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, CT 06269, USA

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

[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, USA

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

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, CT 06269, USA. Email: jill.wegrzyn@uconn.edu and bernard.goffinet@uconn.edu

Running head: *Physcomitrellopsis* genome

Journal: G3 (Genome Report)

Abstract

The first chromosome-scale reference genome of the rare narrow-endemic African moss *Physcomitrellopsis africana* is presented here. Assembled from 73x nanopore long reads and 163x BGI-seq short reads, the 414 Mb reference comprises 26 chromosomes and 22,925 protein-coding genes (BUSCO: C:94.8%[D:13.9%]). This genome holds two genes that withstood rigorous filtration of microbial contaminants, have no homolog in other land plants and are thus interpreted as resulting from two unique horizontal gene transfers from microbes. Further, *Physcomitrellopsis africana* shares 176 of the 273 published HGT candidates identified in *Physcomitrium patens*, but lacks 98 of these, highlighting that perhaps as many as 91 genes were acquired in *P. patens* in the last 40 million years following its divergence from its common ancestor with *P. africana*. These observations suggest rather continuous gene gains via HGT followed by potential losses, during the diversification of the Funariaceae. Our findings showcase both dynamic flux in plant HGTs over evolutionarily "short" timescales, alongside enduring impacts of successful integrations, like those still functionally maintained in extant *Physcomitrellopsis africana*. Furthermore, this study describes the informatic processes employed to distinguish contaminants from candidate HGT events.

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

Article Summary

The first draft genome of the rare South African endemic moss *Physcomitrellopsis Africana* is presented. The 414 Mb assembly contains 22,925 genes, including two uniquely horizontally transferred genes, but lacks 97 of the microbial genes previously identified in the closely related model, *Physcomitrium patens* - highlighting the dynamic role of HGT in the evolution of these moss genomes and loss. This study presents best practices for contamination detection and new insights into HGT identification.

Introduction

Horizontal gene transfer (HGT), also referred to as lateral gene transfer (Zhaxybayeva and Doolittle 2011) is the lateral movement of genetic material between distant 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 than in prokaryotes, it plays a role in shaping eukaryotic evolution with about 0.04-6.49% of eukaryotic genes originating from HGT from microbes (Van Etten and 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 ascomycete 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* and may play a role in defenses against microbial pathogens (Sun *et al*. 2020). Van Etten and Bhattacharya (2020) suggested that rather than being merely anecdotal, HGT has been an important evolutionary force driving the adaptations of land plants to new habitats and stressors throughout their evolution, as previously already hypothesized by Yue *et al.* (2012).

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 been proven to be artifacts resulting from contamination that had gone undetected. For example, the initial report that 17% of the genes in the genome of the tardigrade *Hypsibius dujardini* originated from HGT (Boothby *et al.* 2015), has subsequently been revised to only 0.2%

(Koutsovoulos *et al.* 2016). Similarly, claims of HGT in mammals also turned out to be erroneous due to bacterial contamination (Douvlataniotis *et al.* 2020), highlighting the need for careful analyses and experimental validation. Systematic screening of 43 arthropod genomes by Francois *et al*. (2020) revealed extensive bacterial contaminants, often outnumbering true horizontal gene transfer 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 that of 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).

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). Its genome of 414 MB, the first genome of an African bryophyte, is presented here. This resource complements those available for *Funaria hygrometrica* and *Physcomitrium patens* (Rensing *et al.* 2008; Kirbis *et al.* 2022) and thereby constitutes a fundamental resource to further the reconstruction of the evolution of the genome of the model, *Physcomitrium patens* (Rensing *et al.* 2020). Through rigorous contamination screening and verification, two genes emerge as candidates resulting from unique HGT events. Informatically validating 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 J. Goffinet, 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 long read sequencing, whereas 10329 (RNA and BGI-Seq DNA #5075) provided RNA and DNA for short read 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 consisting of 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 (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.

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).

Genome and transcriptome library preparation and sequencing

DNA was prepared for long-read 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. The HMW DNA library was sequenced on an Oxford Nanopore PromethION (Center for Genome Innovation, UConn) using a FLO-PRO0002 flow cell. . The two short-read DNA libraries were prepared following the methods used by Yu et al (2020), and was sequenced (150 bp PE) on two lanes of a BGISEQ-500 (BGI-Shenzhen, China). .

Approximately 1 µg of RNA was used to generate a paired-end library with an insert fragment size of 200–300 bp of the corresponding cDNA. RNA purification, reverse transcription, library construction and sequencing were performed at WuXi NextCode (Shanghai,

China). The captured coding regions of the transcriptome from total RNA were prepared using the TruSeq® RNA Exome Library Preparation Kit. The two RNA libraries were sequenced on one lane of an Illumina HiSeq 2000 (100 bp PE) at the WuXi NextCode (Shanghai, China).

Quality control of genomic and transcriptomic reads

The genomic short reads were first assessed with FASTQC v.0.11.7 (Andrews 2010). In preparation for assembly with Haslr and Wengan, Sickle v.1.33 with a minimum quality score threshold of 30 (-q) and a minimum length of 50 bp (-l) was employed for read trimming. Nanoplot v.1.21.0 was used to quantify and assess the quality of the nanopore reads. To detect potential contamination, the long reads were aligned against the pre-indexed bacterial, human and viral databases with a metagenomic classifier, Centrifuge v.1.0.4-beta (p+h+v; min-hit length was increased to 50 bp) (Kim *et al.* 2016). Reads aligning to the database were removed. The quality of the transcriptomic short reads was assessed with FASTQC v.0.11.7 (Andrews 2010). The reads were trimmed with Sickle v1.33 (Joshi and Fass 2011) with a minimum quality score threshold of 30 (-q) and a minimum length of 40 bp (-l).

Genome size estimation

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

Transcriptome assembly

The transcriptome was independently assembled to provide protein-level evidence for the structural annotation of the genome, using Trinity v2.6.6 (Grabherr *et al.* 2011), with a minimum contig length of 300bp. Contigs with minimal read support, post assembly, were removed (FPKM > 0.5) with RSEM v1.3.0 (Li and Dewey 2011). Transdecoder v3.0.1 (Haas and Papanicolaou 2011) was used to translate the remaining contigs into Open Reading Frames (ORFs) and remove sequences without a viable frame. To aid Transdecoder in the identification of ORFs, searches against the Pfam database were performed with HMMER v3.1b2 (Zhang and Wood 2003). Transdecoder annotated the putative transcripts as complete, partial, and internal. Those without a defined start and stop codon (defined as internals) were removed

(split_frames.py). The final set of peptide sequences were functionally annotated with EnTAP v0.8.0 (Hart *et al.* 2020) against NCBI's nr protein database and the UniProt/Swiss-prot reference databases. EnTAP was run with contaminant filters that included bacteria, archaea, fungi, and insecta. Transcripts with high confidence alignments to these organisms were removed (contam_removal.py).

Genome assembly

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

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

The Purge Haplotigs pipeline categorized the assembled scaffolds into four coverage levels based on the distribution of mapped reads. This categorization enabled the identification and removal of redundant sequences exhibiting low coverage, presumed to represent erroneous duplicates given the haploid genome. Specific cutoff values of 0, 7, and 65 for the coverage levels were selected to delineate scaffolds to be retained or discarded, based on read depths (Supplementary Fig. S1). The term "allele" is used to describe these redundant sequences for convenience, although technically inaccurate for a haploid genome. The coverage analysis and purging allowed isolation of the primary genome sequence from duplications and artifacts generated during assembly.

To evaluate the quality of the assemblies, QUAST v5.2.0 (Gurevich *et al*. 2013) and BUSCO v4.1.2 (viridiplantae_odb10:425 single copy orthologs) (Manni *et al*. 2021) were employed. Each assembly was also evaluated with Merqury v1.3 (Rhie *et al*. 2020).

Genome annotation

Repeat library construction and masking

The repeat library for the final MaSuRCA assembly was generated using RepeatModeler v.2.01 (Flynn *et al.* 2020) with the long terminal repeat (LTR) discovery pipeline enabled. The genome was then soft-masked with RepeatMasker v.4.0.9-p2 using the consensus repeat library from RepeatModeler (Smit *et al.* 2013-2015).

Structural and functional genome annotation

The RNA reads were aligned to the soft-masked MaSuRCA assembly with HISAT2 v.2.1.0 (Kim *et al.* 2019) to provide evidence for protein-coding gene prediction. Two gene prediction analyses were run on the soft-masked assembly using BRAKER v.2.1.5 (Brůna *et al.* 2021), one with RNA-Seq alignment evidence and one with protein evidence originating from *de novo* assembled transcriptome. Gene predictions from both BRAKER runs were integrated with TSEBRA v.1.0.3 (Gabriel *et al.* 2021). From this point, separate assessments were conducted on the RNA-Seq evidence gene predictions (BRAKER) and the final TSEBRA gene predictions to select the best approach. Putative genes were removed from both sets if they did not contain a complete protein domain. This filter was applied with Interproscan v.5.35-74.0 (Jones *et al.* 2014) using the Pfam database v32.0 (Finn *et al*. 2014). It is worth noting that mono-exonic genes can be the result of fragmented annotations and the target metric of 0.2 (mono:multi-exon gene ratio) is often achieved through protein domain filters (Vuruputoor *et al.* 2022). Metrics for the gene predictions were generated with AGAT (Dainat 2020) and BUSCO. After assessment, the filtered BRAKER gene predictions were selected for functional annotation with EnTAP v.0.10.8 (Hart *et al.* 2020). Functional annotation reports from EnTAP (both sequence similarity search and EggNog taxonomy scope classifications) allowed for the identification of non-target species scaffolds in the assembly (Huerta-Cepas *et al*. 2019).

Assembly-level contaminant filtering

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

Horizontal gene transfer candidate identification

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

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

To assess the validity of the two proposed HGT candidates, the target proteins were independently aligned to the reference genome with Miniprot v0.7. To examine independent transcriptomic support, the RNA reads were aligned to the genome with HISAT2 v2.1.0 and assembled via StringTie2 v2.2.1 (Kovaka *et al*. 2019).

Analyzing HGT candidates from Physcomitrium patens

The 273 putative horizontally transferred genes (HGTs) previously identified in *Physcomitrium patens* (Yue *et al*. 2012; Ma *et al.* 2022) were independently searched against the *Physcomitrellopsis africana* and *F. hygrometrica* protein sets using DIAMOND v2.1.8 (Buchfink *et al.* 2021). DIAMOND searches were conducted with an E-value cutoff of 1e-5 and max target sequences set to 1. Hits against *Physcomitrellopsis africana* and *F. hygrometrica* were collected and merged to generate a summary table with *Physcomitrium patens* HGTs and the respective top hits in each species.

Scaffolding the genome to chromosome scale

A single HiC library was prepared for *Physcomitrellopsis africana* using Proximo Hi-C kit (Phase Genomics) and sequenced as paired-end reads on an Illumina NovaSeq platform (Center for Genome Innovation, UConn). The reads were aligned to the MaSuRCA genome assembly using BWA v0.7.17 but did not provide sufficient coverage due to contamination. Instead, the *Physcomitrium patens* T2T reference genome assembly (Bi *et al.* 2024) was used to scaffold the *Physcomitrellopsis africana* assembly to chromosome-scale with RagTag v2.1.0 under default parameters, except a minimum fragment size threshold of 1 Kb was specified.

Comparative genome analyses

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

GOSeq (Young *et al.* 2010) enrichment analysis was performed in R v4.2.0. GO terms were extracted for each gene from the EnTAP run. Enrichment analysis was investigated separately for the *Biological Process* and *Molecular Function* GO categories. Paralogs of Light Harvesting Complex (LHC), Serine/Threonine-protein kinase (STN) 7, and STN8 were identified with Diamond v2.8.1 (E-value <1e-5).

Whole genome duplication analysis

Chromosome-scale genomes of *F. hygrometrica* and *Physcomitrium patens* were assessed with the reference genome generated for Ph*yscomitrellopsis africana*, with wgd (v.2.0.19) to characterize whole genome duplication events (Chen and Zwaenepoel 2023). Each species was

compared against itself. Nucleotide sequences (CDS) were used as input to Blast & Markov clustering (MCL) (Altschul *et al.* 1997; van Dongen 2000). A Ks distribution was constructed using the 'ksd' subcommand and 'mcl' output (Yang 2007; Katoh and Toh 2008; Price *et al*. 2010). Next, a collinearity analysis was performed with 'syn' using the structural genome annotations (Proost *et al*. 2011). Gaussian mixture models, via 'mix', were used to generate Ks distributions to aid in component interpretation. Model fit was evaluated with the Bayesian and Akaike information criterion (BIC/AIC).

Results and Discussion

Sequencing and quality control of genomic reads

The single ONT run generated 16 million reads (N50: 7,518 bp; 127X coverage: 404 Mb; 116X coverage: 440 Mb; Supplementary Table S1). Centrifuge filtering reduced this set to approximately 14M reads (N50: 4,561 bp; 80X coverage: 404 Mb; 73X coverage: 440 Mb). The primary contaminants included bacteria from *Xanthomonadaceae* (10.68%) followed by *Bradyrhizobiaceae* (0.7%). The short-read genomic libraries (100 bp PE) generated 529 M reads. Following trimming with Sickle, 360 M reads remained (178X coverage: 404 Mb; 163X coverage: 440 Mb; Supplementary Table S2). Coverage estimates are provided for both the original estimate (404 Mb) and the final assembled genome size (440 Mb).

Transcriptome assembly

A total of 36.58M RNA short reads were generated. Following quality filtering using Trimmomatic, the dataset was reduced to 31.84M reads. The *de novo* assembly process with Trinity yielded 143,150 contigs. After expression filtering via RSEM, the number of contigs was reduced to 123,373. Identifying open reading frames (ORFs) in the contigs resulted in 110,852 successfully translated transcripts. The average sequence length of these translated ORFs was 803 bp, with an N50 value of 1,038 bp. To enhance the quality of the transcriptome assembly, internal sequences and putative contaminants were removed, resulting in 74,997 total transcripts. After removing internal sequences and contamination, the total unique sequences with a sequence similarity search alignment are 51,982 (69.3%). An assessment of contamination revealed that 8,720 transcripts, 16.78% of the transcriptome, originated from various sources such as amoeba (0.06%), bacteria (0.64%), fungi (15.93%), or insecta (0.12%), as identified by EnTAP. A further 15,500 (18.5%) sequences remained unannotated. The final BUSCO score of

the remaining transcripts was C:82%[D:35.7%]. A total of 30,657 (36.6%) transcripts aligned to *Physcomitrium patens* proteins.

Genome assembly

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

Polishing and improving genome assembly:

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

Polishing with Pilon had very minimal influence on the completeness, as expected, and substantial impact on the accuracy of the MaSuRCA assembly. The polished MaSuRCA hybrid assembly size increased slightly to 507.10 Mb across 3,590 scaffolds, with a slightly decreased N50 of 378.43 Kb. The BUSCO completeness remained the same at C:96.4%[D:15.5%], and the QV score increased to 35.6.

Refinement of genome assemblies with Purge-haplotigs: The MaSuRCA assembly was selected for further refinement. This decision was based on the overall quality assessed by Merqury,

BUSCO completeness, and overall contiguity. The MaSuRCA assembly was refined with Purge-haplotigs, which reduced the assembly length to 502.34 Mb across 3,237 scaffolds with an N50 value of 382.25 Kb. The BUSCO completeness score remained the same (e.g., 96.4%) and the QV score was minimally reduced to 35.5. At 502.34 Mb, the assembled genome is ~100 Mb longer than the k-mer based estimate (440 Mb, Supplementary Fig. S2).

Genome annotation

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

Protein-coding gene identification: The *Physcomitrellopsis africana* RNA-Seq library had an overall alignment rate of 78.43%, likely due to contaminant content in the source extraction. Regardless, a substantial set of 37M reads were retained, exceeding the minimum needed for prediction. The first BRAKER2 predictions, using RNA-Seq evidence alone, generated 60,917 protein-coding genes with a BUSCO completeness score of C:95.8%[D:15.3]. The mono:multi exonic ratio was 0.52. With only protein evidence, BRAKER2 generated 37,752 gene predictions with a BUSCO score of C:46.1%[D:16.0]. Merging these predictions, as recommended by TSEBRA, resulted in a set of 45,737 genes with a BUSCO score of C:91.3%[D:15.1], and a mono:multi exonic ratio of 1.06. The gene prediction set generated with RNA-Seq alignment evidence exclusively was selected for further refinement due to its higher BUSCO score and lower mono:multi ratio compared to the merged TSEBRA transcripts. Through protein domain filtering (InterProScan filter), the number of mono-exonic genes was further reduced from

20,081 to 12,696, producing a total of 23,561 genes (BUSCO: C:88.5%[D:13.8%]; mono:multi ratio: 0.08).

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

The annotated protein-coding gene space in the contaminant-filtered assembly included 1,708 mono-exonic and 21,853 multi-exonic genes. A BUSCO score of C:93.8%[D:13.6%] and mono:multi ratio of 0.08 was reported. Reciprocal BLAST conducted by EnTAP produced an annotation rate of 78%, of which 93.6% aligned to *Physcomitrium patens* (Table 1; File S1). This contrasts with the annotation rate of 50% in *F. hygrometrica* (Kirbis *et al*. 2022), attributed to significant divergence time from *Physcomitrium patens,* and the lack of closely related species in public genomic databases (Kirbis *et al.* 2020; Rahmatpour *et al.* 2021). The higher annotation rate in this study suggests that many genes found in *Physcomitrium patens,* but not *F. hygrometrica,* may have been acquired in the ancestor of the *Physcomitrellopsis-Entosthodon-Physcomitrium* clade (Medina *et al.* 2019).

Scaffolding the genome assembly to chromosome-scale

The T2T assembly of *Physcomitrium patens* (V4; Bi *et al*. 2024) was used to scaffold the *Physcomitrellopsis africana* genome assembly. The final assembly contained 440.65 Mb in 583 scaffolds with an N50 length of 15.04 Mb. The 26 largest pseudomolecules, corresponding to the *Physcomitrium patens* V4 haploid chromosome number, covered 414.21 Mb or 94% of the 440 Mb assembly. From here on, we will refer to these pseudomolecules as chromosomes, numbered according to their lengths. The final genome assembly had a BUSCO completeness of C:94.8% [S:80.9%, D:13.9%], F:1.2%, M:4.0%, n:425, and a total number of 22,925 protein coding genes. While most of the annotated gene space was retained, the BUSCO duplication score remained at 13%, compared to 11–15% for the *F. hygrometrica* and *Physcomitrium patens* genomes (Kirbis *et al.* 2022). Of the 557 contigs/scaffolds that did not contribute to the 26 chromosomes, 185 (33.34%) contained at least one protein coding gene annotation from the first annotation. Among

those, three (1.6%) contained at least one gene that was annotated as a contaminant but did not meet the original thresholds that removed full scaffolds.

Comparative genome analysis

The comparison of the genomes of *Physcomitrellopsis africana*, *F. hygrometrica*, and *Physcomitrium patens*, reveals that they share 13.5K orthogroups, but also hold 187, 939, and 958 unique orthogroups, respectively. *F. hygrometrica* and *Physcomitrellopsis africana* exclusively share more (1053) orthogroups than *Physcomitrellopsis africana* and *Physcomitrium patens* (745) (Fig. 3A), despite the latter sharing a more recent unique common ancestor (Fig. 1). Thus, whereas Funariaceae share a rather conserved architecture of their vegetative body even after at least 60 MY of divergence, their gene space varies considerably (Kirbis *et al.* 2022). Such differentiation was previously noted based on highly diverging transcriptomes (Rahmatpour *et al.* 2019), and interpreted as reflective of strong ecophysiological adaptations, which may be the driving force in bryophyte evolution (Glime 1990). Signatures of two ancient whole genome duplication (WGD) events were found in both *Physcomitrium patens* and *F. hygrometrica*. Previous studies have shown that genes related to regulation were preferentially retained after the first WGD in *Physcomitrium patens* (Lang *et al.* 2018). Our analysis of *Physcomitrellopsis africana* also retrieved two peaks, in agreement with the two WGD events in *F. hygrometrica* and *Physcomitrium patens*. The Ks peaks were at 0.56 and 0.92, and the AIC and BIC values accorded with two WGD events. To assess the gene expansion and contraction analysis, of the putative gene families from the *Physcomitrellopsis africana* annotation, 809 and 1,694 were, respectively, categorized as expanded and contracted (File S2). Enrichment analysis examined through Gene Ontology's *Biological Process* category revealed 167 expanded terms and six contracted terms. In contrast, within the *Molecular Function* GO category, 62 were contracted, and no terms were significantly expanded.

Although expanded *Biological Process* GO terms were excessively broad, orthogroup analysis revealed a pattern of contraction in GO terms pertaining to oxidoreductase activity, FMN binding, and serine protease activity (Fig. 3B). Whether this unique suite of downregulated categories diagnoses photosynthetic properties of *Physcomitrellopsis africana* only or of the expanded genus sensu Wilding (2015) remains to be tested. The observed changes in GO term enrichment related to oxidoreductase activity, FMN binding, and serine protease activity suggest

potential differences in photosynthetic processes between *Physcomitrellopsis africana* and other moss species.

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

Comparing the genomes of *Physcomitrellopsis africana, F. hygrometrica* and *Physcomitrium patens* reveals both conservation and divergence of LHC genes. For example, *Physcomitrellopsis africana* has eight distinct Lhca genes, whereas *Physcomitrium patens* has 12 and *F. hygrometrica* has seven. Similarly, whereas *Physcomitrellopsis africana* has seven distinct Lhcb genes, *Physcomitrium patens* has 12, and *F. hygrometrica* has nine. Finally, *Physcomitrellopsis africana* has two distinct Lhcbm genes, *Physcomitrium patens* has 14, and *F. hygrometrica* has eight. However, echoing Kirbis *et al*. (2022), more light-harvesting complex (LHC) gene duplicates were retained in *Physcomitrium patens* than in *F. hygrometrica* and *Physcomitrellopsis africana* (Supplementary Fig. S3, Kirbis *et al*. 2022). Whereas *Physcomitrium patens* retained multiple LHC paralogs of possible WGD origin or gene duplications, *F. hygrometrica* and *Physcomitrellopsis africana* may have lost some of this gene redundancy. This pattern of differential retention is further supported by assessing the number of paralogs and orthologs for each LHC gene family across the three genomes (Supplementary Table S4). Although some copies of LHC genes were lost in *F. hygrometrica* and *Physcomitrellopsis africana*, key STN7 and STN8 kinases involved in photosynthetic

acclimation are conserved and retained in all three genomes, suggesting retention of core light signaling components.

Contaminant filtering and the identification of horizontal gene transfer events

In plants, horizontal transfer of genes, can occur at all phylogenetic depths, via introgression following hybridization between plant species (Pfennig 2021) to acquisition of genetic material from other domains (i.e., bacteria) or phyla of eukaryotes (e.g., fungi) (Soucy *et al*. 2015; Husnik and McCutcheon 2018). Identifying such transfers and hence the donors, is challenged by the occurrence of the plant's microbiome (Koutsovoulos *et al*. 2016; Francois *et al*. 2020) and hence requires critical screening of sequencing outputs. Filtering before and after the assembly of *Physcomitrellopsis africana* was employed, illustrating that iterative approaches at all stages enhanced the final quality of the reference. Here, metagenomic tools (Centrifuge; Kim *et al*. 2016), optimized for long reads, identified contaminants before assembly. Since this initial filter did not include the short reads or assess fungal contributions in either sequence set, further filtering was conducted after assembly and annotation. An estimated 22% of the genes were of fungal origin and 31% were of bacterial origin. Scaffolds that contained numerous genes originating from algae, bacteria, or fungi (ABF) were removed. Separately, the annotated gene space was aligned to a set of "donor" and "recipient" databases. This identified 31 potential horizontal gene transfer (HGT) events. Following the set of best practices outlined by Francois *et al*. (2020), an additional set of 10 scaffolds was found to be contaminated as the flanking genes of the candidates were not of plant origin. This resulted in their removal (HGT filter). Two unique HGT candidates survived these filters, and the final scaffolding (Fig. 4).

The first HGT candidate, *Pa1_19801.1* (alignment to C-type lectin mannose-binding isoform-like XP_032077963.1). C-type lectins are the most frequent binding sites across all plants and animals. In vertebrates, CTLs have a function of pathogen recognition. HGTs in the context of plant defense and immunity are not new and have been identified before (Yue *et al.* 2012; Ma *et al*. 2020). In fact, Ma *et al*. (2020) identified a lectin that was acquired from bacteria. The second candidate, *Pa1_12964.1*, aligns to a hypothetical protein found in *Endogone* sp. FLAS-F59071 (RUS16308.1). While there is no information on the function of this gene, transfers from fungi are not surprising. As mentioned previously, fungal killer proteins KP4 have

been acquired by mosses and play an important role in protonemal development (Guan *et al*. 2023).

These two candidates were manually validated by comparing alignments of de novo assembled transcriptomes. Both candidates are directly flanked by well-annotated plant (moss) genes. Additionally, aligning the full set of *Physcomitrium patens* and *Ceratodon purpureus* proteins to the *Physcomitrellopsis africana* scaffolds produced no alignments in proximity to the candidates, further suggesting that these HGT candidates are specific to *Physcomitrellopsis africana* (Fig. 5).

Analyzing HGT candidates from Physcomitrium patens: The genomes of *Physcomitrellopsis africana* and *F. hygrometrica* were screened for the 273 putative horizontally transferred genes (HGTs) previously identified in *Physcomitrium patens* (Yue *et al*. 2012; Ma *et al*. 2022). Ninety-one of these genes (33%) were not found in the other two species, whereas 15 (5%) were shared only by *Physcomitrellopsis africana* and *Physcomitrium patens*, and seven (3%) only by *Physcomitrium patens* and *F. hygrometrica* (Supplementary Fig. S4). The greater number of shared HGTs between *Physcomitrium patens* and *Physcomitrellopsis africana* likely reflects their more recent divergence (at least 40 MYA) compared to that of *Physcomitrium patens* and *F. hygrometrica* (at least 60 MYA) (Medina *et al*. 2018; Bechteler *et al*. 2023). These findings suggest that horizontal gene transfer may have played a role in the evolution of mosses, potentially facilitating their adaptation to diverse environments over their long evolutionary history dating back 500 million years (Bechteler *et al*. 2023). The life cycle of mosses, with stages where cells are exposed to microbes, such as the egg or zygote within the open archegonium or motile sperm cells, could provide opportunities for the uptake of foreign genetic material from bacteria or fungi (Yue *et al*. 2012; Huang 2013). These vulnerable stages raise the hypothesis that putative microbial symbionts or associates of mosses may have been sources of horizontally transferred genes, potentially aiding mosses in adapting to harsh environmental conditions encountered as some of the earliest terrestrial plants.

Data availability

All scripts and data are described in doi.org/10.5281/zenodo.11094703. NCBI BioProject ID PRJNA1020579 contains the genomic short reads and nanopore long reads (SRR26596311, SRR26596310 and SRR26666216) and the RNA reads (SRR26586950), as well as the whole

genome assembly (GCA_036785485.1). The *de novo* assembled transcripts and annotation files are available at dx.doi.org/10.6084/m9.figshare.25724079.

Acknowledgements

The authors would like to thank the Institute for Systems Genomics (ISG) at the University of Connecticut, including the Center for Genome Innovation for sequencing support and the Computational Biology Core for software support and access to high-performance computing.

Funding

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

Conflict of interest

The authors declare no conflict of interest.

Author contributions

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

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. doi: 10.1371/journal.pone.0002033.

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. doi: 10.1093/nar/25.17.3389.

Aminov, R. I. 2011. Horizontal gene exchange in environmental microbiota. Front Microbiol. 2:158. doi: 10.3389/fmicb.2011.00158.

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. Am J Bot. 110(11): e16249. doi: 10.1002/ajb2.16249

Bi, G., Zhao, S., Yao, J., Wang, H., Zhao, M., Sun, Y., Hou, X., Hass, F. B., Varshney, D., Prigge, M., Rensing, S., Jiao, Y., Ma, Y., Yan, J., Dai, J. 2024. Near telomere-to-telomere genome of the model plant Physcomitrium patens. Nat Plants. 10:327–343. doi: 10.1038/s41477-023-01614-7.

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. Proc Natl Acad Sci U S A. 112(52):15976–15981. doi: 10.1073/pnas.1510461112.

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 Genom Bioinform. 3(1):lqaa108. doi: 10.1093/nargab/lqaa108.

Buchfink, B., Reuter, K., Drost, H.-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 18(4), 366–368. doi: 10.1038/s41592-021-01101-x.

Chen, H., Zwaenepoel, A. 2023. Inference of Ancient Polyploidy from Genomic Data. *In Polyploidy: Methods and Protocols* (pp. 3–18). New York, NY: Springer US.

Dainat J. 2020. Another Gff analysis toolkit to handle annotations in any GTF/GFF format (v0.8.0). Zenodo. doi:10.5281/zenodo.3552717

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

Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B., Nestor, C. E. 2020. No evidence for DNA N 6-methyladenine in mammals. Sci Adv. 6(12):eaay3335. doi: 10.1126/sciadv.aay3335.

Emms, D. M., Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238. doi: 10.1186/s13059-019-1832-y.

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. doi: 10.1093/nar/gkt1223.

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. Proc Natl Acad Sci U S A. 117(17):9451–9457. doi: 10.1073/pnas.1921046117.

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. doi: 10.1534/g3.119.400758.

Husnik, F., McCutcheon, J.P. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. 16(2):67–79. doi: 10.1038/nrmicro.2017.137.

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., Stanke, M. 2021. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 22(1):566. doi: 10.1186/s12859-021-04482-0.

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. Nat Biotechnol. 29(7): 644–652. doi: 10.1038/nbt.1883.

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. Plant J. 113(4):665–676. doi: 10.1111/tpj.16060.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072–1075. doi: 10.1093/bioinformatics/btt086.

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. doi: 10.1016/j.isci. 2020.101389.

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. Mol Ecol Resour. 20(2):591–604. doi: 10.1111/1755-0998.13106.

Huang, J. 2013. Horizontal gene transfer in eukaryotes: the weak-link model. Bioessays. 35(10):868–875. doi: 10.1002/bies.201300007.

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. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 47(D1):D309–D314. doi: 10.1093/nar/gky1085.

Iwai, M., Grob, P., Iavarone, A. T., Nogales, E., Niyogi, K. K. 2018. A unique supramolecular organization of photosystem I in the moss Physcomitrella patens. Nat Plants. 4(11):904–909. doi: 10.1038/s41477-018-0271-1.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236–1240. doi: 10.1093/bioinformatics/btu031.

Joshi, N. A. Fass J. N. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). https://github.com/najoshi/sickle.

Katoh, K., Toh, H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9(4):286–298. doi: 10.1093/bib/bbn013.

Kim, D., Song, L., Breitwieser, F. P., Salzberg, S. L. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26(12):1721–1729. doi: 10.1101/gr.210641.116.

Kim, D., Paggi, J. M., Park, C., Bennett, C., Salzberg, S. L. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37(8):907–915. doi: 10.1038/s41587-019-0201-4.

Kirbis, A., Rahmatpour, N., Dong, S., Yu, J., van Gessel, N., Waller, M., Reski, R., Lang, D., Rensing, S. A., Temsch, E. M., Wegrzyn, J. L., Goffinet, B., Liu, Y., Szövényi, P. 2022. Genome dynamics in mosses: Extensive synteny coexists with a highly dynamic gene space. bioRxiv p. 2022.05.17.492078. doi: 10.1101/2022.05.17.492078.

Kirbis, A., Waller, M., Ricca, M., Bont, Z., Neubauer, A., Goffinet, B., Szövényi, P. 2020.

Transcriptional landscapes of divergent sporophyte development in two mosses, Physcomitrium (Physcomitrella) patens and Funaria hygrometrica. Front Plant Sci. 11:747. doi: 10.3389/fpls.2020.00747.

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5):540–546. doi: 10.1038/s41587-019-0072-8

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. Proc Natl Acad Sci U S A. 113(18):5053–5058. doi: 10.1073/pnas.1600338113.

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 Biol. 20(1):278. doi: 10.1186/s13059-019-1910-1.

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. Plant J. 93(3):515–533. doi: 10.1111/tpj.13801.

Li, B., Dewey, C. N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. doi: 10.1186/1471-2105-12-323.

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. doi: 10.1016/j.molp.2022.02.001.

Manni, M., Berkeley, M. R., Seppey, M., Zdobnov, E. M. 2021. BUSCO: Assessing genomic data quality and beyond. Current Protocols. 1(12):e323. doi: 10.1002/cpz1.323.

Marçais, G., Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27(6):764–770. doi: 10.1093/bioinformatics/btr011.

Martin, F. M., Uroz, S., Barker, D. G. 2017. Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. Science. 356(6340):ea.ad4501. doi: 10.1126/science.aad4501.

Medina R, Johnson MG, Liu Y, Wilding N, Hedderson TA, Wickett N, Goffinet B. 2018. Evolutionary dynamism in bryophytes: Phylogenomic inferences confirm rapid radiation in the moss family Funariaceae. Mol. Phylogenet Evol. 120:240–247. doi:

10.1016/j.ympev.2017.12.002.

Medina, R., M. G. Johnson, Y. Liu, N.J. Wickett, A.J. Shaw B. Goffinet. 2019. Phylogenomic delineation of Physcomitrium (Bryophyta: Funariaceae) based on nuclear targeted exons and their flanking regions rejects the retention of Physcomitrella, Physcomitridium and Aphanorrhegma. J Syst Evol. 57:404–417. doi: 10.1111/jse.12516.

Pfennig KS. 2021. Biased hybridization and its impact on adaptive introgression. Trends Ecol Evol. 36(6): 488–497. doi: 10.1016/j.tree.2021.02.010.

Price, M. N., Dehal, P. S., Arkin, A. P. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. PloS One. 5(3):e9490. doi: 10.1371/journal.pone.0009490.

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., Vandepoele, K. 2011. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Research. 40(2):e11–e11. doi: 10.1093/nar/gkr955.

Rahmatpour, N., Perera, N. V., Singh, V., Wegrzyn, J. L., Goffinet, B. 2021. High gene space divergence contrasts with frozen vegetative architecture in the moss family Funariaceae. Mol Phylogenet Evol. 154:106965. doi: 10.1016/j.ympev.2020.106965.

Ranallo-Benavidez, T. R., Jaron, K. S., Schatz, M. C. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Comm. 11(1):1432. doi: 10.1038/s41467-020-14998-3.

Rensing, S. A., Goffinet, B., Meyberg, R., Wu, S.-Z., Bezanilla, M. 2020. The moss Physcomitrium (Physcomitrella) patens: A model organism for non-seed plants. Plant Cell. 32(5):1361–1376. doi: 10.1105/tpc.19.00828.

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S.-I., … Boore, J. L. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science. 319(5859):64–69. doi: 10.1126/science.1150646.

Rhie, A., Walenz, B. P., Koren, S., Phillippy, A. M. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21(1):245. doi: 10.1186/s13059-020-02134-9.

Roach, M. J., Schmidt, S. A., Borneman, A. R. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19(1):460.

doi: 10.1186/s12859-018-2485-7.

Smit, A. F. A., Hubley, R., Green, P. 2013-2015. RepeatMasker Open-4.0. RepearMasker. http://www.repeatmasker.org

Soucy, S.M., Huang, J., Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet. 16(8):472–482. doi: 10.1038/nrg3962.

Sun, G., Bai, S., Guan, Y., Wang, S., Wang, Q., Liu, Y., Liu, H., Goffinet, B., Zhou, Y., Paoletti, M., Hu, X., Haas, F. B., Fernandez-Pozo, N., Czyrt, A., Sun, H., Rensing, S. A., Huang, J. 2020. Are fungi-derived genomic regions related to antagonism towards fungi in mosses? New Phytol. 228(4):1169–1175. doi: 10.1111/nph.16776.

Sun, H., Shang, H., Pan, X., Li, M. 2023. Structural insights into the assembly and energy transfer of the Lhcb9-dependent photosystem I from moss Physcomitrium patens. Nat Plants. 9(8):1347–1358. doi: 10.1038/s41477-023-01463-4.

van Dongen, S. M. 2000. Graph clustering by flow simulation. Doctoral dissertation. Utrecht University, Utrecht, The Netherlands.

Van Etten, J., Bhattacharya, D. 2020. Horizontal gene transfer in Eukaryotes: Not if, but how much? Trends Genet. 36(12):915–925. doi: 10.1016/j.tig.2020.08.006.

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., Zaman, S., Bennett, J., McEvoy, S. L., Caballero, M., Wegrzyn, J. L. 2022. Welcome to the big leaves: best practices for improving genome annotation in non-model plant genomes. Applications in Plant Sciences. 11(4):e11533. doi: 10.1002/aps3.11533

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., Earl, A. M. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One. 9(11):e112963. doi: 10.1371/journal.pone.0112963.

Wilding, N. 2015. Systematics, biogeography and morphological evolution in Entosthodon Schwägr. (Bryopsida, Funariaceae) with a revision of the genus in Africa. Doctoral dissertation. University of Cape Town, South Africa.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 24(8):1586–1591. doi: 10.1093/molbev/msm088.

Young, L. A. 2022. Relationships among AA-genome Chenopodium diploids and a whole genome assembly of the North American species, C. watsonii. Doctoral dissertation.

Brigham Young University, Utah, USA.
Young, M. D., Wakefield, M. J., Smyth, G. K., Oshlack, A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11(2):R14. doi: 10.1186/gb-2010-11-2-r14.
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. doi: 10.46471/gigabyte.8.
Yue, J., Hu, X., Sun, H., Yang, Y. and Huang, J., 2012. Widespread impact of horizontal gene transfer on plant colonization of land. Nat Commun. 3(1):1152. doi: 10.1038/ncomms2148.
Zhaxybayeva, O., Doolittle, W. F. 2011. Lateral gene transfer. Curr Biol. 21(7):R242-R246.
Zhang, Z., Wood, W. I. 2003. A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics 19(2):307–308. doi: 10.1093/bioinformatics/19.2.307.
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. doi: 10.1093/bioinformatics/btt476.
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.a. doi: 10.1186/1471-2164-14-498.

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. The assembly was then scaffolded against the *Physcomitrium patens* V4 genome using RagTag, capturing 94% of the total assembly across 26 chromosomes. Summary statistics derived from Quast (N50 - scaffold/contig length at which 50% of the genome is contained in scaffolds/contigs of this size or greater), BUSCO, and Merqury (QV - Quality Value, an assessment of assembly contiguity) are displayed for each process.

Fig. 3. a) The total shared and unique orthogroups among *Funaria 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 *F. 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. a) Contamination versus horizontal gene transfer (HGT) in the *Physcomitrellopsis africana* genome. 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 because of this filtering process. Sixty-three scaffolds contained at least one gene annotated as a contaminant but did not meet the original thresholds for removing full scaffolds. When the chromosome-scale assembly was generated with RagTag scaffolding, three of these scaffolds were not placed in the 26 chromosomes. b) Identification of HGT candidates via sequence similarity comparisons of *Physcomitrellopsis africana* proteins against

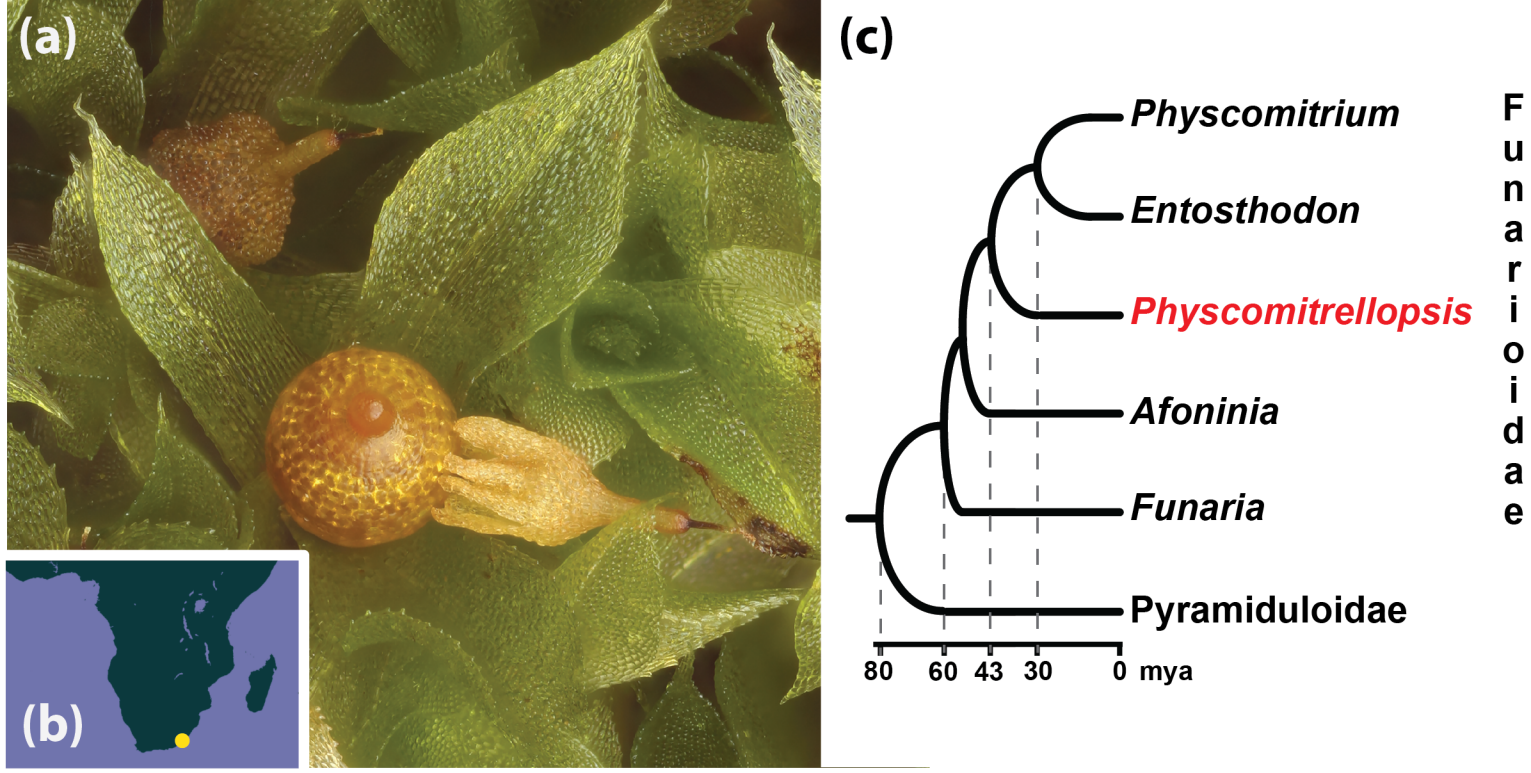
both "donor" databases (archaea, bacteria, fungal, and metazoan) and "recipient" databases (Streptophyta, Tracheophyta, and Spermatophyta). Proteins with >1 and <4 hits against all donor databases, and no hit against recipient databases were labeled as HGT candidates. This analysis was conducted on the 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.

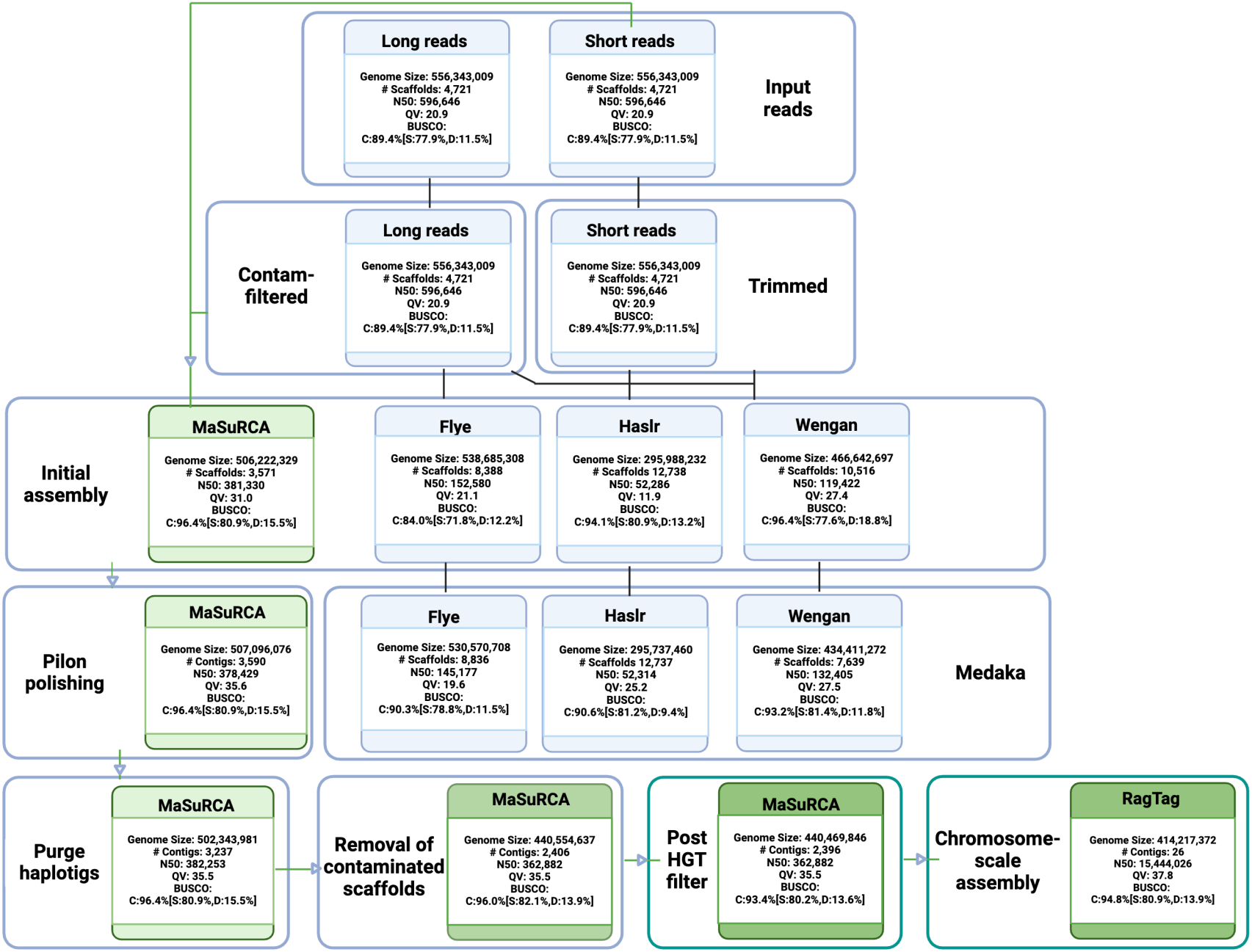
Fig. 5. Integrated Genome Viewer (IGV) screens depicting five tracks of the *Physcomitrellopsis africana* genome. Track 1 shows the protein-coding structural annotation in context to the genome. Track 2 displays genome-guided transcript assemblies via StringTie2. Track 3 illustrates the alignment of the Horizontal Gene Transfer (HGT) candidate (nr database) to the genome. Tracks 4 and 5 show the alignments of *Physcomitrium patens* and *Ceratodon purpureus* proteins onto chromosomes 2 and 26. a) and b) show the HGT candidates *Pa1_04828.1* and *Pa1_22336.1*, where the HGT candidate alignment validates the presence of a protein. In both cases, independent RNA assemblies, via StringTie2, were not able to generate a supporting model. None of the other moss proteins align to the HGT candidates. c) highlights the example of *Pa1_15362.1*, where the StringTie2 transcript and the protein 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.

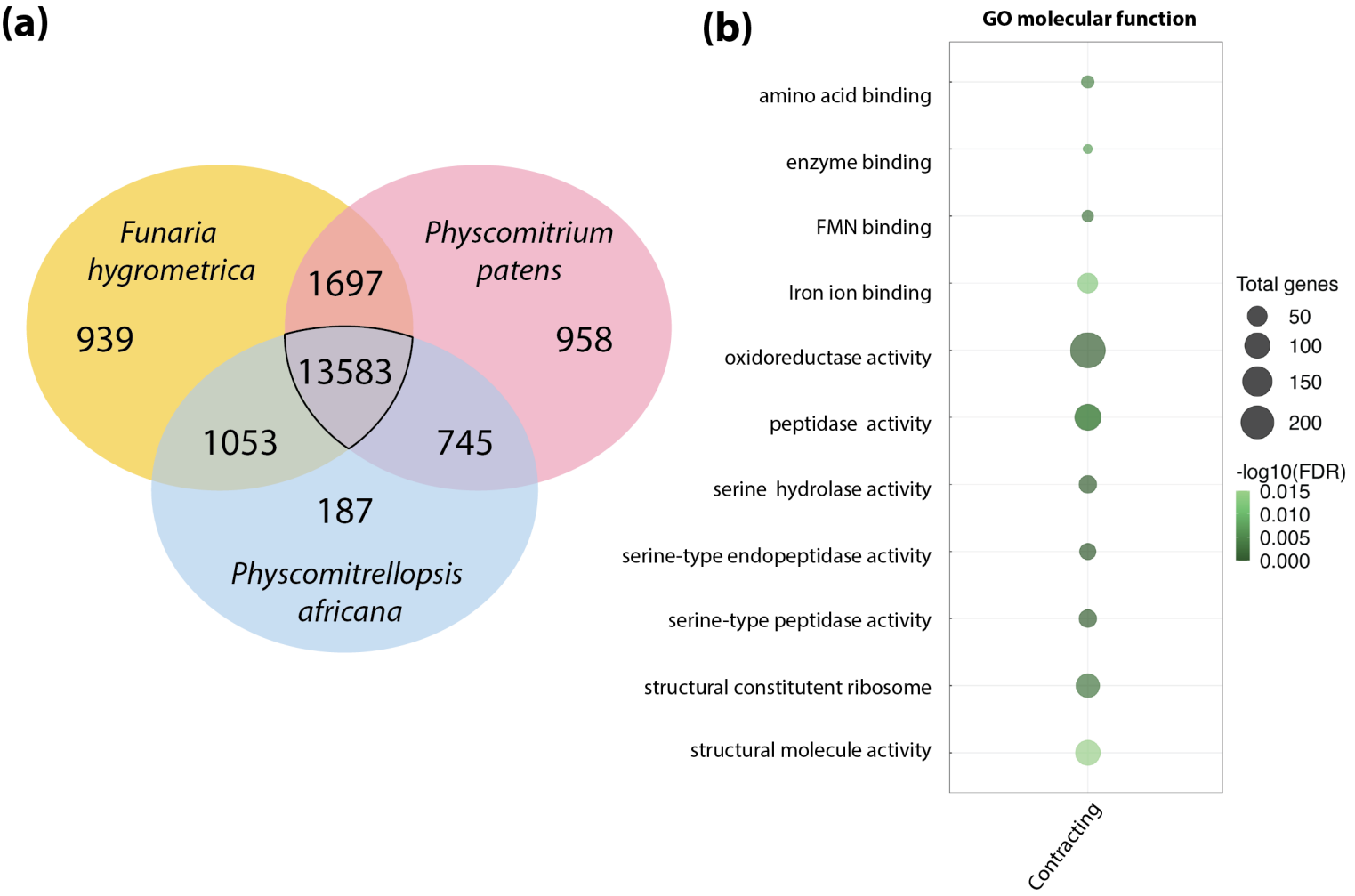
Table 1. Genome annotation statistics for *Physcomitrellopsis africana*.

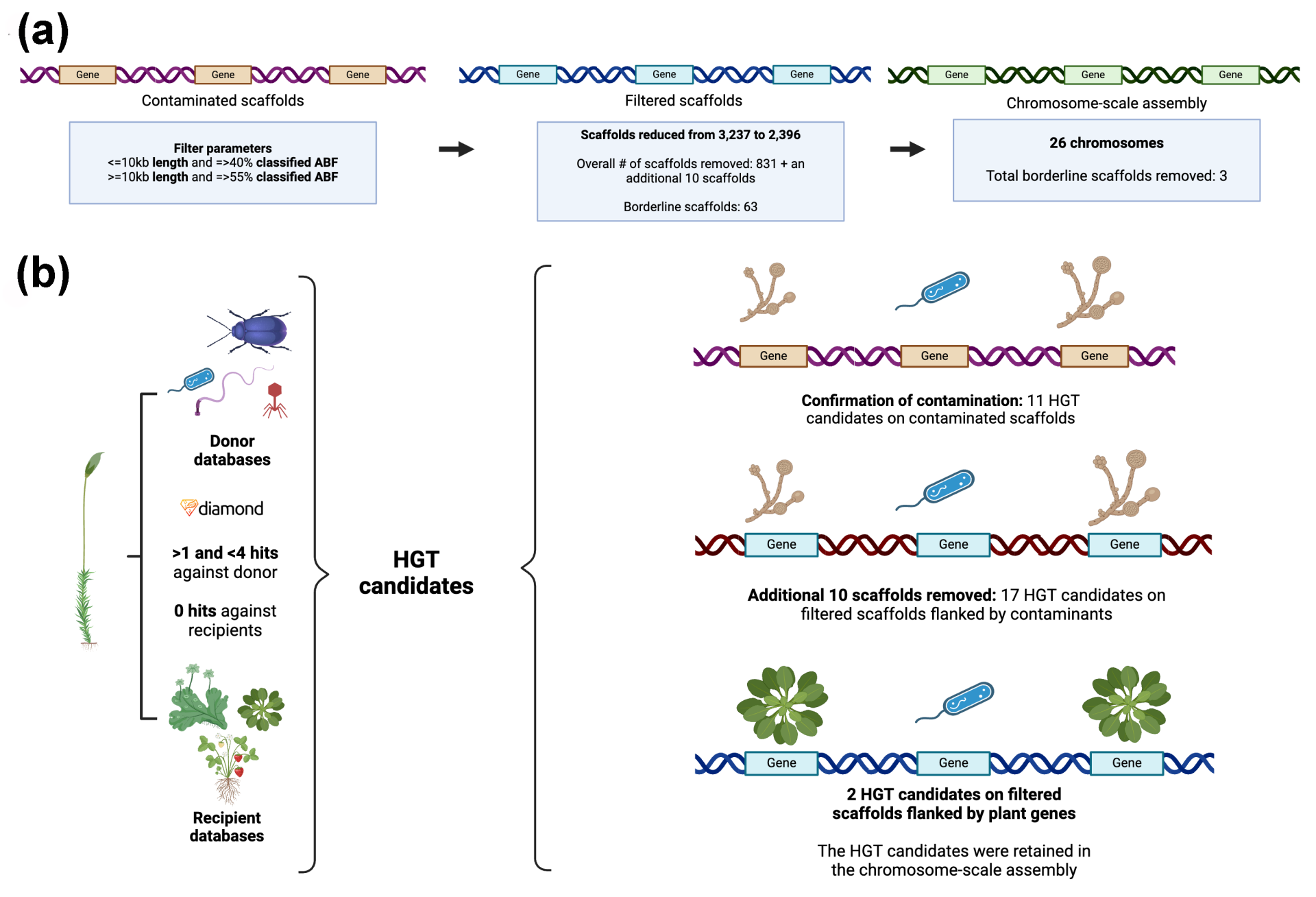
Annotation Method	Total genes	BUSCO (viridiplantae)	Mono: multi ratio	Annotation rate
BRAKER (RNA)	60,917	C:95.8%[D:15.3]	0.52	65%
BRAKER (protein)	37,752	C:46.1%[D:16.0]	0.95	73%
TSEBRA (BRAKER RNA + BRAKER protein)	45,737	C:91.3%[D:15.1]	1.02	74%

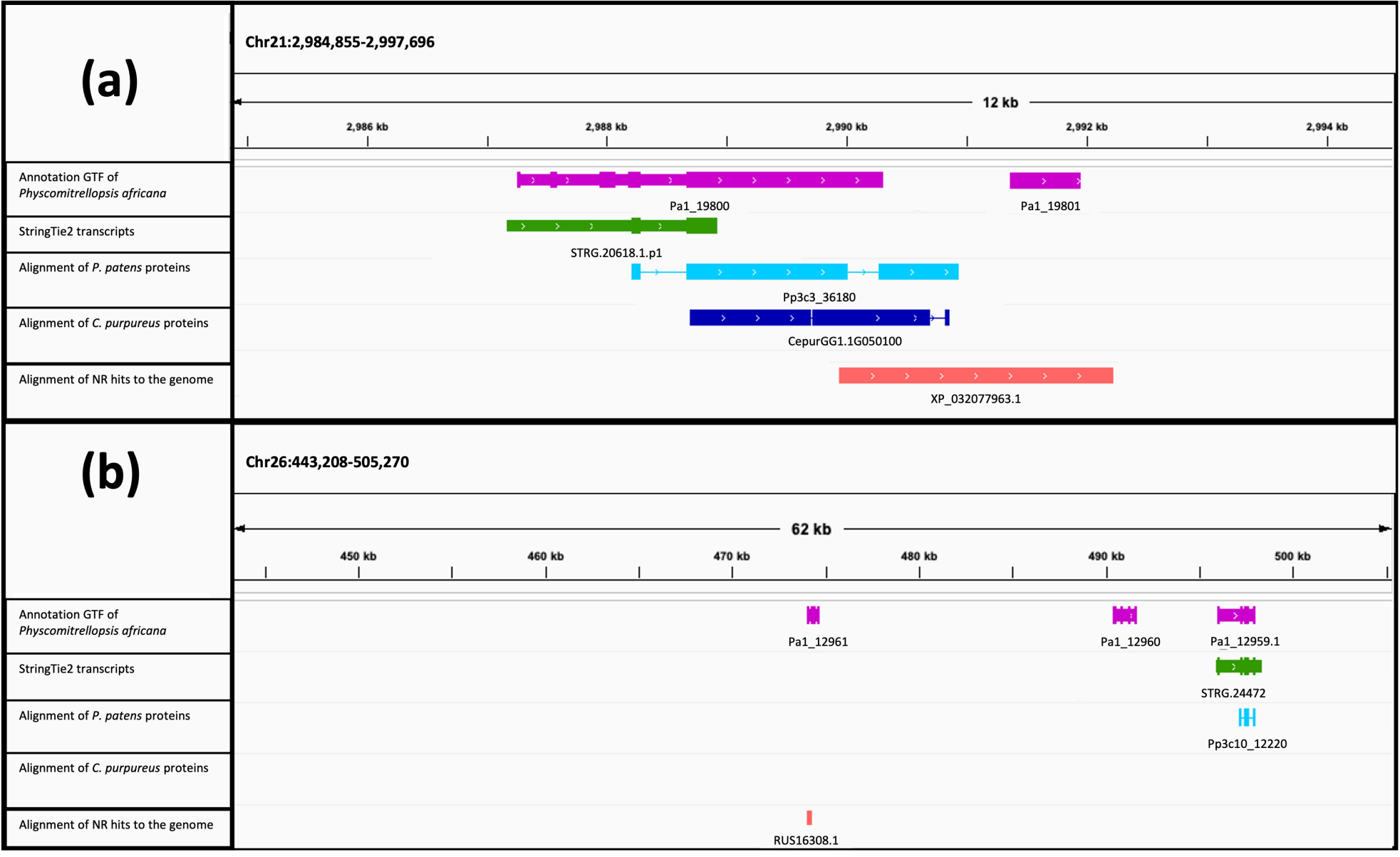
BRAKER (RNA) + InterProScan filter + scaffold contam filter)	23,561	C:93.8%[D:13.6%]	0.08	78%
BRAKER (RNA) + InterProScan filter + scaffold contam filter + HGT filter)	23,535	C:88.5%[D:11.3%]	0.07	83%
Scaffolded to 26 chromosomes with RagTag	22,925	C:86.4%[D:11.1%]	0.07	75%

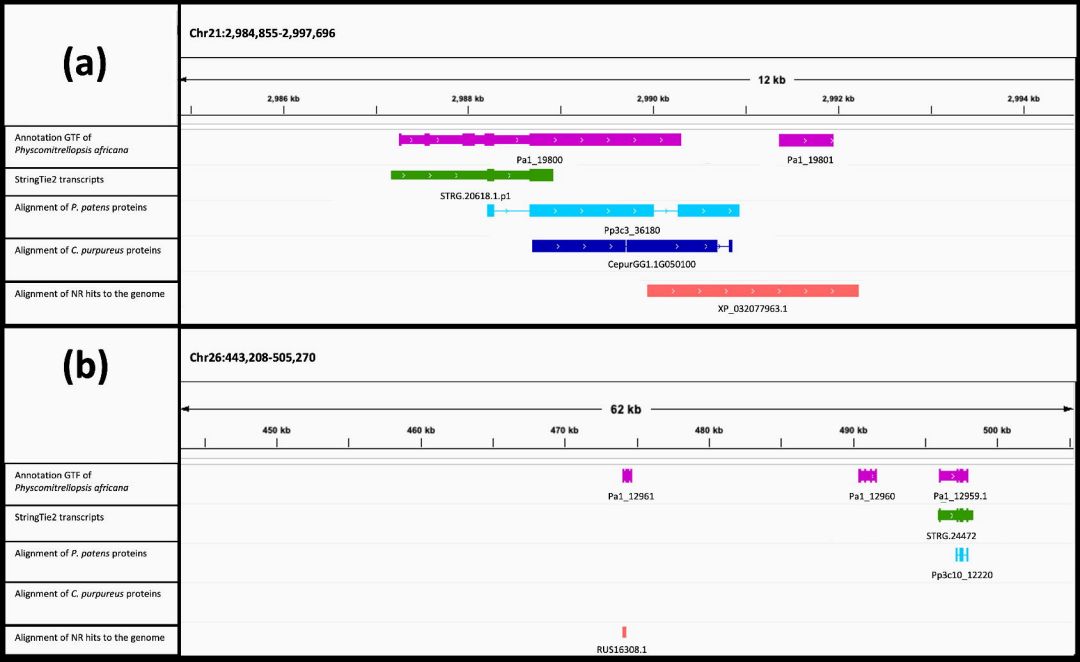

(a)
(b)
(c)
Physcomitrium
Entosthodon
Physcomitrellopsis
Afoninia
Funaria
Pyramiduloidae
Funarioidae
80
60
43
30
0
mya

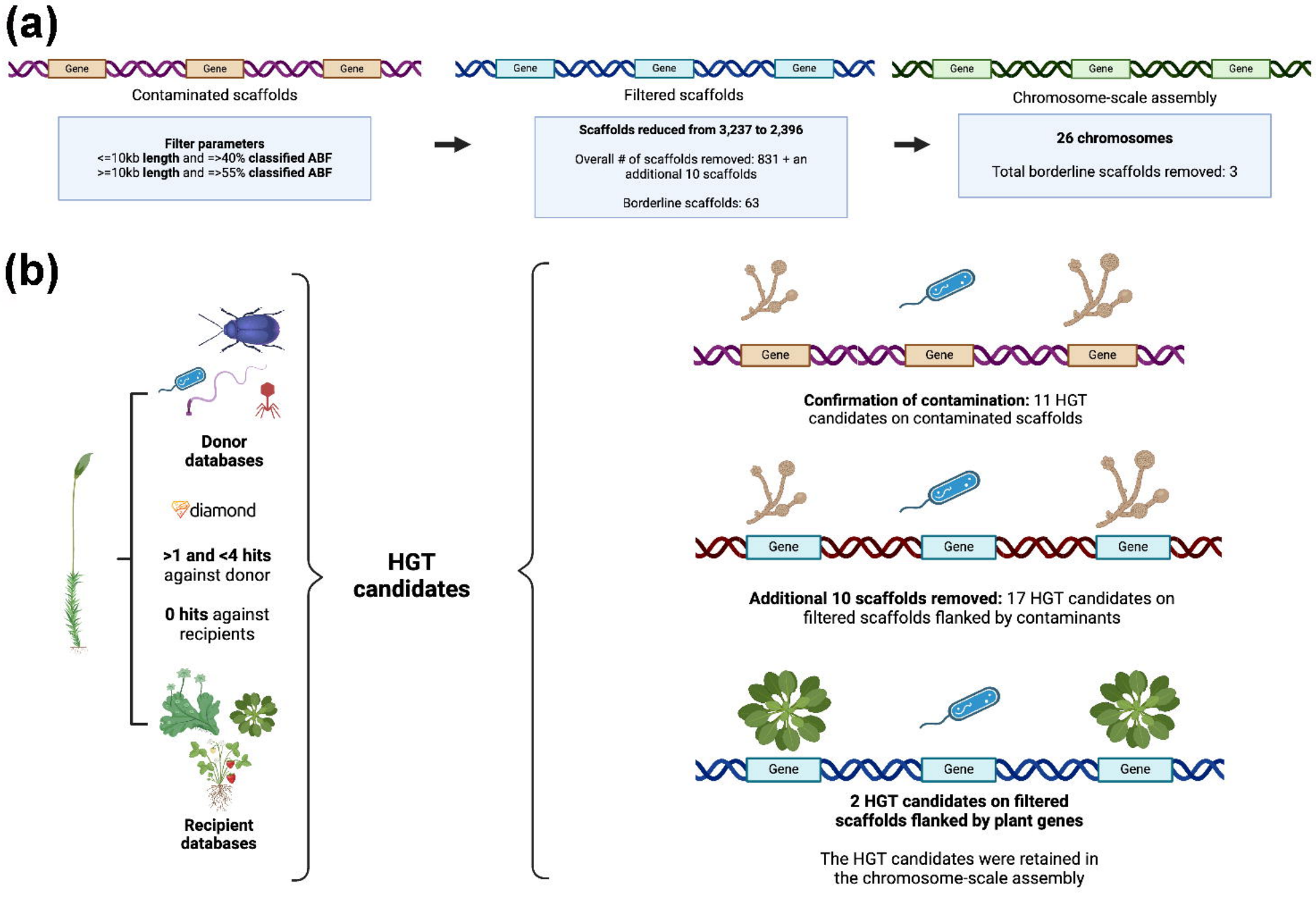

Input reads
Long reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Short reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Contam-filtered
Long reads
Genome Size: 556,343,009
Scaffolds: 4,721
N50: 596,646
QV: 20.9
BUSCO:
C:89.4%[S:77.9%,D:11.5%]
Trimmed
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%]
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%]
Medaka
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%]
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%]
Chromosome-scale assembly
RagTag
Genome Size: 414,217,372
Contigs: 26
N50: 15,444,026
QV: 37.8
BUSCO:
C:94.8%[S:80.9%,D:13.9%]

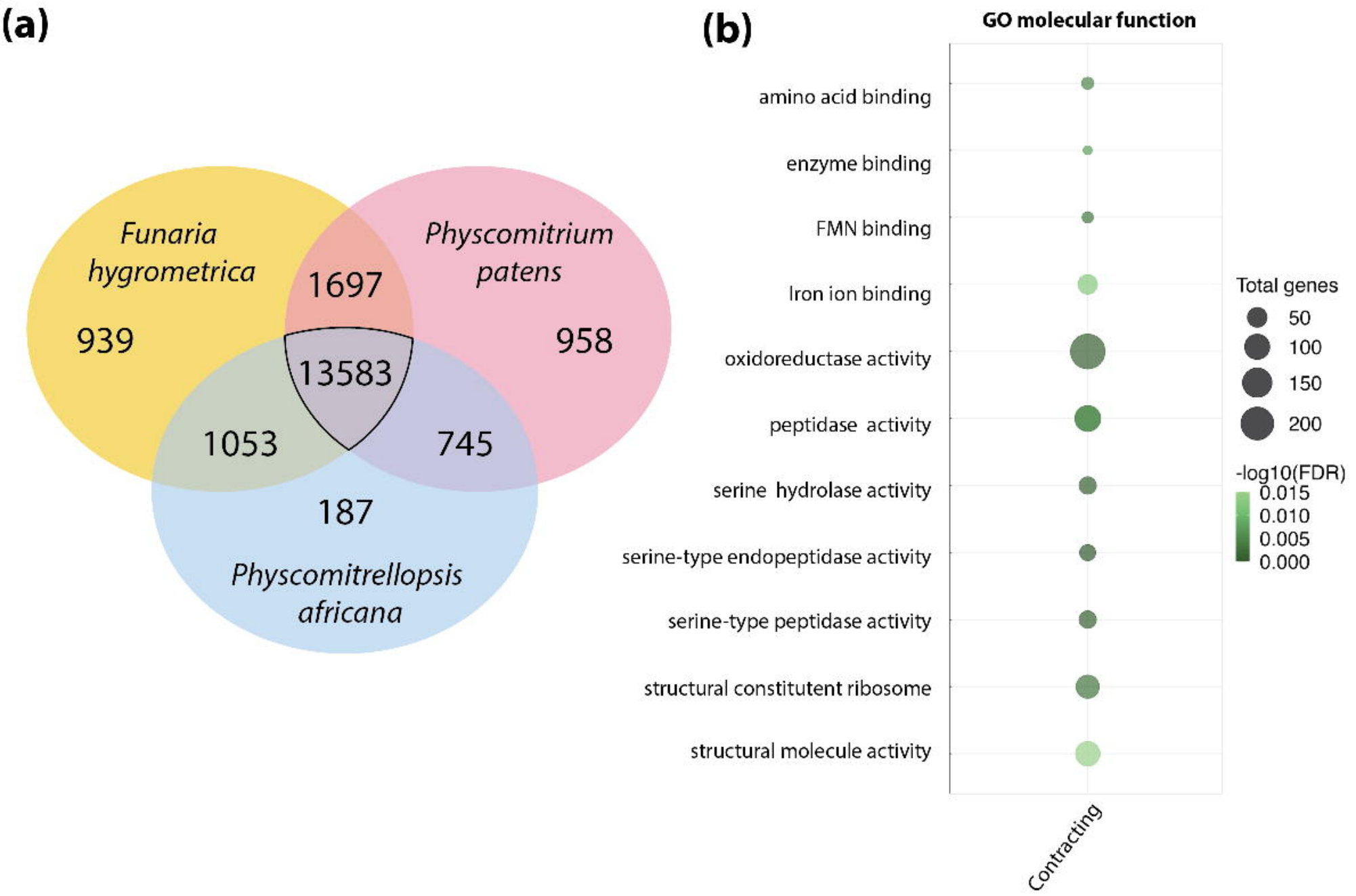

(a)
Funaria hygrometrica
Physcomitrium patens
Physcomitrellopsis africana
939
1697
958
13583
1053
745
187
(b)
GO molecular function
amino acid binding
enzyme binding
FMN binding
Iron ion binding
oxidoreductase activity
peptidase activity
serine hydrolase activity
serine-type endopeptidase activity
serine-type peptidase activity
structural constitutent ribosome
structural molecule activity
Contracting
Total genes
50
100
150
200
-log10(FDR)
0.015
0.010
0.005
0.000

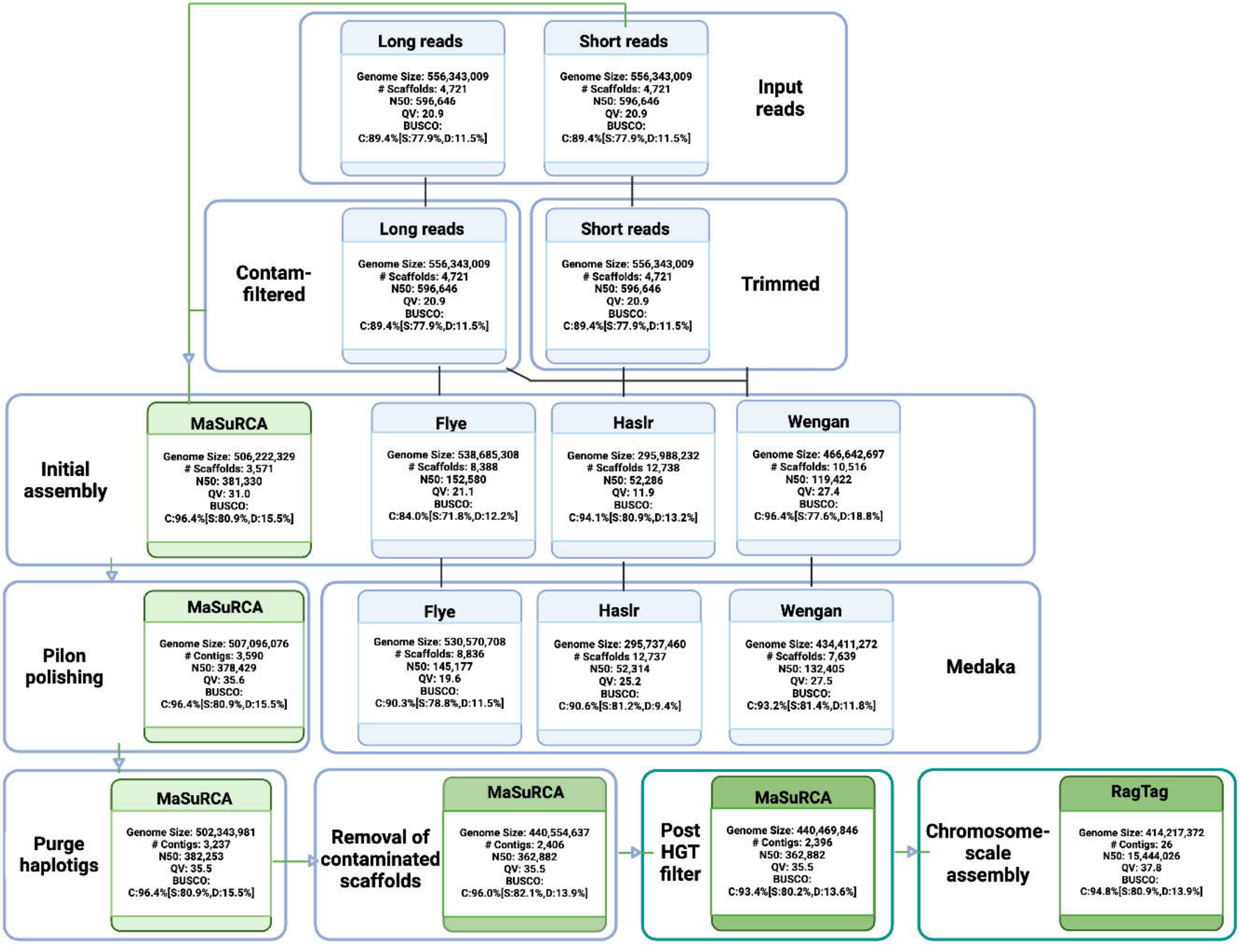

(a)
Gene
Contaminated scaffolds
Filter parameters
<=10kb length and =>40% classified ABF
>=10kb length and =>55% classified ABF
Filtered scaffolds
Scaffolds reduced from 3,237 to 2,396
Overall # of scaffolds removed: 831 + an additional 10 scaffolds
Borderline scaffolds: 63
Chromosome-scale assembly
26 chromosomes
Total borderline scaffolds removed: 3
(b)
Donor databases
diamond
>1 and <4 hits against donor
0 hits against recipients
Recipient databases
HGT candidates
Confirmation of contamination: 11 HGT candidates on contaminated scaffolds
Additional 10 scaffolds removed: 17 HGT candidates on filtered scaffolds flanked by contaminants
2 HGT candidates on filtered scaffolds flanked by plant genes
The HGT candidates were retained in the chromosome-scale assembly

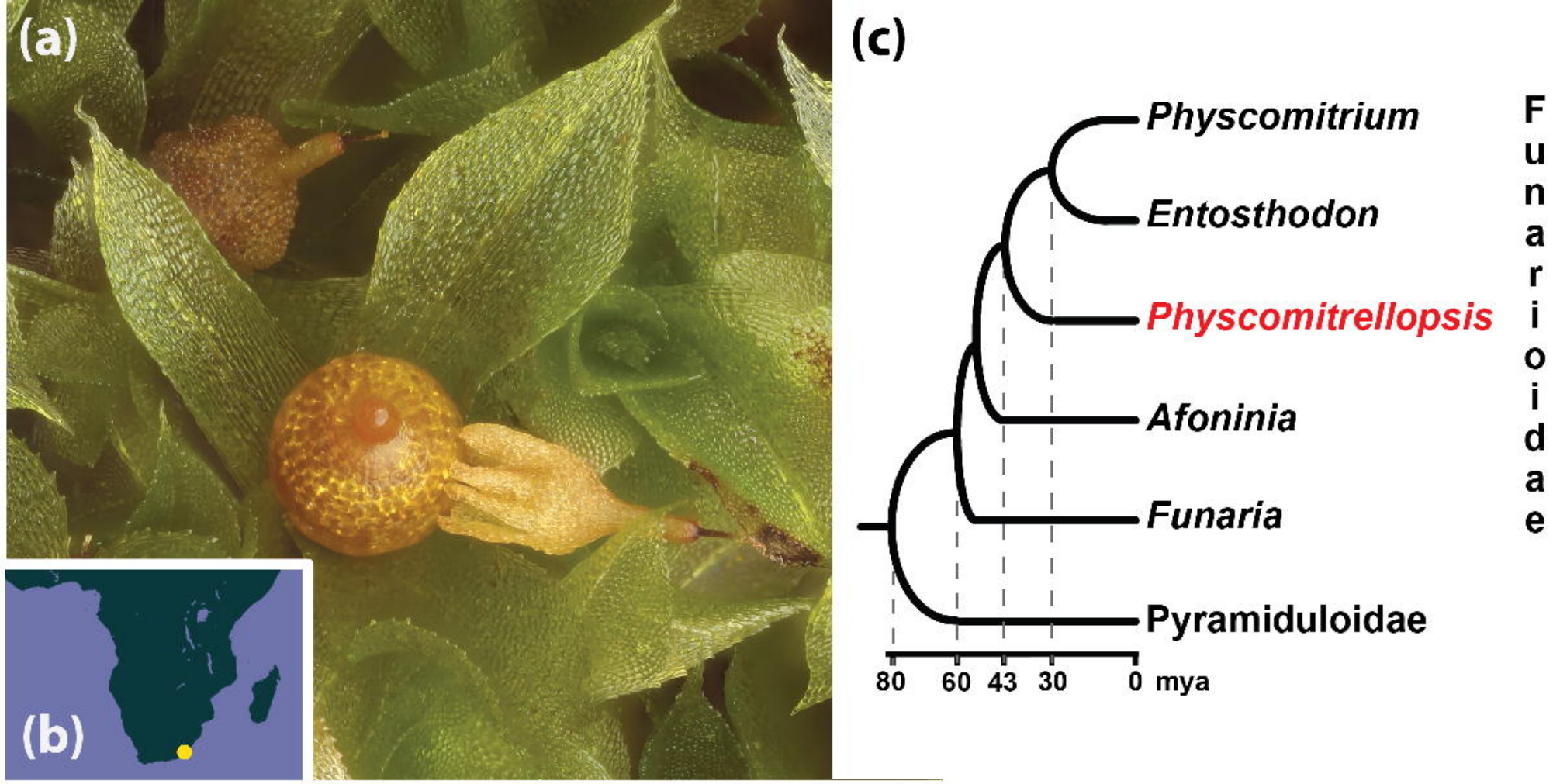

(a)
Chr21:2,984,855-2,997,696
12 kb
2,986 kb
2,988 kb
2,990 kb
2,992 kb
2,994 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_19800
Pa1_19801
StringTie2 transcripts
STRG.20618.1.p1
Alignment of P. patens proteins
Pp3c3_36180
Alignment of C. purpureus proteins
CepurGG1.1G050100
Alignment of NR hits to the genome
XP_032077963.1
(b)
Chr26:443,208-505,270
62 kb
450 kb
460 kb
470 kb
480 kb
490 kb
500 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_12961
Pa1_12960
Pa1_12959.1
StringTie2 transcripts
STRG.24472
Alignment of P. patens proteins
Pp3c10_12220
Alignment of C. purpureus proteins
Alignment of NR hits to the genome
RUS16308.1


(a)
Chr21:2,984,855-2,997,696
12 kb
2,986 kb
2,988 kb
2,990 kb
2,992 kb
2,994 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_19800
Pa1_19801
StringTie2 transcripts
STRG.20618.1.p1
Alignment of P. patens proteins
Pp3c3_36180
Alignment of C. purpureus proteins
CepurGG1.1G050100
Alignment of NR hits to the genome
XP_032077963.1
(b)
Chr26:443,208-505,270
62 kb
450 kb
460 kb
470 kb
480 kb
490 kb
500 kb
Annotation GTF of Physcomitrellopsis africana
Pa1_12961
Pa1_12960
Pa1_12959.1
StringTie2 transcripts
STRG.24472
Alignment of P. patens proteins
Pp3c10_12220
Alignment of C. purpureus proteins
Alignment of NR hits to the genome
RU516308.1


(a)
Gene
Contaminated scaffolds
Filter parameters
<=10kb length and =>40% classified ABF
>=10kb length and =>55% classified ABF
Filtered scaffolds
Scaffolds reduced from 3,237 to 2,396
Overall # of scaffolds removed: 831 + an additional 10 scaffolds
Borderline scaffolds: 63
Chromosome-scale assembly
26 chromosomes
Total borderline scaffolds removed: 3
(b)
Donor databases
diamond
>1 and <4 hits against donor
0 hits against recipients
Recipient databases
HGT candidates
Confirmation of contamination: 11 HGT candidates on contaminated scaffolds
Additional 10 scaffolds removed: 17 HGT candidates on filtered scaffolds flanked by contaminants
2 HGT candidates on filtered scaffolds flanked by plant genes
The HGT candidates were retained in the chromosome-scale assembly


(a)
Funaria hygrometrica
Physcomitrium patens
Physcomitrellopsis africana
939
1697
958
13583
1053
745
187
(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 constitutent ribosome
structural molecule activity
Contracting
Total genes
50
100
150
200
-log10(FDR)
0.015
0.010
0.005
0.000


Input reads
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%]
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%]
Trimmed
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%]
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%]
Medaka
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%]
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%]
Chromosome-scale assembly
RagTag
Genome Size: 414,217,372
Contigs: 26
N50: 15,444,026
QV: 37.8
BUSCO:
C:94.8%[S:80.9%,D:13.9%]


(a)
(b)
(c)
Physcomitrium
Entosthodon
Physcomitrellopsis
Afoninia
Funaria
Pyramiduloidae
80
60
43
30
0 mya
Funarioidae