaSuRCA assembly was generated using RepeatModeler v.2.01
210 (Flynn et al., 2020) with the long terminal repeat (LTR) discovery pipeline enabled. The genome
211 was then soft-masked with RepeatMasker v.4.0.9-p2 using the consensus repeat library from
212 RepeatModeler (Smit, AFA, Hubley, R & Green, P., 2013-2015).

213 *Structural and functional genome annotation*

214 The Illumina RNA reads were aligned to the soft-masked MaSuRCA assembly with HISAT2
215 v.2.1.0 (Kim et al., 2019) to provide evidence for protein-coding gene prediction. Two gene
216 prediction analyses were run on the soft-masked assembly using BRAKER v.2.1.5 (Brůna et al.,
217 2021), one with RNA-Seq alignment evidence and one with protein evidence originating from *de*
218 *novo* assembled transcriptome. Gene predictions from both BRAKER runs were integrated with
219 TSEBRA v.1.0.3 (Gabriel et al., 2021). From this point, separate assessments were conducted on
220 the RNA-Seq evidence gene predictions (BRAKER) and the final TSEBRA gene predictions to
221 select the best approach. Putative genes were removed from both sets if they did not contain a
222 complete protein domain. This filter was applied with Interproscan v.5.35-74.0 (Jones et al.,
223 2014) using the Pfam database v32.0 (Finn et al., 2014). It is worth noting that mono-exonic
224 genes can be the result of fragmented annotations and the target metric of 0.2 (mono:multi-exon
225 gene ratio) is often achieved through protein domain filters (Vuruputoor et al., 2022). Metrics for
226 the gene predictions were generated with gFACs v1.1.3 (Caballero & Wegrzyn, 2019) and
227 BUSCO. After assessment, the filtered BRAKER gene predictions were selected for functional
228 annotation with EnTAP v.0.10.8 (Hart et al., 2020). Functional annotation reports from EnTAP

229 (both sequence similarity search and EggNog taxonomy scope classifications) allowed for the
230 identification of non-target species scaffolds in the assembly (Huerta-Cepas et al., 2019).

231 *Assembly level contaminant filtering*

232 Using the functional annotation results from EnTAP, contaminated scaffolds were removed.
233 Scaffolds with a length of 10 Kb or less, and with 40% or more of their total genes classified as
234 archaea, bacteria, or fungi, were removed. Additionally, scaffolds greater than or equal to 10 Kb
235 with 55% or more genes classified as archaea, bacteria, or fungi, were also excluded. The final
236 annotation was then assessed for the annotation rate using EnTAP, the mono:multi ratio using
237 gFACs, and BUSCO completeness.

238 *Horizontal gene transfer candidate identification*

239 To identify candidate HGTs in *Physcomitrellopsis africana*, protein sequence similarity searches
240 were conducted with Diamond v2.1.8 (Buchfink et al., 2021). The protein sequences of
241 *Physcomitrellopsis africana* were aligned against "donor" databases, which included sequences
242 from bacteria, fungi, archaea, and metazoa from NCBI's nr database. Additionally, the same
243 proteins were aligned to "recipient" databases containing sequences from Streptophyta,
244 Tracheophyta, Embryophyta, Viridiplantae, and Spermatophyta (Ma et al., 2022). Although these
245 categories are not fully exclusive, each database was utilized separately to systematically assess
246 presence across plants at different evolutionary divergence points.

247 To identify candidate genes representing putative horizontal gene transfer (HGT) events
248 unique to *Physcomitrellopsis africana*, the following criteria were utilized: genes were required
249 to have between one and four significant sequence alignments (E-value <1e-5) to microbial
250 donor databases, while exhibiting no significant sequence similarity to plant recipient databases.
251 This range of one to four microbial alignments was selected to capture potential HGTs while
252 avoiding ubiquitous domains shared across many microbes. The lack of hits to plant databases
253 was intended to enrich for *Physcomitrellopsis africana*-specific sequences, rather than those
254 conserved across plants through vertical inheritance. At this stage, any scaffolds containing only
255 bacterial or fungal genes, without any plant-related genes, were removed from the assembly.
256

257 To assess the validity of the two proposed HGT candidates, the genes were visualized on the

258 Integrated Genome Viewer (IGV). The alignments of the HGT candidates from the 'nr' database,
259 as well as the transcripts assembled with StringTie2 (Kovaka et al., 2019), were visualized.

260 *Analyzing HGT candidates from Physcomitrium patens*

261 The 264 putative horizontally transferred genes (HGTs) previously identified in *Physcomitrium*
262 *patens* (J. Zhang et al., 2020) were independently searched against the *Physcomitrellopsis*
263 *africana* and *F. hygrometrica* protein sets using DIAMOND v2.1.8 (Buchfink et al. 2021).
264 DIAMOND searches were conducted with an E-value cutoff of 1e-5 and max target sequences
265 set to 1. Hits against *Physcomitrellopsis africana* and *F. hygrometrica* were collected and
266 merged to generate a summary table with *Physcomitrium patens* HGTs and the respective top
267 hits in each species.
268

269 **Comparative genome analyses**

270 A comparative analysis of the protein-coding gene space was conducted with OrthoFinder
271 v.2.5.1 (Emms & Kelly, 2019) with *F. hygrometrica* (Kirbis et al., 2022) and *Physcomitrium*
272 *patens* v3 (Lang et al., 2018). To provide a preliminary estimate of gene family size dynamics,
273 gene counts from each species in the assembled orthogroups were categorized as neutral,
274 expanded, or contracted. The first and third quartiles were calculated for the distinct gene counts
275 within a gene family for each species. If the number of genes from a species was lower than the
276 first quartile or higher than the third quartile, then the gene family was categorized as
277 "contracted" or "expanded", respectively. If the number of genes did not fit with either of these
278 two criteria, then the gene family was considered "neutral". The longest gene for each
279 orthogroup was used to assign functional attributes to all genes in the group from the original
280 EnTAP annotation. If the longest gene did not originate from *Physcomitrellopsis africana*, then
281 the functional annotation was derived from either *Physcomitrium patens* or *F. hygrometrica*.

282 GOSeq (Young et al., 2010) enrichment analysis was performed in R v4.2.0. GO terms
283 were extracted for each gene from the EnTAP run. Enrichment analysis was investigated
284 separately for *Biological Process* and *Molecular Function* GO categories. Paralogs of LHC,
285 STN7, and STN8 were identified with Diamond v2.8.1 (E-value <1e-5).

286 *Whole genome duplication analysis*

287 Chromosome-scale genomes of *F. hygrometrica* and *Physcomitrium patens* were assessed with
288 the reference genome generated for *Physcomitrellopsis africana,* with wgd (v.1.0.1) to
289 characterize whole genome duplication events (Zwaenepoel & Van de Peer, 2018). Each species
290 was compared against itself. Nucleotide sequences (CDS) were used as input to Blast & Markov
291 clustering (MCL) (Altschul et al., 1997; van Dongen, 2000). A Ks distribution was constructed
292 using the 'ksd' subcommand and 'mcl' output (Katoh & Toh, 2008; Price et al., 2010; Yang,
293 2007). Next, a collinearity analysis was performed with 'syn' using the structural genome
294 annotations (Proost et al., 2011). Gaussian mixture models, via 'mix', were used to generate Ks
295 distributions to aid in component interpretation. Model fit was evaluated with the Bayesian and
296 Akaike information criterion (BIC/AIC).

297

298 **Results and Discussion**

299 **Sequencing and quality control of genomic reads**

300 The single ONT PromethION run generated 16 million reads (N50: 7,518 bp; 127X coverage:
301 404 Mb; 116X coverage: 440 Mb; Table S1). Centrifuge filtering reduced this set to
302 approximately 14M reads (N50: 4,561 bp; 80X coverage: 404 Mb; 73X coverage: 440 Mb). The
303 primary contaminants included bacteria from *Xanthomonadaceae* (10.68%) followed by
304 *Bradyrhizobiaceae* (0.7%). The BGISEQ-500 short-read genomic libraries (100 bp PE)
305 generated 529 M reads. Following trimming with Sickle, 360 M reads remained (178X coverage:
306 404 Mb; 162X coverage: 440 Mb; Table S2). Coverage estimates are provided for both the
307 original estimate (404 Mb) and the final assembled genome size (440 Mb).

308 **Transcriptome assembly**

309 A total of 36.58M Illumina short reads were generated. Following quality filtering using
310 Trimmomatic, the dataset was reduced to 31.84M reads. The *de novo* assembly process with
311 Trinity yielded 143,150 contigs. After expression filtering via RSEM, the number of contigs was
312 reduced to 123,373. Identifying open reading frames (ORFs) in the contigs resulted in 110,852
313 successfully translated transcripts. The average sequence length of these translated ORFs was
314 803 bp, with an N50 value of 1,038 bp. To enhance the quality of the transcriptome assembly,

315 internal sequences and putative contaminants were removed, resulting in 74,997 total transcripts.
316 After removing internal sequences and contamination, the total unique sequences with an
317 alignment is 51,982 (69.3%; File S1). The final BUSCO score of the remaining transcripts was
318 C:82%[D:35.7%]. A total of 30,657 (36.6%) transcripts aligned to *Physcomitrium patens*
319 proteins. An assessment of contamination revealed 8,720 (10.41%) of sequences potentially
320 contaminated, with contributions from various sources such as amoeba (0.40%), bacteria (3.8%),
321 fungi (94.99%), and insecta (0.77%). A total of 15,500 (18.5%) sequences remained
322 unannotated.

323 **Genome assembly**

324 *Initial assembly:* Multiple genome assembly approaches were employed to generate
325 comprehensive draft assemblies of the *Physcomitrellopsis africana* genome (Fig. 2). The long-
326 read approach, Flye, assembled 538.68 Mb in 8,388 scaffolds, with an N50 of 152.58 Kb. The
327 BUSCO completeness score was C:84.0%[D:12.2%] and the Merqury QV score was 21.1.
328 Among the hybrid approaches, Haslr produced a 295.98 Mb reference distributed across 12,738
329 scaffolds, with an N50 of 52.29 Kb. The BUSCO score was C:94.1%[D:13.2%], and the QV
330 score was 11.9. Wengan assembled a total of 466.64 Mb across 10,516 scaffolds, with an N50 of
331 119.42 Kb and a BUSCO score of C:96.4%[D:18.8%], and a QV score of 27.4. Finally,
332 MaSuRCA assembled 506.22 Mb across 3,571 scaffolds, with an N50 of 381.33 Kb, (BUSCO:
333 C:96.4%[15.5%], and a QV score of 31.0).

334 *Polishing and improving genome assembly:*

335 The Medaka polished Flye assembly had a genome size of 530.57 Mb in 8,386 scaffolds, and the
336 N50 was decreased to 145.17 Kb. The BUSCO dropped to C:90:3%[D:11.5%]. The QV score
337 decreased to 19.6. Polishing of the Haslr assembly resulted in a genome size of 295.73 Mb,
338 across 12,737 scaffolds. The N50 remained almost unchanged at 52.31 Kb, and the BUSCO
339 score reduced to C:90.6%[D:9.4%]. The QV score more than doubled to 25.2. Filtering for
340 scaffolds less than 3 Kb, followed by Medaka, resulted in a smaller genome size for the Wengan
341 assembly at 434.41 Mb across 7,639 scaffolds. The N50 increased to 132.40 Kb. The BUSCO
342 score was reduced to C:93.2%[D:11.8%], and the QV increased slightly to 27.5.

343 Polishing with Pilon had very minimal little influence on the completeness, as expected,
344 and substantial impact on the accuracy of the MaSuRCA assembly. The polished MaSuRCA

345 hybrid assembly size increased slightly to 507.10 Mb across 3,590 scaffolds, with a slightly
346 decreased N50 of 378.43 Kb. The BUSCO completeness remained the same at
347 C:96.4%[D:15.5%], and the QV score increased to 35.6.

348 *Refinement of genome assemblies with Purge-haplotigs*: The MaSuRCA assembly was selected
349 for further refinement. This decision was based on the overall quality assessed by Merqury,
350 BUSCO completeness, and overall contiguity. The MaSuRCA assembly was refined with Purge-
351 haplotigs which reduced the assembly length to 502.34 Mb across 3,237 scaffolds with an N50
352 value of 382.25 Kb. The BUSCO completeness score remained the same (e.g., 96.4%) and the
353 QV score was minimally reduced to 35.5. At 502.34 Mb, the assembled genome is ~100 Mb
354 longer than the k-mer based estimate (440 Mb, Fig S1).

355

356 **Genome annotation**

357 *Repeat identification:* RepeatModeler produced a library containing 580 unique repeats that were
358 used to softmask 50.22% of the final assembly (Table S3; File S2). Of the repeats, 36.53% was
359 composed of *Ty3/Gypsy* and 1.46% of *Ty1/Copia*. This is similar to the pattern in *Physcomitrium*
360 *patens*, wherein approximately 57% of the genome is composed of repeat elements, with long
361 terminal repeats (LTRs), particularly the *Gypsy* family, accounting for 48% of the masked
362 genome (Lang et al., 2018). These findings align with observations by Kirbis et al. (2022),
363 suggesting a common pattern across these species regarding the activation of *Gypsy* elements in
364 *Physcomitrellopsis africana* and *Physcomitrium patens*. In contrast, *F. hygrometrica*, which
365 diverged from *Physcomitrium patens* 60 to 80 MYA (Bechteler, Peñaloza-Bojacá, Bell,
366 Burleigh, McDaniel, et al., 2023; Medina et al., 2018), exhibits a lower overall repeat estimate of
367 35%. Here, *Gypsy* elements contribute less to the LTR content, i.e., roughly 10%, whereas *Copia*
368 elements contribute 17%.

369 *Protein-coding gene identification:* The *Physcomitrellopsis africana* RNA-Seq reads had an
370 overall alignment rate of 78.43%, likely due to contaminant content, but a substantial set of 37M
371 reads were retained, exceeding the minimum needed for prediction. The first BRAKER2
372 predictions, using RNA-Seq evidence alone, generated 60,917 protein-coding genes with a
373 BUSCO completeness score of C:95.8%[D:15.3]. The mono:multi exonic ratio was 0.52. With
374 only protein evidence, BRAKER2 generated 37,752 gene predictions with a BUSCO score of

375 C:46.1%[D:16.0]. Merging these predictions, as recommended by TSEBRA, resulted in a set of
376 45,737 genes with a BUSCO score of C:91.3%[D:15.1], and the mono:multi exonic ratio was
377 1.06. The gene prediction set generated with RNA-Seq alignment evidence exclusively was
378 selected for further refinement due to its higher BUSCO score and lower mono:multi ratio
379 compared to the merged TSEBRA transcripts. Through protein domain filtering (InterProScan
380 filter), the number of mono-exonic genes was further reduced from 20,081 to 12,696, producing
381 a total of 23,561 genes (BUSCO: C:93.8%[D:13.8%]; mono:multi ratio: 0.08).

382 A total of 831 scaffolds, a total of 1250 genes associated with contaminants were
383 removed, based on the functional annotations of the gene space. This also reduced the assembly
384 length to 440 Mb in 2,406 scaffolds with an N50 of 363 Kb, and an assembly BUSCO score of
385 C:96%[D:13.9%]. This substantial reduction led to an assembly within ~40 Mb of the original k-
386 mer-based estimate (e.g., 404 Mb).

387 The annotated protein-coding gene space in the contaminant-filtered assembly included
388 1,708 mono-exonic and 21,853 multi-exonic genes. The annotated gene space has a BUSCO
389 score of C:93.8%[D:13.6%], with a mono:multi ratio of 0.08. Reciprocal BLAST conducted by
390 EnTAP produced an annotation rate of 78%, of which 93.6% aligned to *Physcomitrium patens*
391 (Table 1; File S3). This contrasts with the annotation rate of 50% in *F. hygrometrica* (Kirbis et
392 al., 2022), likely attributed to significant divergence time from the model *Physcomitrium patens,*
393 and the lack of closely related species in public genomic databases (Kirbis et al., 2020;
394 Rahmatpour et al., 2021). The higher annotation rate in this study suggests that many genes
395 found in *Physcomitrium patens* but not *F. hygrometrica* may have been acquired in the ancestor
396 of the *Physcomitrellopsis-Entosthodon-Physcomitrium* clade (Medina et al., 2019).

397

398 **Comparative genome analysis**

399 In comparison of the three genomes of *Physcomitrellopsis africana*, *F. hygrometrica*, and
400 *Physcomitrium patens*, a total of 14.2K orthogroups are shared among the species, and 959, 936,
401 and 179 are, respectively, unique to each species. *F. hygrometrica* and *Physcomitrellopsis*
402 *africana* exclusively share more (1092) orthogroups than *Physcomitrellopsis africana* and
403 *Physcomitrium patens* (809) (Fig. 3A), despite the latter sharing a more recent unique common
404 ancestor (Fig. 1). Thus, whereas Funariaceae share a rather conserved architecture of their

405 vegetative body even after at least 60 MY of divergence, their gene space varies considerably,
406 reflecting significant innovation (Kirbis et al. 2022) perhaps driven by ecophysiological
407 adaptations (Glime, 1990). Of the putative gene families from the *Physcomitrellopsis africana*
408 annotation, 809 and 1,694, respectively, were categorized as expanded and contracted (File S4).
409 Enrichment analysis examined through Gene Ontology's *Biological Process* category revealed
410 167 expanded terms and 6 contracted terms. In contrast, within the *Molecular Function* GO
411 category, 62 were contracted, and no terms were significantly expanded.

412 Although expanded *Biological Process* GO terms were excessively broad, orthogroup
413 analysis revealed a pattern of contraction in GO terms pertaining to oxidoreductase activity,
414 FMN binding, and serine protease activity (Fig. 3B). Whether this unique suite of downregulated
415 categories diagnoses photosynthetic properties of *Physcomitrella africana* only or of the
416 expanded genus sensu Wilding (2015) remains to be tested.

417 The complement of light-harvesting complex (LHC) genes are expanded in the moss
418 *Physcomitrium patens* compared to algae and vascular plants (Alboresi et al., 2008; Iwai et al.,
419 2018). LHC proteins bind chlorophylls and carotenoids to facilitate light absorption and energy
420 transfer to the reaction centers of Photosystems I and II. The LHC genes are classified into two
421 groups: Lhca encodes antenna proteins for PSI (LHCI) while Lhcb encodes antenna proteins for
422 PSII (LHCII). Although ancestral land plants contain several LHC homologs, further expansion
423 and redundancy occurred in *Physcomitrium patens* after whole genome duplication events
424 (Alboresi et al., 2008; Rensing et al., 2008; H. Sun et al., 2023; Zimmer et al., 2013). This led to
425 a larger repertoire of LHC genes compared to the alga *Chlamydomonas reinhardtii* and the
426 vascular plant *Arabidopsis thaliana*. Specific Lhca and Lhcb paralogs are represented in multiple
427 copies in *Physcomitrium patens* compared to one to two copies in the other species. The major
428 antenna proteins encoded by Lhcbm also show greater redundancy and diversity in
429 *Physcomitrium patens* (Iwai et al., 2018; H. Sun et al., 2023).

430 Comparing the genomes of *F. hygrometrica* and *Physcomitrellopsis africana* to the
431 reference genome of *Physcomitrium patens* reveals both conservation and divergence of LHC
432 genes. For example, while *Physcomitrium patens* has 12 distinct Lhca compared to 8 distinct
433 Lhca in *Physcomitrellopsis africana*, and 7 distinct Lhca in *F. hygrometrica*. Similarly, while
434 *Physcomitrium patens* has 12 distinct Lhcb, *Physcomitrellopsis africana* has 8, and *F.*

435 *hygrometrica* has 9. Finally, *Physcomitrium patens* has 14 distinct Lhcbm, *Physcomitrellopsis*
436 *africana* has 2, and *F. hygrometrica* has 8. Two rounds of whole genome duplication (WGD)
437 occurred in the most recent common ancestor of the Funariaceae, as evidenced by shared
438 duplication signatures across all three genomes (Fig. S3). However, echoing Kirbis et al. (2022),
439 more duplicates from these WGDs were retained in *Physcomitrium patens* than in *F.*
440 *hygrometrica* and *Physcomitrellopsis africana*. Whereas *Physcomitrium patens* retained multiple
441 LHC paralogs as a result of the ancestral WGDs, *F. hygrometrica* and *Physcomitrellopsis*
442 *africana* seem to have lost some of this redundancy. This pattern of differential retention is
443 further supported by assessing the number of paralogs and orthologs for each LHC gene family
444 across the three genomes (Table S4). Although some copies of LHC genes were lost in *F.*
445 *hygrometrica* and *Physcomitrellopsis africana*, key STN7 and STN8 kinases involved in
446 photosynthetic acclimation are conserved and retained in all three genomes, suggesting retention
447 of core light signaling components.

448 *Contaminant filtering and the identification of horizontal gene transfer events:*

449 Horizontal gene transfer (HGT) involves the exchange of genetic material between different
450 organisms, extending beyond prokaryotes, with instances observed within eukaryotes as well
451 (Guo et al., 2023; Kirsch et al., 2022; Xi et al., 2012). While lateral transfer can occur between
452 related plant lineages through hybridization, HGT in mosses involves exchange between
453 evolutionarily distant microbes and land plants. Mosses, which originated 500 million years ago
454 (Bechteler, Peñaloza-Bojacá, Bell, Burleigh, Mcdaniel, et al., 2023), have formed symbiotic
455 relationships with microbes, aiding in nutrient acquisition and stress resistance (Berg et al., 2016;
456 Hornschuh et al., 2006). In addition, genes of putative fungal origin regulate developmental
457 transitions in mosses (Wang et al., 2020).

458 Several reports of HGT in eukaryotes were later reclassified as contamination. The
459 assembly of *Physcomitrellopsis africana* employed filtering processes before and after genome
460 assembly, illustrating that comprehensive approaches at all stages significantly enhanced the
461 final genome assembly. Here, metagenomic tools optimized for long-reads identified
462 contaminants prior to assembly. Since this initial filter did not include the short reads or assess
463 fungal contributions in either sequence set, further filtering was conducted after assembly and
464 annotation. The protein-coding genes were functionally characterized through sequence

465 similarity. An estimated 22% of the genes were of fungal origin and 31% were of bacterial
466 origin. Scaffolds that contained a majority of genes likely originating from algae, bacteria, or
467 fungi (ABF) were removed. Separately, the annotated gene space was aligned to a set of "donor"
468 and "recipient" databases. This identified 31 potential horizontal gene transfer (HGT) events.
469 Following the set of best practices outlined by François et al. (2020), an additional set of 10
470 scaffolds were found to be contaminated as the flanking genes of the potential HGT candidates
471 were not of plant origin. This resulted in their removal (HGT filter). Two unique HGT
472 candidates survived these filters (Fig. 4).

473 The first HGT candidate, *Pa1_22336.1* (alignment to *Paracoccaceae bacterium*),
474 contains a methyltransferase domain. A recent phylogenetic analysis conducted among
475 bryophytes revealed that DNA methyltransferases identified in *Marchantia polymorpha*,
476 *Physcomitrium patens*, and *Anthoceros angustus* form clades with their bacterial homologs. This
477 suggests that these genes were likely acquired via horizontal gene transfer (HGT) from bacteria
478 (J. Zhang et al., 2020). The second HGT candidate, *Pa1_04828* (alignment to *Pedobacter sp.*
479 SYSU D00873), encodes a glycosyl hydrolase, functioning as an antibacterial defense. This gene
480 family is widely distributed across various taxonomic groups as a result of independent transfers
481 from bacteria to plants, fungi, animals, and archaea. In the hornwort *Anthoceros angustus*, the
482 HGT-derived glycosyl hydrolases are thought to enhance the metabolic adaptability in response
483 to changing environments, particularly in cell wall synthesis and modification (Haimlich et al.,
484 2022; J. Zhang et al., 2020).

485 These two candidates were manually assessed comparing alignments of *de novo*
486 assembled transcriptomes. *Pa1_22336.1* had transcriptomic support, tentatively increasing
487 confidence in it being a high-confidence HGT event. *Pa1_04828* lacked this transcriptomic
488 support. Both candidates are directly flanked by well annotated plant (moss) genes. Additionally,
489 aligning the full set of *Physcomitrium patens* and *Ceratodon purpureus* proteins to the
490 *Physcomitrellopsis africana* scaffolds produced no alignments in proximity to the candidates,
491 further suggesting that the HGT candidates may be specific to *Physcomitrellopsis africana* (Fig.
492 5).

493 *Analyzing HGT candidates from Physcomitrium patens:* The genomes of *Physcomitrellopsis*
494 *africana* and *F. hygrometrica* were screened for the 264 putative horizontally transferred genes

495 (HGTs) previously identified in *Physcomitrium patens* (Ma et al., 2022). Ninety of these genes
496 (34%) were not found in either species,16 (6%) were found in *Physcomitrellopsis africana* and
497 *Physcomitrium patens*, and 7 (3%) were shared between *Physcomitrium patens* and *Funaria*
498 *hygrometrica*. The greater number of shared HGTs between *Physcomitrium patens* and
499 *Physcomitrellopsis africana* likely reflect their more recent divergence (at least 20 MYA)
500 compared to that of *Physcomitrium patens* and *Funaria hygrometrica* (at least 60 MYA
501 (Bechteler, Peñaloza-Bojacá, Bell, Burleigh, McDaniel, et al., 2023; Medina et al., 2018) (Fig.
502 S4).

503

504 **Data availability**

505 All scripts and data used are available through
506 https://gitlab.com/PlantGenomicsLab/physcomitrellopsis-africana-genome. Illumina short and
507 Nanopore long genomic reads, RNA-Seq reads, de novo assembled transcripts, whole genome
508 shotgun assembly, and annotation files will be uploaded to NCBI BioProject: PRJNA1020579.

509 **Acknowledgements**

510 The authors would like to thank the Institute for Systems Genomics (ISG) and Computational
511 Biology Core at the University of Connecticut for high-performance computing services.

512 **Funding**

513 This study was made possible through the US National Foundation grants DEB-0919284
514 (fieldwork), DEB-1753811 to BG, and DBI-1943371 to JW.

515 **Conflict of interest**

516 The authors declare no conflict of interest.

517 **Author contributions**

518 J.W. & B.G. designed the study. B.G., N.W. & T.H. conducted fieldwork to sample the wild
519 population. Y. C. & Y. L. generated genomic and transcriptomic data. A.S. & V.S.V. conducted
520 all analyses. V.S.V., J.W. & B.G. wrote the paper. All authors approved of the final version.

521

522

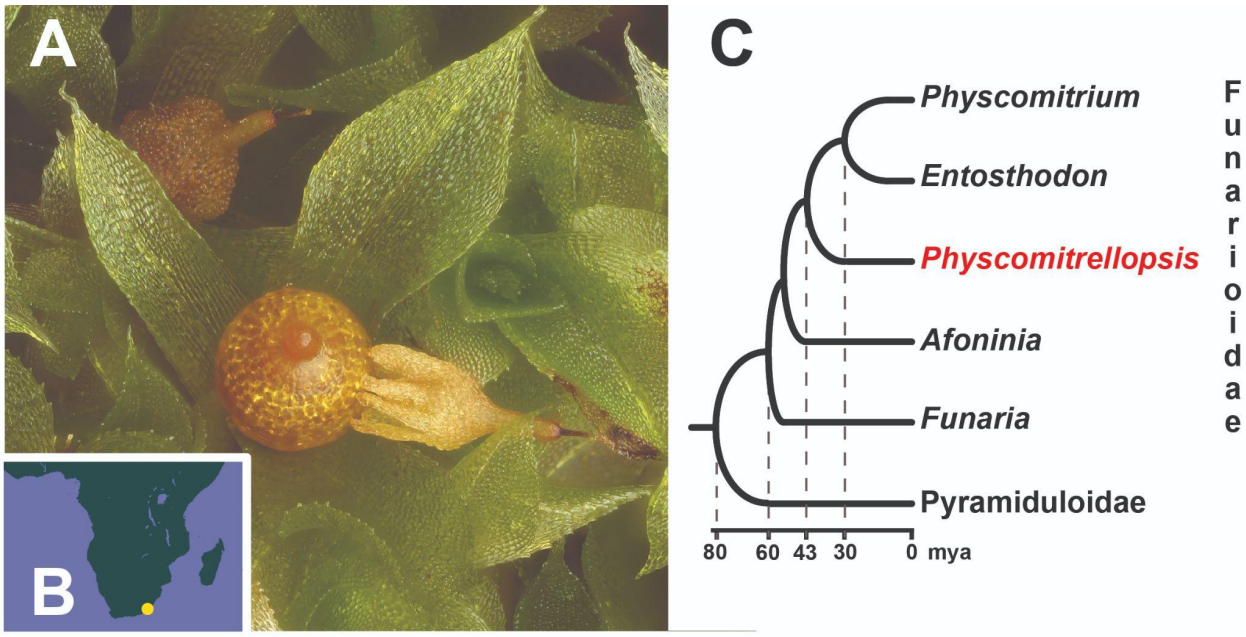


Fig. 1. A *Physcomitrellopsis africana* exhibits a reduced architectural complexity of the sporophyte similar to that observed in *Physcomitrium patens*, namely a sessile, aperistomate and cleistocarpous sporangial capsule. **B** The known geographic distribution of *Physcomitrellopsis africana*, a rare narrow endemic to the Eastern Cape Region in South Africa. **C** Phylogenetic relationships and chronology of the evolution of *Physcomitrellopsis* based on Medina et al. (2018).

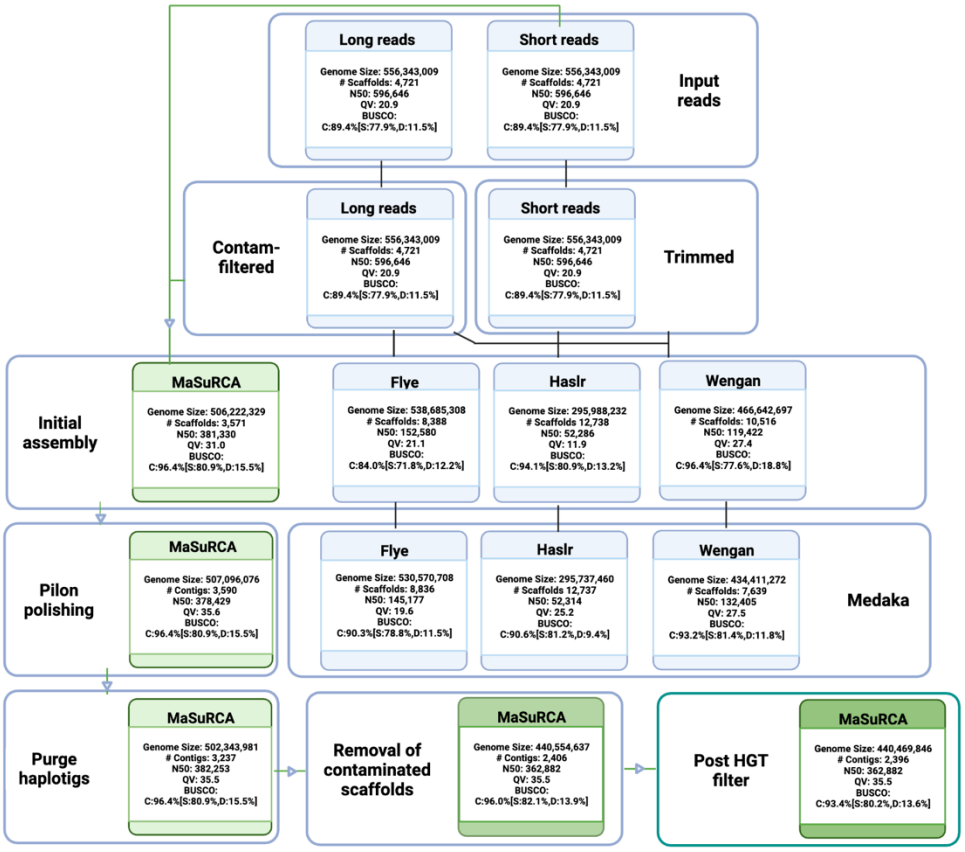


Fig. 2. Workflow and statistics for read QC, genome assembly, polishing, and haplotype phasing. Short reads generated from BGI-seq were subject to trimming and the long reads (ONT) were filtered for contaminants with Centrifuge. Filtered and trimmed reads were utilized as input for four different genome assembly software tools: Flye (long-read only), Haslr, Wengan, and MaSuRCA. The green box indicates the selection of the MaSuRCA assembly for further analysis. The final assembly was polished using Pilon and phased with Purge haplotigs. A total of 831 scaffolds were removed as a result of contaminant filtering using EnTAP following structural genome annotation. Ten additional scaffolds were removed after HGT candidate assessment. Summary statistics derived from Quast, BUSCO, and Merqury are displayed for each process.

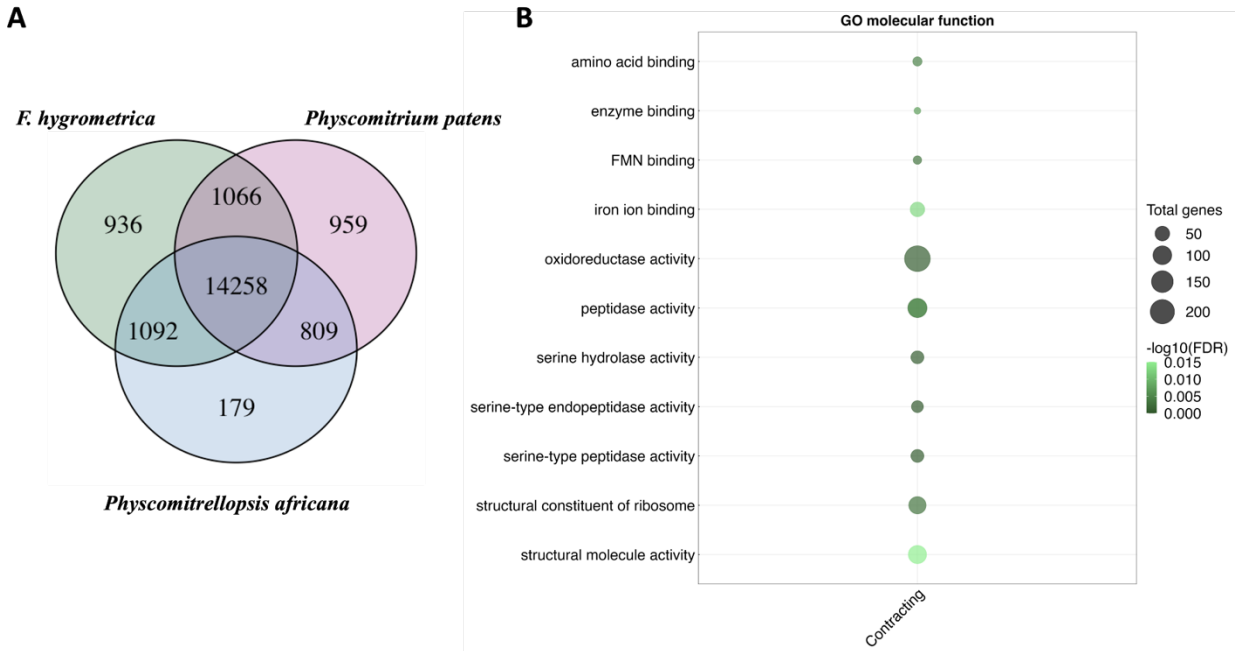


Fig. 3. A Number of shared and unique orthogroups among *F. hygrometrica*, *Physcomitrellopsis africana*, and *Physcomitrium patens*. **B** Enriched *Molecular Function* GO terms for gene families in a comparative analysis between *Physcomitrellopsis africana*, *Physcomitrium patens*, and *Funaria hygrometrica* showed contraction in the GO terms related to oxidoreductase activity, as well as serine peptidase activity and FMN binding for *Physcomitrellopsis africana*. The size of each bubble represents the number of gene families within an orthogroup, and the gradient of the color denotes the significance level of enrichment- the darker green denotes more significance.

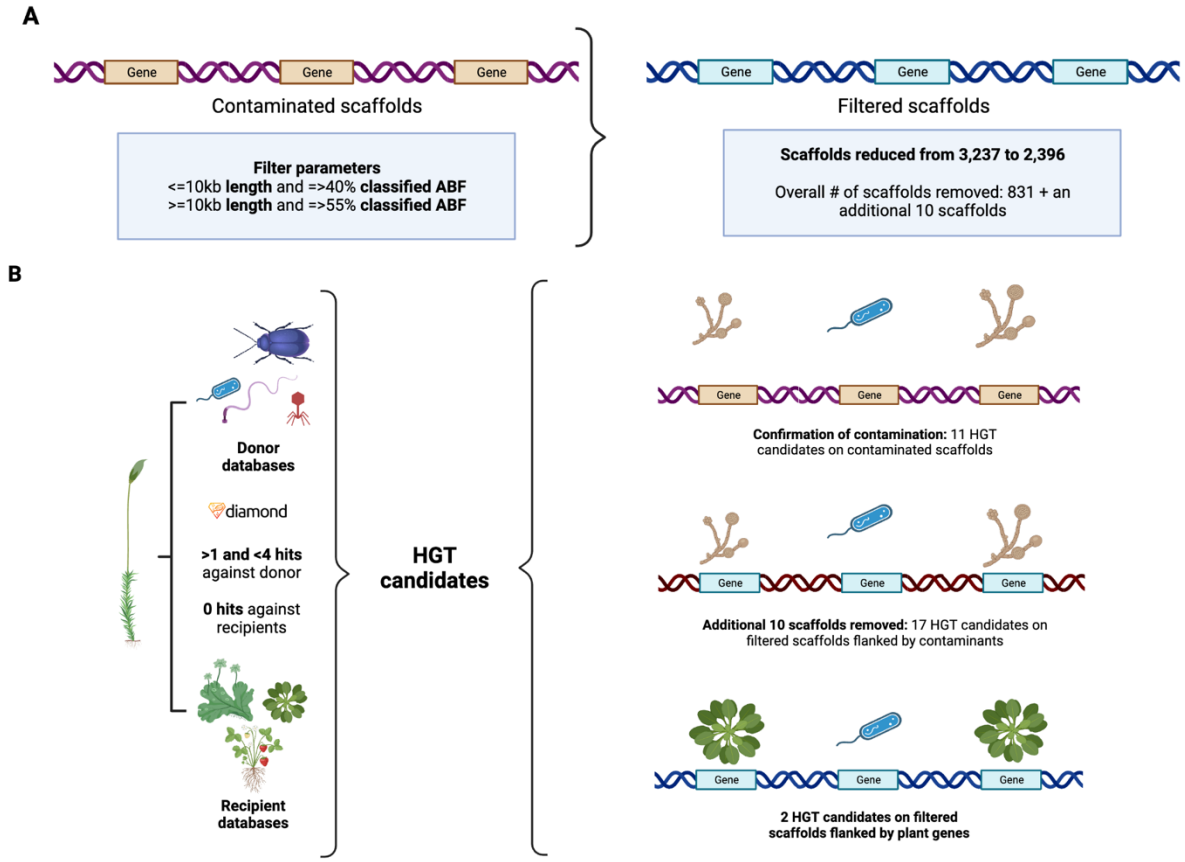


Fig. 4. Contamination versus horizontal gene transfer (HGT) in the *Physcomitrellopsis africana* genome. **A** Parameters used for removing a set of contaminated scaffolds from the draft genome based on functional characterization of the annotated gene space. Scaffolds with a length of 10 kb or less, and those with 40% or more of their total genes classified as archaea, bacteria, or fungi were removed. Additionally, scaffolds with a length greater than or equal to 10 Kb and having 55% or more genes classified as archaea, bacteria, or fungi were also excluded. In total, 831 scaffolds were removed as a result of this filtering process. **B** Identification of HGT candidates via sequence similarity comparisons of *Physcomitrellopsis africana* proteins against both "donor" databases (archaea, bacteria, fungal, and metazoan) and "recipient" databases (Streptophyta, Viridiplantae, Tracheophyta, and Spermatophyta). Proteins with >1 and <4 hits against all of the donor databases, and no hit against recipient databases were labeled as HGT candidates. This analysis was conducted on the contaminated scaffolds removed in A, confirming their contamination status. On the post-filtered scaffolds, there were some putative HGTs that were flanked by contaminants. These scaffolds were also removed from the assembly (additional 10), resulting in the retention of two HGT candidates in the final analysis.

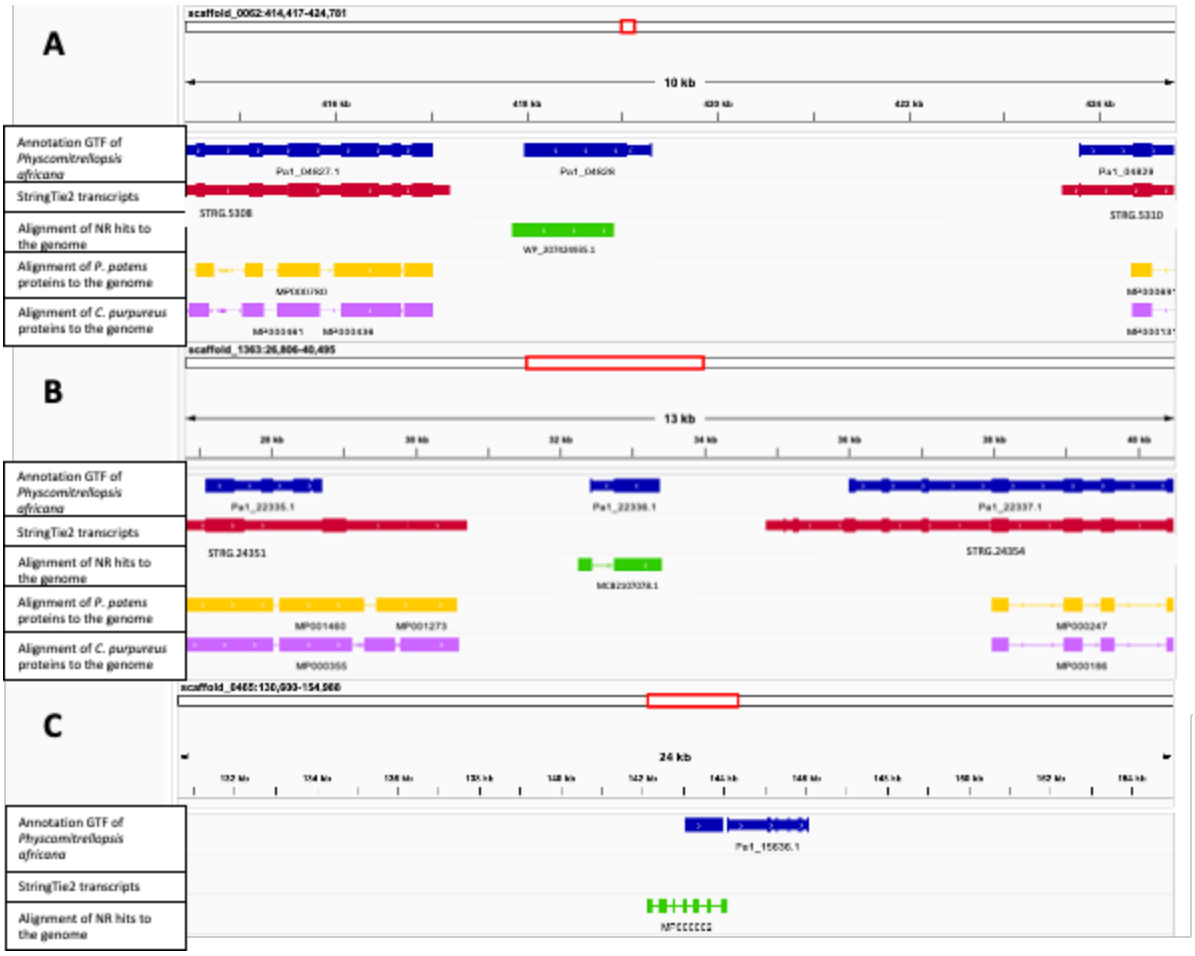


Fig. 5. Integrated Genome Viewer (IGV) screens depicting five tracks of the *Physcomitrellopsis africana* genome. Track 1 shows the protein-coding structural annotation in context to the genome. Track 2 displays genome-guided transcript assemblies via StringTie2. Track 3 illustrates the alignment of the Horizontal Gene Transfer (HGT) candidate (NR database) to the genome. Track 4 and 5 show the alignments of *Physcomitrium patens* and *Ceratodon purpureus* proteins onto the scaffolds 62 and 1363. **A and B** show the HGT candidates *Pa1_04828* and *Pa1_22336.1*, where the HGT candidate alignment validates the presence of a protein, while the StringTie2 transcript does not align with the HGT placement. None of the moss proteins align to the HGT candidates. **C** highlights the example of *Pa1_15636.1,* where the StringTie2 transcript and the MiniProt alignment reveal that the gene spans two gene models, indicating a false identification as an HGT candidate. Further transcriptomic evidence is required to modify the annotation model and establish the validity of these gene models.

572 **Table 1.** Genome annotation statistics for *Physcomitrellopsis africana*

Annotation Method	Total genes	BUSCO (viridiplantae)	Mono: multi ratio	Annotation rate
BRAKER (RNA)	60,917	C:95.8%[D:15.3]	0.52	65%
BRAKER (protein)	37,752	C:46.1%[D:16.0]	0.95	73%
TSEBRA (BRAKER RNA + BRAKER protein)	45,737	C:91.3%[D:15.1]	1.02	74%
BRAKER (RNA) + InterProScan filter + scaffold contam filter)	23,561	C:93.8%[D:13.6%]	0.08	78%
BRAKER (RNA) + InterProScan filter + scaffold contam filter + HGT filter)	23,535	C:93.8%[D:13.6%]	0.07	83%

573
574

575 **Table S1.** Pre-QC Read Stats.

Instrument	Total Reads	N50 (bp)	Average Read Length (bp)	Coverage (404 Mb)	Coverage (440 Mb)
Nanopore PromethION	16,640,338	7,518	3,080	127	116
Illumina Genomic Reads	529,496,757	N/A	100 PE	262	240

576

577 **Table S2:** Post-QC Read Stats

578

Read Type	Total Reads	N50 (bp)	Average Read Length (bp)	Coverage (404 Mb)	Coverage (440 Mb)
Nanopore PromethION	14,495,188	4,561	2,233	80	73
Illumina Genomic Reads	359,619,533	N/A	100 PE	178	163

579

580 **Table S3.** Repeat content of the MaSuRCA genome assembly.

Sequences:	3237		
Total Length:	502343981 bp		
GC Level:	36.07%		
Bases Masked:	253596074 bp	50.48%	
	Number of elements	**Length (bp)**	**Percentage of sequence**
Retroelements	182967	203774766	40.56
SINEs	116	50840	0.01
Penelope	44	43124	0.01
LINEs	5635	3528322	0.7
CRE/SLACS	25	87739	0.02
L2/CR1/Rex	0	0	0
R1/LOA/Jockey	0	0	0
R2/R4/NeSL	0	0	0
RTE/Bov-B	0	0	0
L1/CIN4	1520	1352049	0.27
LTR elements	177216	200195604	39.85
BEL/Pao	0	0	0
Ty1/Copia	7771	7313798	1.46
Gypsy/DIRS1	146546	183504589	36.53
Retroviral	0	0	0
DNA transposons	3075	2020920	0.4
hobo-Activator	161	85415	0.02
Tc1-IS630-Pogo	828	502079	0.1
En-Spm	0	0	0
MuDR-IS905	0	0	0
PiggyBac	0	0	0
Tourist/Harbinger	576	270022	0.05
Other	0	0	0
Rolling-circles	7857	3140001	0.63
Unclassified	70182	35763560	7.12
Total interspersed repeats		241559246	48.09
Small RNA	454	382891	0.08
Satellites	0	0	0
Simple repeats	177543	7162961	1.43
Low complexity	26138	1350975	0.27

581

Table S4. Comparison of light-harvesting complex (LHC) gene complements in *Physcomitrellopsis africana*, *Funaria hygrometrica*, *Physcomitrium patens*, *Chlamydomonas reinhardtii**, and *Arabidopsis thaliana** *(Alboresi et al., 2008).* The number of paralogs is shown for key LHC genes including antenna proteins (Lhca, Lhcb) and major light-harvesting proteins (Lhcbm).

Gene	*Physcomitrellopsis*	*Funaria*	*Physcomitrium*	*Chlamydomonas*	*Arabidopsis*
Lhca1	3	2	3	1	1
Lhca2	2	3	4	1	1
Lhca3	2	1	4	1	1
Lhca5	1	1	1	1	1
Lhcbm	2	8	14	9	0
Lhcb3	1	1	1	0	3
Lhcb4	2	2	2	1	4
Lhcb5	2	1	2	1	5
Lhcb6	1	2	2	0	2
Lhcb7	1	1	1	1	1
Lhcb9	1	2	2	9	2

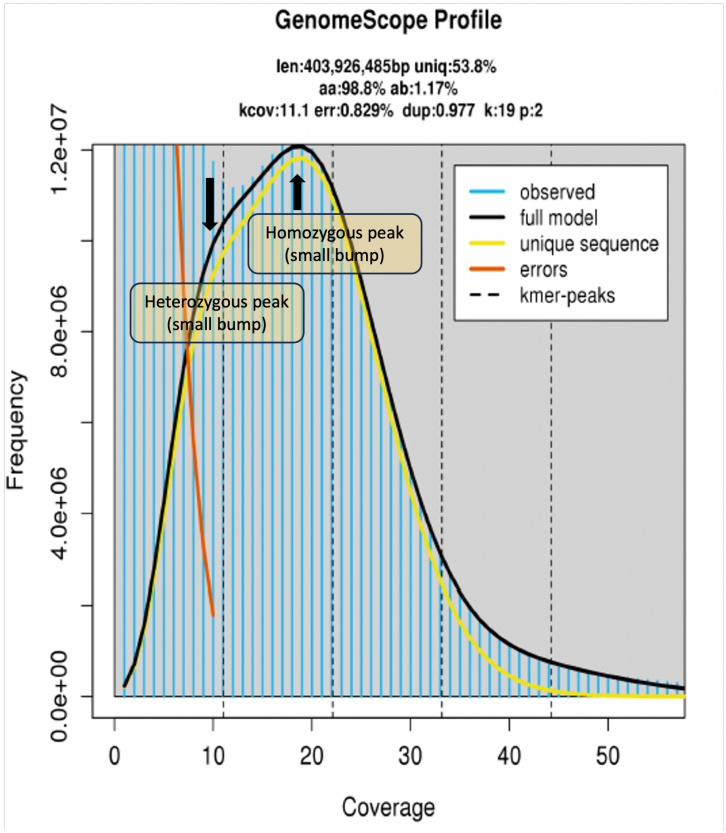


Fig S1: GenomeScope output for preliminary analysis of *Physccomitrellopsis africana*. GenomeScope generates coverage plots to estimate genomic characteristics including len (inferred total genome length), uniq (percent of unique genome), het (heterozygosity rate), kcov (kmer coverage), err (read error rate), and dup (rate of read duplications). The het value of ~1.17% represents a low-end approximation of heterozygosity based on unfiltered short reads. The peak at 9X corresponds to the estimated heterozygous portion (ab) while the peak at 19X corresponds to the estimated homozygous portion (aa). Further data quality control and analyses were performed to reduce heterozygosity, as this is a haploid genome.

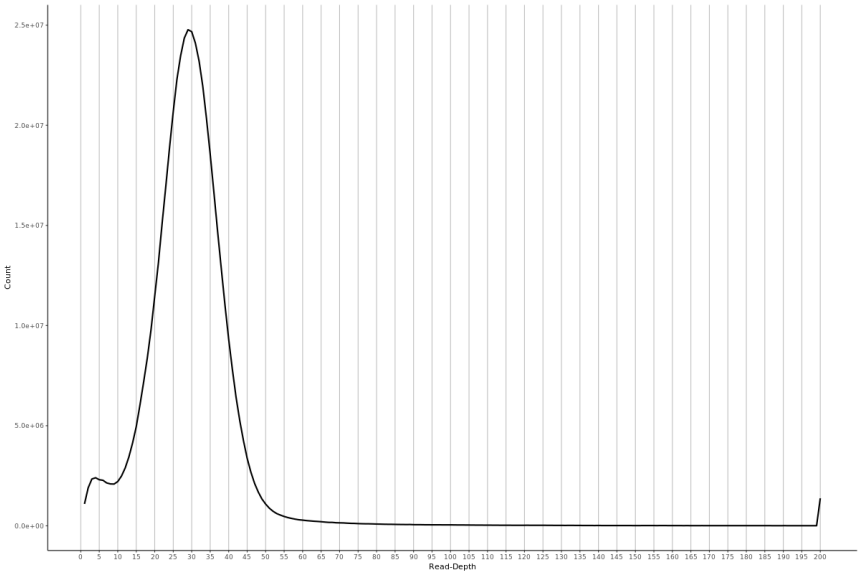


Fig S2: Histogram of k-mer coverage distribution for the *Physcomitrellopsis africana* genome assembly. The distribution shows two prominent peaks, with a smaller peak at ~7x coverage representing redundant haplotig sequences and a larger peak at ~65x representing the primary haploid genome coverage. The Purge Haplotigs pipeline used the local minima at 0x and 7x coverage as cutoffs to categorize and remove low coverage scaffolds presumed to be haplotigs or assembly artifacts. The 65x peak informed the coverage threshold for retaining the primary genome.

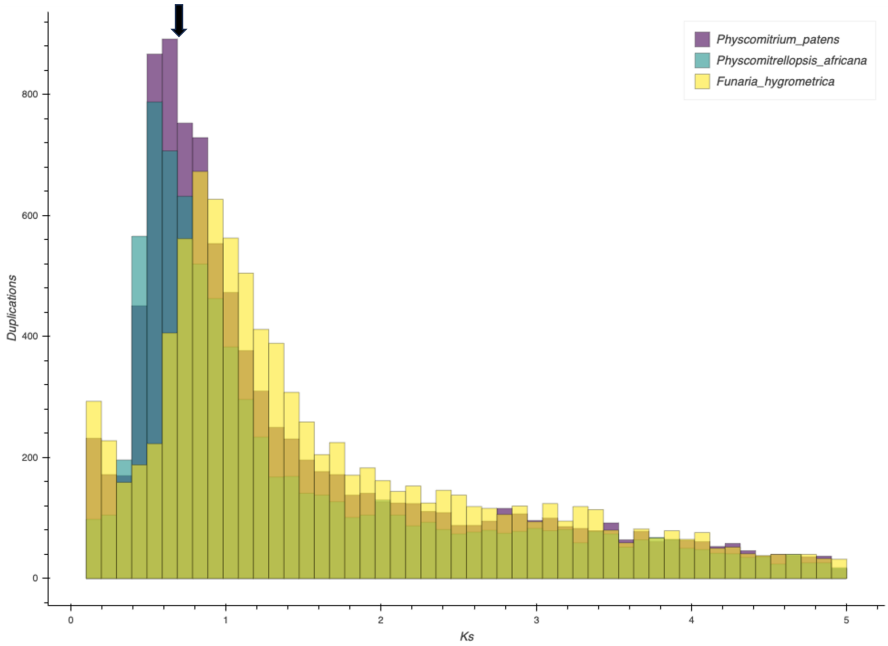


Fig S3: Ks density plots for *Physcomitrium patens*, *Funaria hygrometrica*, and *Physcomitrellopsis africana* genomes. Signatures of the two ancestral whole genome duplication (WGD) events are evident as shared peaks across all three species, as denoted by the black arrow. *Physcomitrium patens* exhibits a more prominent peak indicating greater retention of WGD duplicates compared to *Funaria hygrometrica* and *Physcomitrellopsis africana*.

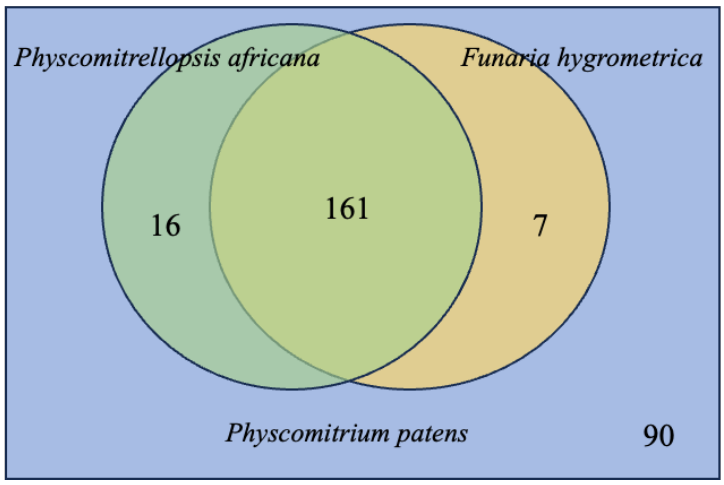


Fig. S4. Horizontally transferred genes (HGTs) identified in *Physcomitrium patens* assessed in both *Physcomitrellopsis africana* and *Funaria hygrometrica* genomes.

Literature Cited

Alboresi, A., Caffarri, S., Nogue, F., Bassi, R., & Morosinotto, T. (2008). In silico and biochemical analysis of Physcomitrella patens photosynthetic antenna: identification of subunits which evolved upon land adaptation. *PloS One*, *3*(4), e2033.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Aminov, R. I. (2011). Horizontal gene exchange in environmental microbiota. *Frontiers in Microbiology*, *2*, 158.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online. *Retrieved May*, *17*, 2018.

Bechteler, J., Peñaloza-Bojacá, G., Bell, D., Burleigh, J. G., McDaniel, S. F., Christine Davis, E., Sessa, E. B., Bippus, A., Christine Cargill, D., Chantanoarrapint, S., Draper, I., Endara, L., Forrest, L. L., Garilleti, R., Graham, S. W., Huttunen, S., Jauregui Lazo, J., Lara, F., Larraín, J., … Villarreal A, J. C. (2023). Comprehensive phylogenomic time tree of bryophytes reveals deep relationships and uncovers gene incongruences in the last 500 million years of diversification. *American Journal of Botany*. https://doi.org/10.1002/ajb2.16249

Berg, G., Rybakova, D., Grube, M., & Köberl, M. (2016). The plant microbiome explored:

implications for experimental botany. *Journal of Experimental Botany*, *67*(4), 995–1002.

Boothby, T. C., Tenlen, J. R., Smith, F. W., Wang, J. R., Patanella, K. A., Nishimura, E. O., Tintori, S. C., Li, Q., Jones, C. D., Yandell, M., Messina, D. N., Glasscock, J., & Goldstein, B. (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(52), 15976–15981.

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108.

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368.

Caballero, M., & Wegrzyn, J. (2019). gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. *Genomics, Proteomics & Bioinformatics*, *17*(3), 305–310.

Di Genova, A., Buena-Atienza, E., Ossowski, S., & Sagot, M.-F. (2021). Efficient hybrid de novo assembly of human genomes with WENGAN. *Nature Biotechnology*, *39*(4), 422–430.

Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B., & Nestor, C. E. (2020). No evidence for DNA N 6-methyladenine in mammals. *Science Advances*, *6*(12), eaay3335.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–D230.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457.

Francois, C. M., Durand, F., Figuet, E., & Galtier, N. (2020). Prevalence and Implications of Contamination in Public Genomic Resources: A Case Study of 43 Reference Arthropod Assemblies. *G3*, *10*(2), 721–730.

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*, *22*(1), 566.

Glime, J. M. (1990). The ecology column: Introduction. *The Bryological Times*, *55*, 5–7.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Guan, Y., Ma, L., Wang, Q., Zhao, J., Wang, S., Wu, J., Liu, Y., Sun, H., & Huang, J. (2023). Horizontally acquired fungal killer protein genes affect cell development in mosses. *The Plant Journal: For Cell and Molecular Biology*, *113*(4), 665–676.

Guo, X., Hu, X., Li, J., Shao, B., Wang, Y., Wang, L., Li, K., Lin, D., Wang, H., Gao, Z., Jiao, Y., Wen, Y., Ji, H., Ma, C., Ge, S., Jiang, W., & Jin, X. (2023). The Sapria himalayana genome provides new insights into the lifestyle of endoparasitic plants. *BMC Biology*, *21*(1), 134.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Haas B, Papanicolaou A (2012) TransDecoder (Find Coding Regions within Transcripts) [WWW Document]. https://transdecoder.github.io/.

Haghshenas, E., Asghari, H., Stoye, J., Chauve, C., & Hach, F. (2020). HASLR: Fast Hybrid Assembly of Long Reads. *iScience*, *23*(8), 101389.

Haimlich, S., Fridman, Y., Khandal, H., Savaldi-Goldstein, S., & Levy, A. (2022). Widespread horizontal gene transfer between plants and their microbiota. In *bioRxiv* (p. 2022.08.25.505314). https://doi.org/10.1101/2022.08.25.505314

Hart, A. J., Ginzburg, S., Xu, M. S., Fisher, C. R., Rahmatpour, N., Mitton, J. B., Paul, R., & Wegrzyn, J. L. (2020). EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Molecular Ecology Resources*, *20*(2), 591–604.

Hornschuh, M., Grotha, R., & Kutschera, U. (2006). Moss-associated methylobacteria as phytosymbionts: an experimental study. *Naturwissenschaften*, *93*(10), 480–486.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019).

693 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource
694 based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314.
695 Iwai, M., Grob, P., Iavarone, A. T., Nogales, E., & Niyogi, K. K. (2018). A unique
696 supramolecular organization of photosystem I in the moss Physcomitrella patens. *Nature*
697 *Plants*, *4*(11), 904–909.
698 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J.,
699 Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M.,
700 Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function
701 classification. *Bioinformatics*, *30*(9), 1236–1240.
702 Joshi NA, F. J. N. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for*
703 *FastQ files* (Version Version 1.33). https://github.com/najoshi/sickle
704 Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment
705 program. *Briefings in Bioinformatics*, *9*(4), 286–298.
706 Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome
707 alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*,
708 *37*(8), 907–915.
709 Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive
710 classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729.
711 Kirbis, A., Rahmatpour, N., Dong, S., Yu, J., van Gessel, N., Waller, M., Reski, R., Lang, D.,
712 Rensing, S. A., Temsch, E. M., Wegrzyn, J. L., Goffinet, B., Liu, Y., & Szövényi, P.
713 (2022). Genome dynamics in mosses: Extensive synteny coexists with a highly dynamic
714 gene space. In *bioRxiv* (p. 2022.05.17.492078). https://doi.org/10.1101/2022.05.17.492078
715 Kirbis, A., Waller, M., Ricca, M., Bont, Z., Neubauer, A., Goffinet, B., & Szövényi, P. (2020).
716 Transcriptional landscapes of divergent sporophyte development in two mosses,
717 Physcomitrium (Physcomitrella) patens and Funaria hygrometrica. *Frontiers in Plant*
718 *Science*, *11*, 747.
719 Kirsch, R., Okamura, Y., Haeger, W., Vogel, H., Kunert, G., & Pauchet, Y. (2022). Metabolic
720 novelty originating from horizontal gene transfer is essential for leaf beetle survival.
721 *Proceedings of the National Academy of Sciences of the United States of America*, *119*(40),
722 e2205857119.
723 Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone

reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A. A., & Blaxter, M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini [Review of *No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini*]. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(18), 5053–5058.

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278.

Lang, D., Ullrich, K. K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B., Piednoel, M., Gundlach, H., Van Bel, M., Meyberg, R., Vives, C., Morata, J., Symeonidi, A., Hiss, M., Muchero, W., Kamisugi, Y., Saleh, O., Blanc, G., Decker, E. L., … Rensing, S. A. (2018). The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *The Plant Journal*, *93*(3), 515–533.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., Eily, A., Koppers, N., Kuo, L.-Y., Li, Z., Simenc, M., Small, I., Wafula, E., Angarita, S., Barker, M. S., Bräutigam, A., dePamphilis, C., Gould, S., Hosmani, P. S., … Pryer, K. M. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, *4*(7), 460–472.

Ma, J., Wang, S., Zhu, X., Sun, G., Chang, G., Li, L., Hu, X., Zhang, S., Zhou, Y., Song, C.-P., & Huang, J. (2022). Major episodes of horizontal gene transfer drove the evolution of land plants. *Molecular Plant*, *15*(5), 857–871.

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12), e323.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Martin, F. M., Uroz, S., & Barker, D. G. (2017). Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. *Science*, *356*(6340), ea.ad4501

755 https://doi.org/10.1126/science.aad4501
756 Medina, R., Johnson, M., Liu, Y., Wilding, N., Hedderson, T. A., Wickett, N., & Goffinet,
757 B. (2018). Evolutionary dynamism in bryophytes: Phylogenomic inferences confirm rapid
758 radiation in the moss family Funariaceae. *Molecular Phylogenetics and Evolution*, *120*,
759 240–247.
760 Medina, R., M. G. Johnson, Y. Liu, N.J. Wickett, A.J. Shaw & B. Goffinet. 2019. Phylogenomic
761 delineation of Physcomitrium (Bryophyta: Funariaceae) based on nuclear targeted exons
762 and their flanking regions rejects the retention of Physcomitrella, Physcomitridium and
763 Aphanorrhegma. *Journal of Systematics and Evolution, 57*, 404–417.Price, M. N., Dehal, P.
764 S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for
765 Large Alignments. *PloS One*, *5*(3), e9490.
766 Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., & Vandepoele,
767 K. (2011). i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely
768 large data sets. *Nucleic Acids Research*, *40*(2), e11–e11.
769 Rahmatpour, N., Perera, N. V., Singh, V., Wegrzyn, J. L., & Goffinet, B. (2021). High gene
770 space divergence contrasts with frozen vegetative architecture in the moss family
771 Funariaceae. *Molecular Phylogenetics and Evolution*, *154*, 106965.
772 Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and
773 Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*,
774 *11*(1), 1432.
775 Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S.-Z., & Bezanilla, M. (2020). The moss
776 Physcomitrium (Physcomitrella) patens: A model organism for non-seed plants. *The Plant*
777 *Cell*, *32*(5), 1361–1376.
778 Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T.,
779 Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T.,
780 Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S.-I., … Boore, J. L.
781 (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land
782 by plants. *Science*, *319*(5859), 64–69.
783 Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality,
784 completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245.
785 Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig

786 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460.
787 Smit, AFA, Hubley, R & Green, P. (2013-2015). *RepeatMasker Open-4.0*. RepearMasker.
788 http://www.repeatmasker.org
789 Sun, G., Bai, S., Guan, Y., Wang, S., Wang, Q., Liu, Y., Liu, H., Goffinet, B., Zhou, Y., Paoletti,
790 M., Hu, X., Haas, F. B., Fernandez-Pozo, N., Czyrt, A., Sun, H., Rensing, S. A., & Huang,
791 J. (2020). Are fungi-derived genomic regions related to antagonism towards fungi in
792 mosses? *The New Phytologist*, *228*(4), 1169–1175.
793 Sun, H., Shang, H., Pan, X., & Li, M. (2023). Structural insights into the assembly and energy
794 transfer of the Lhcb9-dependent photosystem I from moss Physcomitrium patens. *Nature*
795 *Plants*, *9*(8), 1347–1358.
796 Van Dongen, S. M. (2000). Graph clustering by flow simulation. Doctoral dissertation. Utrecht
797 University, Netherlands.
798 Van Etten, J., & Bhattacharya, D. (2020). Horizontal gene transfer in Eukaryotes: Not if, but how
799 much? *Trends in Genetics*, *36*(12), 915–925.
800 Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., Zaman,
801 S., Bennett, J., McEvoy, S. L., Caballero, M., & Wegrzyn, J. L. (2022). Welcome to the big
802 leaves: best practices for improving genome annotation in non-model plant genomes. In
803 *bioRxiv* (p. 2022.10.03.510643). https://doi.org/10.1101/2022.10.03.510643
804 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A.,
805 Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for
806 comprehensive microbial variant detection and genome assembly improvement. *PloS One*,
807 *9*(11), e112963.
808 Wang, S., Guan, Y., Wang, Q., Zhao, J., Sun, G., Hu, X., Running, M. P., Sun, H., & Huang, J.
809 (2020). A mycorrhizae-like gene regulates stem cell and gametophore development in
810 mosses. *Nature Communications*, *11*(1), 2030.
811 Wilding, N. (2015). Systematics, biogeography and morphological evolution in Entosthodon
812 Schwägr. (Bryopsida, Funariaceae) with a revision of the genus in Africa. Doctoral
813 dissertation. University of Cape Town, South Africa.
814 Xi, Z., Bradley, R. K., Wurdack, K. J., Wong, K., Sugumaran, M., Bomblies, K., Rest, J. S., &
815 Davis, C. C. (2012). Horizontal transfer of expressed genes in a parasitic flowering plant.
816 *BMC Genomics*, *13*, 227.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, *11*(2), R14.

Young, LA, (2022). *Relationships Among AA-Genome Chenopodium Diploids and a Whole Genome Assembly of the North American Species, C. watsonii* (Doctoral dissertation, Brigham Young University)

Yu, J., Li, L., Wang, S., Dong, S., Chen, Z., Patel, N., Goffinet, B., Chen, H., Liu, H., & Liu, Y. (2020). Draft genome of the aquatic moss Fontinalis antipyretica (Fontinalaceae, Bryophyta). *GigaByte*, *2020*, gigabyte8.

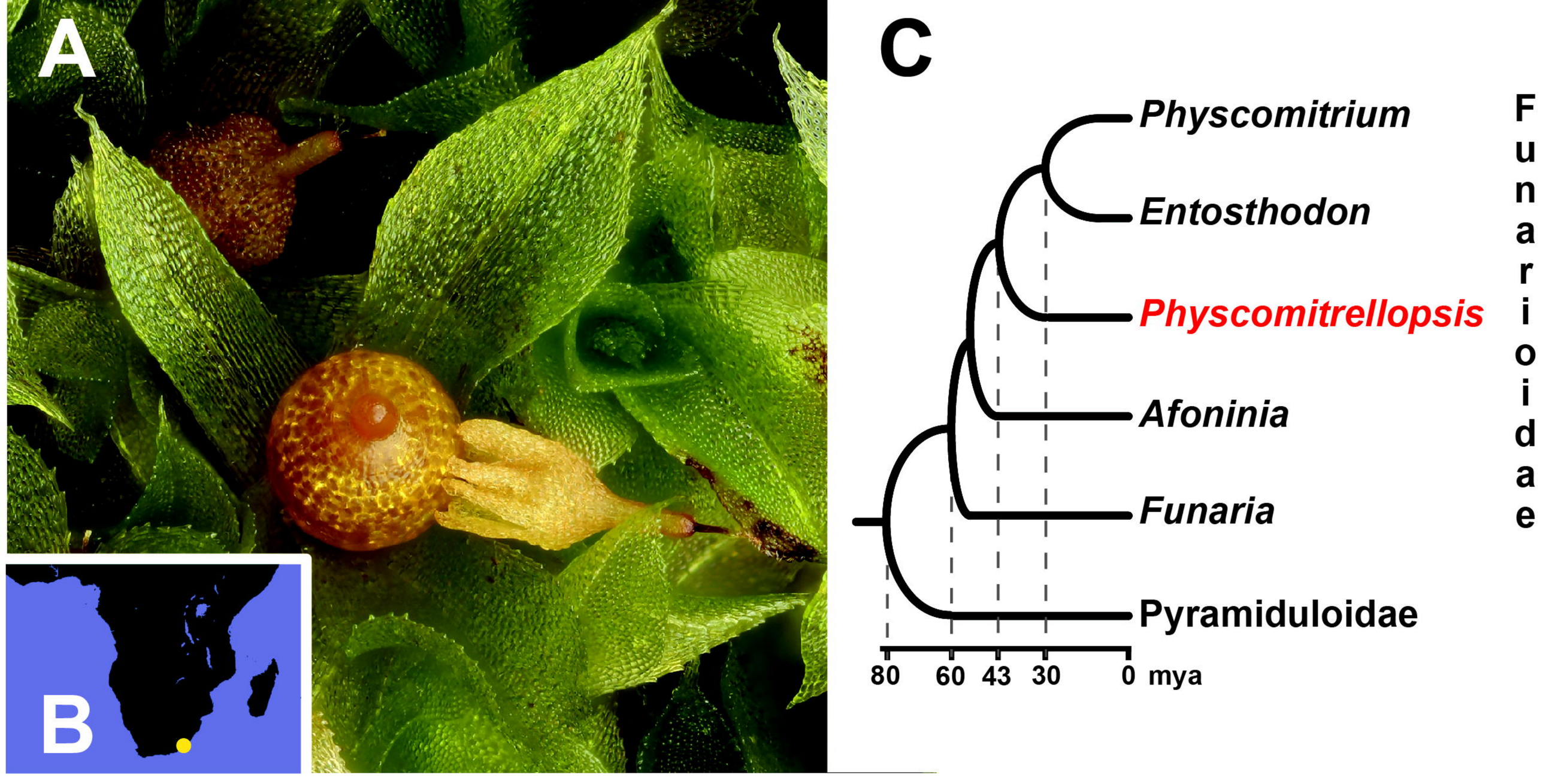
Zhang, J., Fu, X.-X., Li, R.-Q., Zhao, X., Liu, Y., Li, M.-H., Zwaenepoel, A., Ma, H., Goffinet, B., Guan, Y.-L., Xue, J.-Y., Liao, Y.-Y., Wang, Q.-F., Wang, Q.-H., Wang, J.-Y., Zhang, G.-Q., Wang, Z.-W., Jia, Y., Wang, M.-Z., … Chen, Z.-D. (2020). The hornwort genome and early land plant evolution. *Nature Plants*, *6*(2), 107–118.

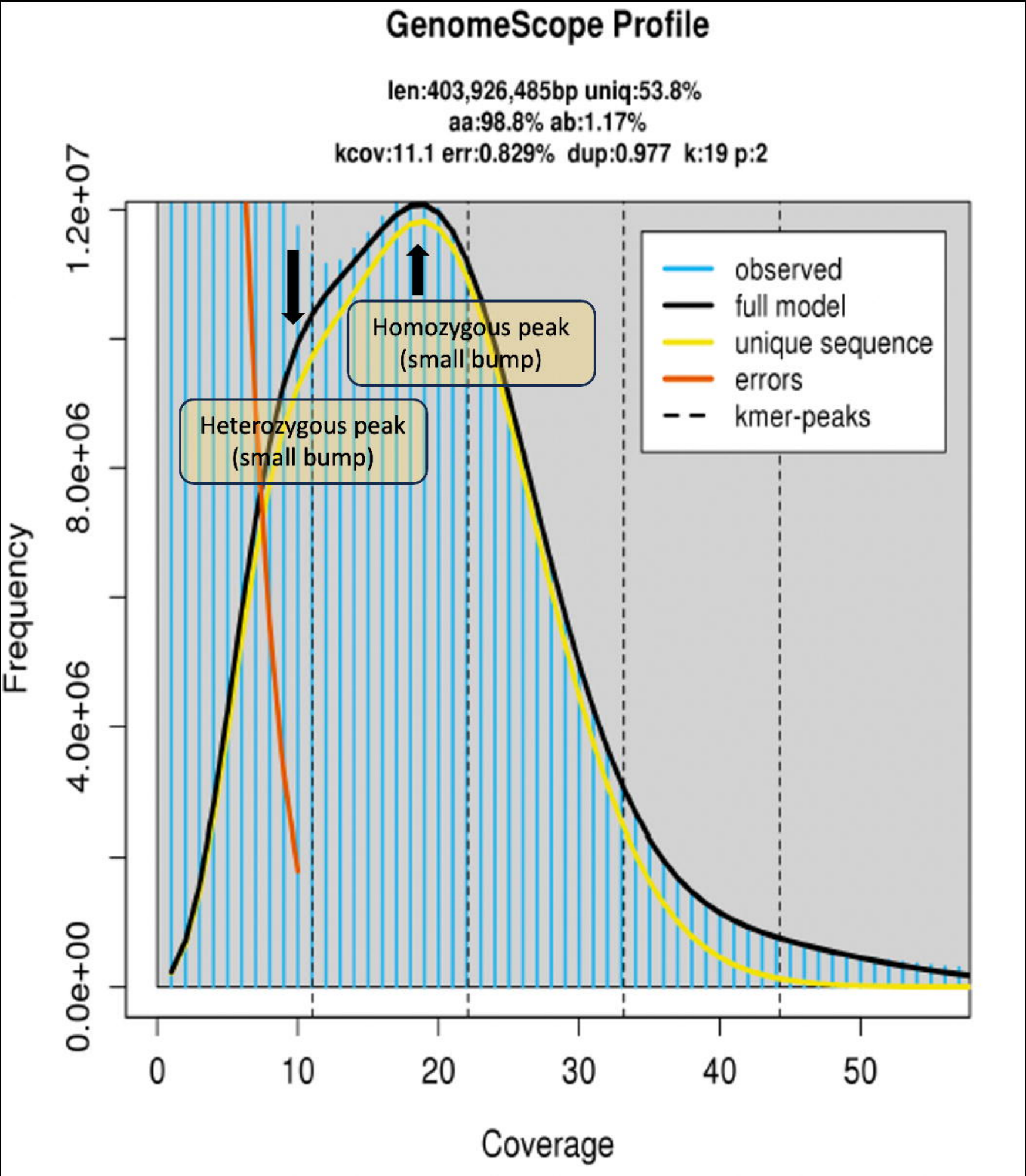
Zhang, Z., & Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics*, *19*(2), 307–308.

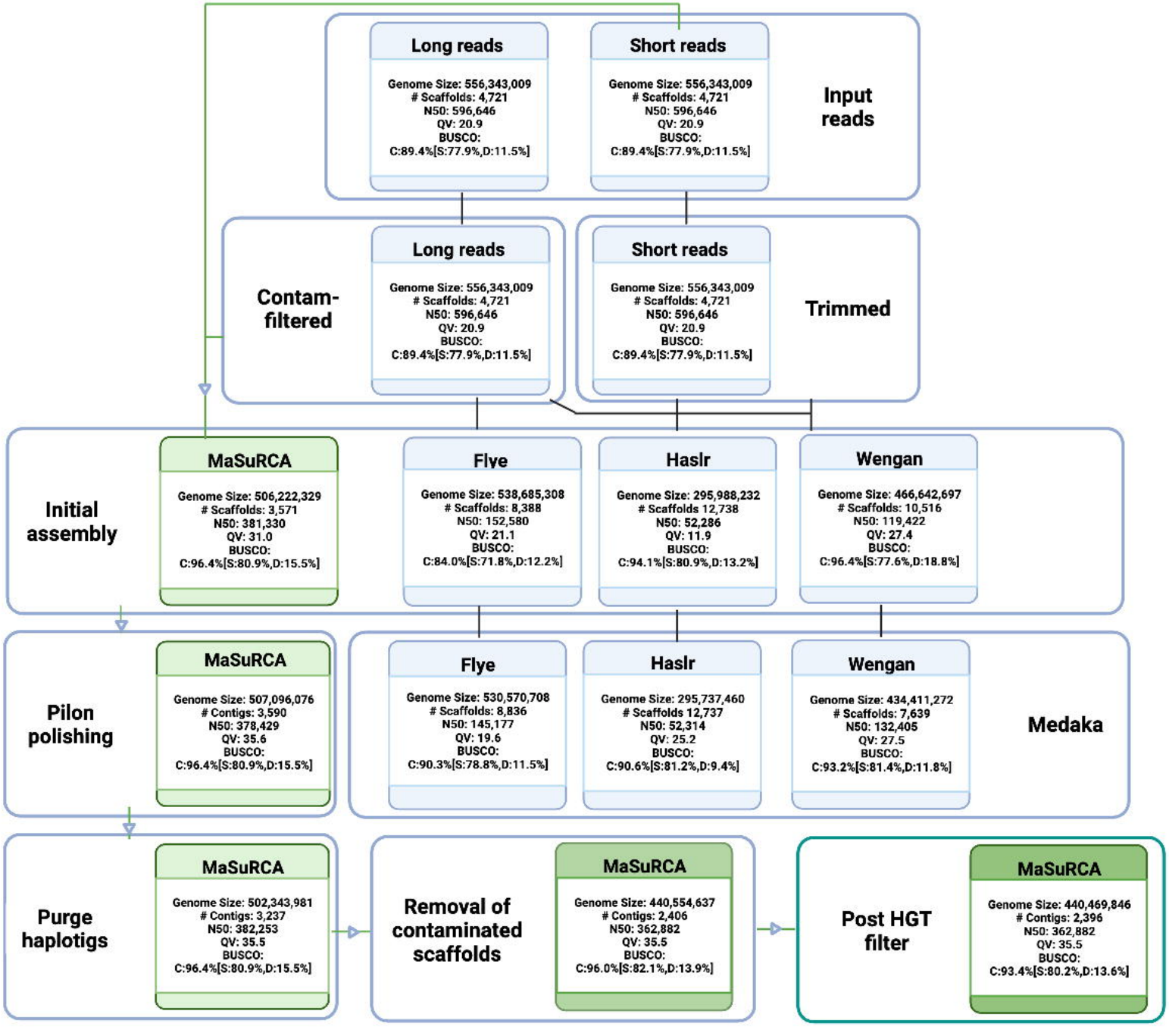
Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677.

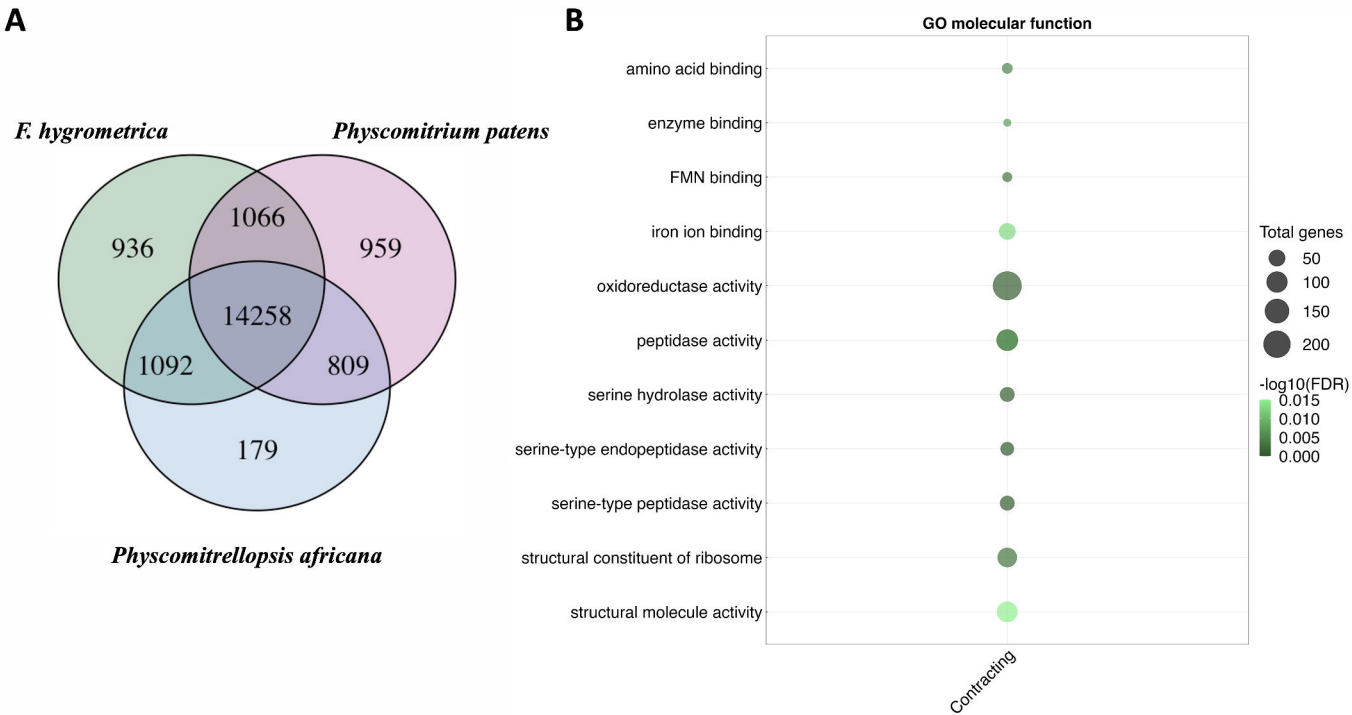
Zimmer, A. D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., ... & Reski, R. (2013). Reannotation and extended community resources for the genome of the non-seed plant Physcomitrella patens provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, *14*, 1–20.

Zwaenepoel, A., & Van de Peer, Y. (2018). wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*, *35*(12), 2153–2155.

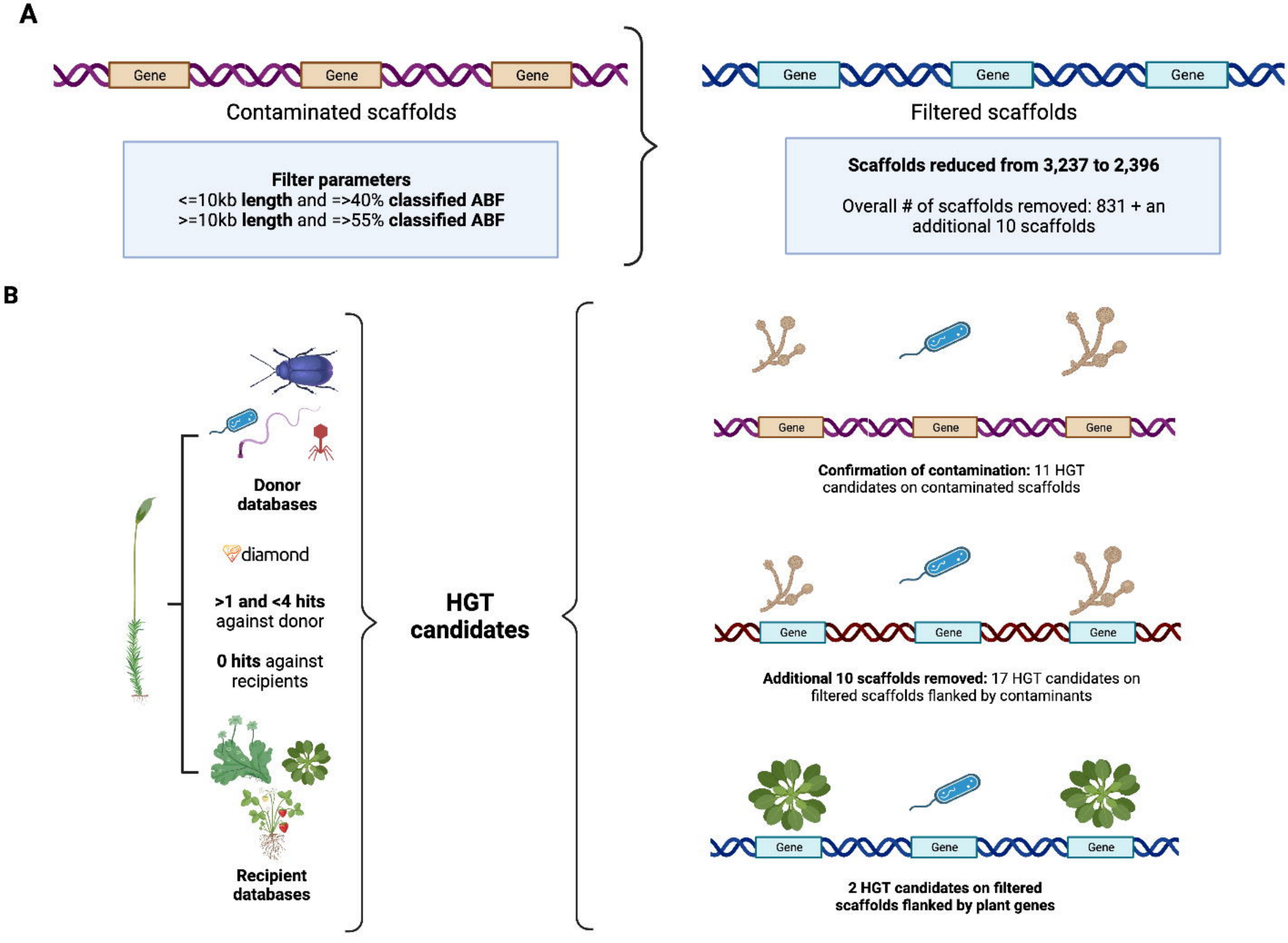

A
B
C
Physcomitrium
Entosthodon
Physcomitrellopsis
Afoninia
Funaria
Pyramiduloidae
80
60
43
30
0 mya
Funarioidae

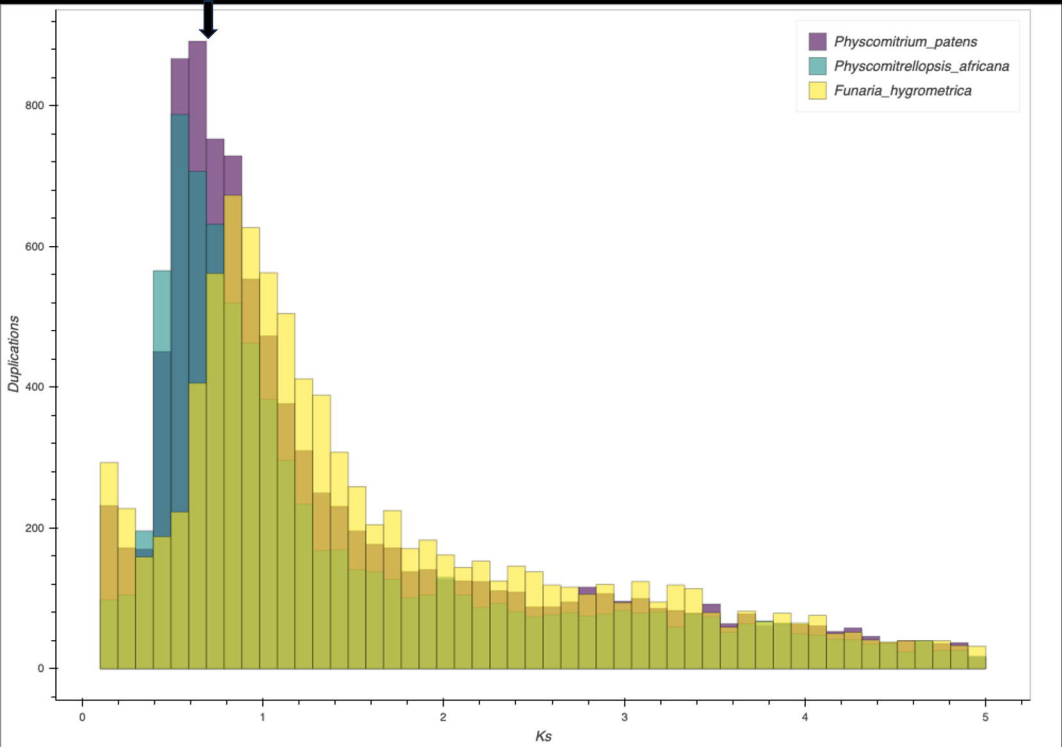

GenomeScope Profile
len:403,926,485bp uniq:53.8%
aa:98.8% ab:1.17%
kcov:11.1 err:0.829% dup:0.977 k:19 p:2
Homozygous peak
(small bump)
Heterozygous peak
(small bump)
observed
full model
unique sequence
errors
kmer-peaks
Frequency
0.0e+00
4.0e+06
8.0e+06
1.2e+07
0
10
20
30
40
50
Coverage

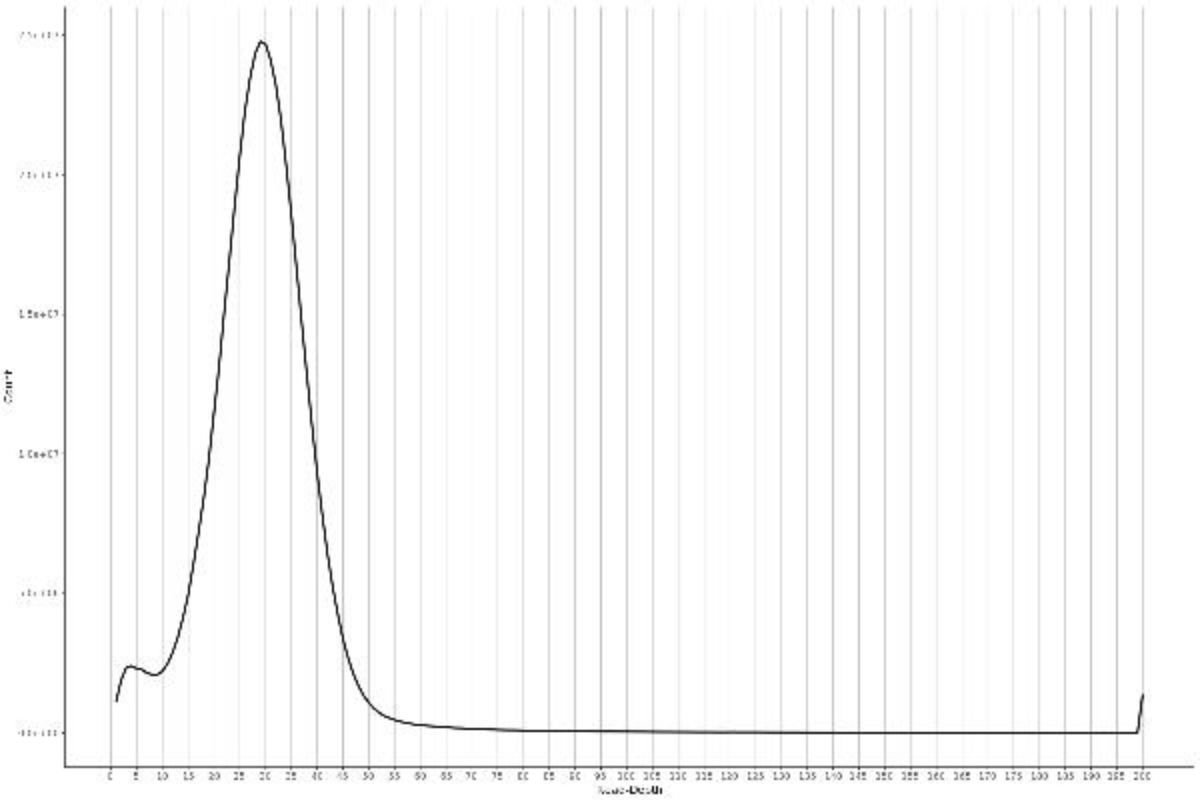

Long reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Short reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Input reads
Contam-filtered
Long reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Short reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Trimmed
Initial assembly
MaSuRCA
Genome Size: 506,222,329
Scaffolds: 3,571
N50: 381,330
QV: 31.0
BUSCO:
C:96.4%[S:80.9%,D:15.5%]
Flye
Genome Size: 538,685,308
Scaffolds: 8,388
N50: 152,580
QV: 21.1
BUSCO:
C:84.0%[S:71.8%,D:12.2%]
Haslr
Genome Size: 295,988,232
Scaffolds 12,738
N50: 52,286
QV: 11.9
BUSCO:
C:94.1%[S:80.9%,D:13.2%]
Wengan
Genome Size: 466,642,697
Scaffolds: 10,516
N50: 119,422
QV: 27.4
BUSCO:
C:96.4%[S:77.6%,D:18.8%]
Pilon polishing
MaSuRCA
Genome Size: 507,096,076
Contigs: 3,590
N50: 378,429
QV: 35.6
BUSCO:
C:96.4%[S:80.9%,D:15.5%]
Flye
Genome Size: 530,570,708
Scaffolds: 8,836
N50: 145,177
QV: 19.6
BUSCO:
C:90.3%[S:78.8%,D:11.5%]
Haslr
Genome Size: 295,737,460
Scaffolds 12,737
N50: 52,314
QV: 25.2
BUSCO:
C:90.6%[S:81.2%,D:9.4%]
Wengan
Genome Size: 434,411,272
Scaffolds: 7,639
N50: 132,405
QV: 27.5
BUSCO:
C:93.2%[S:81.4%,D:11.8%]
Medaka
Purge haplotigs
MaSuRCA
Genome Size: 502,343,981
Contigs: 3,237
N50: 382,253
QV: 35.5
BUSCO:
C:96.4%[S:80.9%,D:15.5%]
Removal of contaminated scaffolds
MaSuRCA
Genome Size: 440,554,637
Contigs: 2,406
N50: 362,882
QV: 35.5
BUSCO:
C:96.0%[S:82.1%,D:13.9%]
Post HGT filter
MaSuRCA
Genome Size: 440,469,846
Contigs: 2,396
N50: 362,882
QV: 35.5
BUSCO:
C:93.4%[S:80.2%,D:13.6%]


A
F. hygrometrica
Physcomitrium patens
Physcomitrellopsis africana
936
1066
959
14258
1092
809
179
B
GO molecular function
amino acid binding
enzyme binding
FMN binding
iron ion binding
oxidoreductase activity
peptidase activity
serine hydrolase activity
serine-type endopeptidase activity
serine-type peptidase activity
structural constituent of ribosome
structural molecule activity
Contracting
Total genes
50
100
150
200
-log10(FDR)
0.015
0.010
0.005
0.000

A
scaffold_0062:414,417-424,781
10 kb
416 kb
418 kb
420 kb
422 kb
424 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_04827.1
Pa1_04828
Pa1_04829
StringTie2 transcripts
STRG.5308
STRG.5310
Alignment of NR hits to the genome
WP_207424935.1
Alignment of P. patens proteins to the genome
MP000780
Alignment of C. purpureus proteins to the genome
MP000461
MP000436
B
scaffold_1363:26,806-40,495
13 kb
28 kb
30 kb
32 kb
34 kb
36 kb
38 kb
40 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_22335.1
Pa1_22336.1
Pa1_22337.1
StringTie2 transcripts
STRG.24351
STRG.24354
Alignment of NR hits to the genome
MCB2107078.1
Alignment of P. patens proteins to the genome
MP001460
MP001273
MP000247
Alignment of C. purpureus proteins to the genome
MP000355
MP000186
C
scaffold_0465:130,600-154,988
24 kb
132 kb
134 kb
136 kb
138 kb
140 kb
142 kb
144 kb
146 kb
148 kb
150 kb
152 kb
154 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_15636.1
StringTie2 transcripts
Alignment of NR hits to the genome
MP000002


A
Gene
Gene
Gene
Contaminated scaffolds
Filter parameters
<=10kb length and =>40% classified ABF
>=10kb length and =>55% classified ABF
Gene
Gene
Gene
Filtered scaffolds
Scaffolds reduced from 3,237 to 2,396
Overall # of scaffolds removed: 831 + an additional 10 scaffolds
B
Donor databases
diamond
>1 and <4 hits against donor
0 hits against recipients
Recipient databases
HGT candidates
Gene
Gene
Gene
Confirmation of contamination: 11 HGT candidates on contaminated scaffolds
Gene
Gene
Gene
Additional 10 scaffolds removed: 17 HGT candidates on filtered scaffolds flanked by contaminants
Gene
Gene
Gene
2 HGT candidates on filtered scaffolds flanked by plant genes


Physcomitrium_patens
Physcomitrellopsis_africana
Funaria_hygrometrica
Duplications
800
600
400
200
0
0
1
2
3
4
5
Ks


Count
Water-Depth