nature genetics

Supplementary information

https://doi.org/10.1038/s41588-025-02325-9

Bryophytes hold a larger gene family space than vascular plants

In the format provided by the authors and unedited

Supplementary Notes

Bryophytes hold a larger gene family space than vascular plants

Content

1. Bryophyte genomes and the monophyly of bryophytes.....3

2. Structural evolution of the bryophyte genomes 3

3. Comparative genomics and gene family evolution of bryophytes 5

4. Unique resistance (R) genes in bryophytes 6

5. WGD analysis.....7

6. Identification of NLR genes 9

7. Stress experiments and differential gene expression analysis.....10

8. Evaluation of the efficacy of genome decontamination strategy 11

9. Assessment of the correlation between genome quality, gene numbers and different sequencing platforms.....14

10. Assessment of the impact of genome sampling on gene family clustering results 15

11. Evaluation of the reliability of HGT analyses and identification 16

10. Supplementary Figures ……...25-71

1. Bryophyte genomes and the monophyly of bryophytes

We sequenced, assembled, and annotated 123 new genomes of bryophytes (compiled in a centralized platform: www.bryogenomes.org), including 37 liverworts, 82 mosses, and four hornworts, representing 47 out of 55 bryophyte orders (Fig. 1 and Extended Data Fig. 1), generated by a combination of Nanopore, 10X genomics, stLFR technologies, and Hi-C approaches (details see Methods, Supplementary Data 1). Our assemblies are comparable in quality (Supplementary Data 2) to previously published bryophyte genomes [1-4]. These genomes hold an average of 94% of BUSCO genes, suggesting a high level of completeness of the captured gene space (Supplementary Data 2). Genome annotation predicted on average 28,462 gene models in bryophyte species (17,508, 28,092, and 29,163 in hornworts, liverworts, and mosses, respectively) (Supplementary Data 2). This is generally less than the number of genes found in vascular plants (34,794 genes, Supplementary Fig. 16). We developed a bryophyte lineage-specific benchmark gene set (bryophyte_odb10) based on 27 chromosome-level bryophyte genomes [5].

Bryophyte genomes are relatively small among land plants, with an average size of 110 Mb, 414 Mb, and 1,031 Mb in hornworts, mosses, and liverworts, respectively (Fig. 1 and Supplementary Data 2), and range from 87 Mb for the hornwort *Folioceros fuciformis* to 5,039 Mb for the liverwort *Treubia lacunosa*. Repeat content contributes much to the variation in genome size, with the lowest average proportion observed in hornworts (31%) and the highest in liverworts (58%). Although liverworts possess larger genomes, they encode shorter genes (1,468 bp on average) than mosses (1,710 bp on average) or hornworts (1,732 bp on average). These differences are likely the result of liverwort genomes carrying fewer introns (on average one per gene) than those of mosses and hornworts (on average two or three per gene). Hornworts possess the least genes and smallest introns (Supplementary Data 3), with an average length of almost half that of liverworts and mosses (143 bp in the former *vs* 253 bp and 255 bp in the latter), in line with their compact genomes (Fig. 1, Supplementary Data 3).

The comprehensive phylogenetic sampling of bryophyte genomes provided an opportunity to revisit the phylogeny of land plants (Supplementary Data 1 & 4), and resolved bryophytes as monophyletic and sister to extant tracheophytes (Fig. 1 and Extended Data Fig. 2). Within bryophytes, hornworts are sister to the setaphytes (i.e., liverworts plus mosses) [6,7]. The relationships within each major lineage of bryophytes are largely congruent with recent phylogenomic studies [6-10].

2. Structural evolution of the bryophyte genomes

Our pseudo-chromosome level genome assemblies revealed that bryophytes share a general genome organization with genes and transposons distributed evenly across chromosomes (Supplementary Fig.

1 & 38). This extends the earlier observations based on genomes from bryophyte model species [3,4,11] to genomes from across the entire bryophyte tree of life (Supplementary Fig. 1a), and strengthens the hypothesis that this feature represents the ancestral state in green plants [12,13]. In angiosperms, chromosomes are typically characterized by a gene-poor centromere flanked by transposon-rich pericentromeric regions [14,15]. In contrast, the centromeres of mosses, e.g., *Ceratodon purpureus*, *Physcomitrium patens*, and *Sphagnum angustifolium* are enriched in *RLC5*, a specific *copia* element [4,16,17], although this structure is not seen in all mosses (Supplementary Fig. 38).

DNA methylation is distributed quite evenly across chromosomes of bryophytes (see Supplementary Fig. 1a and Supplementary Data 28) as formerly seen in *Physcomitrium patens* [4], rather than concentrated in centromeric regions as in angiosperms [18]. In bryophytes, all three types of DNA methylation (CG, CHG, and CHH) levels are lower in the gene body than in its flanking regions (Supplementary Fig. 1a), a pattern also consistent with that of *Physcomitrium patens* [4]. This contrasts with most angiosperms where CG DNA methylation largely increases toward the center of the gene body [19]. It is notable that the overall CHH DNA methylation level is significantly higher in bryophytes (especially in mosses) reaching an average level of 23% (Supplementary Data 28) as compared to only 1.3% in angiosperms [19] and 7.4% in gymnosperms [19].

Whole genome duplications (WGD) occurred, and often repeatedly, in many plant lineages and are considered a major driving force during the evolution of land plants [20,21]. *Ks* age plots and phylogenomic analyses (see Methods) indicate that both liverworts (despite their large genomes) and hornworts are devoid of signatures of past WGDs (Supplementary Figs. 1b, 19 & 26–31), which is in accordance with recent analyses based on genomic [1,2,4] and transcriptomic data [22]. Thus, we find no evidence supporting the proposed WGD event in the most recent common ancestor (MRCA) of jungermanniopsid liverworts [20] (Supplementary Fig. 1b). Our phylogenomic analyses (Supplementary Figs. 19 & 32) identified significant gene duplications, and hence likely WGD, at five deep nodes in the evolution of mosses (Supplementary Data 29), which is consistent with inferences from transcriptomic data [23]. Three (i.e., FTDB, FUNA, and SPHA) of these five WGD events are supported by evidence from *Ks* plots, phylogenomics, and chromosome collinearity (Supplementary Figs. 1D, 19 & 32), whereas the remaining two (OTPB and BRYO) lack the support from chromosome collinearity. Mechanisms triggering autopolyploidy in bryophytes may include diplospory, syndiplospory, somatic polyploidy and apospory [24]. The latter may be more likely in mosses than in liverworts or hornworts, as the stalk of mosses supporting the sporangium typically develops over longer period of time during which it remains totipotent.

As previously shown [3,11], collinearity among the three bryophyte lineages is very limited (Supplementary Fig. 1c) and may reflect their ancient divergence. Although, pairwise collinearities can be seen within both clades of mosses and hornworts at varying degrees (Supplementary Fig. 1c), they are largely absent among liverworts, possibly because they possess more abundant (Fig. 1) and ancient (Supplementary Fig. 39) repeat elements in their genomes that facilitate intragenomic rearrangements.

3. Comparative genomics and gene family evolution of bryophytes

Algae, bryophytes, and tracheophytes (see Methods for taxon sampling) contain 2,853, 6,457, and 6,926 core gene families (orthogroups), respectively, with 2,711 that are shared by all three groups (Supplementary Data 30). Ancestral reconstruction of the gene families of 344 Archaeplastida genomes revealed that a significant burst of gene acquisition (7,146 gene families) occurred in the most recent common ancestor (MRCA) of embryophytes (Extended Data Fig. 3). Among those annotated with known function, prominent gene families involve in signal transduction, secondary metabolism, and post translational modification (Supplementary Data 7). The MRCA of bryophytes holds a larger set of gene families than the MRCA of tracheophytes (19,328 *vs.* 17,354) (Extended Data Fig. 3). An obvious reduction in the number of gene families was observed in the MRCA of hornworts among that of the seven major land plant lineages (hornworts, liverworts, mosses, lycophytes, ferns, gymnosperms, and angiosperms) (Extended Data Fig. 3). These gene losses are primarily represented in signal transduction in COG functional terms for the four focal nodes (the MRCA of bryophytes, hornworts, liverworts, and mosses) in bryophytes (Supplementary Fig. 40).

We investigated the evolutionary history of key gene families related to developmental (e.g., cuticle, root, stomata, and water conducting tissue) and physiological/biochemical networks (e.g., UV protection, rhythm regulation, and secondary metabolite biosynthesis) that may have facilitated adaptation of plants to the terrestrial environment (Supplementary Data 31, mostly based on studies of *Arabidopsis*). Ancestral reconstruction of these gene families revealed that the ancestor of embryophytes experienced significant gains of these gene families and that the respective ancestors of embryophytes, bryophytes, and tracheophytes also held a rich repertoire of these genes (Supplementary Fig. 41 and Supplementary Data 31). Subsequently, the three major lineages of bryophytes, in particular liverworts and hornworts, incurred significant losses in many of these pathways (Supplementary Fig. 41) which may explain or result from their simpler gametophytic vegetative body. In support of this, the early diverging lineages within each of the three bryophyte clades tend to contain a larger set of the aforementioned genes than later-diverging ones

(Supplementary Data 31).

Phytohormones and transcription factors (TFs) play essential roles in the regulation of growth, development, and response to environmental stimuli. Most (74/76) of the investigated gene families related to the biosynthesis and signaling of phytohormones can be traced back to the MRCA of bryophytes (Supplementary Fig. 42). Additionally, tracheophytes gained two families (i.e., *JAZ* and *BKI1*). Some of the phytohormone genes were partially and independently lost in bryophyte lineages (Supplementary Data 32). Among bryophytes, mosses retained a larger set of phytohormone genes than liverworts and hornworts, likely due to mosses having undergone successive WGDs. Liverworts and hornworts conspicuously lost genes involved in the biosynthesis and signaling processes of jasmonic acid (JA), cytokinin (CK), and brassinosteroid (BR) (Supplementary Data 32). Likewise, liverworts and hornworts generally displayed fewer TFs than mosses and tracheophytes (Supplementary Fig. 43), a pattern perhaps best explained by the lack of WGDs during their evolution and diversification, given that TFs may be retained in excess following WGD, as shown in seed plants [25,26]. For the 71 plant TF families in the iTAK database [27], streptophyte algae evolved the majority of them before the divergence of embryophytes. The MRCA of embryophytes acquired four additional TFs, i.e., *FAR1, GeBP, ULT*, and *VOZ* [28] and held a nearly complete repertoire of TFs, except for *SAP* (Supplementary Fig. 42, Supplementary Data 33) that was acquired in the MRCA of tracheophytes. The three major bryophyte lineages experienced sporadic losses of TFs, especially in more recently diverged lineages. However, early-diverging bryophyte major clades tend to contain more TFs. For example, the *YABBY* transcription factor that regulates the polarity of plant leaves, and the *C2C2*-type *YABBY* are present in all *Sphagnum* species but missing in most setaphytes. *DBP*, involved in the regulation of circadian rhythm genes, is present in *Takakia lepidozioides* but is missing in almost all other mosses. *ULT* that acts as a regulator for adaxial-abaxial axis development is missing in all liverworts and many derived mosses. However, profiling the aforementioned phytohormone genes and TFs heavily relies on the current knowledge of their function in flowering plants, and whether bryophyte-specific regulatory factors exist might be underestimated.

4. Unique resistance (R) genes in bryophytes

Plants have evolved innate immunity to defend themselves from invading pathogens. Compared to aquatic organisms, land plants are exposed to more or different pathogenic threats, as a result of their close and direct contact with soil and air, and hence have developed a more advanced and sophisticated immune system in which nucleotide binding site (NBS)-encoding genes play a key role [29]. In angiosperms, two types of NBS-encoding genes - *CC-NBS-LRR* (*CNL*) and *TIR-NBS-LRR* (*TNL*) genes - function as pathogen detectors and activate hypersensitive reactions. *RPW8-NBS-LRR* (*RNL*), the

third type of NBS-encoding genes work relatively downstream by receiving and continuously transducing the activated signal. However, in bryophytes, the *RNL* genes together with the modules (*RNL_{ADR1}-EDS1-PAD4* and *RNL_{NRG1}-EDS1-SAG101*) responsible for immune signal transduction [30] are absent, suggesting a distinct immune system in bryophytes compared to angiosperms. Supporting evidence has been provided by previous studies that reported specific NBS-encoding genes - *Hydro-NBS-LRR* (*HNL*, liverwort specific) and *PK-NBS-LRR* (*PNL*, moss specific) in bryophytes [31,32] (Extended Data Fig. 6, Supplementary Data 34 & 35). Here, we identified three new types, *TIR2-NBS-LRR* (*T_2NL*, shared by mosses and gymnosperms), *NBS-HEAT-ARM-TPR* (*NHAT*, bryophyte specific), and *NEW-NBS-LRR* (*NNL*, shared by bryophytes, lycophytes and gymnosperms) with *NNL* containing an undetermined N-terminal domain (Extended Data Fig. 6).

To test the functional role of the bryophyte NBS-encoding genes, we transiently expressed these bryophyte immune receptor genes in *Nicotiana benthamiana* leaves (see details in Extended Data Fig. 6b). The new *NNL, T_2NL*, and *NHAT* genes and the previously identified bryophyte-specific *HNL* and *PNL* genes all successfully induced hypersensitive reactions in *N. benthamiana*, confirming their roles in plant immunity (Extended Data Fig. 6b). These results also indicate that plant immune signal transduction does not solely rely on the *RNL_{ADR1}-EDS1-PAD4* or *RNL_{NRG1}-EDS1-SAG101* modules, which presumably only evolved in seed plants. Some unidentified immune pathways may act in non-seed plants. In bryophytes, *HNL*, *PNL*, *NNL*, *T_2NL*, and *NHAT* genes may trigger hypersensitive reactions via distinct immune pathways or work as alternative downstream signal receptors like *RNL* genes. Coincidently, all five bryophyte NBS-encoding genes are absent in angiosperms and they were either lost in the MRCA of angiosperms (*NNL* and *T_2NL*) or originated in bryophyte lineages (*NHAT*, *PNL* and *HNL*). Thus, different embryophyte lineages developed divergent immune receptors and the diversity of immune systems may reflect the divergent evolutionary trajectories in terrestrial environments by bryophytes and tracheophytes.

5. WGD analysis

Ten representative moss species were selected to test all candidate WGDs while avoiding redundant species sampling, as shown in Supplementary Fig. 32. Firstly, ORTHOFINDER v2.4.0 [33] with inflation factor set as 3.0 was implemented to delineate orthogroups. To assure at least one gene from both descendants present at the root and avoid large gene family size that may contain much noise and cause computational retardation, 30,758 orthogroups were filtered out and 7,220 orthogroups were retained. MUSCLE v3.8.1551 [34] with default parameters was performed to obtain a peptide multiple sequence alignment (MSA) which was trimmed and back-translated into a codon-level MSA by TRIMAL v1.4.1 [35] with parameter "-automated1". The trimmed nucleotide MSA was then fed into the RAxML v8.2.9

[36] with substitution model GTRGAMMA and 200 rapid bootstrap replicates to acquire a maximum likelihood (ML) gene tree. Eleven orthogroups with no proper alignment after trimming were filtered out. Two phylogenomic programs, PUG v2.0 [37] and WHALE v2.0.3 [38] were called to achieve the gene tree-species tree reconciliation and WGDs probing.

In total, 7,209 ML gene trees were imported into PUG (Supplementary Fig. 33). The option "-estimate_paralogs" was set to estimate all possible unique gene pairs from gene trees. Paralog pairs whose last common ancestral (LCA) nodes had bootstrap support value (BSV) less than 80 were not in counts. A single gene duplication event may be represented by multiple paralog pairs and hereby we only counted it once as "unique" gene duplication event. The inspection of WGDs in PUG was based on the number of gene duplications mapping to each internal branch and branches with higher than 20% of the maximum number of duplications among all internal branches were regarded as potential WGDs. The bootstrap replicate of each retained orthogroup was fed into ALEOBSERVE v1.0 [39] under the default parameter to construct the conditional clade distribution (CCD) and the resultant 7,209 CCD files were then imported into WHALE as input. The Bayesian gene tree-species tree reconciliation analysis was performed and the occurrences of five WGDs were queried under the DL+WGD model, which considers both small-scale gene duplication and loss (DL) and WGDs. The adopted critical branch-specific DL+WGD model incorporates various DL rates across the species tree, wherein the duplication (λ) and loss rates (μ) were assumed to be equal on each branch. A *Beta* (3,1) prior distribution was used for η, the parameter denoting the geometric prior distribution on the number of genes at the root. An improper flat prior for the mean branch rate r was set and the standard deviation σ was set to follow an exponential prior with mean 0.1. The branch rates were set to follow a multivariate Gaussian prior. Another more relaxed DL+WGD model was also adopted, wherein an independent bivariate normal prior with mean 0 and standard deviation 1 for the mean log-DL rate, a Uniform (-1,1) prior for the correlation coefficient of DL rate for each individual branch and an exponential prior with mean 1 for the standard deviation of branch rates were set. The iteration of Markov chain Monte Carlo (MCMC) chains was set as 500. The Bayes Factor (BF) was applied to estimate the significance of WGD models using the Savage-Dickey density ratio [38].

Physcomitrium patens and *Sphagnum girgensohnii* were used for the absolute timing of FUNA and SPHA WGD. Firstly, the paralogous K_S-age distributions were constructed as previously described [40]. Secondly, anchor pairs and other duplicated pairs lying under the WGD peak (so called peak-based duplicates) were collected for phylogenetic dating. The K_S range of *S. girgensohnii* WGD peak for subsequent absolute dating was set between 0.2 and 0.7 on account of the densest aggregation of duplicates, while for *P. patens* was between 0.4 and 1.1. For each WGD paralogous pair, an orthogroup

including the two WGD paralogues and orthologues from other species identified by INPARANOID v.4.1 [41] was constructed, as shown in Supplementary Fig. 32. The orthologue of *P. patens* was replaced with two paralogues when dating itself, as such the *S. girgensohnii*. In total, 22 orthogroups based on anchor pairs and 178 orthogroups based on peak-based duplicates were collected for *S. girgensohnii*, while 20 and 37 were collected for *P. patens*. The orthogroups were then dated using the BEAST v2.4.6 [42] under an uncorrelated relaxed clock model and LG+G model with four site-rate categories. A starting tree with branch lengths satisfying all fossil prior constraints and topology adopted in this paper was created. Fossil calibrations exploited in absolute divergence time estimate were also configured using log-normal calibration priors as shown in Supplementary Data 26 and Supplementary Fig. 34. A run without data was performed to ensure proper placements of the marginal calibration prior. MCMC sampling for each orthogroup was run for 10 million generations with sampling every 1,000 generations. The resultant trace files were inspected using Tracer v.1.5 [43] with a burn-in as 1000 and minimum ESS for all parameters set as 100 to assure sufficient sampling and convergence. In total, 158 and 35 orthogroups were accepted for *S. girgensohnii* and *P. patens*, respectively, and the absolute age distribution were constructed with consensus peak and 90% confidence interval derived from the kernel density estimation (KDE) of 2,500 bootstrap replicates as shown in Supplementary Figs. 35 & 36.

6. Identification of NLR genes

Nucleotide-binding domain leucine-rich repeat (*NLR*) containing genes were identified from all available 138 bryophyte genomes following the methods described previously [44]. A two-step strategy was adopted to identify *NLR* genes. The first step was to conduct simultaneously hidden Markov models search (HMMER) searches and NCBI BLASTP searches using the amino acid sequence of the NB-ARC (NBS) domain (Pfam accession number: PF00931) as a query to identify potential *NLR* genes. Then, the resulted non-redundant candidate genes were subjected to online PFAM analysis with an E-value cut-off of 10^{-4} (http://pfam.sanger.ac.uk/) to confirm the presence of the NBS domain. The identified *NLR* genes were subjected to NCBI's conserved domain database (CDD) (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) using the default settings to determine whether they encoded CC, TIR, RPW8, protein kinase, Hydrolase, HEAT, Arm, TIR2, WD40 or TPR domains, and combining these results with the PFAM analysis to define the R gene types.

Amino acid sequences of the NBS domain were extracted from all identified *NLR* genes and used for multiple sequence alignments using CLUSTALW [45] integrated in MEGA v7.0 [46] with default parameter settings. Sequences that were too short (< 190 amino acids, and less than two-thirds of a regular NBS domain) or too divergent were removed. The resulting alignments were manually

corrected using MEGA v7.0 for further improvement. Additionally, structural motif annotation was performed using the PFAM analysis (http://pfam.janelia.org/) and SMART tools (http://smart.embl-heidelberg.de/). To analyze the evolution of *NLR* genes in bryophytes and other plant lineages (including glaucophytes, rhodophytes, chlorophytes, streptophyte algae, lycophytes, ferns, gymnosperms and angiosperms), a phylogenetic tree was reconstructed based on the aligned amino acid sequences of the conserved NBS domains. Phylogenetic analyses were conducted using IQTREE2 v2.2.0 [47] with the maximum likelihood method and with 1000 ultra-fast bootstrap replicates. The best-fit model was estimated by MODELFINDER [48] as implemented in IQTREE2.

To test whether the newly identified *NLR* genes are able to trigger the hypersensitive response (HR), agrobacterium containing *NLR* gene expression vectors were transformed into *Nicotiana benthamiana* leaves for transient expression. Vectors of full-length *NLR* genes were synthetized by Tsingke Biotechnology (Co., Ltd, Beijing). The primers used for constructing the expression vector of *NLR* genes are listed in Supplementary Table 15. Full-length and truncated fragments of *NLR* genes were cloned into the pCHF1 vector in the KpnI/BamHI restriction enzyme sites, respectively. Wild-type *N. benthamiana* individuals were grown under a 16 h (25 to 28 ℃)/8 h (22 to 25 ℃) light/dark photoperiod condition. The binary constructs were transformed into *Agrobacterium* strain GV3101-pMP90 and stored in a -80℃ freezer with 20% glycerol. Two days before the *Agrobacterium* infiltration, the constructs were streaked out on solid L medium plate (with Ampicillin and rifampicin) and grown in a 28℃ incubator. *Agrobacterium* strains containing the constructs were shaken to an OD600 of 0.8-1.0, harvested, resuspended in infiltration buffer (10 mM MES, 10 mM $MgCl_2$, and 0.1 mM acetosyringone), and infiltrated into leaves of *N. benthamiana* wild-type using a needleless syringe. The infiltrated leaves were wrapped up with black cloth for growth under dark for 24 hours, and were observed 1-, 3-, and 5-days post-infiltration, imaged under white light. Three experimental repeats were performed on different plants.

7. Stress experiments and differential gene expression analysis

To study the stress response of mosses to abiotic stresses, we analyzed transcriptome data from heat treatment, drought treatment, and UV stress for *Physcomitrium patens* (Supplementary Table 16–25). Transcriptome data of heat treatment at 37℃ for *P. patens* were downloaded from the NCBI database under the project No. SRP224458, and used for subsequent analyses. We used 200 mM mannitol to simulate drought stress for *P. patens* growing on soil, and harvested the moss samples after 0 hour, 0.5 hour, 1 hour, 2 hours, and 4 hours, respectively. UV stress was performed under UV light with a dose of 200 mw/cm^2, and harvested samples after an exposure time of 0 hour, 1 hour, 2 hours, 4 hours and 8 hours, respectively. Three experimental replicates were performed for each treatment. Total moss

RNA was extracted from 100 mg of *P. patens* gametophytic tissues from the control or drought-treated/UV treated samples using Nucleospin® RNA plant (Macherey-Nagel-Germany). The RNA was quality-controlled with 1% agarose gel electrophoresis, a Qubit fluorometer (Thermo Fisher Scientific, USA), and a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA). Transcriptome libraries were constructed with a TruSeq RNA Library Prep Kit v2 (Illumina, CA, USA) with insert sizes of 200–400 bp and sequenced for PE 150 bp on an Illumina NovaSeq 6000 platform.

RNA-seq reads were trimmed using TRIMMOMATIC v0.39 [49] and mapped against the *P. patens* annotated gene models using TopHat2 v2.1.1 [50] by retaining the best alignments. TPM (Transcripts per million) were calculated using the EXPRESS program, which was incorporated in TRINITY v2.8.4 [51] package. Furthermore, differentially expressed genes with an adjust p-value ≤ 0.05, FDR ≤ 0.01, and at least a 2-fold expression change compared with the control group were identified using R package edgeR v3.42.2 [52]. The other 80 transcriptome TPM data of *M. polymorpha* were downloaded from the MarpolBase (https://marchantia.info/), and 58 transcriptome TPM data of *P. patens* were downloaded from the Phytozome V13 database.

8. Evaluation of the efficacy of genome decontamination strategy

Bryophytes generally form close symbiotic relationships with microorganisms, leading to varying degrees of contamination in their sequencing data. Even axenic culture initiated from a single spore of the hornwort *Anthoceros angustus*, still experienced data contamination [2]. Drawing on the experiences and lessons from previous publications, we explored and carried out rigorous decontamination process in the early stages of the project. Moreover, chimeric contigs created by the assembly of contaminant and non-contaminant sequences are extremely difficult if not impossible to distinguish from bona-fide contigs scaffolds. We carried out two types of analyses to test the impact of chimeric contigs in our assembly.

First, we downloaded the raw long reads from three bacterial datasets (*Streptomyces* sp. PU10, *Sphingomonas capsulata* HAMBI_0103, and Soil microbiome), two fungal (*Exophiala dermatitidis*, *Rhizophagus irregularis*) and one algal dataset (*Chlorella variabilis*), totaling 3.8 Gb. These lineages are abundantly represented in the metagenomes of bryophytes [53,54]. We then artificially mixed these exogenous read data with the raw long reads of *Marchantia polymorpha* (totaling 10 Gb) obtained from axenic cultures (NCBI bioproject: PRJEB65250). We performed genome assembly using NextDenovo, based on the clean raw data of *Marchantia polymorpha* and a simulation of artificially

adding microbiome contaminant data. This resulted in four assemblies, the authentic *Marchantia* assembly (raw), the authentic *Marchantia* assembly (decontaminated), the mixed assemblies (raw), and the mixed assemblies decontaminated following our decontamination pipeline (Supplementary Fig. 44). To assess the presence of chimeric contigs in the assembly, we assume that a chimeric contig must be covered across a 5%–95% region by clean reads (*Marchantia Polymorpha* only) and contaminated reads (bacterial, fungus, or algal) simultaneously. We mapped the pure *Marchantia* long reads and the contaminated reads to these assemblies, respectively. We examined the read coverage and depth for each contig in these assemblies and identified those contigs that were partially (i.e., with read coverage of 5%–95% regions in one contig) mapped with *Marchantia* reads and contaminated reads, respectively (putative chimeric contigs). To our surprise, none of the assemblies contained such chimeric contigs. All the contigs were exclusively mapped to by either *Marchantia* reads or contaminated reads, never by both. Despite the low coverage (~20×) and the resultant relatively fragmented contig assemblies, the synteny analyses of the chromosome level assembly of the authentic *Marchantia* genome and the mixed genome (decontaminated) show consistency (Supplementary Fig. 45), suggesting that the mixed assembly would also yield reliable results if handled carefully. We further compared the quality of the authentic assembly and the decontaminated assembly, both of which exhibited similarly high values of contig N50, scaffold N50, and Busco completeness (Supplementary Data 36). This simulation analysis suggests that chimeric contigs are not common in assemblies of the nanopore long reads and that our decontamination procedure is robust and effective.

We also generated one nanopore assembly of *Conocephalum conicum* using pre-decontamination procedure, where contaminated reads are removed before genome assembly. To do this, we performed a BLASTN search against the NT database using the self-corrected long nanopore reads from the NextDenovo analysis. We then filtered reads with three different contamination coverage thresholds (>50%: removing reads with contamination hits coverage >50%; >20%: removing reads with contamination hits coverage of >20%; >0%: fully remove the reads with any hits), the resultant clean reads were used for genome assembly. We compared these pre-decontaminated genome assemblies with our routine post-decontaminated assembly (Supplementary Fig. 46, Supplementary Data 37). Our results indicate that the pre-decontamination assembly was more fragmented and with more contaminations that could not be removed based on BLAST searches against available databases

(Supplementary Fig. 46). This is not unexpected, however, especially considering that previous studies have reported a significant amount of horizontal gene transfers (HGTs) in *Marchantia polymorpha* and *Physcomitrium patens* [55]. Adopting a pre-decontamination procedure, would largely remove HGTs, thus reducing genome continuity and fidelity. In addition, many unique sequences in the genome do not match any existing database entries. These sequences need to be assembled in longer contigs containing known sequences to better determine their origin. The strategy of pre-decontamination would prematurely remove these critical contamination tag-containing sequences at an early stage, resulting in a large number of unrecognizable contaminating sequences in the assembled genome. This would lead to a larger residue of contaminating sequences. The situation would be even worse for short-read sequencing data. Moreover, the pre-decontamination consumes a lot of computational resources, making the project unfeasible. These are the reasons why we adopted a post-decontamination strategy in our study. This comparative analysis suggests that our results are robust and reliable.

Therefore, chimeric contigs are not an issue for nanopore long-read assembly, even for the ONT data with only ~20× coverage. The sequencing depth for our genomes are much higher, with average read coverages of 89×, 281×, and 241× for ONT, stLFR, and 10x genomics data, respectively. Hence, our genome assembly should be rarely impacted by chimeric contigs. We evaluated the genome completeness based on the several criteria, and all suggest high levels of completeness (see also point 7 below): 1) Busco completeness was very high, averaging 94% (Supplementary Data 2), 2) the RNA data mapping rate was high, averaging 83% (Supplementary Data 2), and 3) the WGS read depth was similar among three gene categories (core, accessory, and unique) for each species, regardless of the sequencing platform, consistent with model bryophytes and previously published data (Supplementary Data 38).

We also summarized the ratio of single-exon genes for three gene family categories, and as expected, unique gene families had the highest percentage of single-exon genes, while core gene families had the lowest (Supplementary Data 39). This observation is consistent with findings in model bryophytes. Older genes, such as core genes, tend to accumulate more introns, while younger genes, such as accessory and especially unique genes, contain fewer introns. In fact, single-exon are a major feature of new genes, which is consistent as observed in flowering plants [56-58].

9. Assessment of the correlation between genome quality, gene numbers and different sequencing platforms

We use three sequencing technologies (ONT, 10x genomics, stLFR) to assemble our bryophyte genomes. The 10x Genomics microfluidic [59,60] and stLFR (single tube Long Fragment Read) [61] sequencing methods are indeed long read sequencing approaches based on next-generation sequencing technology. For both, high molecule weight (HMW) DNA is used, and long DNA fragments were sheared and barcoded before library construction and sequencing. During the assembly process, these long DNA fragments were assembled by recalling their barcodes, making these methods distinct from traditional short-read next generation sequencing method. Limited by the sequencing technology, the average contig N50 is not long for 10x genomics (32 Kbp) and stLFR (37 Kbp). However, benefiting from HMW DNA library construction techniques, the average scaffold N50 values are much higher, with 119 Kbp and 506 Kbp for 10x and stLFR, respectively. Since bryophytes generally form close symbiotic relationship with various kinds of microorganisms, simply counting the DNA reads mapping rate is not appropriate. We therefore applied other assessment strategies, such as BUSCO assessment (averaging at 94%), RNA mapping rate assessment (83% on average), and gene depth assessment of WGS short reads data. We counted the number of predicted genes, and the number of genes in three major categories (core, accessory, and unique) across three major bryophyte lineages for data generated by different sequencing platforms. We compared these data with previously published data (Supplementary Data 40). In hornworts, the number of genes predicted by the genomes we sequenced with the stLFR platform was 17,507, the number of accessory families was 2,547, and the number of unique families was 1,457, which was significantly lower than the corresponding number of genes published for the three hornworts using the ONT (Oxford Nanopore Technologies) platform [2,3]. In liverworts, the number of gene sets predicted by the three sequencing platforms was similar. The ONT platform predicted the highest number of unique genes, comparable to the published *Marchantia polymorpha* data (www.marchantia.io). The number of unique genes in the 10x genomics and stLFR sequencing results was generally lower than that of the ONT platform and the published data. In mosses, the ONT sequencing platform predicted the highest total number of genes, while the 10x genomics and stLFR platforms predicted a relatively lower total number of genes, close to the published values for mosses [4,16,17]. The ONT platform predicted the highest number of unique genes,

generally exceeding that of the published species. By contrast, the 10x genomics and stLFR platforms predicted notably fewer numbers of unique genes than the published genomes [1-3,17]. These results suggest that the assemblies with better continuity from the ONT platform tend to have a higher total number of predicted genes and a higher number of unique genes, compared to the assemblies with lower continuity produced by the 10x genomics and stLFR platforms (Supplementary Data 40). This may reflect the reliability of the long-length read sequencing results from the 10x genomics and stLFR platforms, and also corroborates that our rigorous decontamination process can effectively filter out contamination.

We also summarized the WGS read depth for all genes from three categories (core, accessory, and unique) for 65 species (Supplementary Data 41). The WGS reads depth is similar among the three categories for each species, consistent with the model bryophytes and previously published data. This indicates that the unique family, mostly single exon genes, are authentic.

10. Assessment of the impact of genome sampling on gene family clustering results

Generally speaking, genomes with poor continuity tend to contain a higher number of fragmented genes and some mis-annotation due to incorrect assembly. The ONT data with a better assembly continuity in our study instead contained a higher number of genes and a higher number of unique genes compared to the other two platforms (10x genomics and stLFR) or the published ones [1-3]. In addition, we newly ran a cluster analysis using Orthofinder, only with the published bryophyte genomes and a complementary number of vascular plants and algae (12 algae outgroups, 12 bryophytes and 12 vascular plants). Species of charophytes contained significantly lower numbers of average core gene families, thus reducing the total number of gene families (Supplementary Data 42). The lower number of core gene families may have been caused by inadequate sampling as well as their ancient divergences. Notably, the high number of unique gene numbers in charophytes is mainly caused by the outlier *P. margaritaceum,* whose genome quality is rather poor with a contig N50 of only 116 kb, Busco completeness of 92%, and a high fragmented gene rate of 21%. Excluding the outlier *P. margaritaceum*, the number of unique orthogroups, and the total number of the orthogroups for charophytes is distinctly smaller than for embryophytes (Supplementary Data 43). Hence, we reached a same conclusion, i.e., bryophytes have larger gene family numbers and higher numbers of unique and accessory gene families (Supplementary Data 42 & 43). Collectively, our conclusion about

bryophytes holding a larger gene family space and with an astonishing high number of unique gene families is solid.

11. Evaluation of the reliability of HGT analyses and identification

The topic of HGT in eukaryotes has indeed long been a subject of considerable debate among scientists, with discussions largely focusing on the extent, mechanisms, and evolutionary implications. Some researchers are basically against all cases of HGT and do not believe that HGT could occur in eukaryotes, arguing that many of the evidence may stem from phylogenetic inconsistencies or artifacts of gene loss [62]. Others, however, contend that advancements in genomic sequencing and improved analytical methods have unveiled numerous genuine cases of HGT, emphasizing that disregarding these HGT events risks overlooking significant adaptive and evolutionary contributions [63,64]. Recent studies have provided strong evidence for HGT in plants [65-70], including mosses [71], and even in animals [72-74], suggesting that HGT may be more common than previously thought.

In fact, HGT can always be interpreted by alternative scenarios, which has been discussed extensively in the literature [75,76]. However, several lines of reasoning support horizontal gene transfer (HGT) as a more plausible explanation in our study. Firstly, if all cases of HGT events were instead attributed to different gene loss from the last common ancestor of eukaryotes, the gene pool of this ancestor would need to be implausibly large, which is possible in theory but unlikely. Moreover, gene loss cannot account for phylogenetic scenarios where genes in the recipient group are derived from donor groups, resulting in donor sequences being paraphyletic while the recipient sequences are monophyletic. For instance, plastid sequences often form a monophyletic group nested within cyanobacteria sequences. This topology cannot be explained by simple gene losses, as even if gene losses occurred across other eukaryotic groups, the cyanobacterial sequences would still form a monophyletic instead of paraphyletic group. In our analyses, we carefully weigh the likelihood of HGT versus gene loss scenarios. In all identified cases of HGT, the gene loss scenario would require numerous, independent loss events across multiple groups, whereas a single HGT event can readily explain the observed tree topology. The greater the number of hypothetical losses needed to explain the topology, the less plausible the gene loss scenario becomes. Therefore, HGT provides the most parsimonious and straightforward explanation for the observed phylogenetic patterns.

Ma et al. [55] conducted an extensive analysis of horizontal gene transfer (HGT) in streptophytes, with a limited sampling of 3 streptophyte algae, 3 bryophytes and 5 tracheophytes, resulting in the identification of 503 HGT events, comprising 467 orthogroups and 36 species-specific HGTs in our 54-species phylogeny. In our study, we expanded the sampling to include 7 streptophyte algae, 26 bryophytes and 16 tracheophytes. The total number of newly identified horizontal gene transfers (HGTs) amounted to 1,306, comprising 1,038 orthogroups and 268 lineage-specific HGTs. These newly identified HGTs were predominantly found within the bryophyte lineages, with the majority of them being orthogroups that included more than one species.

The direction of HGT transfers mainly depends on the ages of the receptor and donor groups where the horizontally transferred genes are located. In our analysis, we first manually examined the distribution of the subject gene among the donor group and subsequently evaluated its potential presence in the ancestor of the donor group or its major sub-groups. This step was intended to confirm that the gene existed in the donor group before its appearance in the plant recipient, as corroborated by age estimates from the Timetree of Life (timetree.org) or other relevant data. The donor taxon was required to be at least as old as the recipient [77]. For instance, many acquired genes in land plants were exclusively detected in multiple major groups of bacteria (e.g., proteobacteria and actinobacteria) or fungi (e.g., both ascomycetes and basidiomycetes), suggesting that they were present in the last common ancestor of these groups. Because the ancestors of these groups emerged much earlier than land plants, the subject gene was most likely transferred from bacteria or fungi to land plants [55].

Our analyses revealed that 'HGT' genes exhibit distinct characteristics compared to other plant genes, such as fewer introns and shorter gene lengths (Supplementary Fig. 14), and resemble the microbial donors instead. These findings provide further supports for the hypothesis that these genes are more likely to have originated through HGT rather than convergent evolution between plants and donor organisms or through gene loss. If convergent evolution were the driving force, similarities would be expected to occur primarily in functional domains, rather in gene structure as well. Likewise, if these genes had been inherited from the common ancestor of prokaryotes and eukaryotes, extensive evolutionary divergence would likely have resulted in structural characteristics resembling those of other plant genes.

We constructed phylogenetic trees for all 1,809 HGT gene families (see methods section and

Supplementary Data 10) and deposited the tree files in the figshare repository (doi: 10.6084/m9.figshare.23528667). We manually inspected the trees and summarized the highest bootstrap support values for the nodes leading to the branches of HGT genes and the donor groups. In this analysis, we define a bootstrap value of 75 or higher as strong, between 60–75 as moderate, and below 60 as weak support. Approximately 89% of the nodes received strong support (bootstrap ≥ 75), 7% received moderate support (60–75), and only 4% received week supports (below 60) (Supplementary Fig. 47 and Data 44). We further compared the bootstrap support values between previously reported HGT events (503 orthogroups) [55] and our newly identified events (1,306 orthogroups). Among the previously reported HGT events, 94% had strong support, 5% had moderate support, and 1% had weak support. Similarly, for the newly identified HGT events, 89% had strong support, 9% had moderate support, and 2% had weak support. These results demonstrate that the quality of our HGT findings is comparable to that of the previously published results (Supplementary Data 44).

It is important to emphasize that the interpretation of the phylogenetic trees is based solely on the currently available data. As more data become available, conclusions about the evolutionary trajectory may need to be revised. For instance, if a HGT event is inferred to occur on at a specific node, it is possible that the event occurred earlier but is only observed at that node due to the limited sampling available today. Similarly, if an HGT gene is currently found only in bryophytes, it is possible that the event happens in the common ancestor of embryophytes, and was subsequently lost in many extant vascular plants. With the sequencing of additional genomes of vascular plants, homologs of the HGT gene may be identified. In such cases, the conclusion that the HGT event occurred in the ancestor of bryophytes would need to be revised accordingly.

References

1 Bowman, J. L. *et al.* Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304.e215 (2017).

2 Zhang, J. *et al.* The hornwort genome and early land plant evolution. *Nat. Plants* **6**, 107–118 (2020).

3 Li, F. W. *et al. Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nat. Plants* **6**, 259–272 (2020).

4 Lang, D. *et al.* The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* **93**, 515–533 (2018).

5 Zhou, X. *et al*. Universal single-copy ortholog benchmark gene set for bryophytes. J. Genet. Genomics. In press. (2025)

6 Renzaglia, K. S., Villarreal Aguilar, J. C. & Garbary, D. J. Morphology supports -the setaphyte hypothesis: mosses plus liverworts form a natural group. *Bryophyte Diversity Evol.* **40**, 11–17 (2018).

7 Puttick, M. N. *et al.* The interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* **28**, 733–745 (2018).

8 Frangedakis, E. *et al.* The hornworts: morphology, evolution and development. *New Phytol.* **229**, 735–754 (2021).

9 Liu, Y. *et al.* Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nat. Commun.* **10**, 1485 (2019).

10 Harris, B. J., Harrison, C. J., Hetherington, A. M. & Williams, T. A. Phylogenomic evidence for the monophyly of bryophytes and the reductive evolution of stomata. *Curr. Biol.* **30**, 2001–2012 e2002 (2020).

11 Diop, S. I. *et al.* A pseudomolecule-scale genome assembly of the liverwort *Marchantia polymorpha*. *Plant J.* **101**, 1378–1396 (2020).

12 Szövényi, P., Gunadi, A. & Li, F. W. Charting the genomic landscape of seed-free plants. *Nat. Plants* **7**, 554–565 (2021).

13 Montgomery, S. A. *et al.* Chromatin organization in early land plants reveals an ancestral association between H3K27me3, transposons, and constitutive heterochromatin. *Curr. Biol.*

30, 573–588 (2020).

14 Perumal, S. *et al.* A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral *Brassica* genome. *Nat. Plants* **6**, 929–941 (2020).

15 Wang, Z. *et al.* A chromosome-level reference genome of *Ensete glaucum* gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae. *Gigascience* **11**, giac027 (2022).

16 Healey, A. L. *et al.* Newly identified sex chromosomes in the *Sphagnum* (peat moss) genome alter carbon sequestration and ecosystem dynamics. *Nat. Plants* **9**, 238–254 (2023).

17 Carey, S. B. *et al.* Gene-rich UV sex chromosomes harbor conserved regulators of sexual development. . *Sci. Adv.* **30**, eabh2488 (2021).

18 Schmitz, R. J., Lewis, Z. A. & Goll, M. G. DNA methylation: shared and divergent features across eukaryotes. *Trends Genet.* **35**, 818–827 (2019).

19 Bewick, A. J. *et al.* The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, 65 (2017).

20 Leebens-Mack, J. H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).

21 Van de Peer, Y., Ashman, T. L., Soltis, P. S. & Soltis, D. E. Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* **33**, 11–26 (2021).

22 Dong, S., Yu, J., Zhang, L., Goffinet, B. & Liu, Y. Phylotranscriptomics of liverworts: revisiting the backbone phylogeny and ancestral gene duplications. *Ann. Bot.* **130**, 951–964 (2022).

23 Gao, B. *et al.* Ancestral gene duplications in mosses characterized by integrated phylogenomic analyses. *J. Syst. Evol.* **60**, 144–159 (2020).

24 Patel, N., Medina, R., Johnson, M. & Goffinet, B. Karyotypic diversity and cryptic speciation: have we vastly underestimated moss species diversity? *Bryophyte Diversity Evol.* **43**, 150–163 (2021).

25 Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).

26 Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci.*

USA **102**, 5454–5459 (2005).

27 Yi, Z. *et al.* Itak: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **9**, 1667–1670 (2016).

28 Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte agae. *Genome Biol. Evol.* **9**, 3384–3397 (2017).

29 Liu, Y. *et al.* An angiosperm *NLR* Atlas reveals that *NLR* gene reduction is associated with ecological specialization and signal transduction component deletion. *Mol. Plant* **14**, 2015–2031 (2021).

30 Sun, X. H. *et al.* Pathogen effector recognition-dependent association of *NRG1* with *EDS1* and *SAG101* in *TNL* receptor immunity. *Nat. Commun.* **12**, 3335 (2021).

31 Xue, J. Y. *et al.* A primary survey on bryophyte species reveals two novel classes of nucleotide-binding site (*NBS*) genes. *PLoS One* **7** (2012).

32 Shao, Z. Q., Xue, J. Y., Wang, Q., Wang, B. & Chen, J. Q. Revisiting the origin of plant *NBS-LRR* genes. *Trends Plant Sci.* **24**, 9–12 (2019).

33 Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 1–14 (2019).

34 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

35 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

36 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

37 McKain, M. R. *et al.* A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales. *Genome Biology and Evolution* **8**, 1150-1164 (2016).

38 Zwaenepoel, A. & Van de Peer, Y. Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates. *Molecular Biology and Evolution* **36**, 1384-

1404 (2019).

39 Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

40 Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).

41 O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).

42 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T. & Drummond, A. J. Beast 2: a software platform for bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014).

43 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973 (2012).

44 Shao, Z. Q., Zhang, Y. M., Wang, B. & Chen, J. Q. Computational identification of microRNA-targeted nucleotide-binding site-leucine-rich repeat genes in plants. *Bio-protocol* **5**, e1637–e1637 (2015).

45 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).

46 Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).

47 Quang, M. B. *et al.* IQTREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

48 Kalyaanamoorthy, S., Minh, B. Q., Wong, T., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

49 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

50 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

51 Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

52 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

53 Kostka, J. E. *et al.* The Sphagnum microbiome: new insights from an ancient plant lineage. *New Phytol* **211**, 57-64 (2016).

54 Lee, E. H., Eo, J. K., Ka, K. H. & Eom, A. H. Diversity of arbuscular mycorrhizal fungi and their roles in ecosystems. *Mycobiology* **41**, 121-125 (2013).

55 Ma, J. *et al.* Major episodes of horizontal gene transfer drove the evolution of land plants. *Mol. Plant* **15**, 857–871 (2022).

56 Zhang, L. *et al.* Rapid evolution of protein diversity by *de novo* origination in *Oryza*. *Nat. Ecol. Evol.* **3**, 679–690 (2019).

57 Cui, X. *et al.* Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol. Plant* **8**, 935–945 (2015).

58 Wu, D. D. *et al.* "Out of pollen" hypothesis for origin of new genes in flowering plants: study from *Arabidopsis thaliana*. *Genome Biol. Evol.* **6**, 2822–2829 (2014).

59 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res* **27**, 757-767 (2017).

60 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021).

61 Wang, O. *et al.* Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res* **29**, 798-808 (2019).

62 Martin, W. F. Too much eukaryote LGT. *Bioessays* **39**, 12 (2017).

63 Boto, L. Are there really too many eukaryote LGTs? A reply to William Martin. *Bioessays* **40** (2018).

64 Leger, M. M., Eme, L., Stairs, C. W. & Roger, A. J. Demystifying eukaryote lateral gene transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). *Bioessays* **40**, e1700242 (2018).

65 Bergthorsson, U., Adams, K. L., Thomason, B. & Palmer, J. D. Widespread horizontal

transfer of mitochondrial genes in flowering plants. *Nature* **424**, 197-201 (2003).

66 Yang, Z. *et al.* Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants. *Nature Plants* **5**, 991-1001 (2019).

67 Cheng, S. *et al.* Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **179**, 1057–1067 e1014 (2019).

68 Wu, J. J. *et al.* Post-transfer adaptation of HGT-acquired genes and contribution to guanine metabolic diversification in land plants. *New Phytologist* **244**, 694-707 (2024).

69 Liu, Y. *et al.* The Cycas genome and the early evolution of seed plants. *Nature Plants* **8**, 389–401 (2022).

70 Wang, K. *et al.* Horizontally acquired *CSP* genes contribute to wheat adaptation and improvement. *BioRxiv* **doi: 10.1101/2024.06.04.597356** (2024).

71 Vuruputoor, V. S. *et al.* Crossroads of assembling a moss genome: navigating contaminants and horizontal gene transfer in the moss *Physcomitrellopsis africana*. *G3: Genes, Genomes, Genetics* **14**, jkae104 (2024).

72 Nowell, R. W. *et al.* Bdelloid rotifers deploy horizontally acquired biosynthetic genes against a fungal pathogen. *Nature Communications* **15**, 5787 (2024).

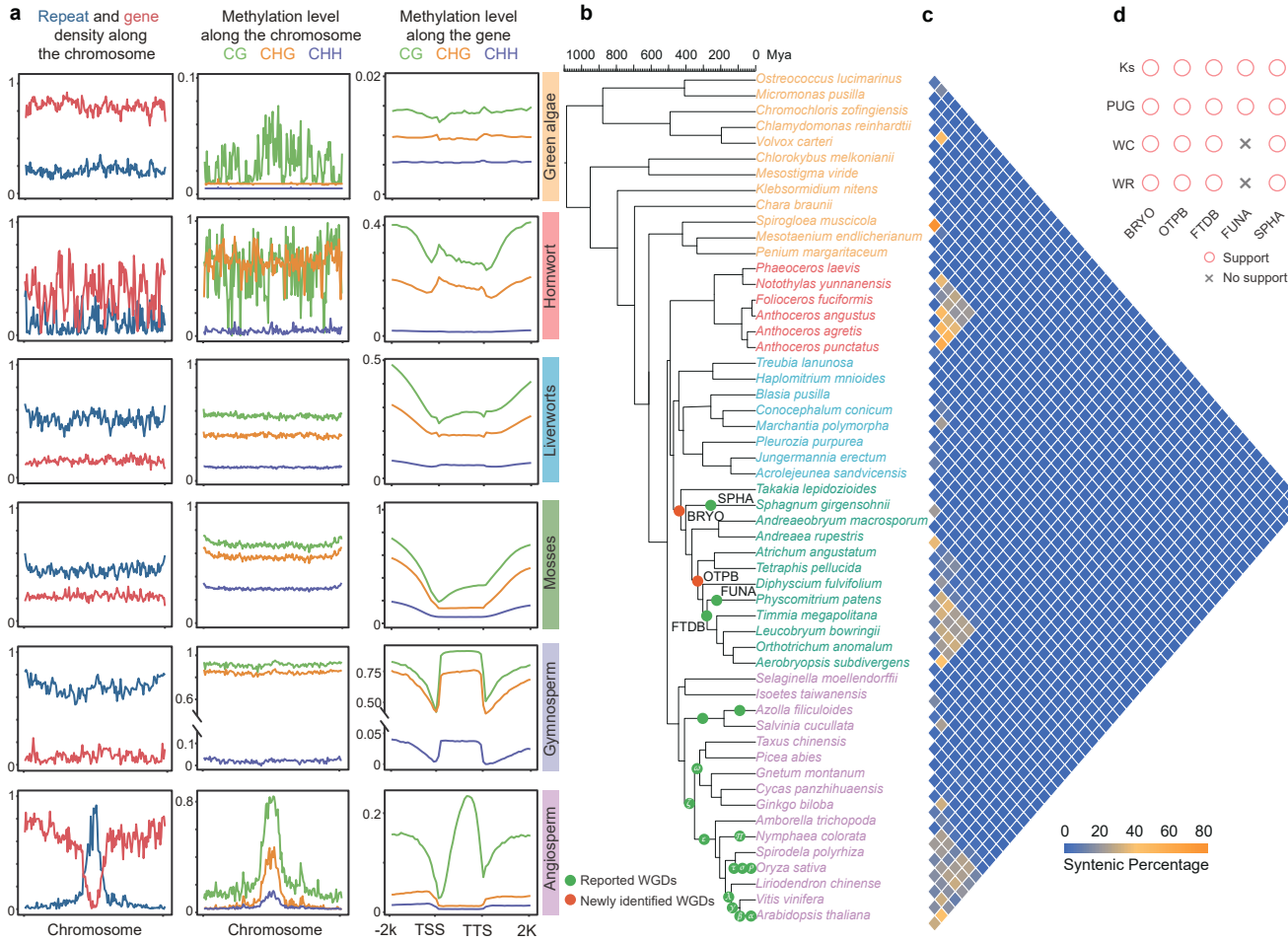
73 Kambayashi, C. *et al.* Geography-Dependent Horizontal Gene Transfer from Vertebrate Predators to Their Prey. *Molecular Biology and Evolution* **39** (2022).

74 Xia, J. *et al.* Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell* **184**, 1693-1705.e1617 (2021).

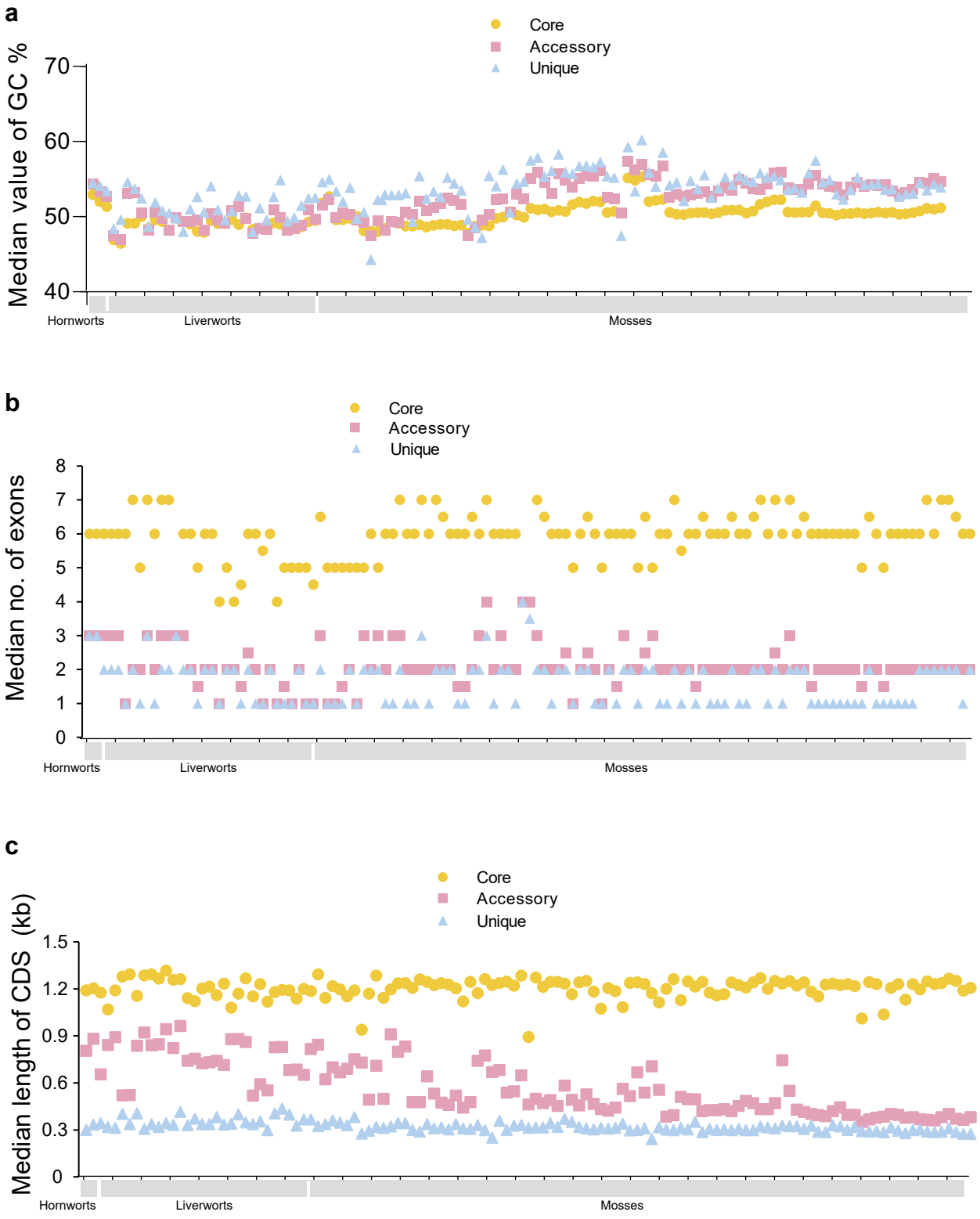
75 Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nature reviews. Microbiology* **3**, 679–687 (2005).

76 Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics* **9**, 605–618 (2008).

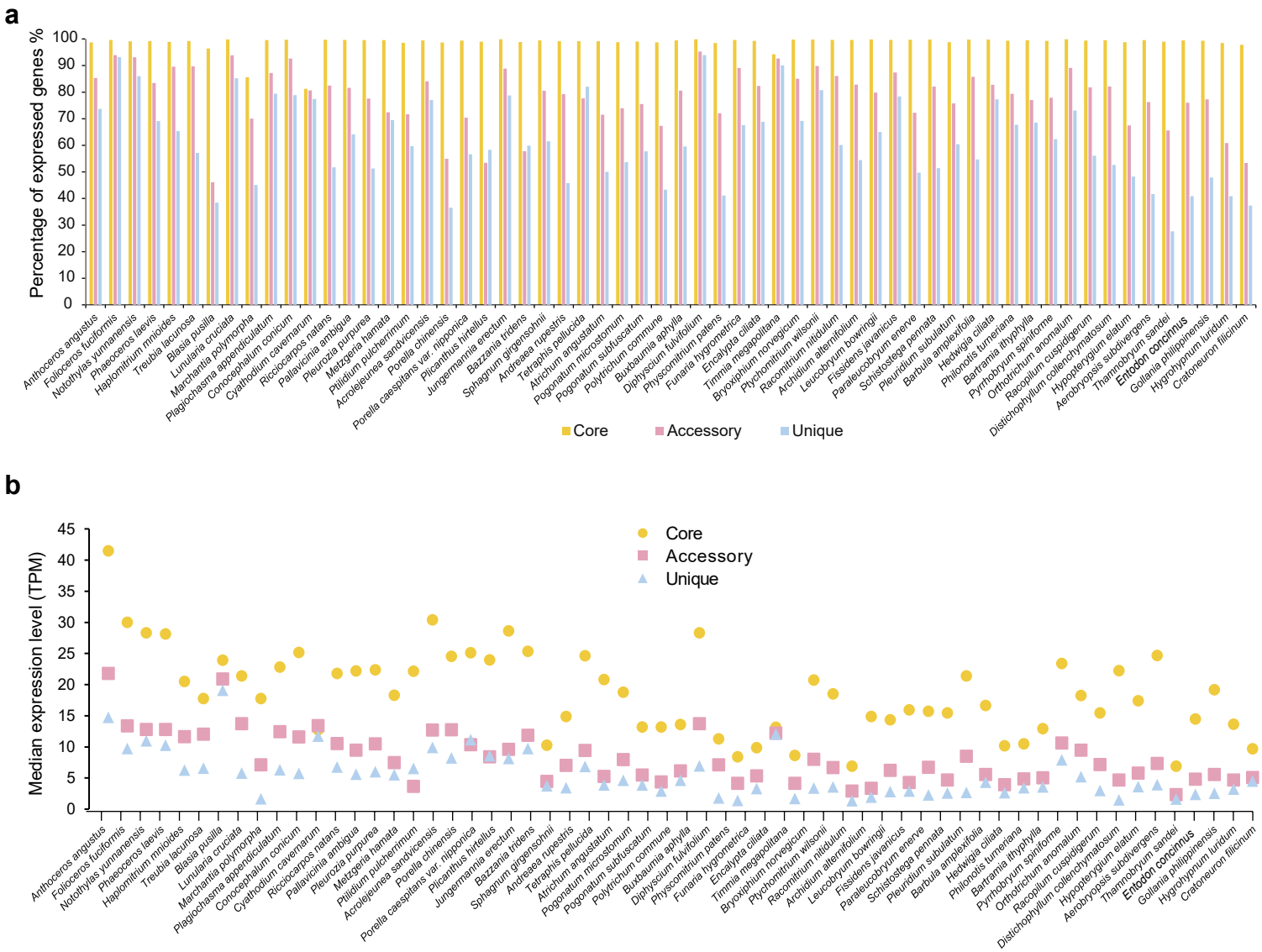
77 Huang, J. & Gogarten, J. P. Concerted gene recruitment in early plant evolution. *Genome Biology* **9**, R109 (2008).



Supplementary Figure 1. Genome characteristics and structural evolution. (**a**) Repeat, gene density, and DNA methylation level along chromosomes, and in genes and the flanking regions in different plant lineages: Green algae, *Chlamydomonas reinhardtii*; Hornwort, *Notothylas yunnanensis*; Liverworts, an average of 9 liverwort species; Mosses, an average of 10 moss species; Gymnosperm, *Pinus tabuliformis;* Angiosperm, *Arabidopsis thaliana* (Supplementary Data 2A). Each bin of repeat density, gene density, and methylation level along the chromosome represents the average percentage of the bins corresponding to each chromosome (scaffolds). The chromosomes or scaffolds (>1 Mb) of each species were split into 200 bins equally. The methylation level along the gene displays the average methylation level of bins corresponding to each gene models (see details in Methods section). (**b**) Ancient polyploidy events showing on the 54-Viridiplantae time tree. BRYO (shared by all Bryophyta), OTPB (shared by Oedipodiopsida, Tetraphidopsida, Polytrichopsida, and Bryopsida) represent newly identified WGDs, while SPHA (shared by Sphagnaceae), FTDB (shared by Funariidae, Timmiidae, Dicranidae, and Bryidae) and FUNA (shared by Funariaceae) represent previously recognized WGDs (Healey et al., 2023; Rensing et al., 2008) and confirmed by this study. (**c**) Matrices of syntenic percentages of pairwise genome comparisons are aligned to the corresponding species. Each cell in the heatmap represents an overall syntenic percentage of a genome comparison, which is calculated by the total number of syntenic genes divided by the total number of gene pairs between two genomes (detailed method see Zhao et al. (2021)). (**d**) Evidences, supporting each WGD event displayed in (**b**), from *Ks* (Ks), and phylogenomic analyses using the programs of PUG (PUG), and WHALE with critical model (WC) and relaxed model (WR).



Supplementary Figure 2. Comparison of gene structure among the core, accessory, and unique genes in bryophytes. (**a**) median percentage of GC content; (**b**) median number of exons per gene; (**c**) median length of coding sequences (CDS). The x-axis represents bryophyte samples sorted phylogenetically according Figure 1. The numbers of core, accessory, and unique genes for each species were indicated in Supplementary Data 6.

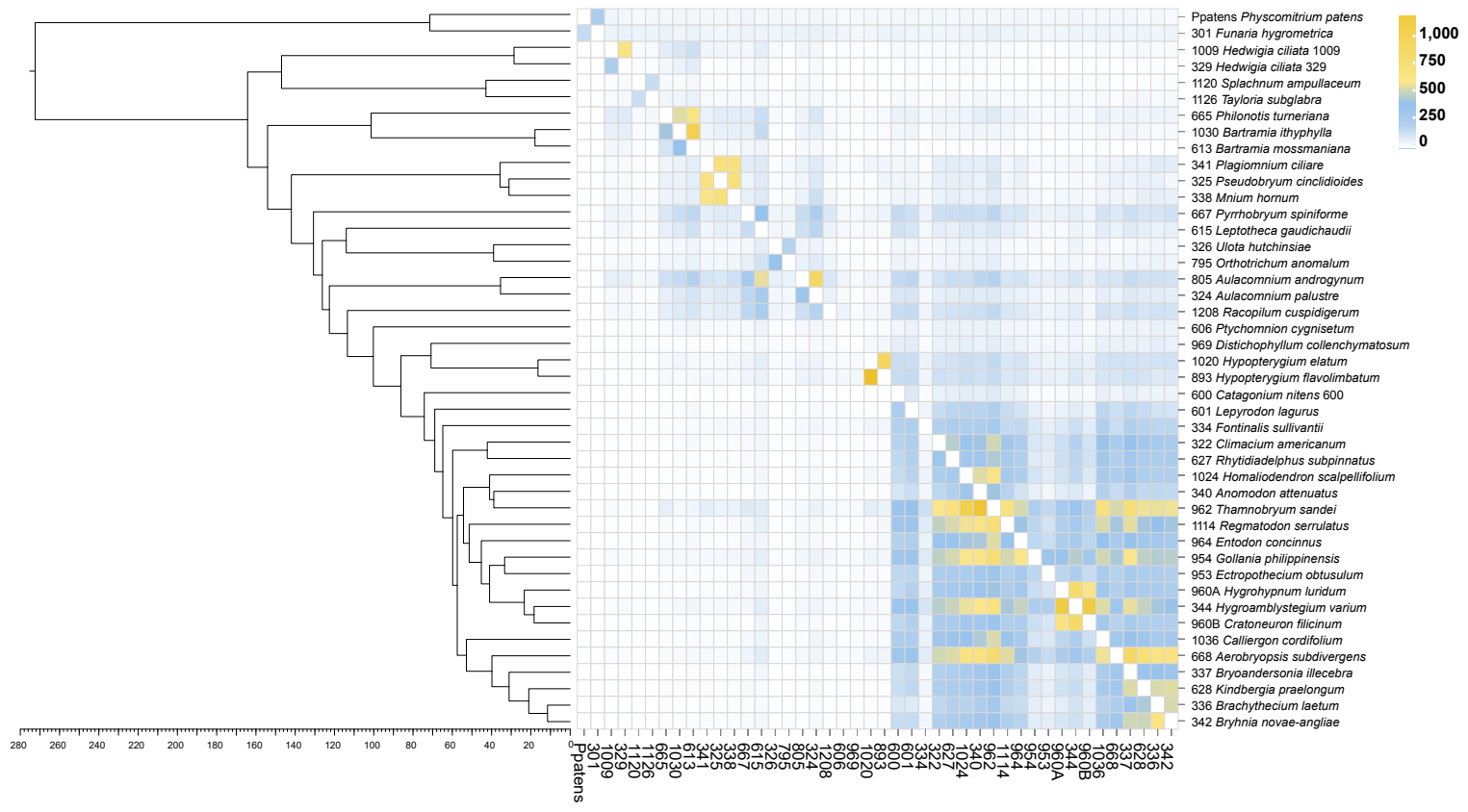


Supplementary Figure 3. Percentage of genes expressed (a) and median expression level of genes (b) in the gene families of three categories in 60 bryophyte species with transcriptome data. The numbers of core, accessory, and unique genes for each species were indicated in Supplementary Data 6.



Supplementary Figure 4. Candidate cases of new genes of *Marchantia polymorpha ssp. ruderalis* (MPO) originated from the non-coding DNA sequences in *Marchantia*. **(a)** Venn diagram showing the number of unique genes of MPO with high-confidence blastn hits (e- value < 1e-10) in the non- coding regions of the other three closely related Marchantia species. **(b)** Alignments showing evolutionary trajectory of the several representative unique genes of MPO. The genome alignments were generated with MAFFT and visualized in Geneious v10.0.2 to trace the origination of unique genes. The gray bar denotes nucleotide sequence, the yellow bar denotes annotation of the coding sequence (CDS), the grey bar above the CDS bar denotes amino acid sequence, and differences from the concensus sequences was highlighted in colored dots, among which, the black dots denote stop codons in the proto-cds. The corresponding gene is in the MP TAK-1/2 std v7.1 genome assembly was indicated above each alignment.

a



b

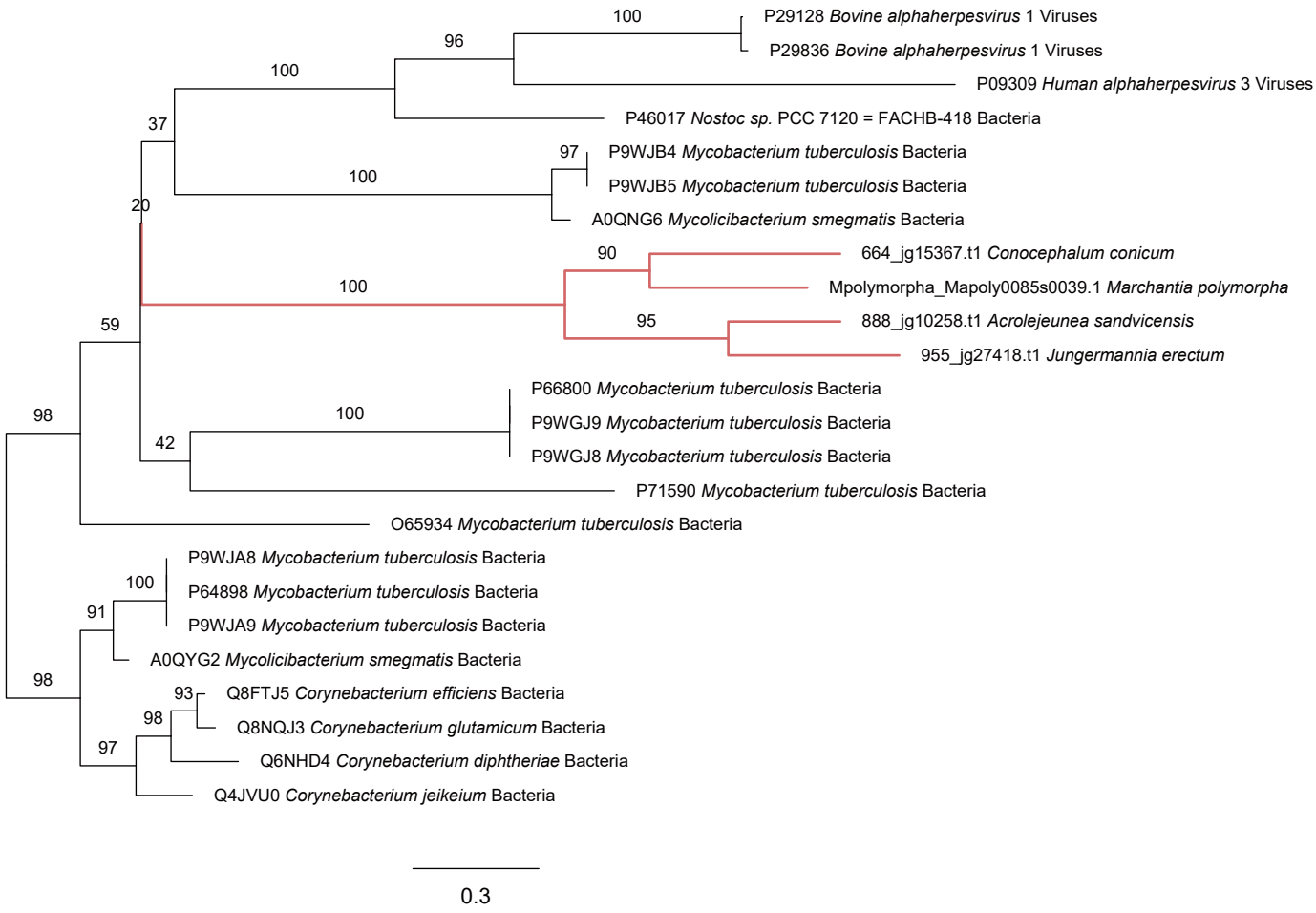
Supplementary Figure 5. Candidate cases of new genes originated from the non-coding DNA sequences in Bryidae mosses. (a) Heatmap showing the number of unique genes with high-confidence blastn hits (e-value < 1e-10) in the non-coding regions of at least 2 other moss species. **(b)** Alignments showing evolutionary trajectory of the several representative unique genes of Bryhnia novae-angliae (DNA# 342). The genome alignments were generated with MAFFT and visualized in Geneious v10.0.2 to trace the origination of unique genes. The gray bar denotes nucleotide sequences, the yellow bar denotes annotation of the coding sequence (CDS), the grey bar above the CDS bar denotes amino acid sequence, and differences from the consensus sequence were highlighted in colored dots, among which, the black dots denote stop codons in the *proto-cds*. The species phylogeny is show on the left.



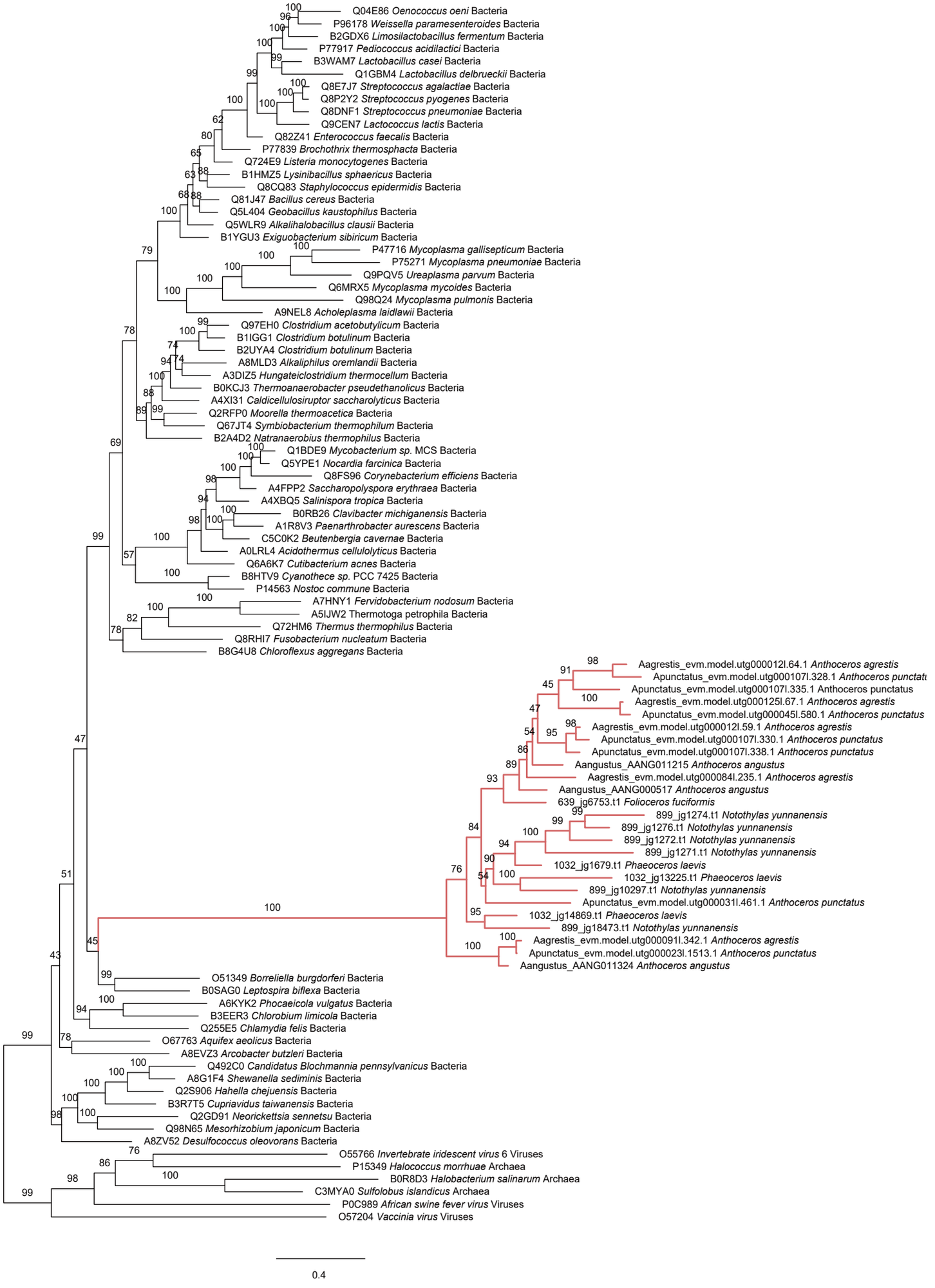
Supplementary Figure 6 The phylogram of the (E)-beta farnesene synthase inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



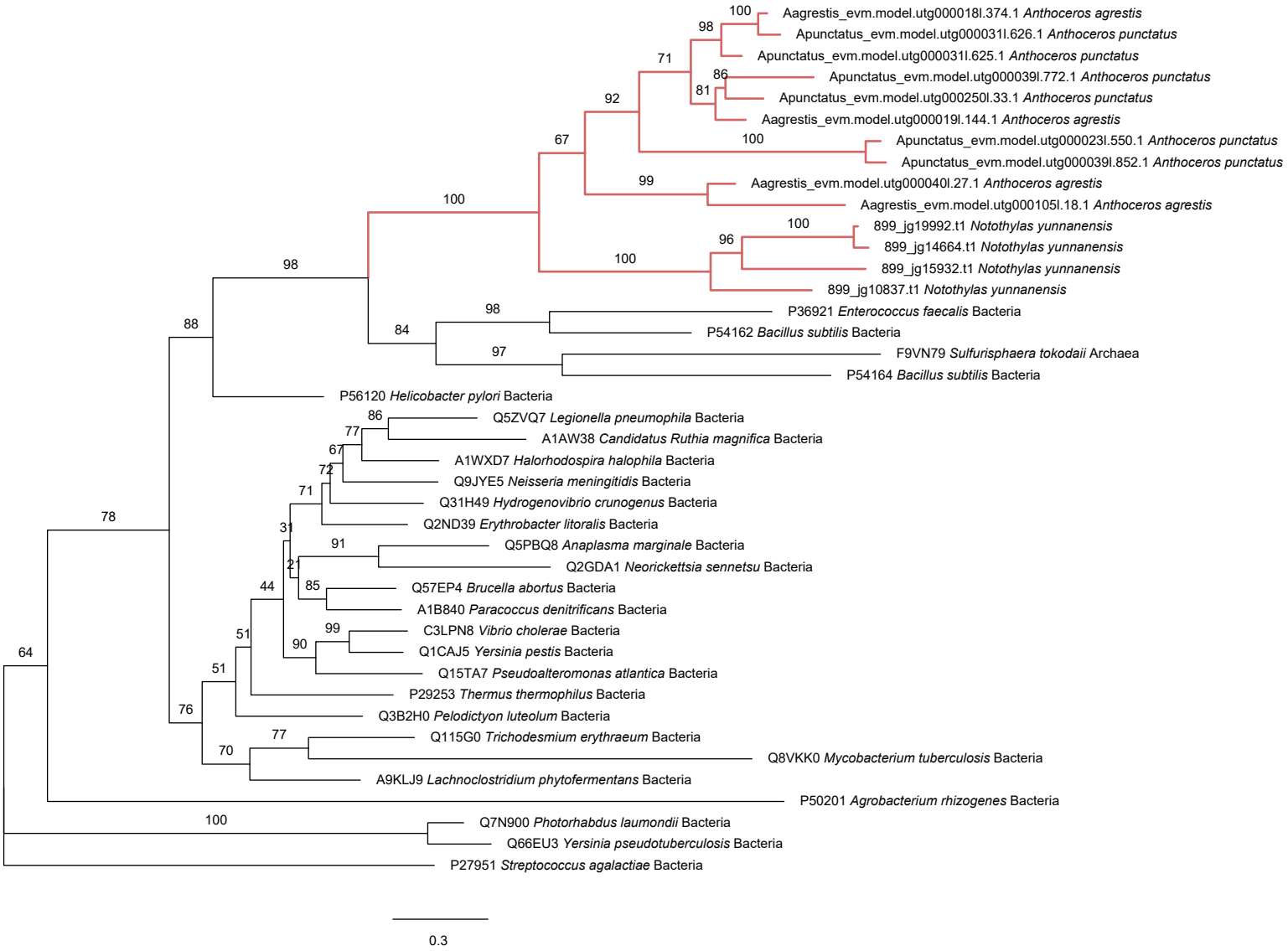
Supplementary Figure 7. The phylogram of the Flavin-containing monooxygenase inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



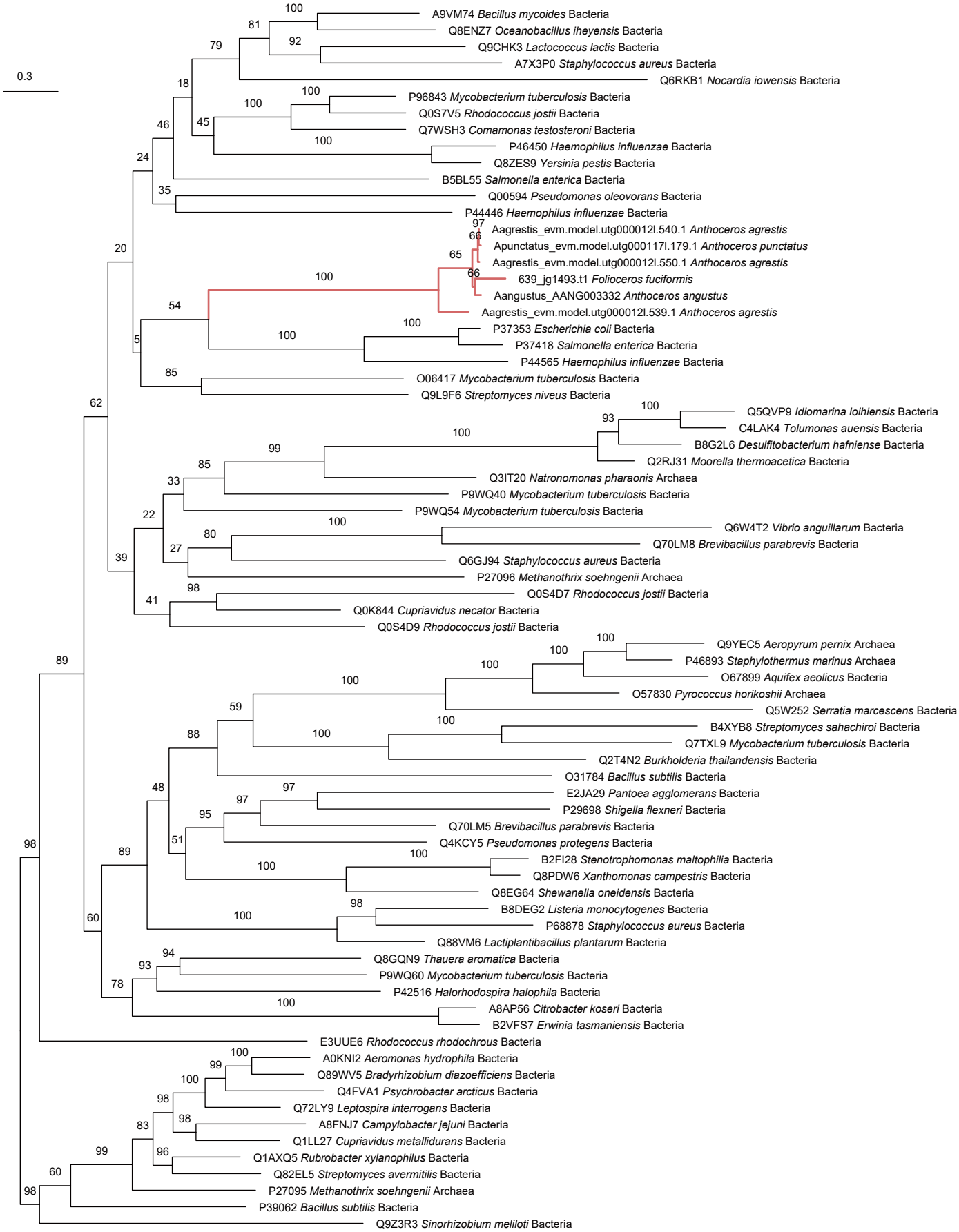
Supplementary Figure 8. The phylogram of the FHA domain (PF00498) containing genes inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



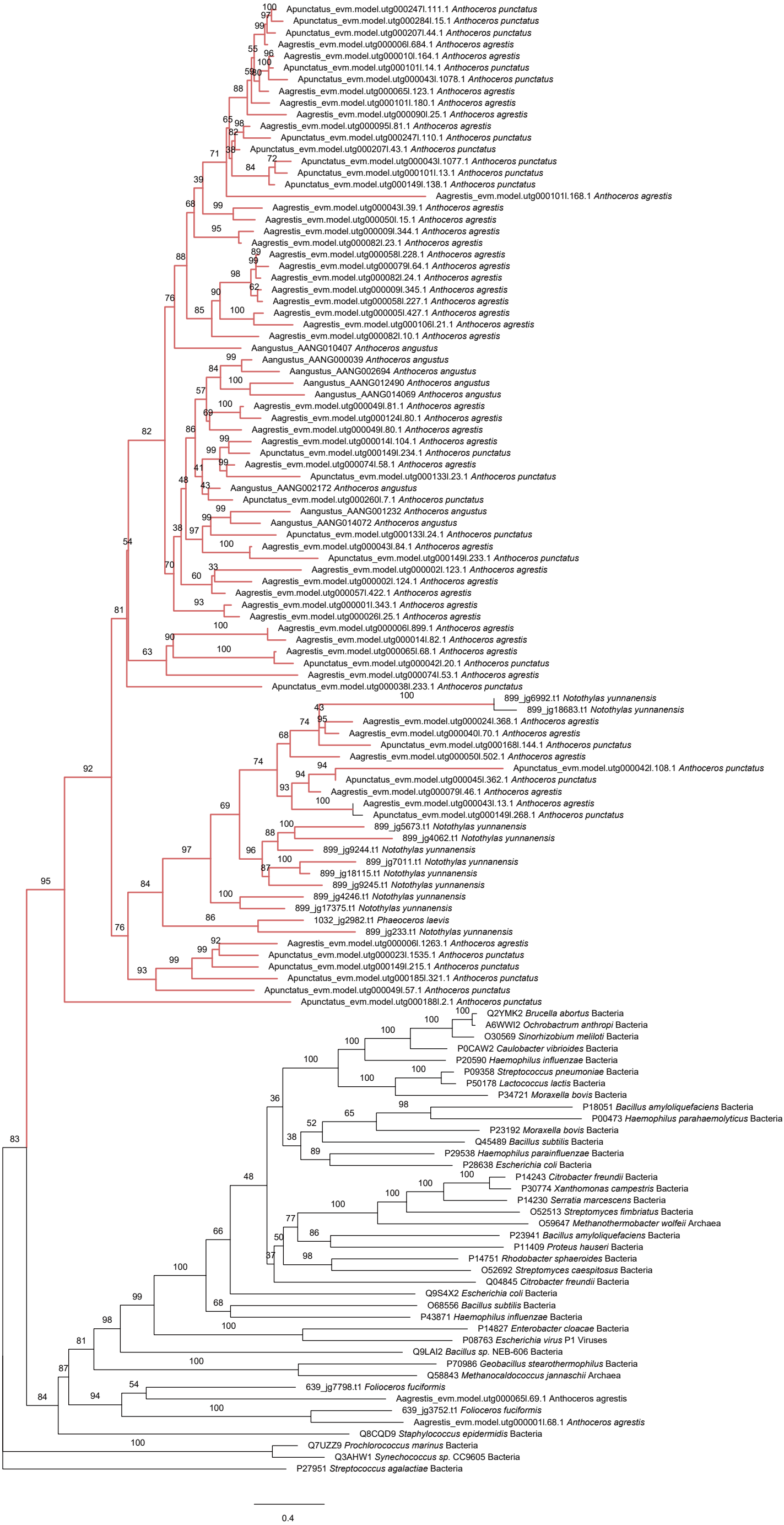
Supplementary Figure 9. The phylogram of an uncharacterized gene inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



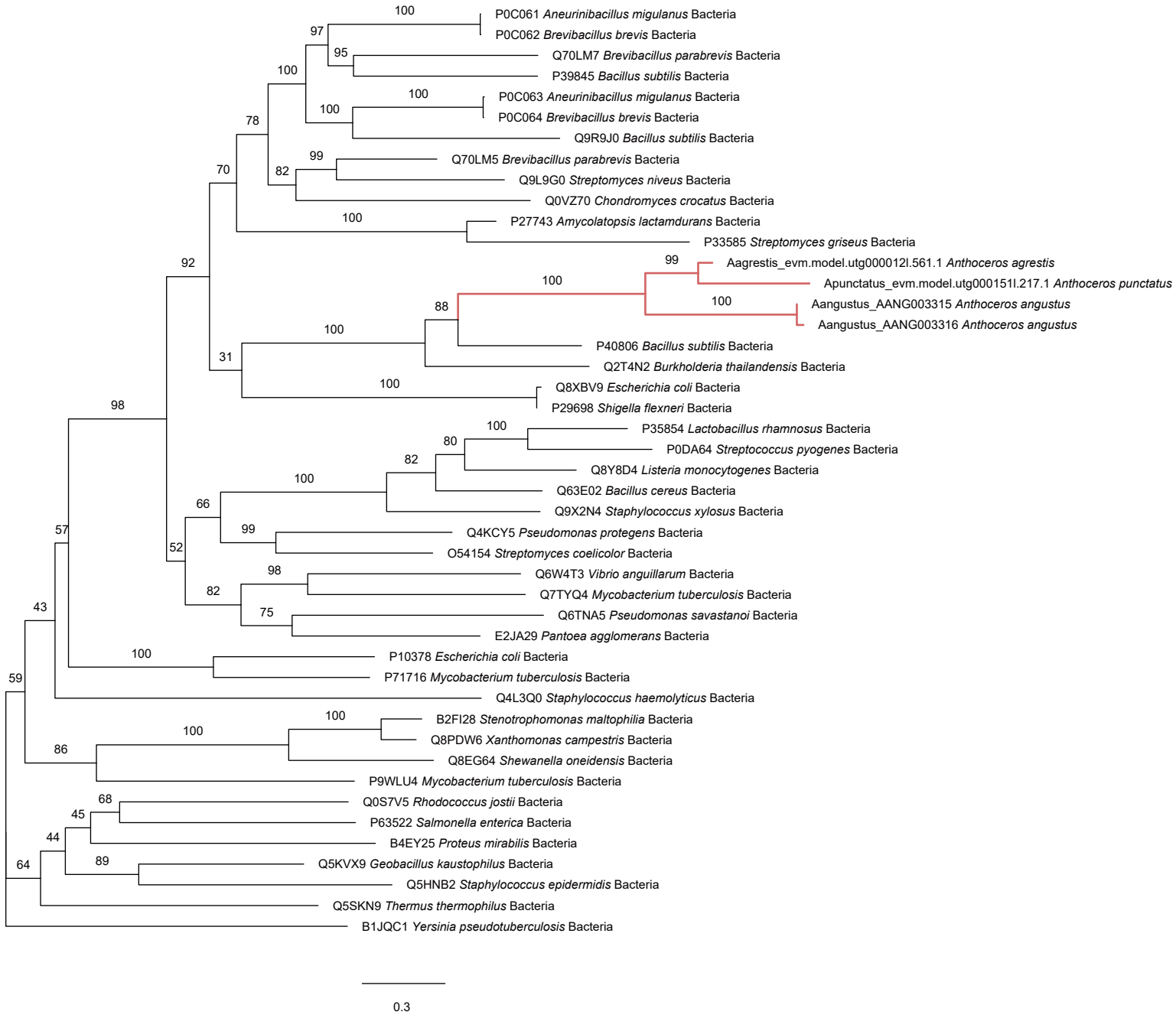
Supplementary Figure 10. The phylogram of retrovirus-related Pol polyprotein inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



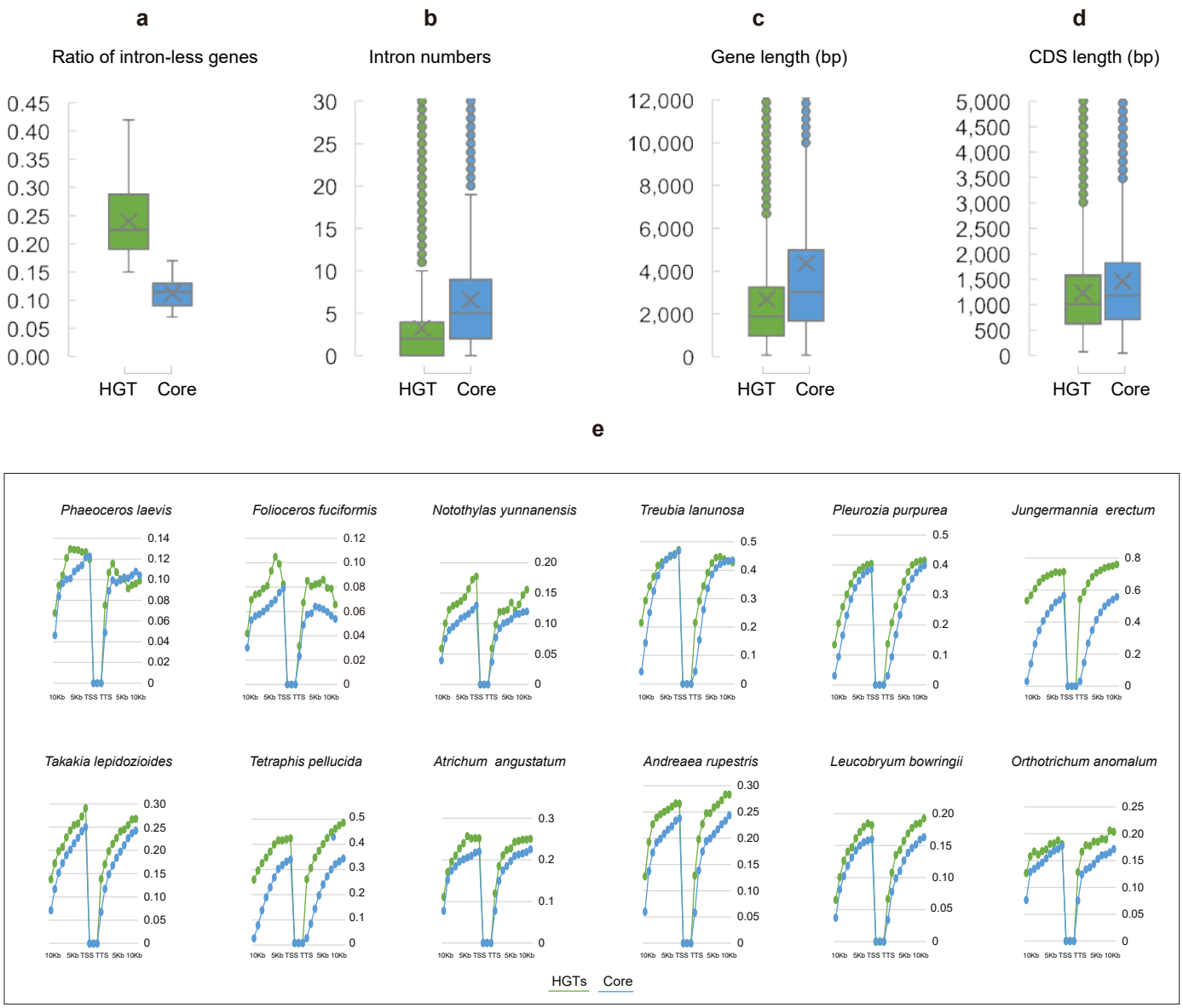
Supplementary Figure 11. The phylogram of L-proline--ligase (Pltf) inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



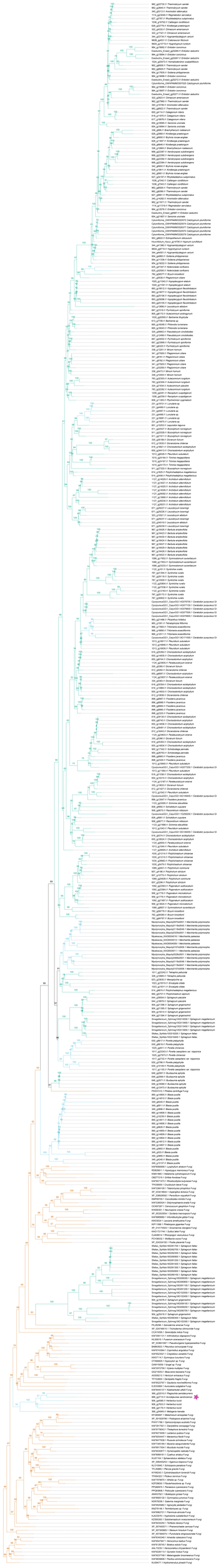
Supplementary Figure 12. The phylogram of DNA methylase N-4/N-6 domain containing gene inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



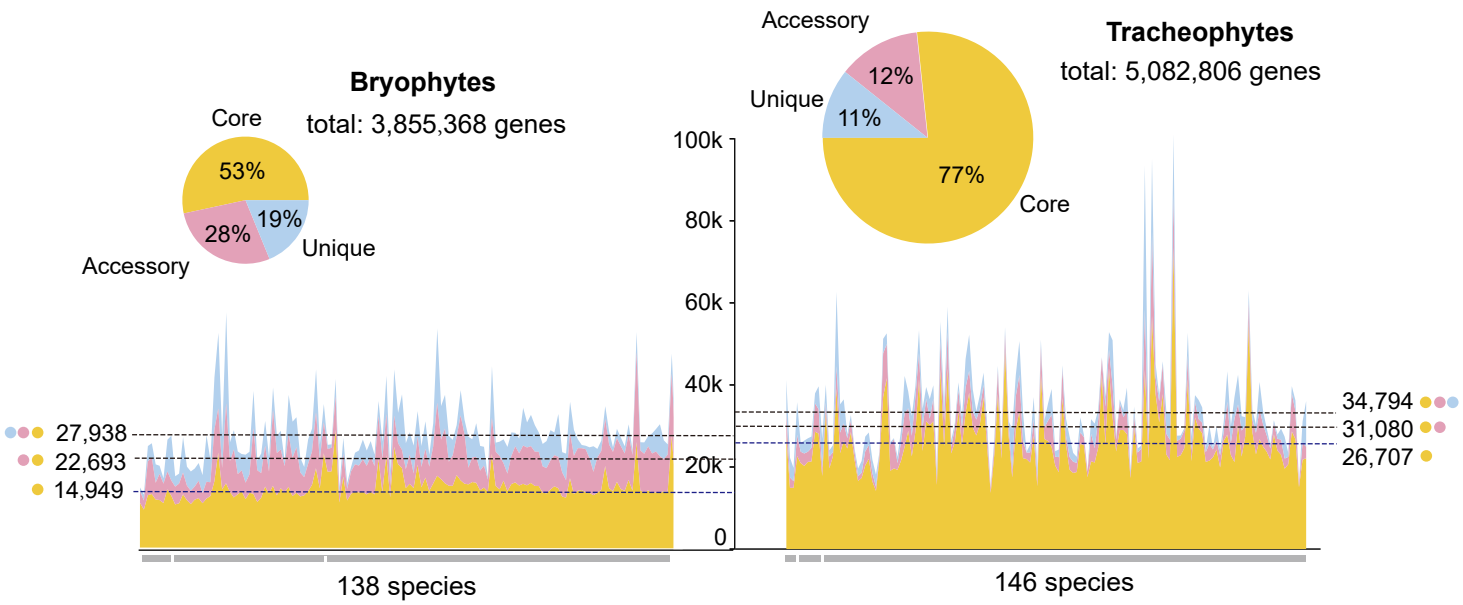
Supplementary Figure 13. The phylogram of 4-coumarate--CoA ligase (4CL) inferred using IQTREE2 with the JTT amino acid substitution model. Ultra-fast bootstrap supports were indicated on branches, bryophyte branches were highlighted in red.



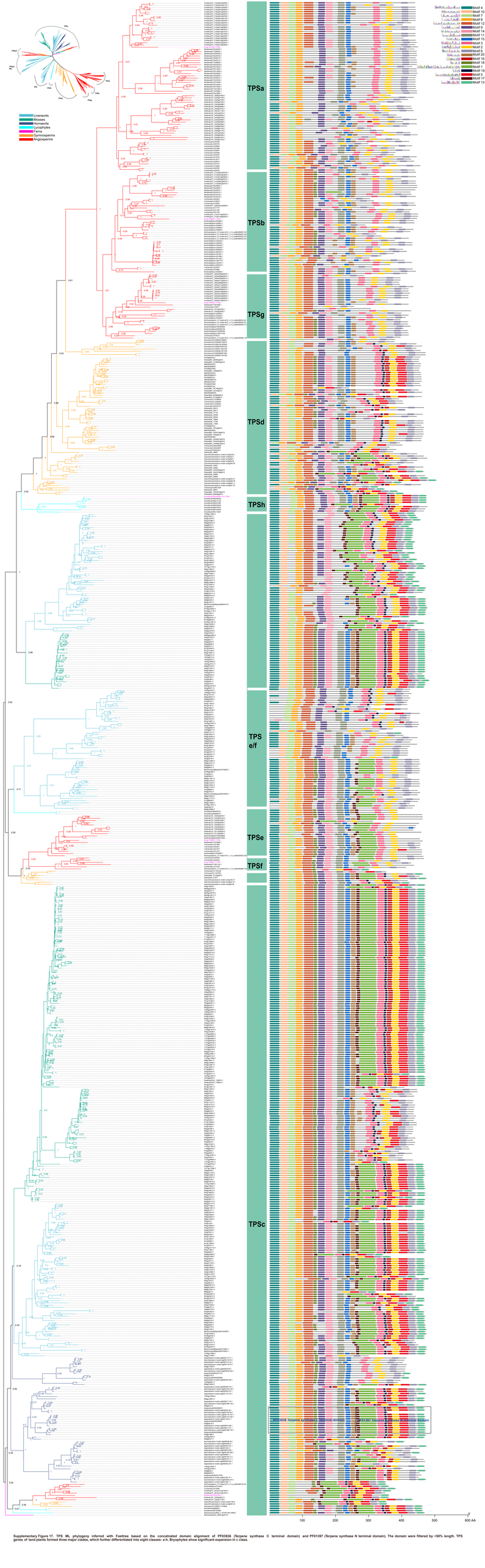
Supplementary Figure 14. Characteristics of HGT genes compared with Viridiplantae core genes. (**a**) intronless gene ratio; (**b**) intron number per gene; (**c**) gene length; (**d**) coding sequence (CDS) length; (**e**) LTR coverage in the 10 kb upstream and downstream region of hgt genes and core genes in different bryophyte species.



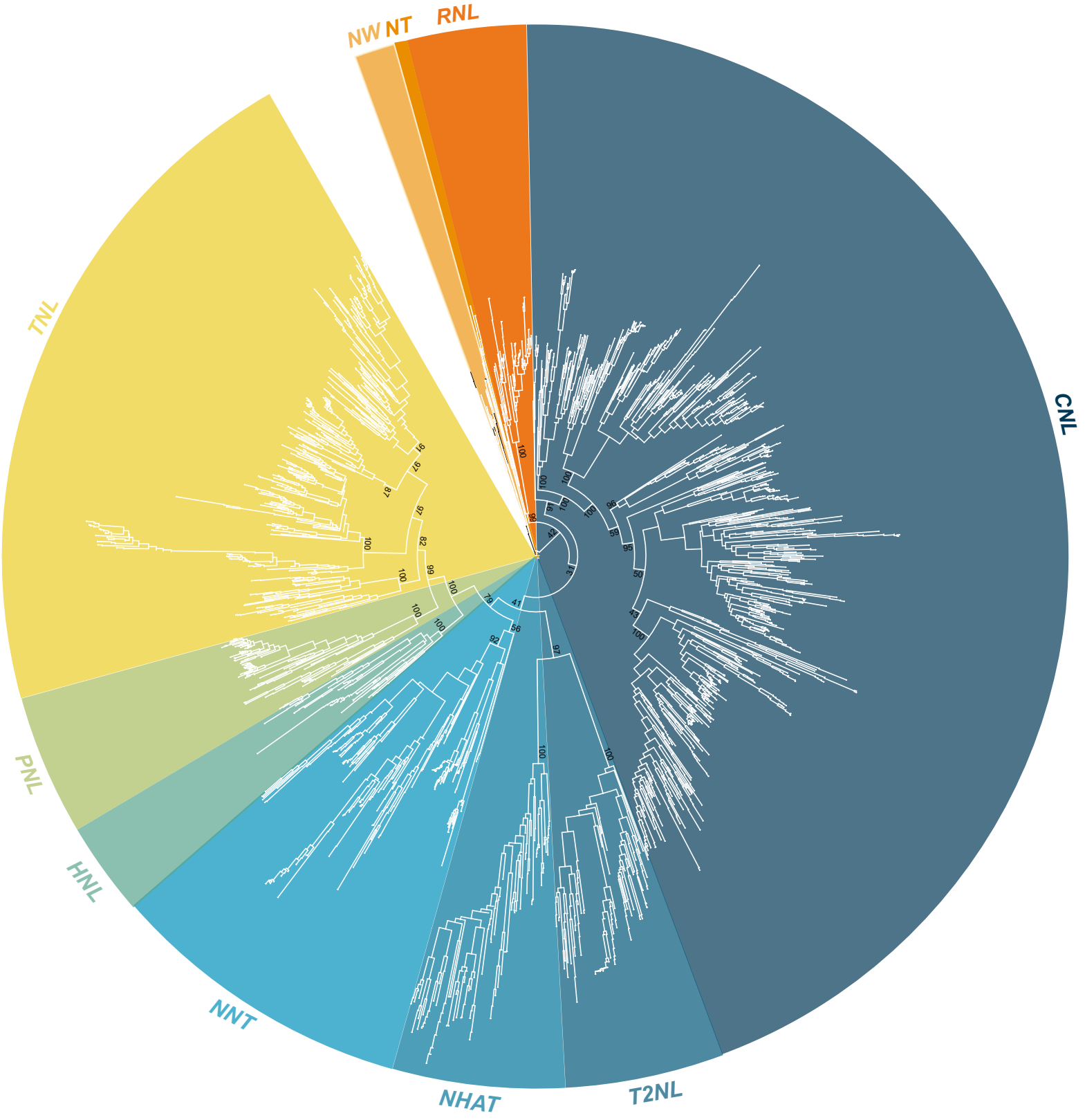
Supplementary Figure 15. Phylogram of FBT (Fungal body lectin domain PF07367 containing) proteins. Maximum likelihood tree of FBT proteins in Boughoutiales inferred with IQTREE2 with maximum likelihood ultrafast bootstrap support. Aligned amino acid sequences. The best-fit model was estimated by ModelFinder and branch support values were assessed with UFBoot2. tests. Branches leading to insects, nematodes and fungi accessions were shaded in green, blue and yellow respectively. The black dashed accession of gene *888_gp27101* from *Acrotejeuna sandunesnes* were used for insect toxicology experiments. Branches leading to insects, nematodes and fungi accessions were shaded in green, blue and yellow respectively.



Supplementary Figure 16. The number of genes of each category (core, accessory, and unique) in individual genomes of bryophytes and tracheophytes. Taxa are sorted phylogenetically based on the tree in Extended Data Fig. 3.

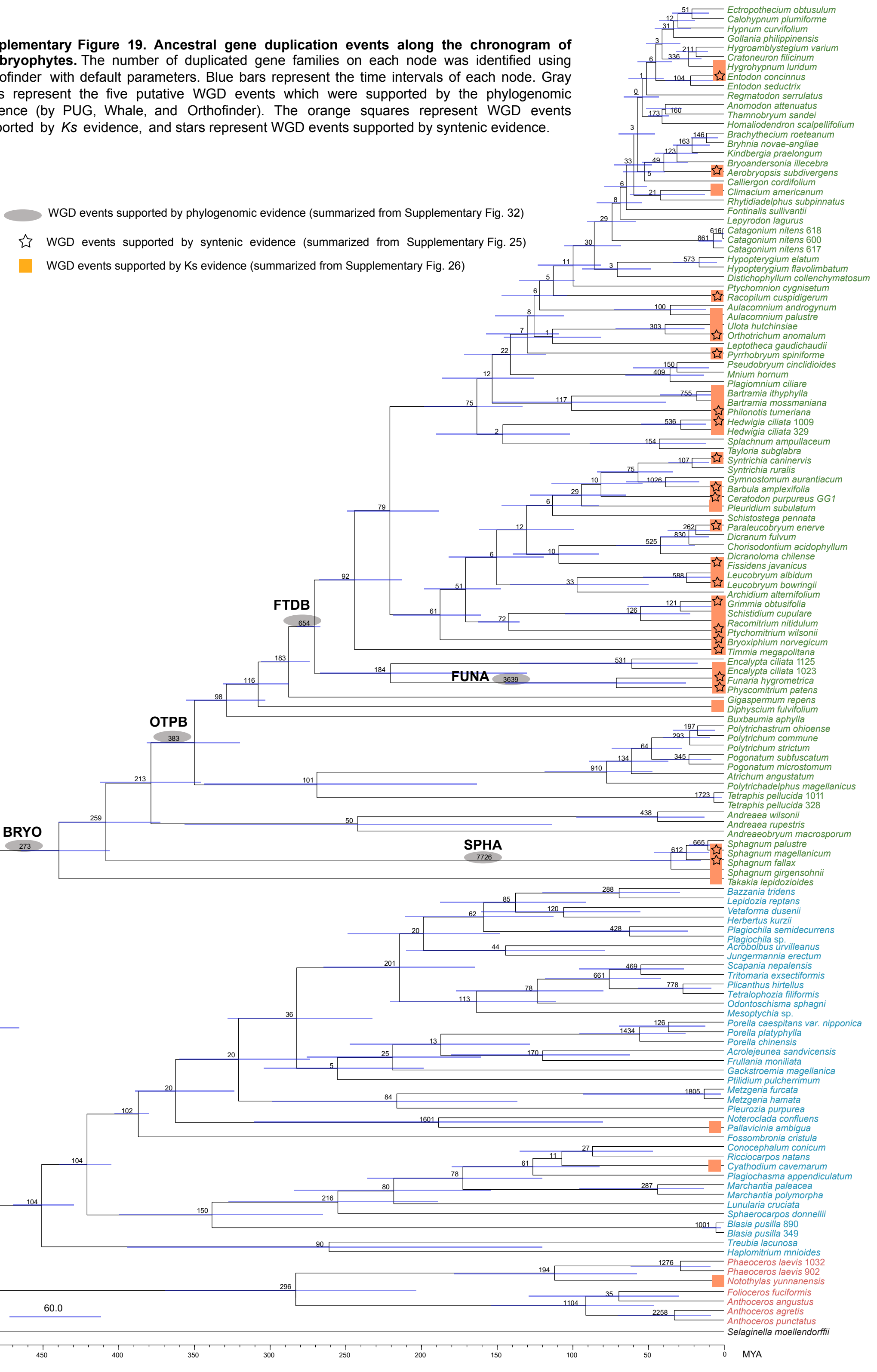


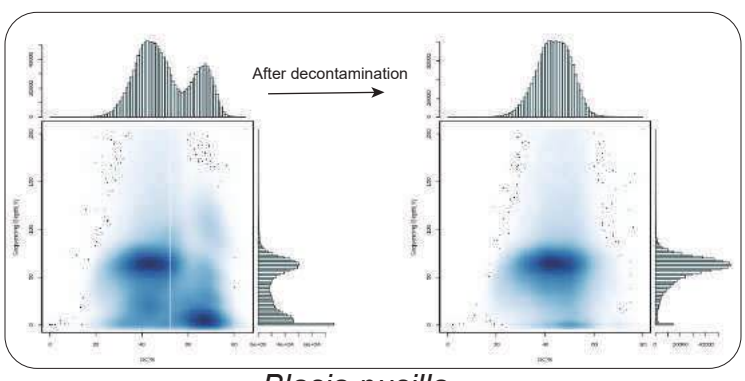
Supplementary Figure 17. TPS ML phylogeny inferred with Fasttree based on the concatenated domain alignment of PF03936 (Terpene synthase C terminal domain) and PF01397 (Terpene synthase N terminal domain). The domain were filtered by >50% length. TPS genes of land plants formed three major clades, which further differentiated into eight classes: a-h. Bryophytes show significant expansion in c class.



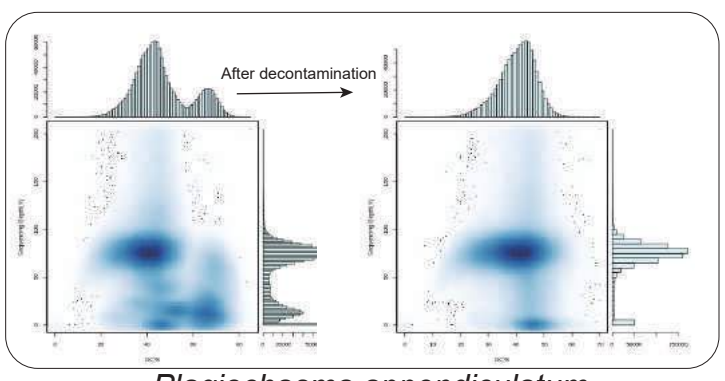
Supplementary Figure 18. Phylogenetic tree of *NLR* genes in green plants inferred with IQTREE2 with maximum likelihood method based on the aligned amino acid sequences of the conserved NBS domains. The best-fit model was estimated by ModelFinder and branch support values were assessed with UFBoot2 tests. Immune receptors of different types were colored accordingly.

Supplementary Figure 19. Ancestral gene duplication events along the chronogram of the bryophytes. The number of duplicated gene families on each node was identified using Orthofinder with default parameters. Blue bars represent the time intervals of each node. Gray ovals represent the five putative WGD events which were supported by the phylogenomic evidence (by PUG, Whale, and Orthofinder). The orange squares represent WGD events supported by *Ks* evidence, and stars represent WGD events supported by syntenic evidence.

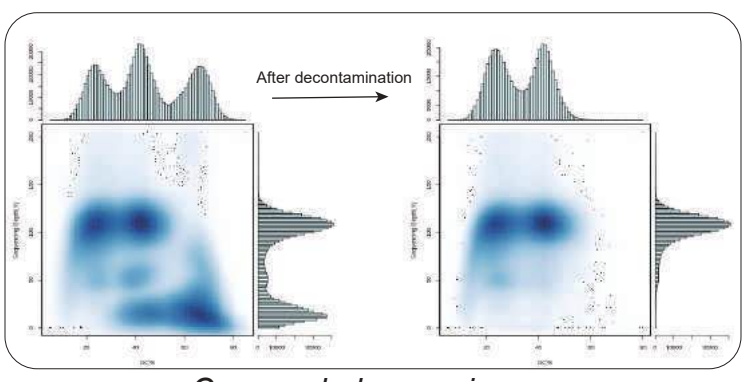




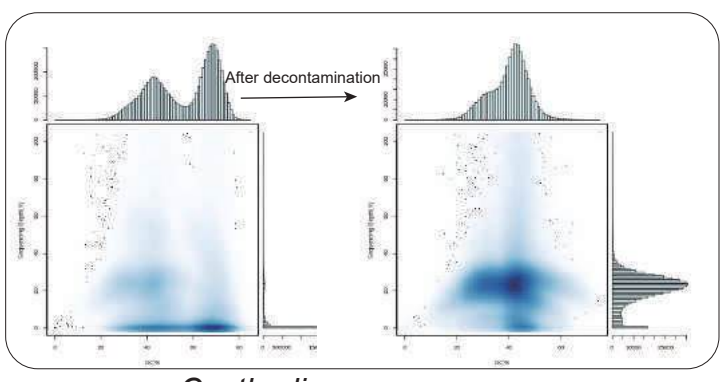
Blasia pusilla



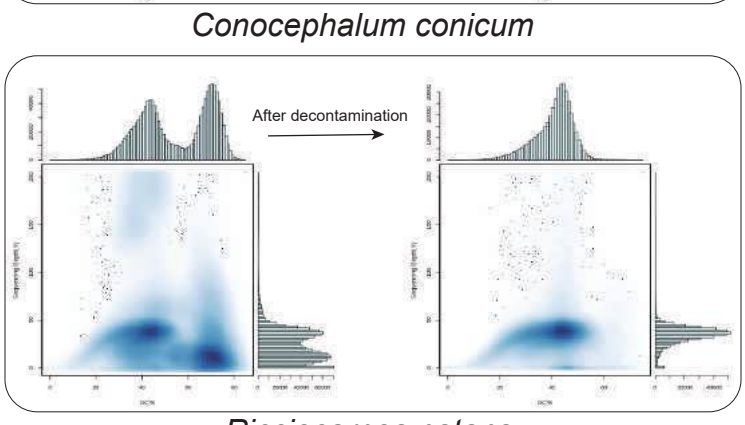
Plagiochasma appendiculatum



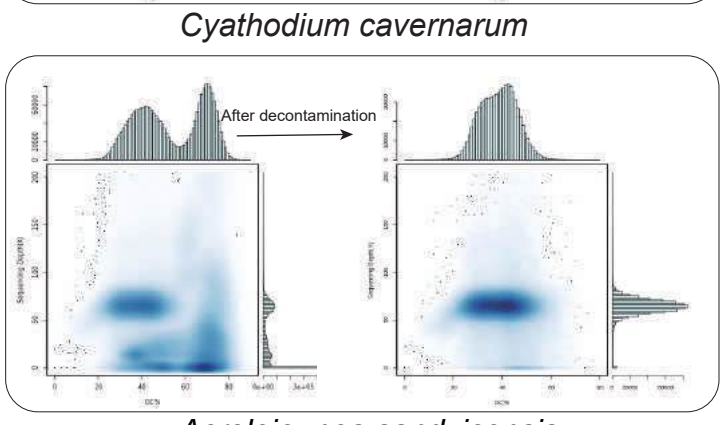
Conocephalum conicum



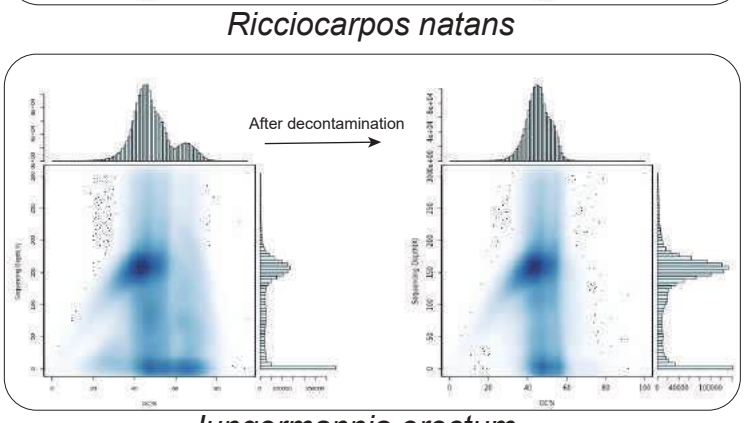
Cyathodium cavernarum



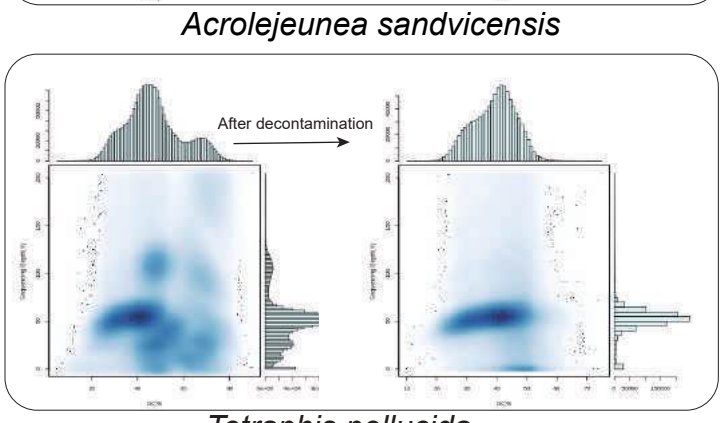
Ricciocarpos natans



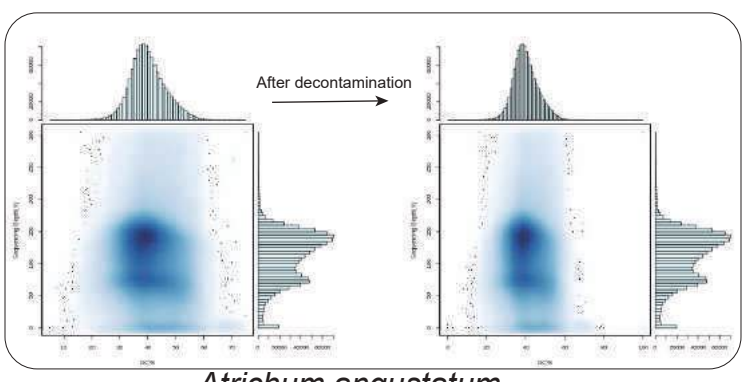
Acrolejeunea sandvicensis



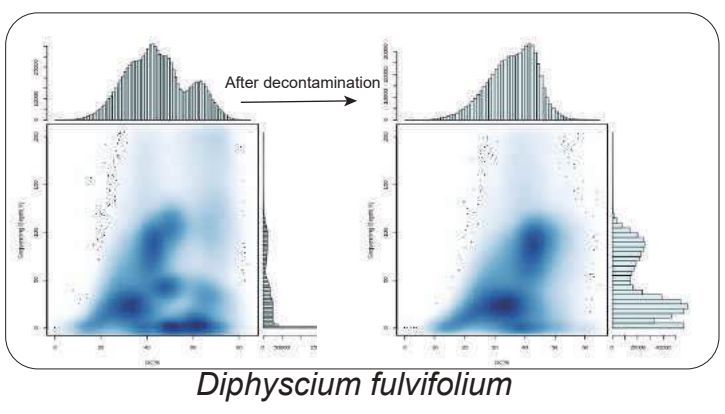
Jungermannia erectum



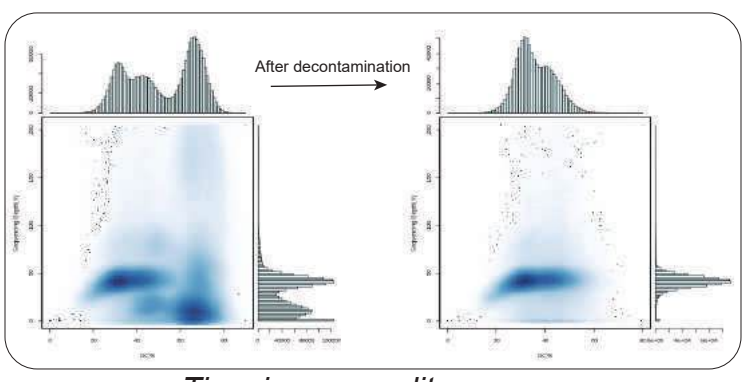
Tetraphis pellucida



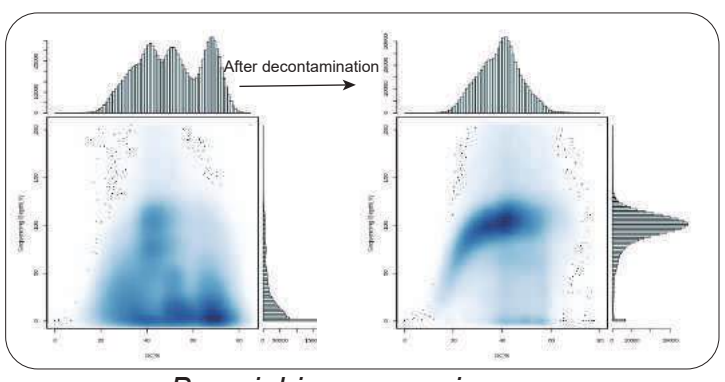
Atrichum angustatum



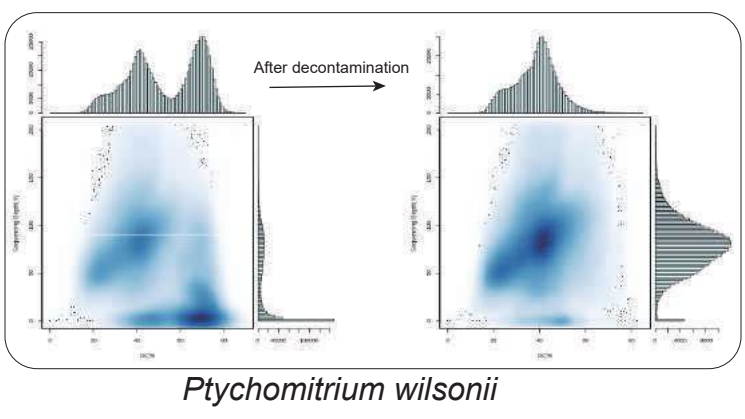
Diphyscium fulvifolium



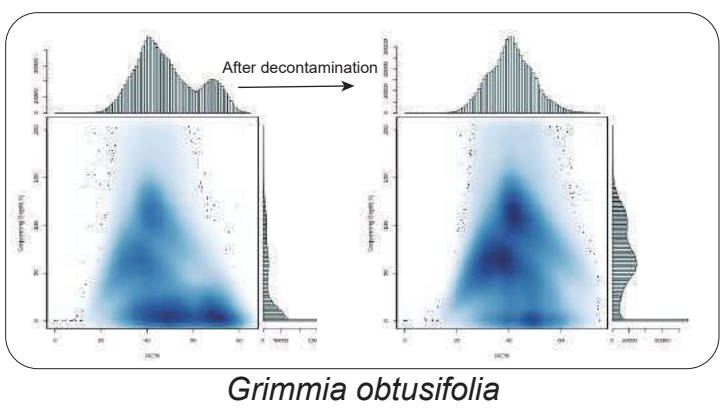
Timmia megapolitana



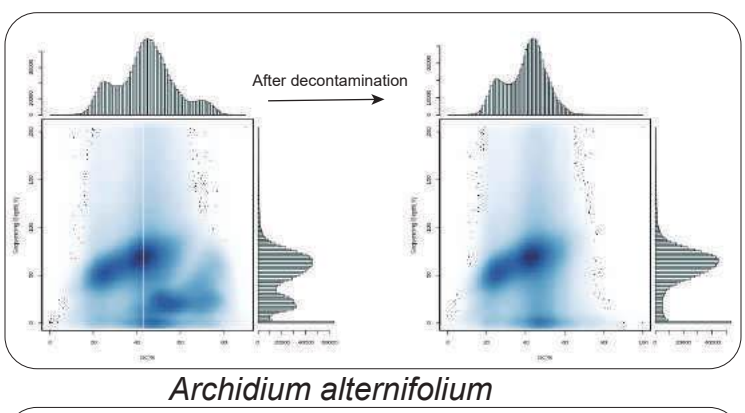
Bryoxiphium norvegicum



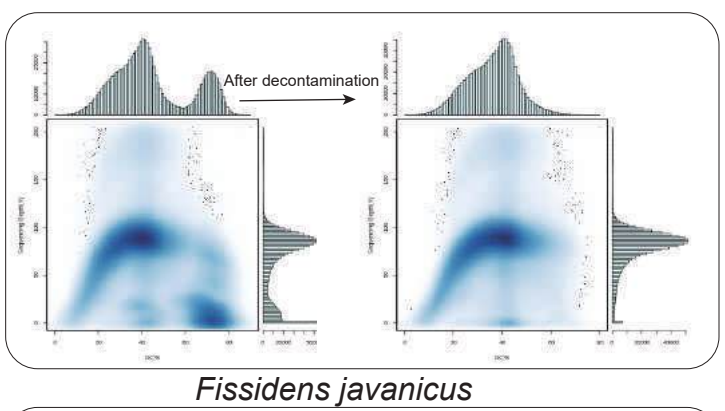
Ptychomitrium wilsonii



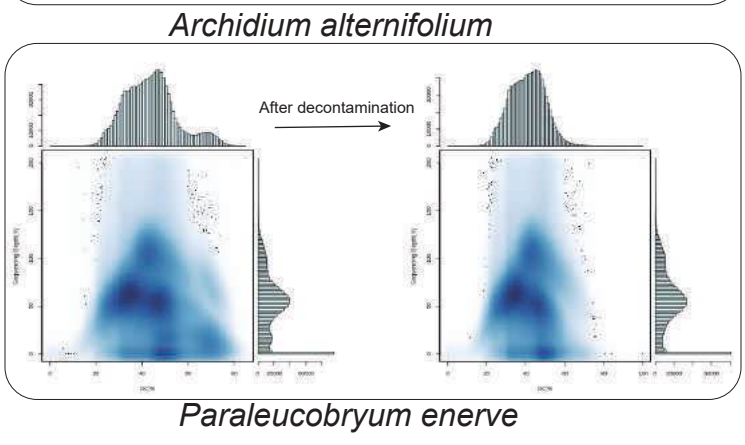
Grimmia obtusifolia



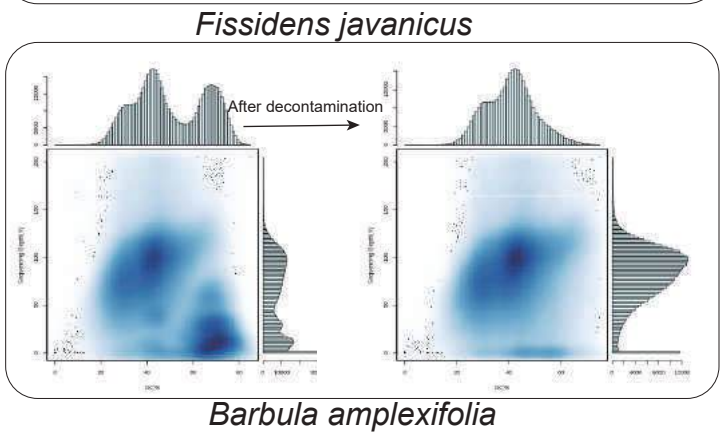
Archidium alternifolium



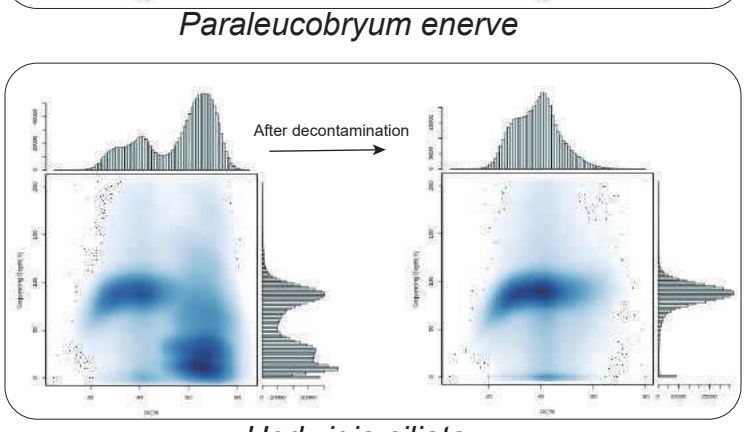
Fissidens javanicus



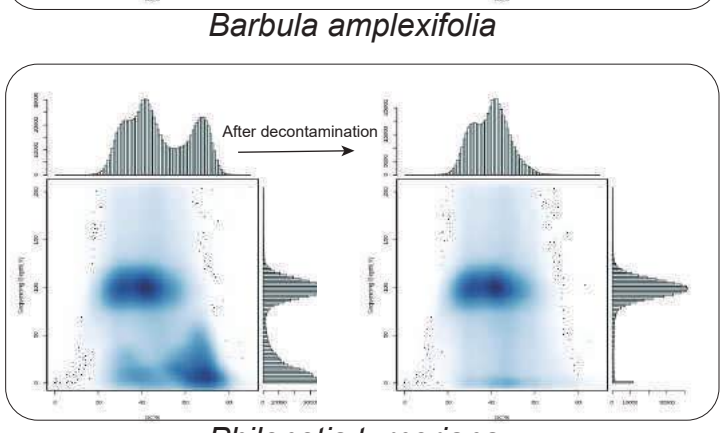
Paraleucobryum enerve



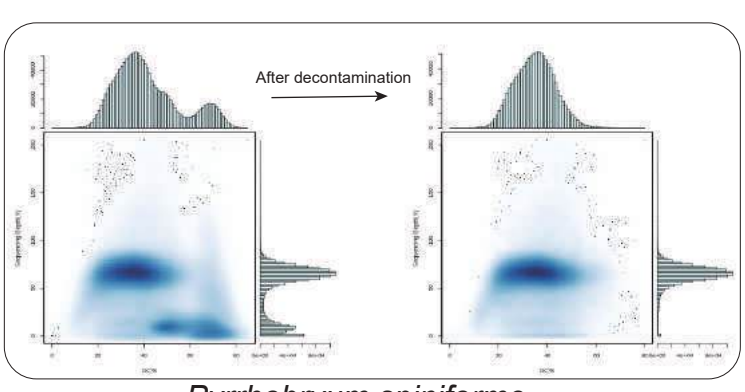
Barbula amplexifolia



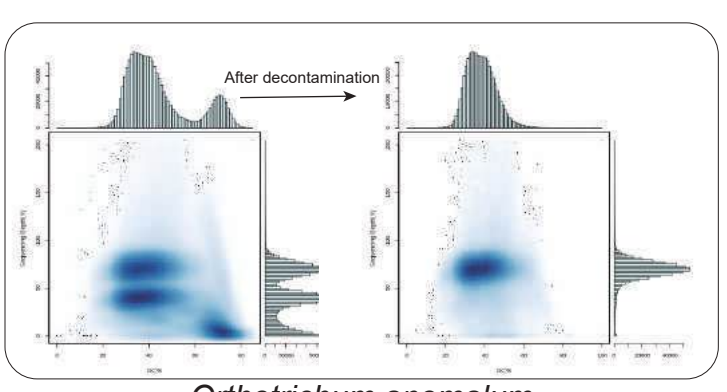
Hedwigia ciliata



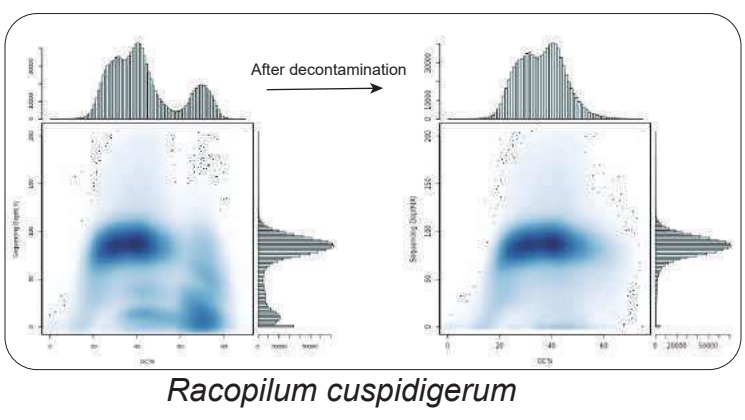
Philonotis turneriana



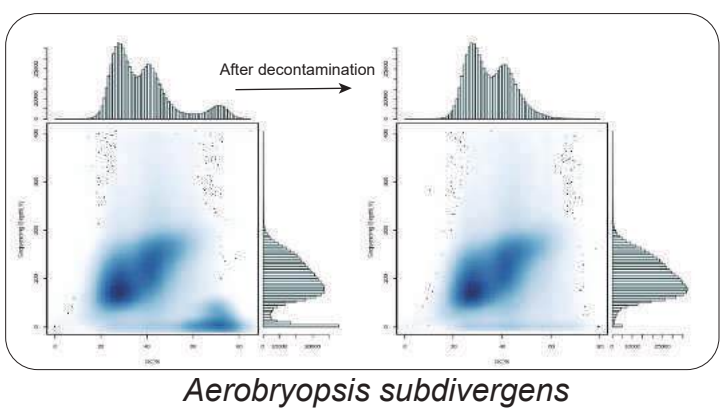
Pyrrhobryum spiniforme



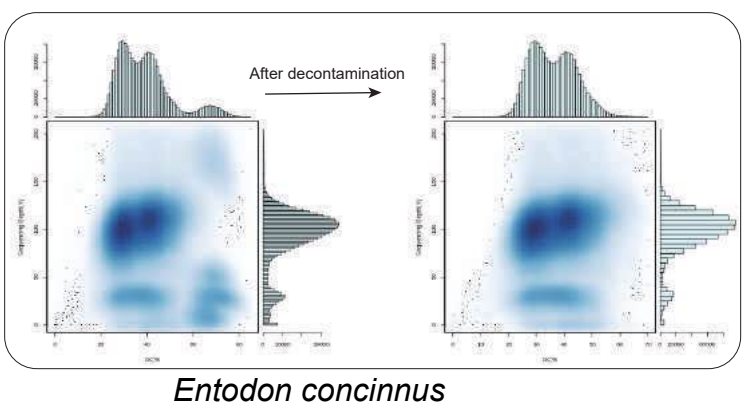
Orthotrichum anomalum



Racopilum cuspidigerum

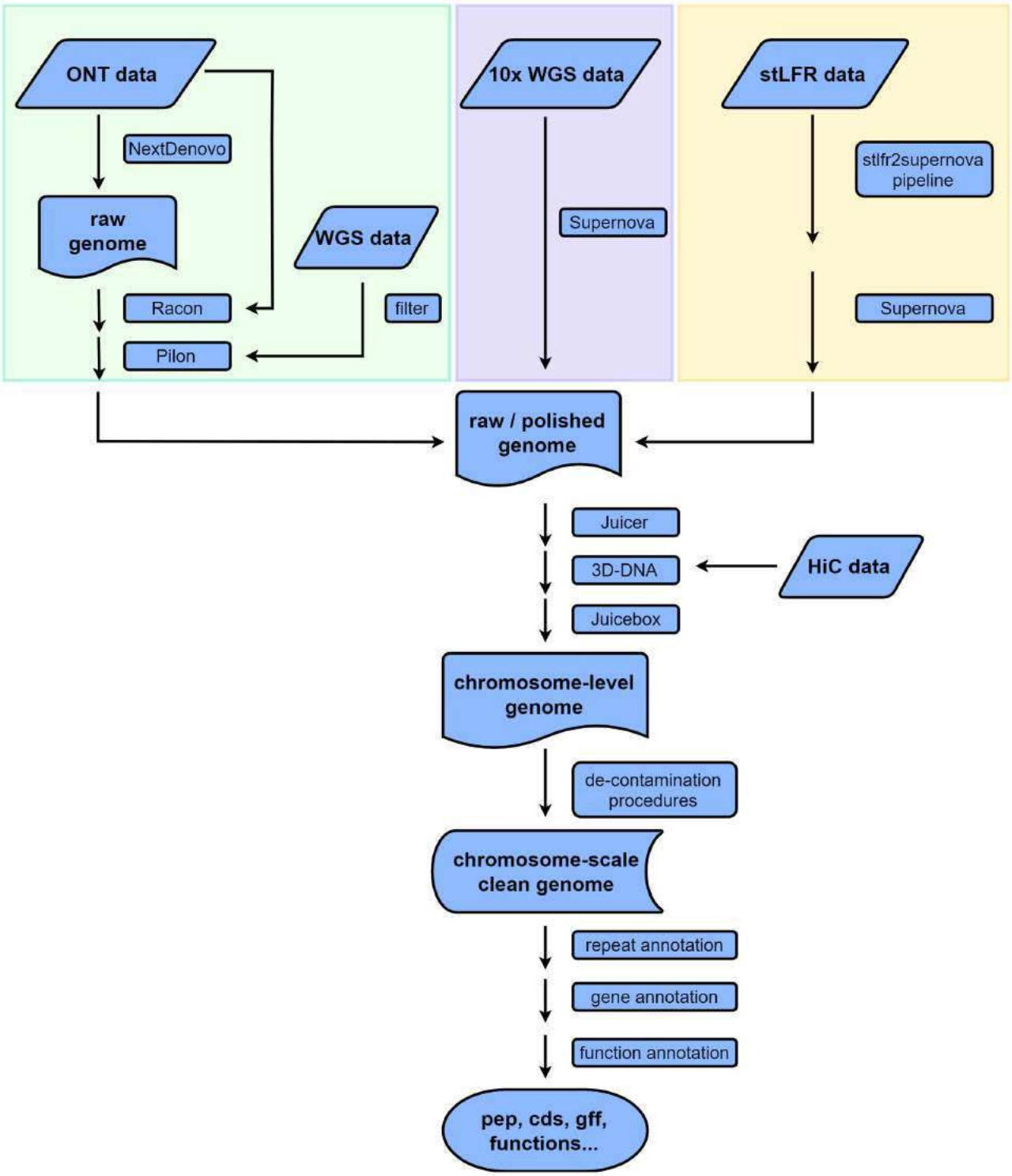


Aerobryopsis subdivergens

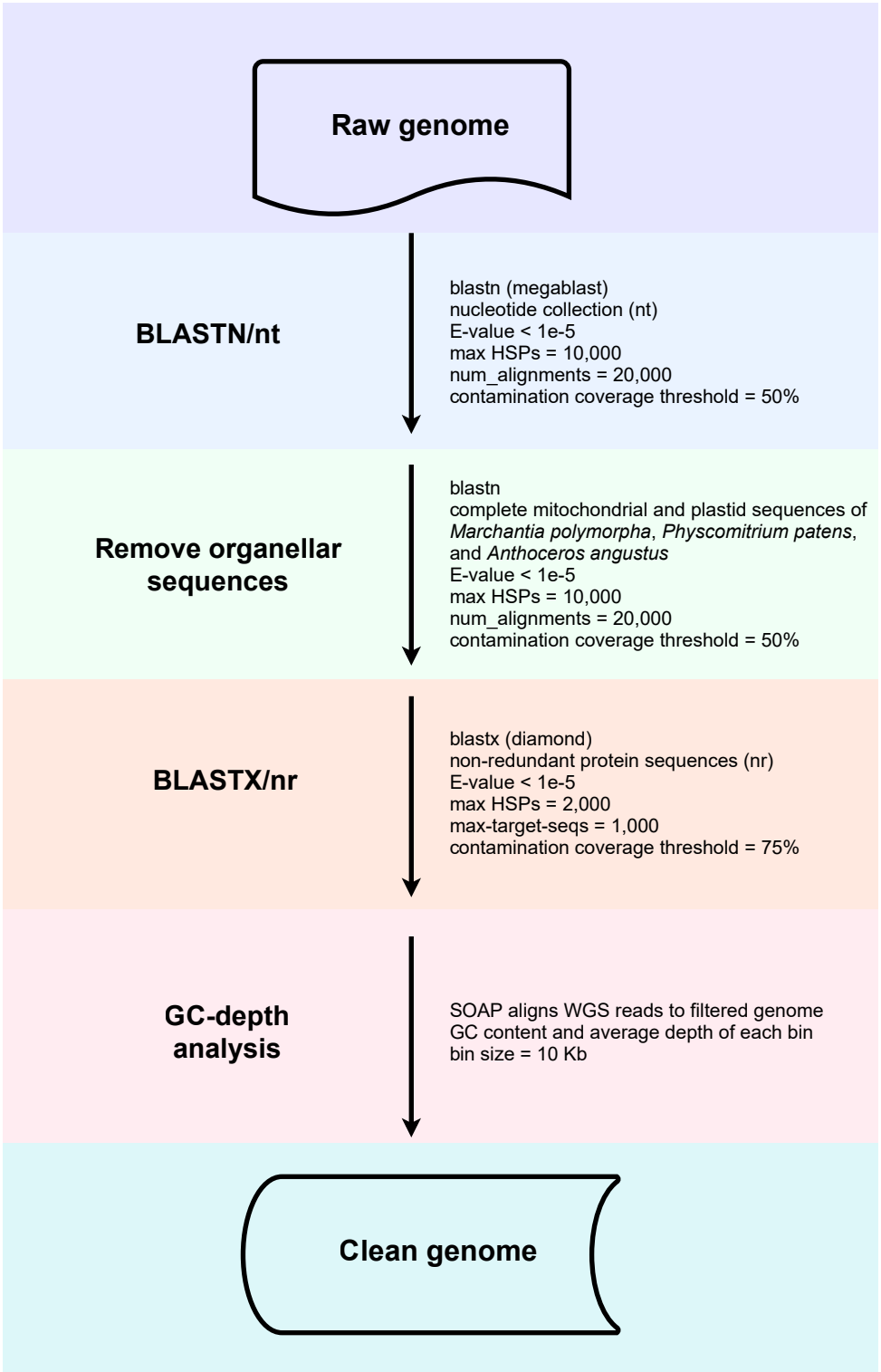


Entodon concinnus

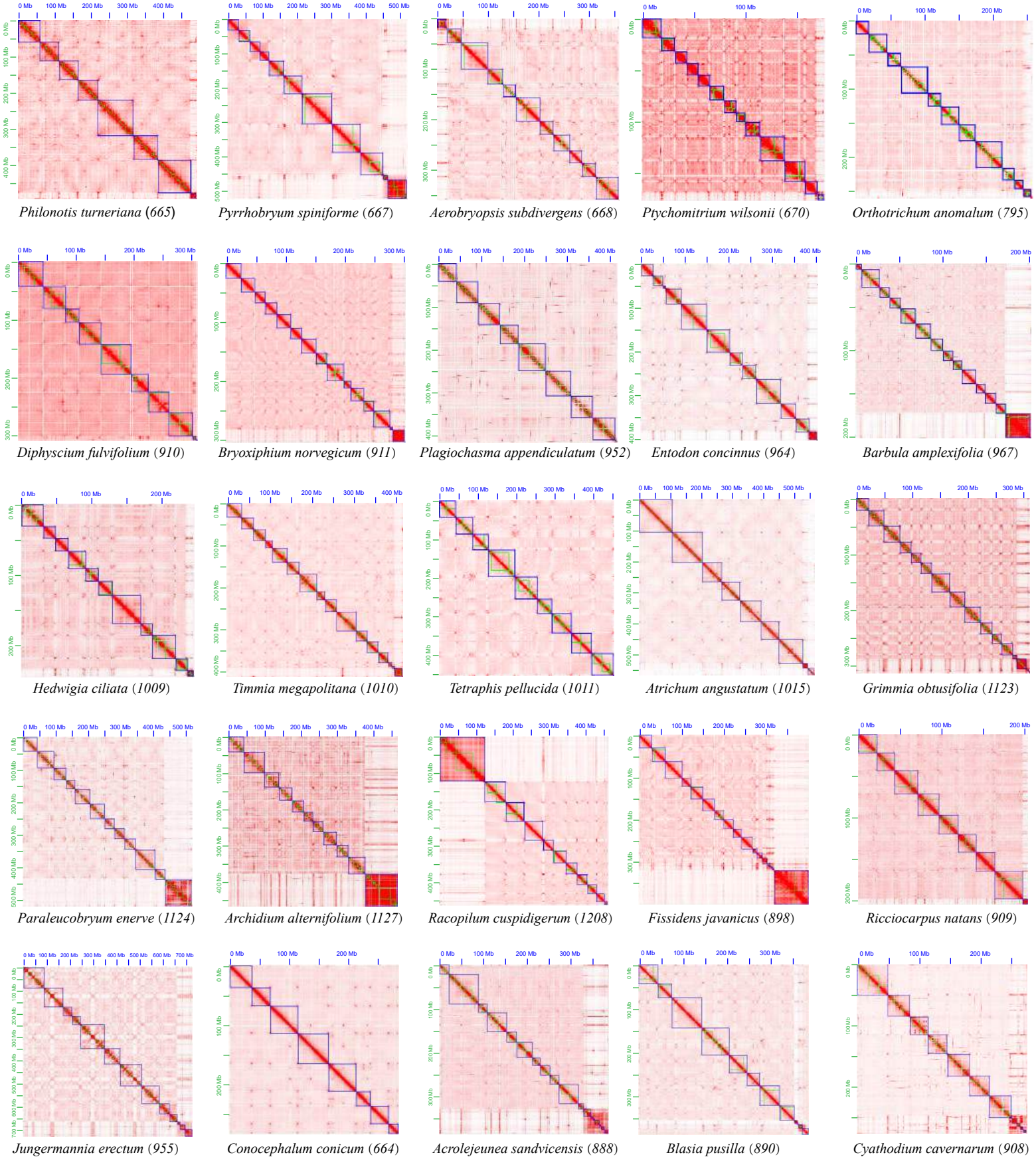
Supplementary Figure 20. Comparative GC depth plot of representative bryophyte genomes for raw and final assemblies showing the effectiveness of decomtanination. The plot of the GC content and sequencing depth of the raw assembly *vs*. final assembly was generated using short read sequencing data. Species were sorted by phylogenetic order in Figure 1.



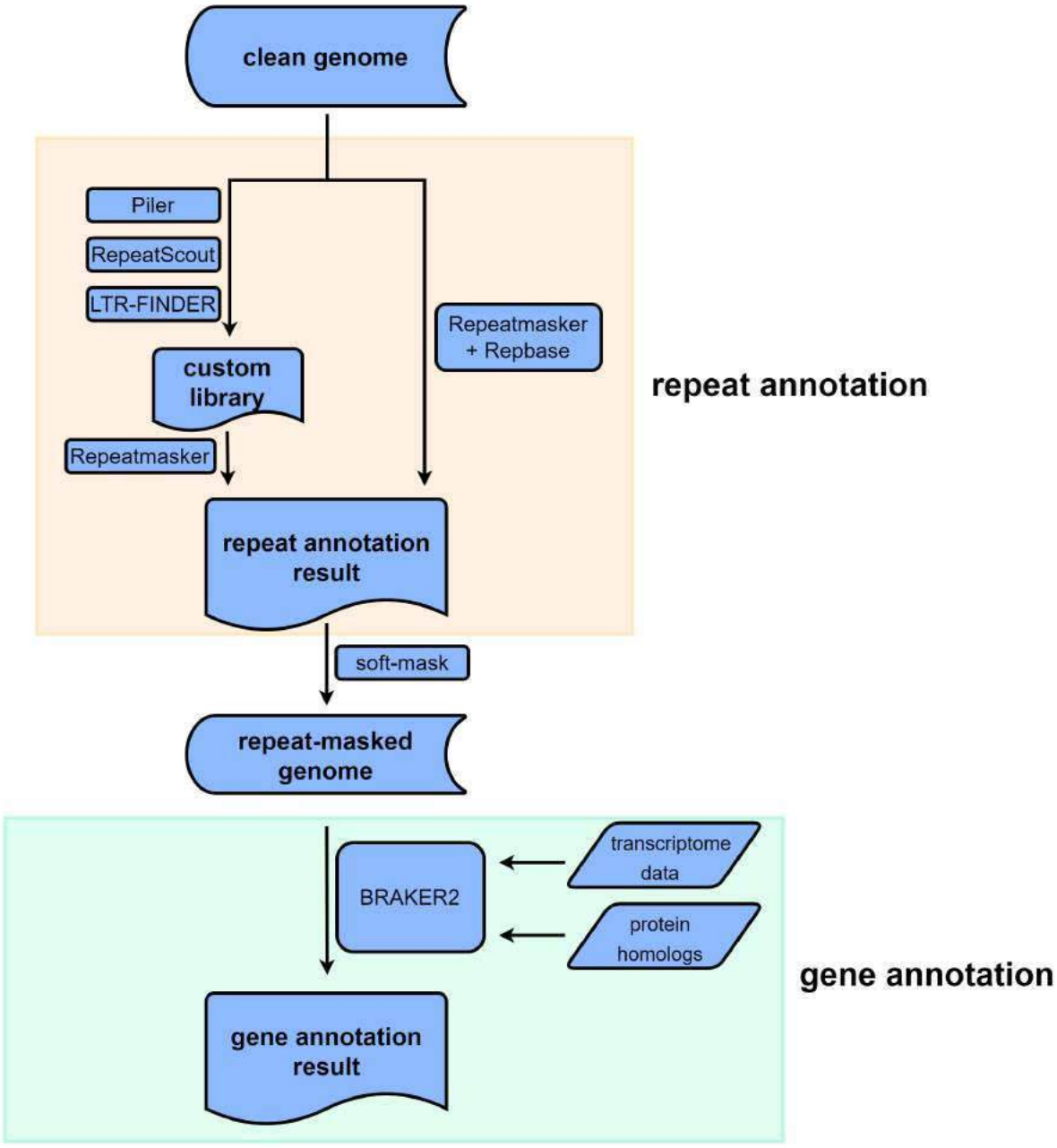
Supplementary Figure 21. Flow chart showing genome assembly and annotation pipeline used in the current study.



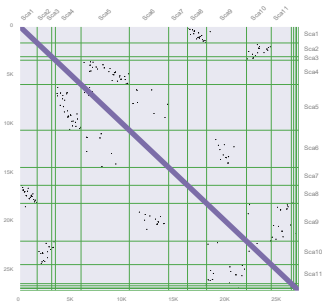
Supplementary Figure 22. Flow chart showing de-contamination pipeline used in the current study.



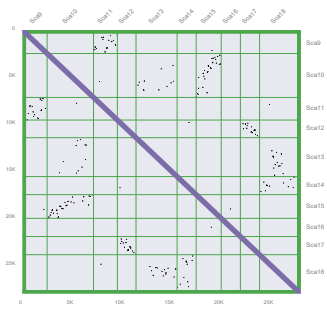
Supplementary Figure 23. Hi-C heatmap generated by Juicebox for 25 pseudo-chromosome level assemblies. Juicer was used to extract valid data, which were fed to 3D-DNA pipeline. Hi-C contacted map binary files were loaded into Juicebox to manually adjust and generate the final Hi-C heatmaps.



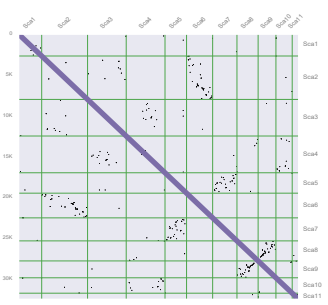
Supplementary Figure 24. Flow chart showing annotation pipeline used in the current study.



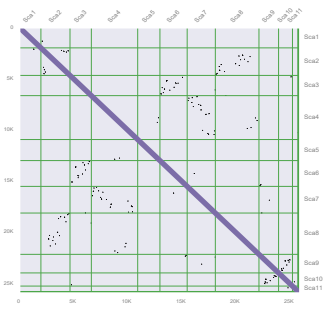
Entodon concinnus (964)



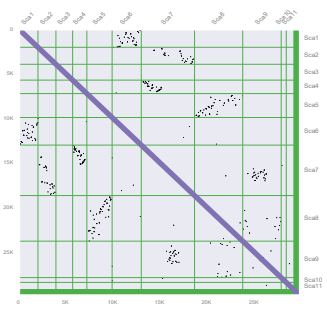
Aerobryopsis subdivergens (668)



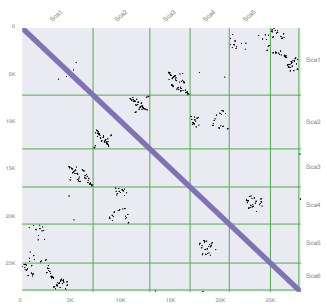
Racopilum cuspidigerum (1208)



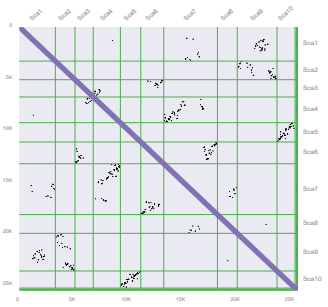
Orthotrichum anomalum (795)



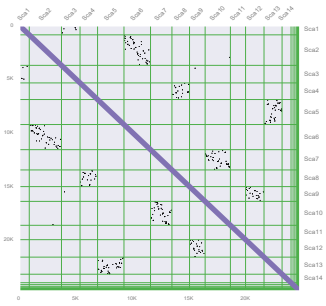
Pyrrhobryum spiniforme (667)



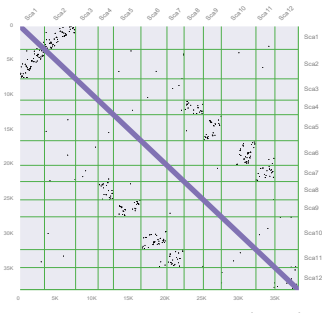
Philonotis turneriana (665)



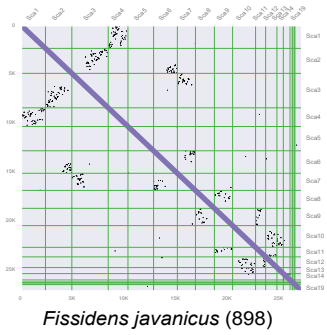
Hedwigia ciliata (1009)



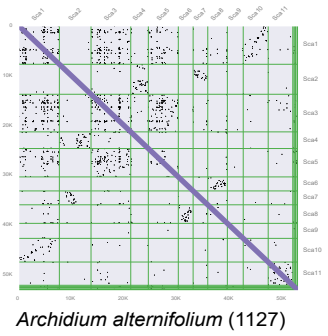
Barbula amplexifolia (967)



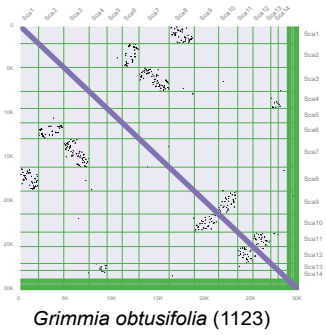
Paraleucobryum enerve (1124)



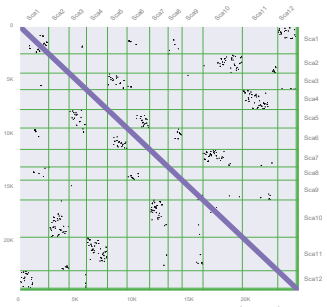
Fissidens javanicus (898)



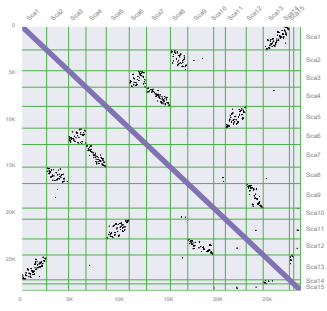
Archidium alternifolium (1127)



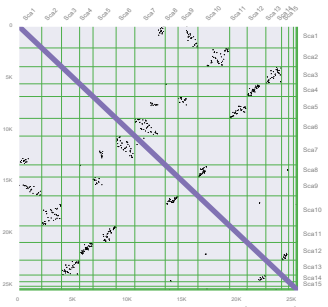
Grimmia obtusifolia (1123)



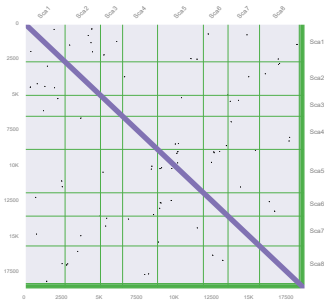
Ptychomitrium wilsonii (670)



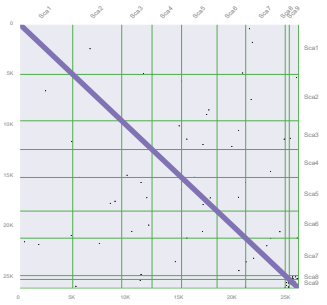
Bryoxiphium norvegicum (911)



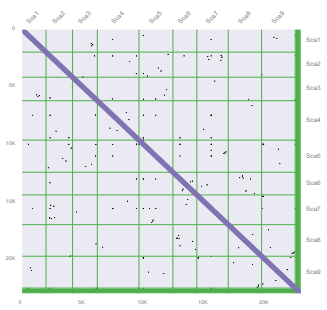
Timmia megapolitana (1010)



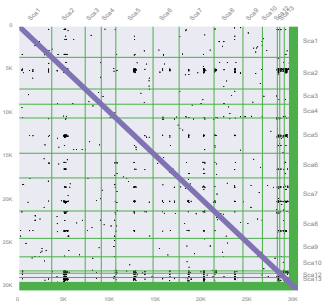
Diphyscium fulvifolium (910)



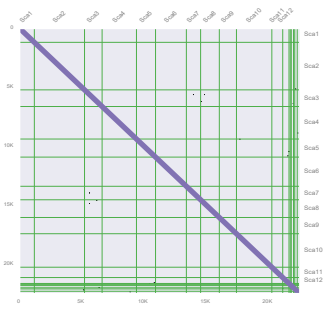
Atrichum angustatum (1015)



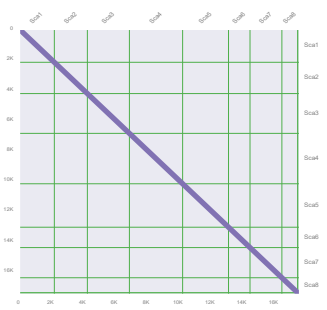
Tetraphis pellucida (1011)



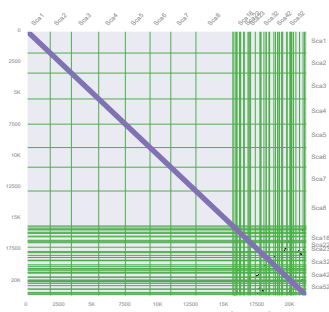
Jungermannia erectum (955)



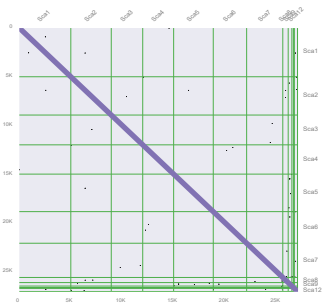
Acrolejeunea sandvicensis (888)



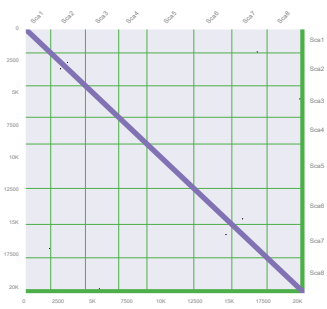
Conocephalum conicum (664)



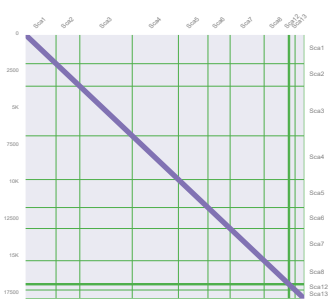
Ricciocarpos natans (909)



Cyathodium cavernarum (908)

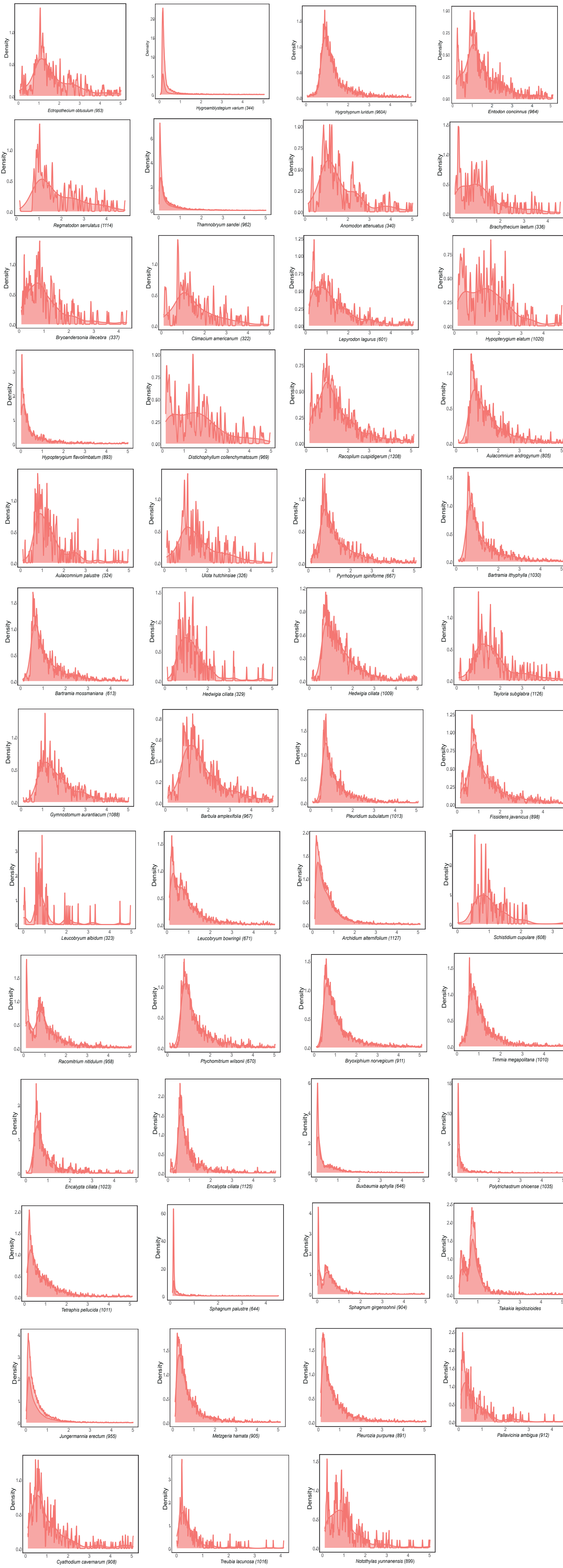


Plagiochasma appendiculatum (952)

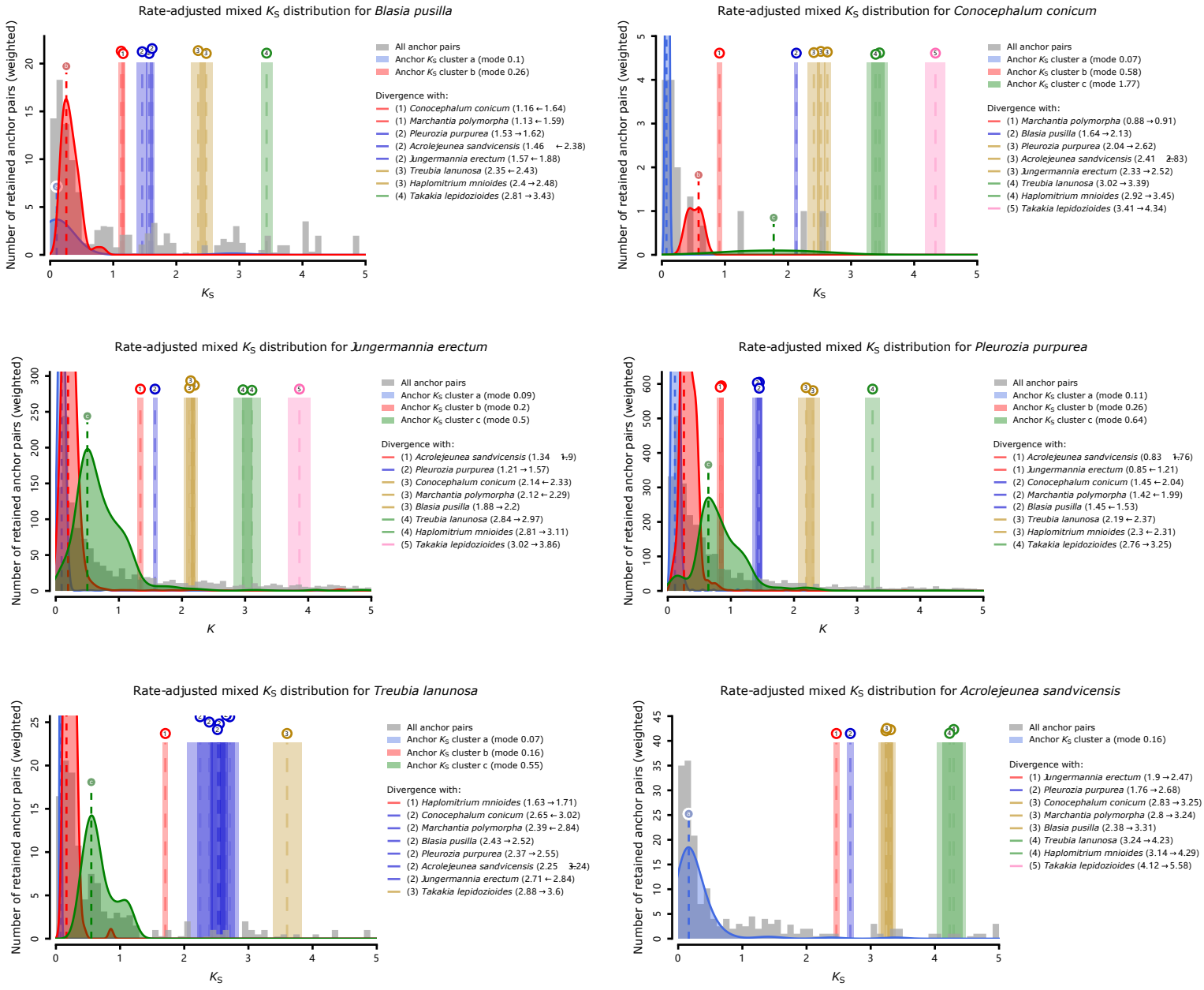


Blasia pusilla (890)

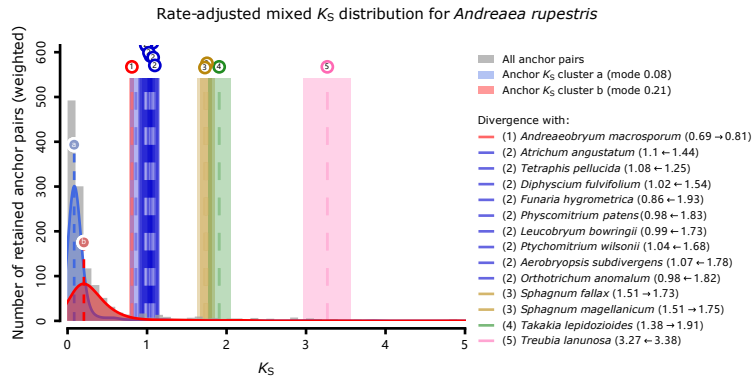
Supplementary Figure 25. Dot plot of the 25 bryophyte species with pseudo-chromosome level assemblies. Genomic synteny analyses were conducted using JCVI with default settings. Species were sorted by the phylogeny in Figure 1.

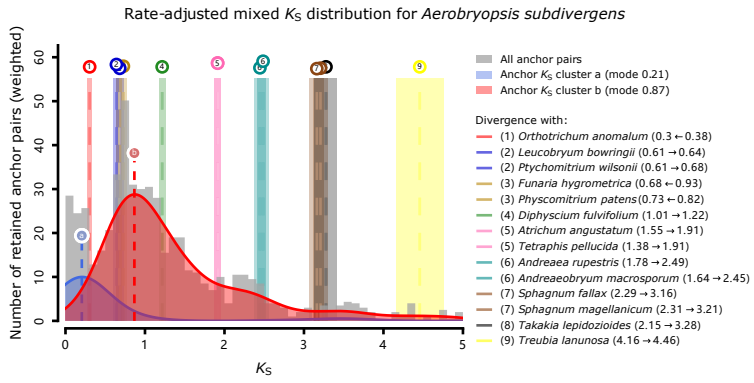


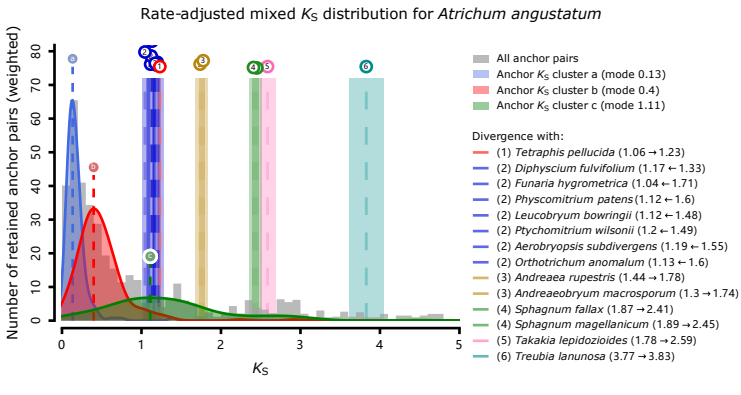
Supplementary Figure 26. *Ks* plots of representative bryophyte species.

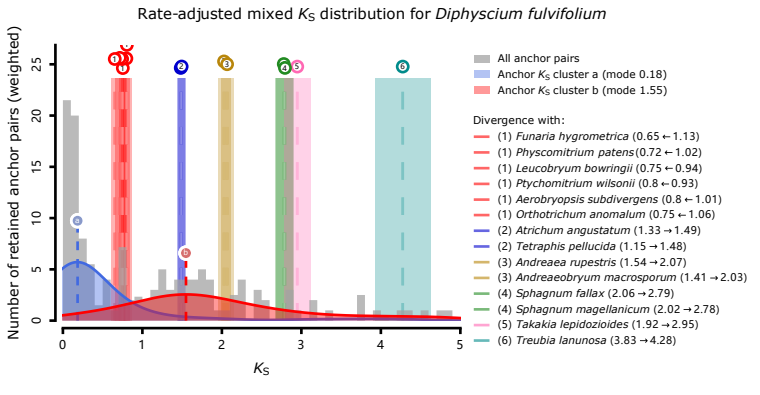


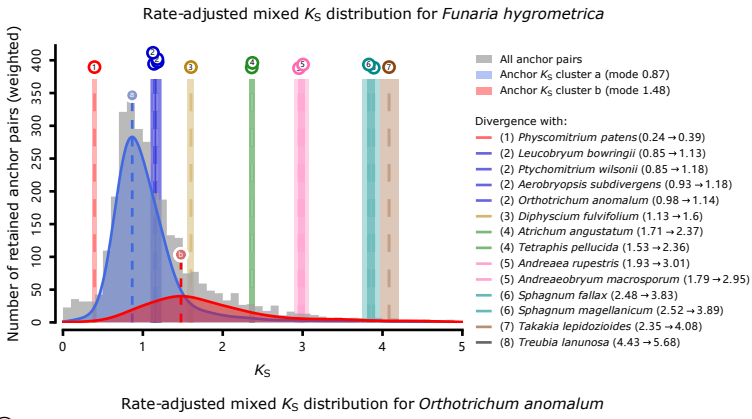
Supplementary Figure 27. The rate-adjusted mixed *Ks*-age distribution of anchor pairs of liverwort genome dataset. The *Ks* distributions of anchor clusters overlaid by colored KDE curves and labelled by letters, with vertical lines denoting associated peaks derived from the lognormal mixture modeling of median *Ks* values for the collinear segment pairs, were shown as a grey histogram. Rate-corrected mode of orthologous *Ks*-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the *Ks* shifts resulted from the substitution rate correction.

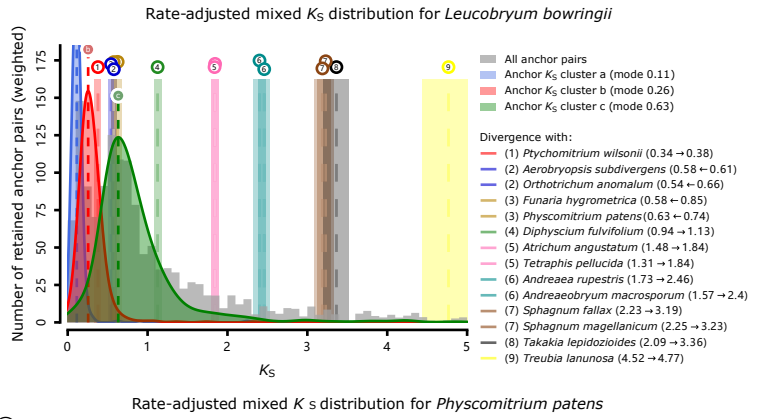


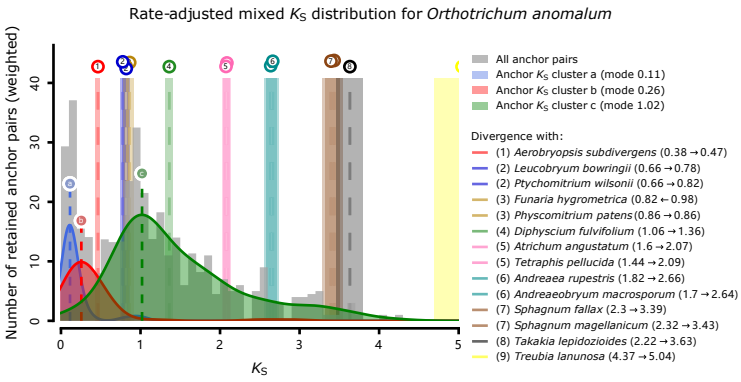


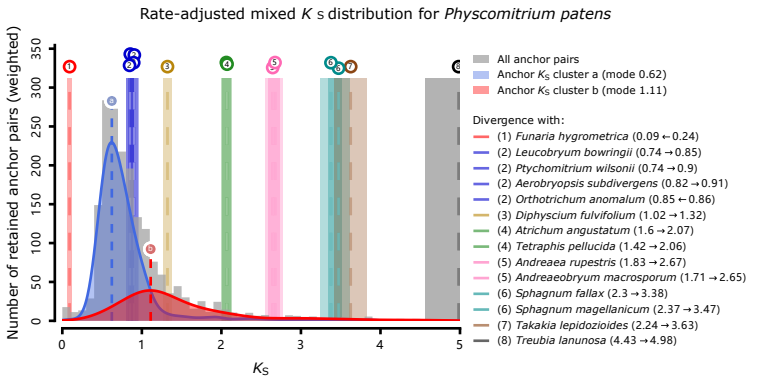


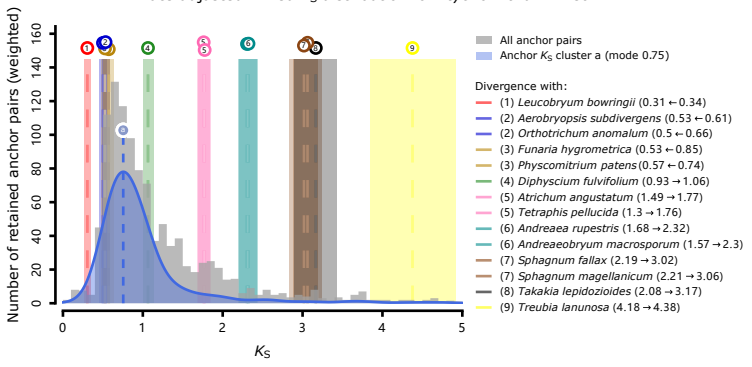


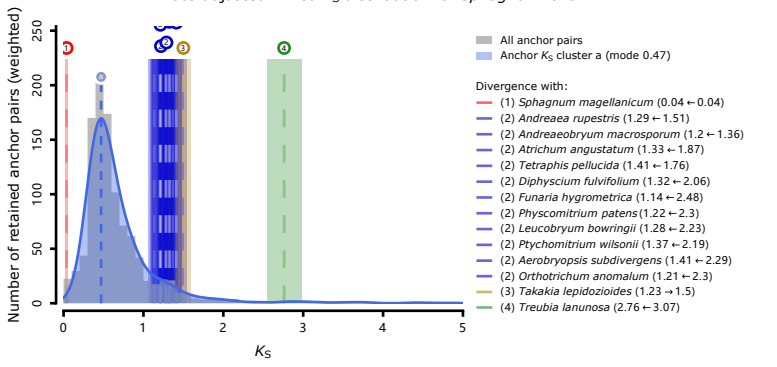


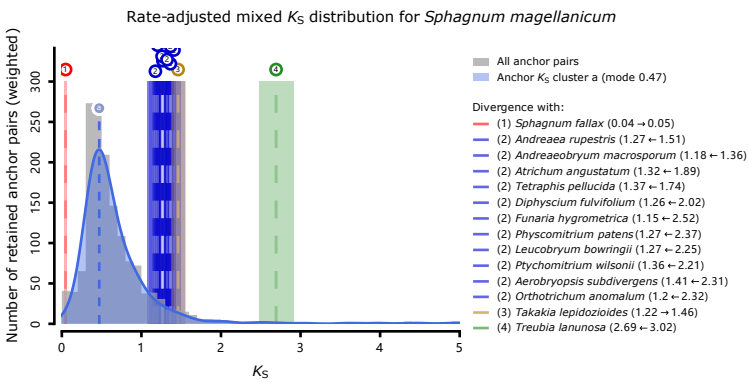


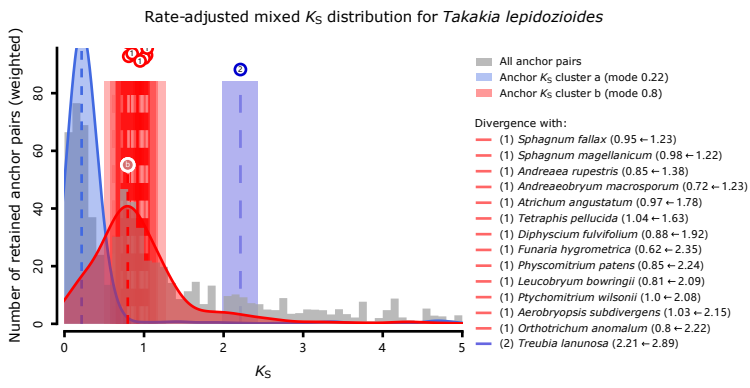


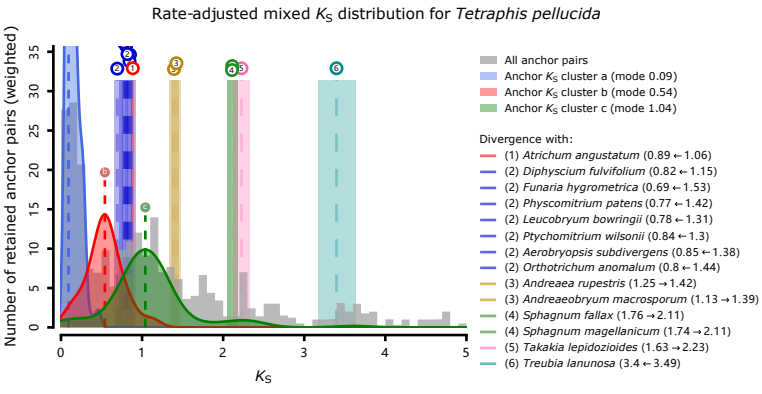




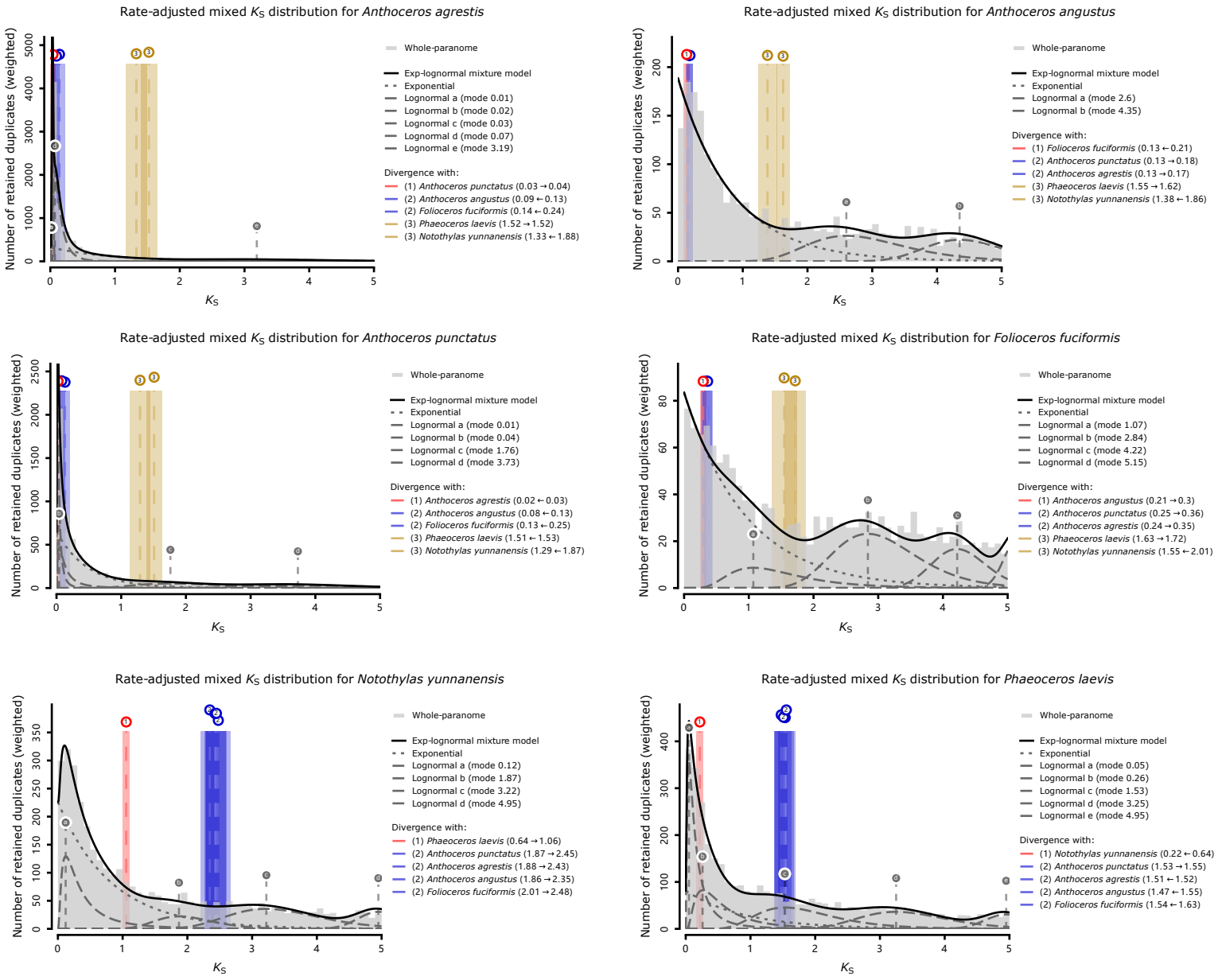




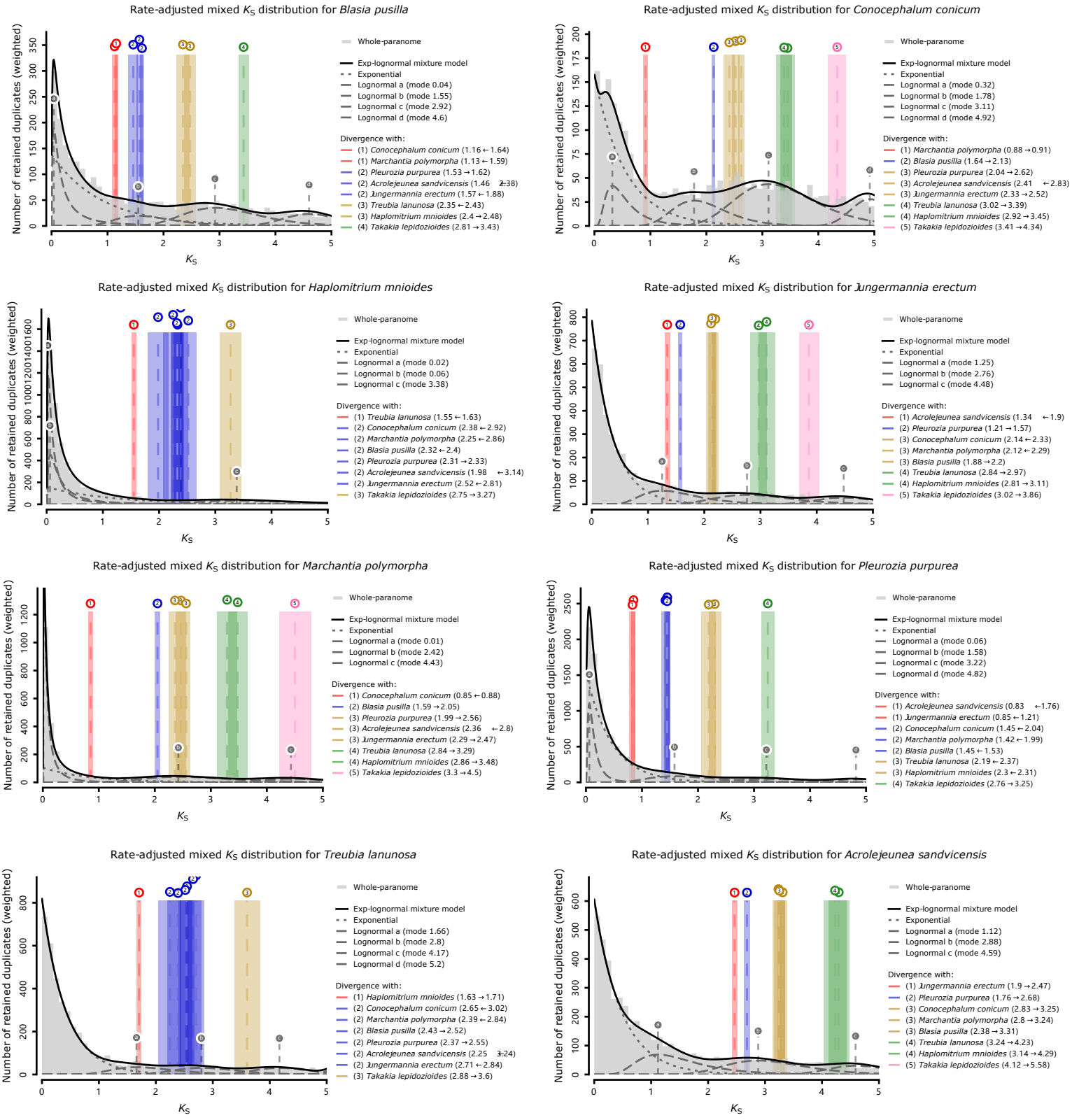




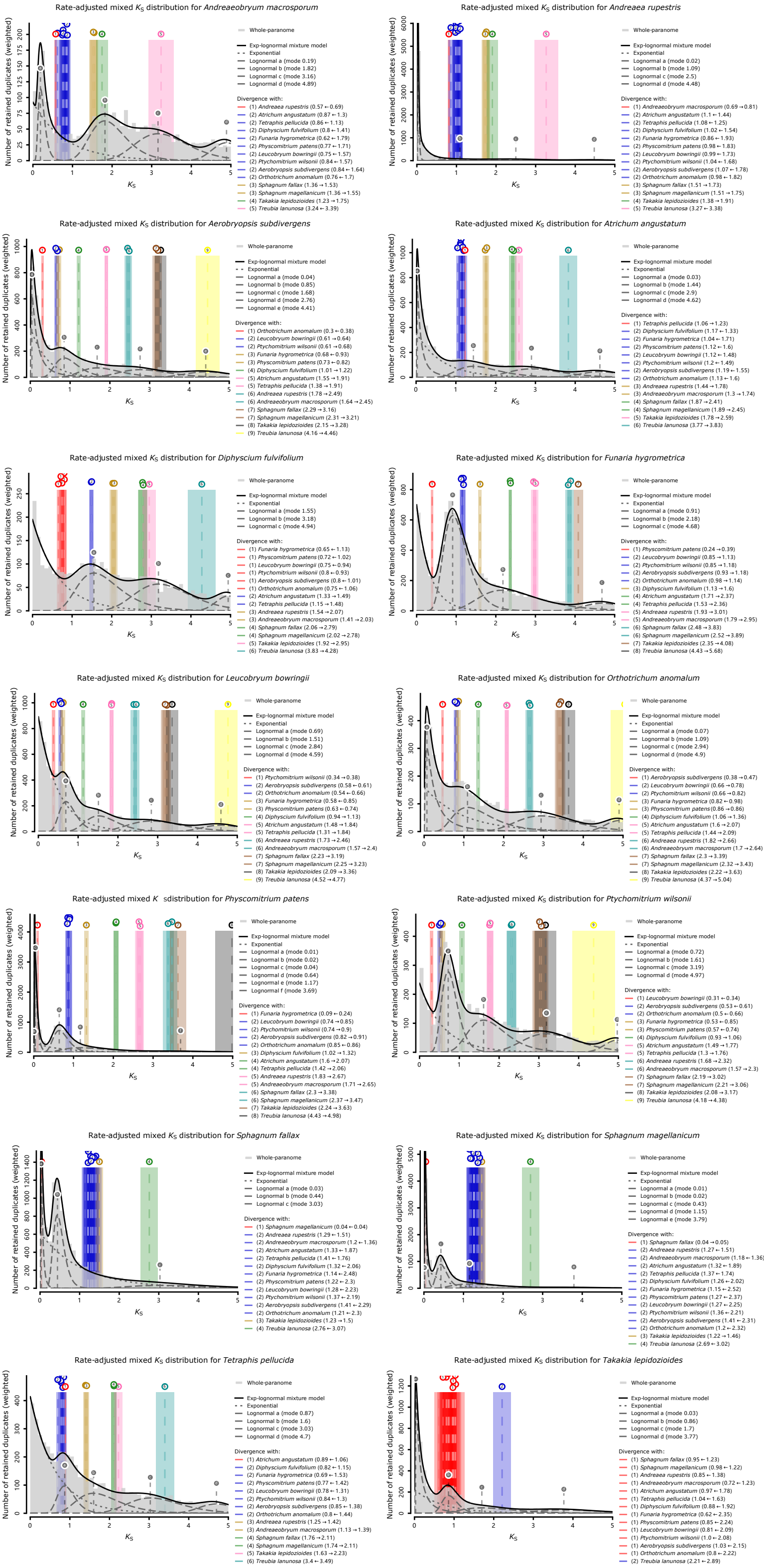
Supplementary Figure 28. The rate-adjusted mixed *Ks*-age distribution of anchor pairs of moss genome dataset. The *Ks* distributions of anchor clusters overlaid by colored KDE curves and labelled by letters, with vertical lines denoting associated peaks derived from the lognormal mixture modeling of median *Ks* values for the collinear segment pairs, were shown as a grey histogram. Rate-corrected mode of orthologous *Ks*-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the Ks shifts resulted from the substitution rate correction.



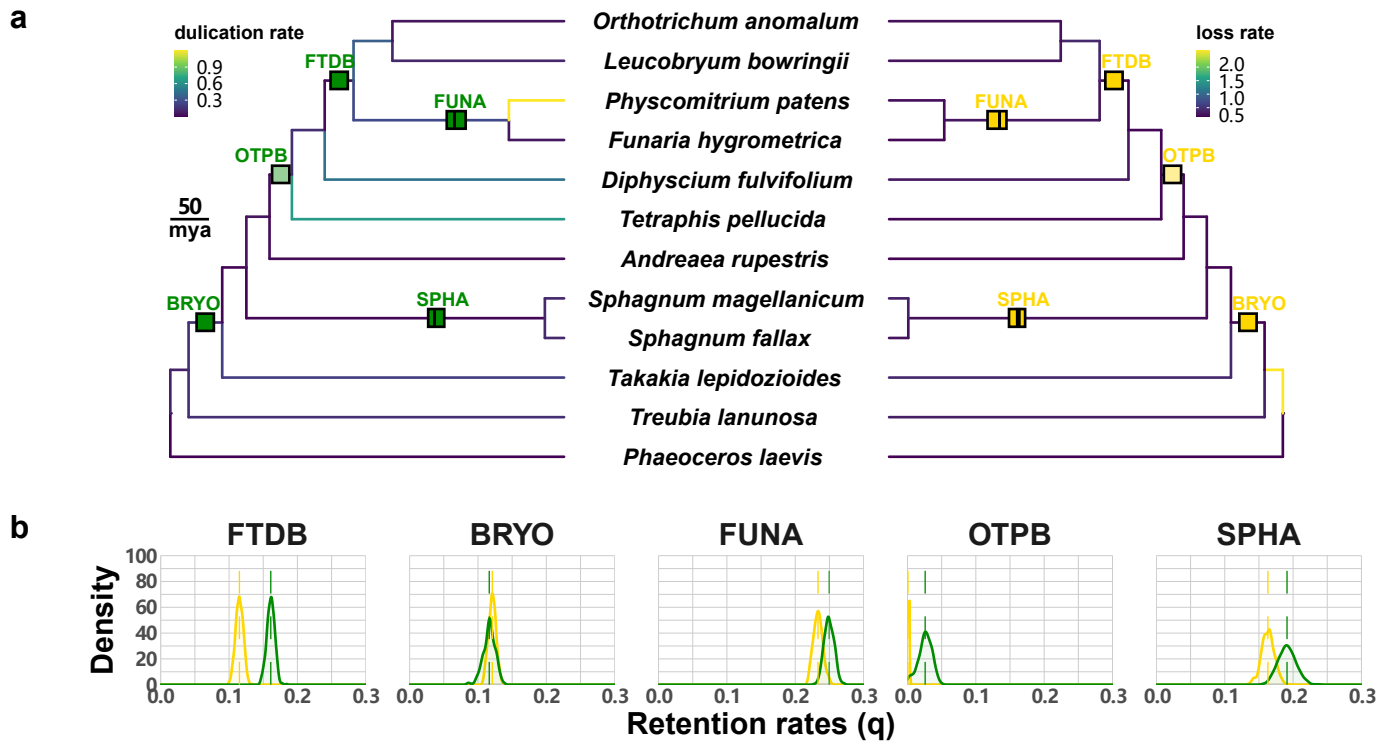
Supplementary Figure 29. The rate-adjusted mixed *Ks*-age distribution of whole paranome for hornwort genome dataset. Ks-age distributions for the whole paranome of different hornworts were overlaid with rate- corrected speciation events in colored vertical lines. The overall mixture model in the dark solid line of each paralogous *Ks*- age distribution consists of an exponential component in dotted gray curve and optimized log-normal components in dashed gray curves. Each log-normal component is labeled with a letter, shown as vertical dashed gray lines with circular labels. Rate-corrected mode of orthologous *Ks*-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the Ks shifts resulted from the substitution rate correction.



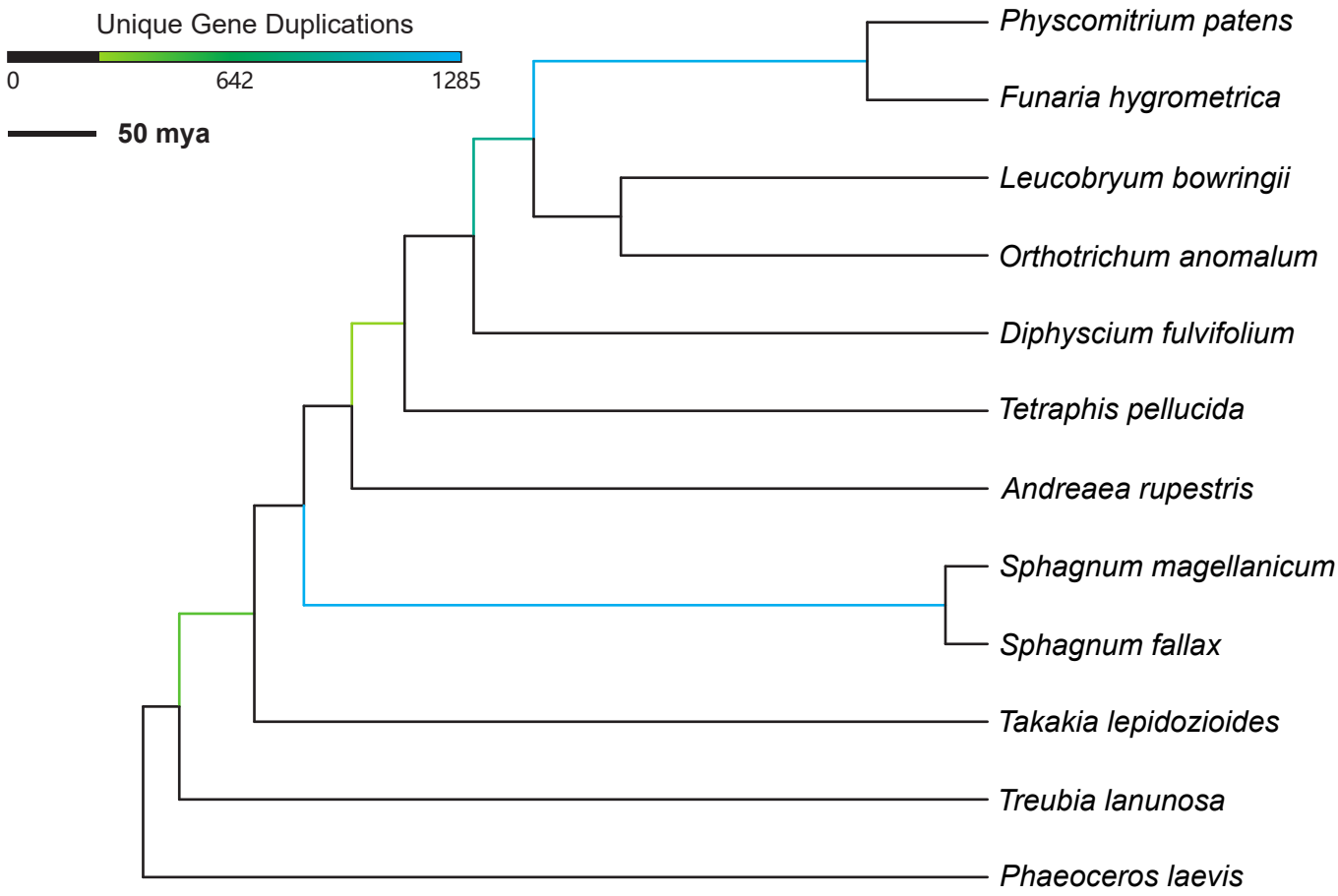
Supplementary Figure 30. The rate-adjusted mixed *Ks*-age distribution of whole paranome for liverwort genome dataset. *Ks*-age distributions for the whole paranome of different liverworts were overlaid with rate-corrected speciation events in colored vertical lines. The overall mixture model in the dark solid line of each paralogous *Ks*-age distribution consists of an exponential component in dotted gray curve and optimized log-normal components in dashed gray curves. Each log-normal component is labeled with a letter, shown as vertical dashed gray lines with circular labels. Rate-corrected mode of orthologous *Ks*-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the Ks shifts resulted from the substitution rate correction.



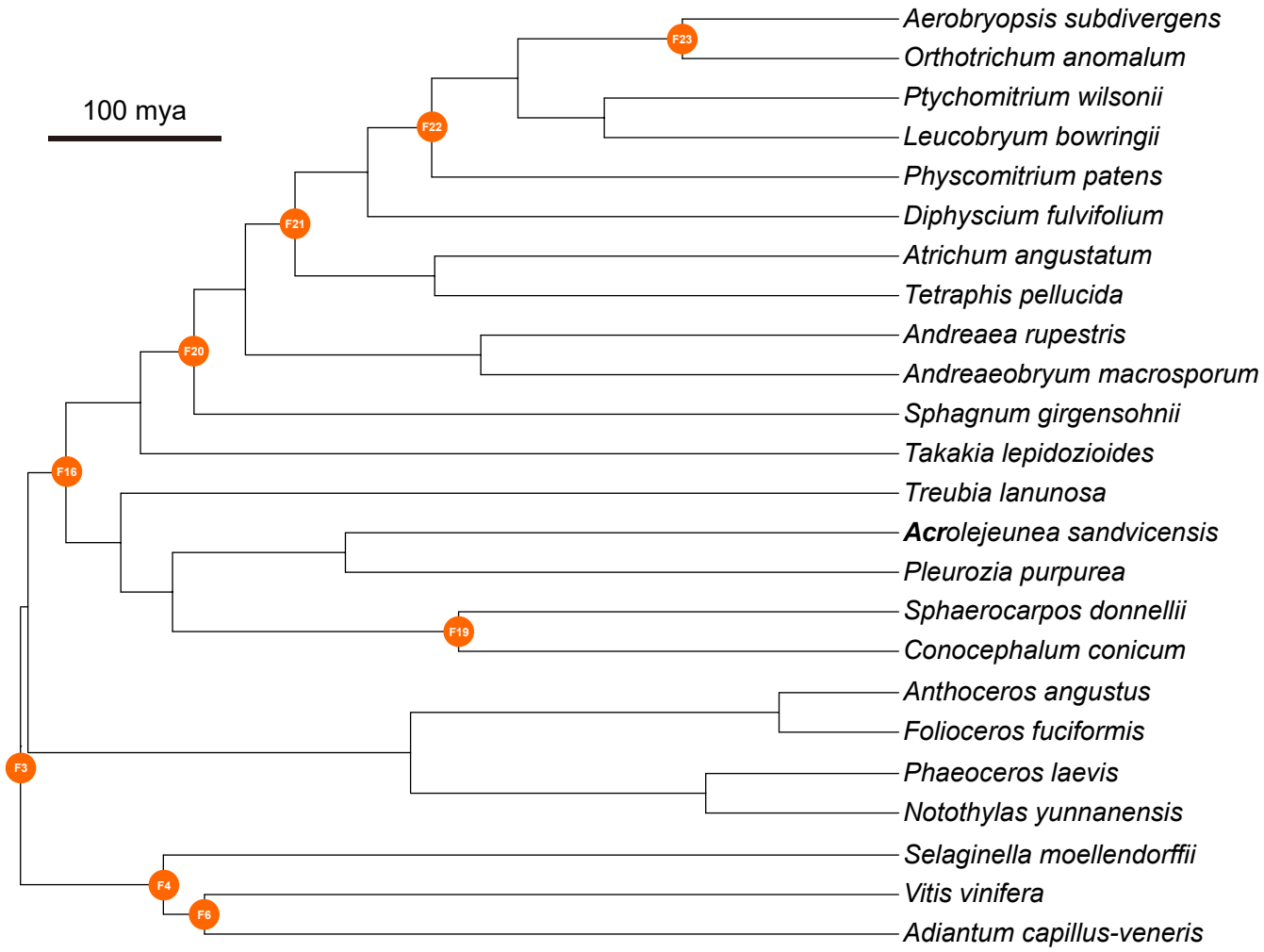
Supplementary Figure 31. The rate-adjusted mixed *Ks*-age distribution of whole paranome for moss genome dataset. *Ks*-age distributions for the whole paranome of different liverworts were overlaid with rate- corrected speciation events in colored vertical lines. The overall mixture model in the dark solid line of each paralogous *Ks*-age distribution consists of an exponential component in dotted gray curve and optimized log-normal components in dashed gray curves. Each log-normal component is labeled with a letter, shown as vertical dashed gray lines with circular labels. Rate-corrected mode of orthologous *Ks*-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the *Ks* shifts resulted from the substitution rate correction.



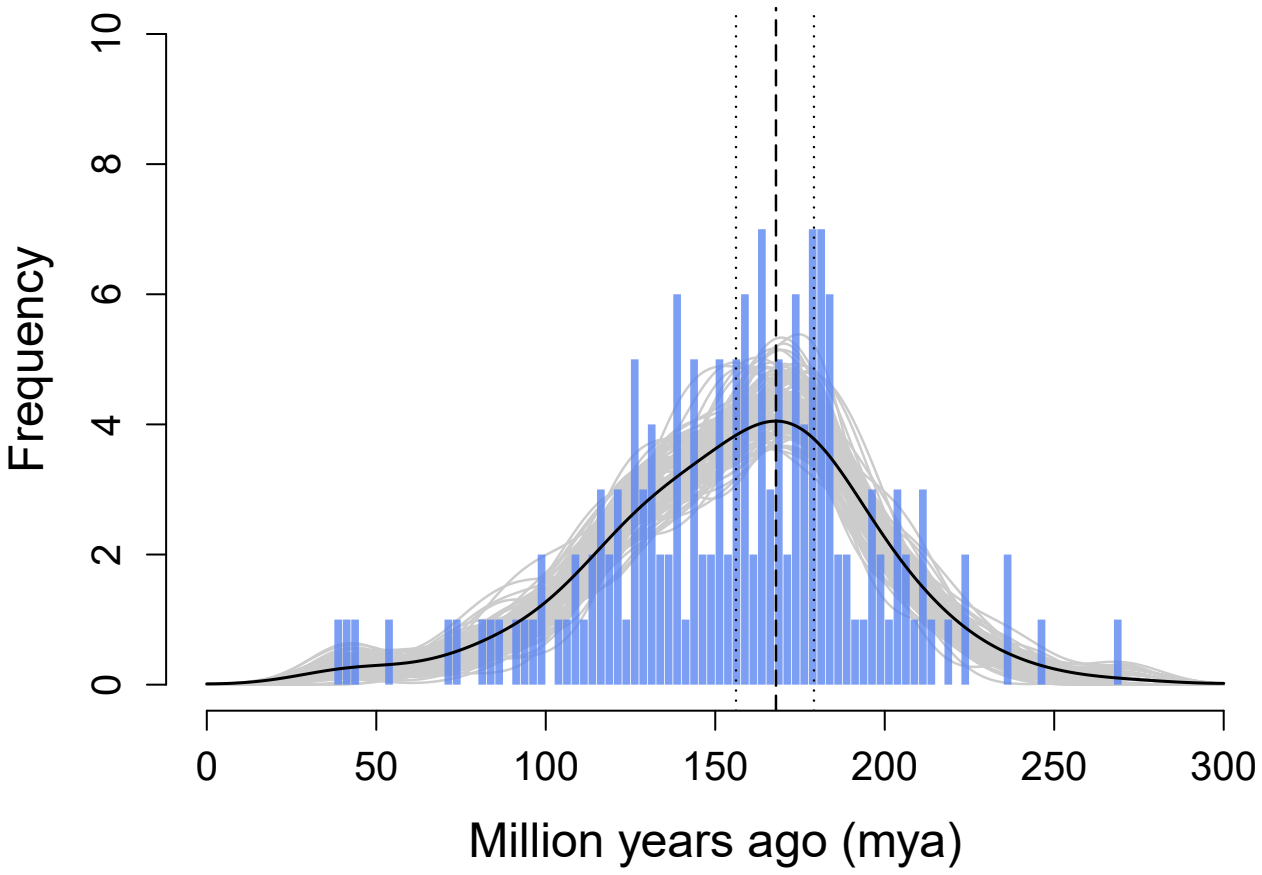
Supplementary Figure 32. Duplication, loss, and WGD retention rates of queried moss phylogeny. (**a**) The reconstructed divergence time tree pinpointing the phylogenetic position of five hypothetical WGDs. Branch lengths represent absolute divergence time and branches are colored based on the posterior mean of duplication and loss rate estimated from Whale (left and right panel, respectively). The filled rectangles denoting WGDs significantly supported by Whale in which green rectangles indicating support from relaxed model and yellow for the critical model. The translucent rectangles depict WGDs lacking of significant support from Whale. WGDs with further absolute timing information are circumscribed with width representing 90% credibility intervals (CI) and black line in between representing consensus peak. (**b**) The posterior distribution of WGD retention rates (q) for the five putative WGDs. Dotted lines indicate the posterior mean of each posterior distribution. Color rule follows A.



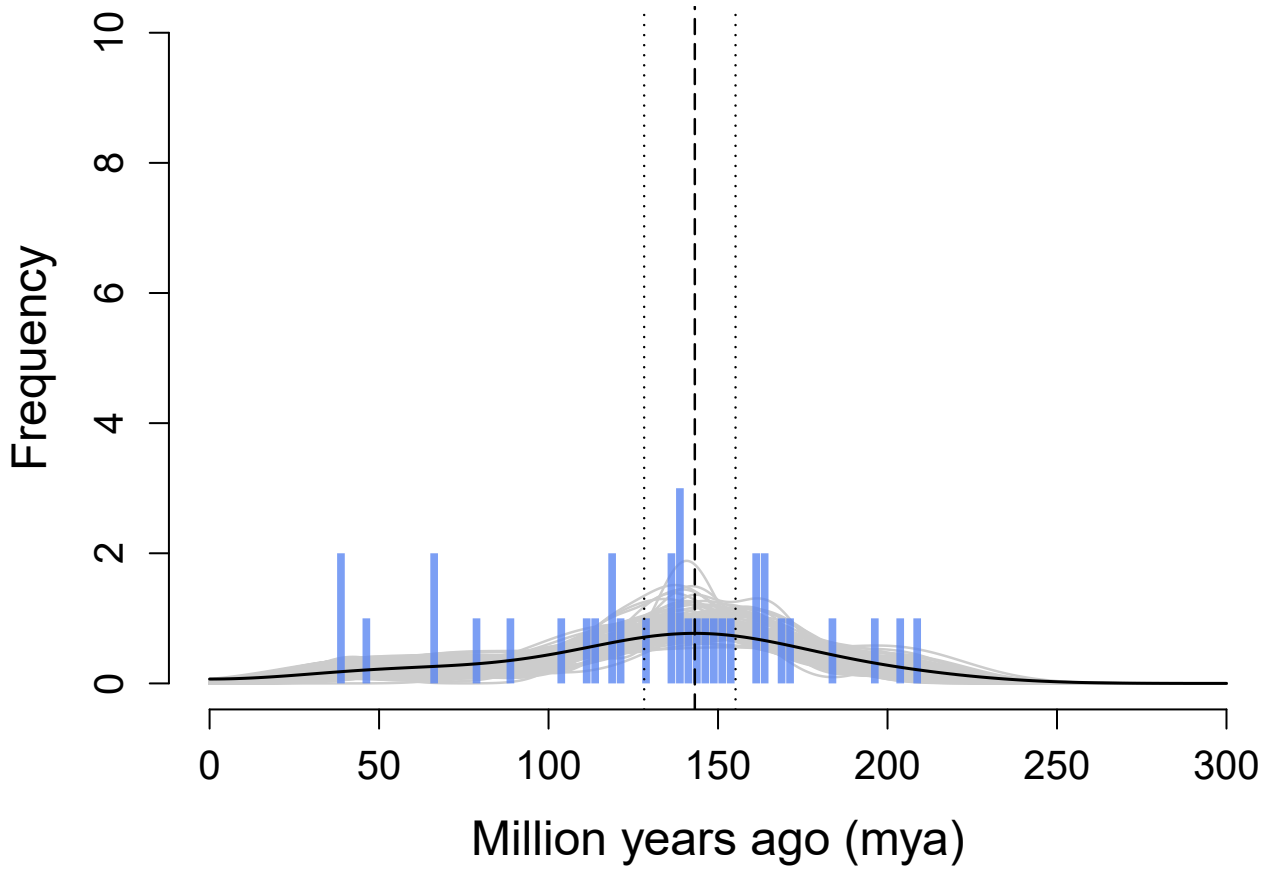
Supplementary Figure 33. WGD inspection using PUG. The color bar represents the number of unique gene duplications reconciled at each branch with at least 80% bootstrap support and five branches with at least 20% of the maximum number of duplication among all internal branches are colored and identified as probable WGDs in the queried species tree by PUG.



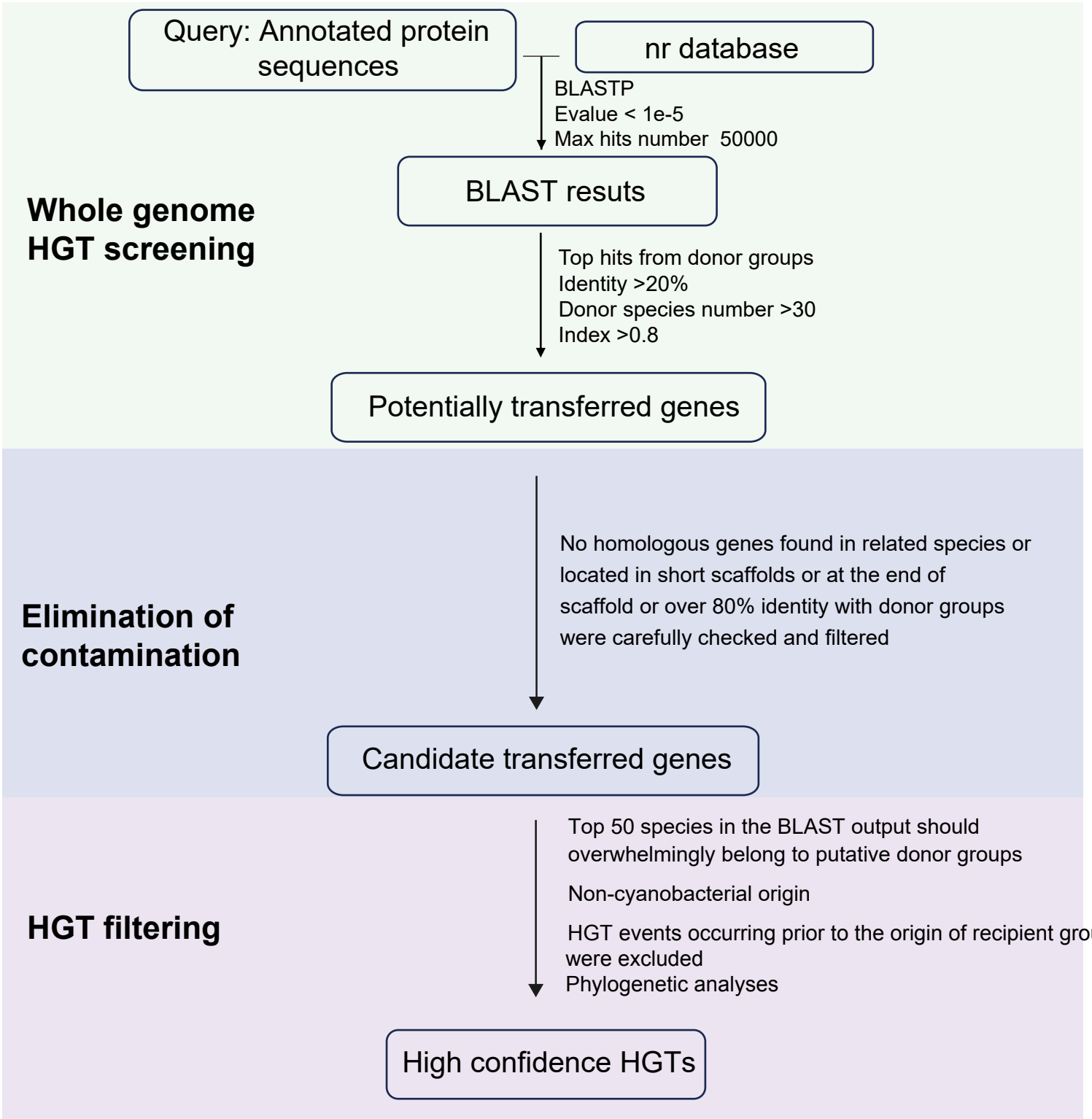
Supplementary Figure 34. Starting tree and fossil calibrations in absolute dating analysis. The starting tree used in Beast for phylogenetic dating with fossil calibrations marked on corresponding nodes symbolized in orange circles with labels.



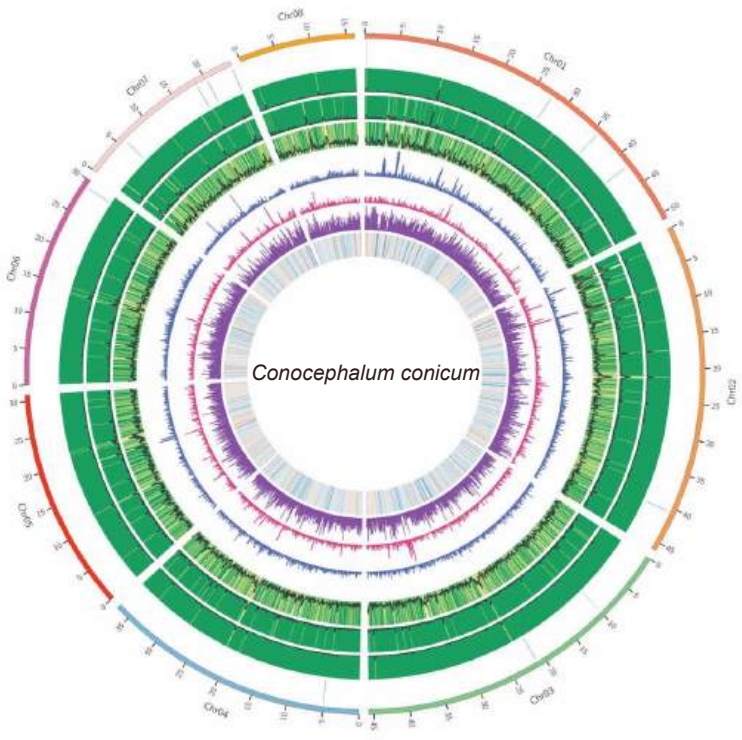
Supplementary Figure 35. Absolute age distribution of SPHA WGD. The solid black line represents the KDE of paralogous date estimates and the vertical dashed black line represents the consensus peak at 167.91 mya. The grey lines represent KDE from 2,500 bootstrap replicates and the vertical black dotted lines represent the corresponding 90% confidence interval for the WGD age estimate, 156.12-179.12 mya. The blue histogram shows the raw distribution of absolute divergence date estimates for *Sphagnum girgensohnii* paralogues.

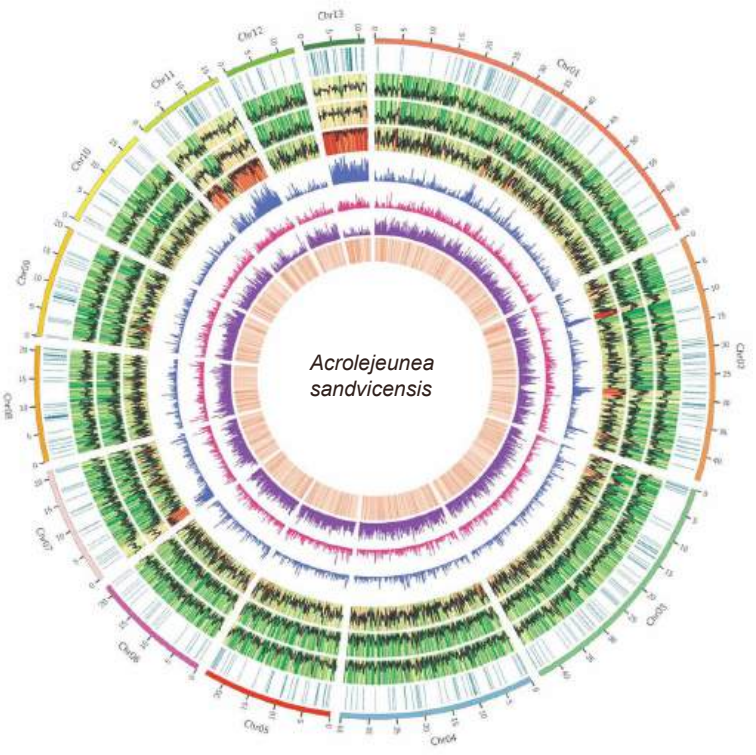


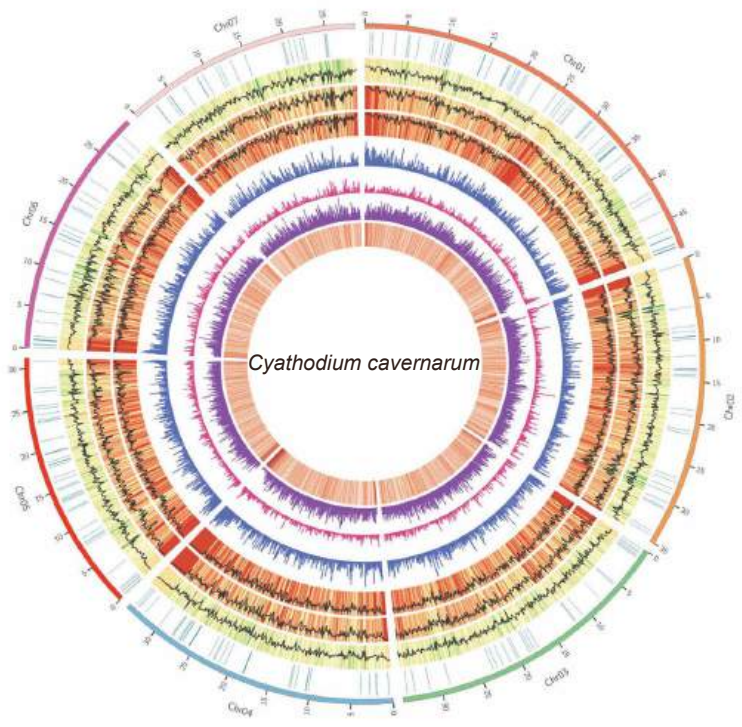
Supplementary Figure 36. Absolute age distribution of FUNA WGD. The solid black line represents the KDE of paralogous date estimates and the vertical dashed black line represents the consensus peak at 143.15 mya. The grey lines represent KDE from 2,500 bootstrap replicates and the vertical black dotted lines represent the corresponding 90% confidence interval for the WGD age estimate, 128.21−155.16 mya. The blue histogram shows the raw distribution of absolute divergence date estimates for P. patens paralogues.

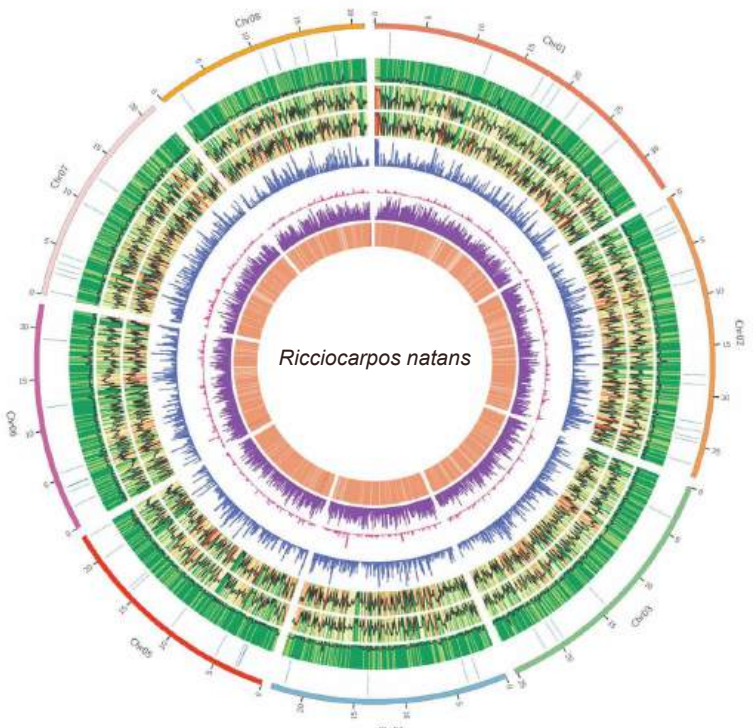


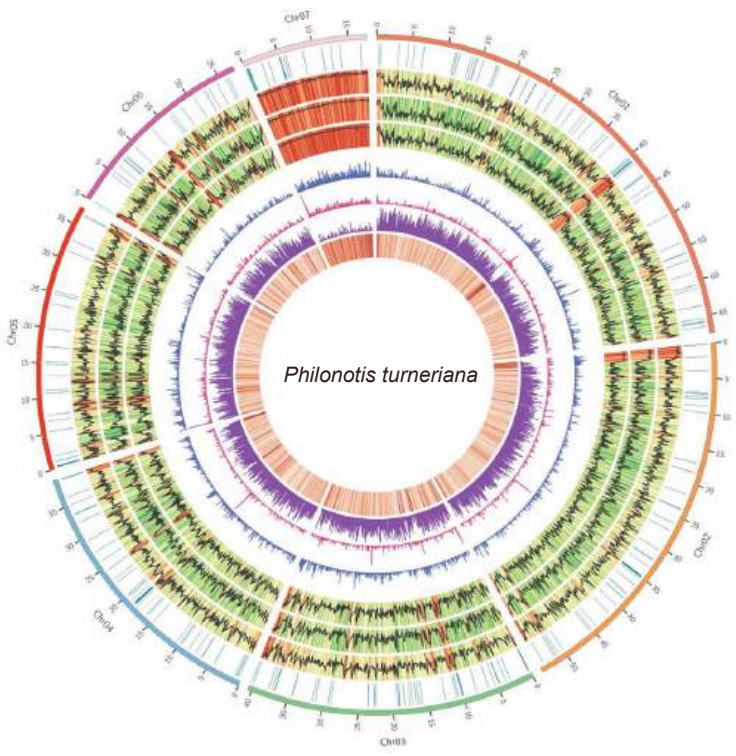
Supplementary Figure 37. Flow chart showing the pipeline for HGT identification used in the current study. HGT identification in this study consists of three main steps: whole genome HGT screening, elimination of contamination, and HGT filtering.

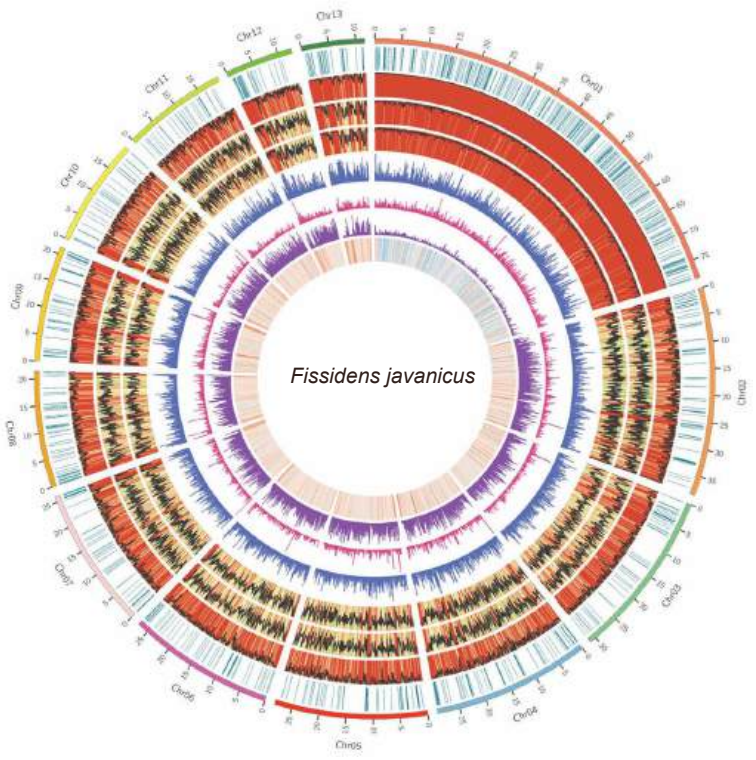


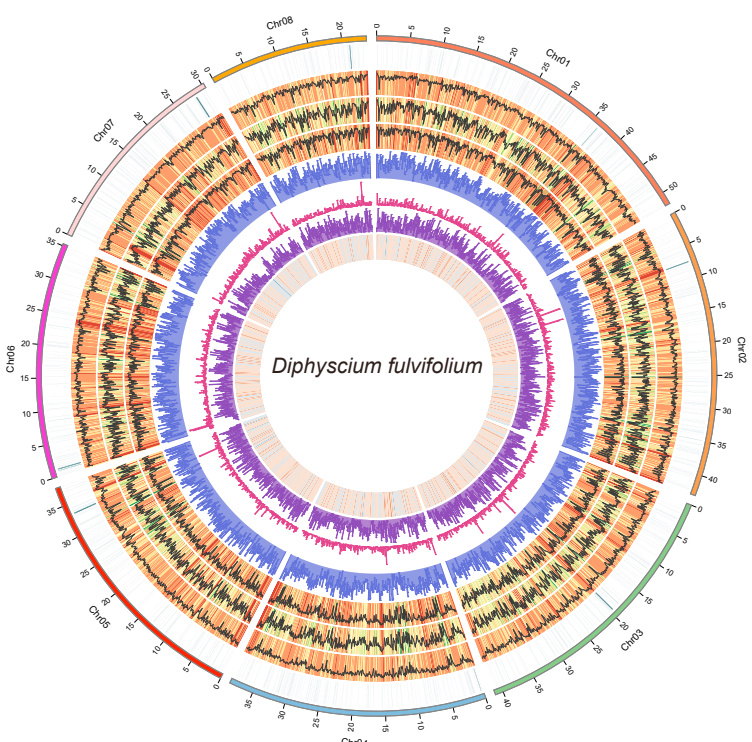


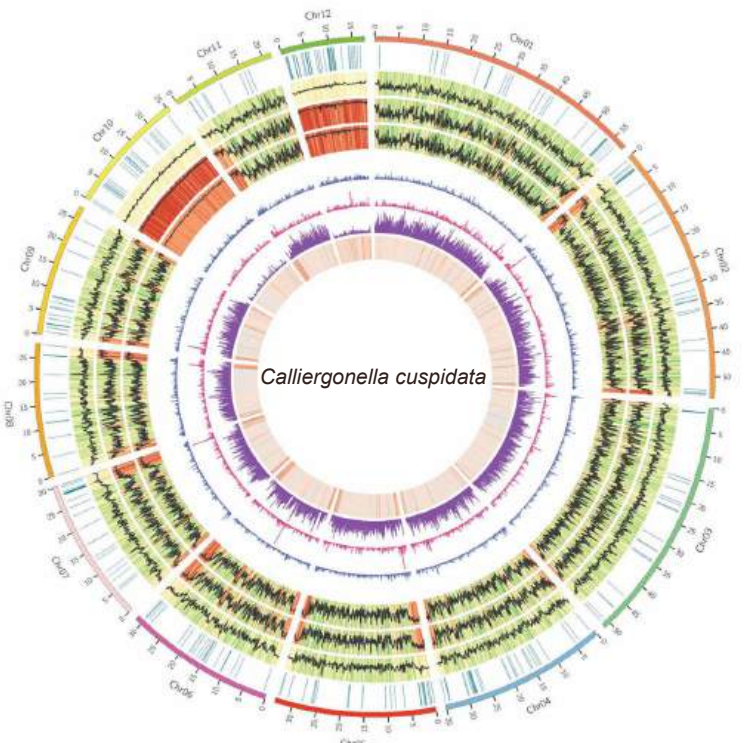




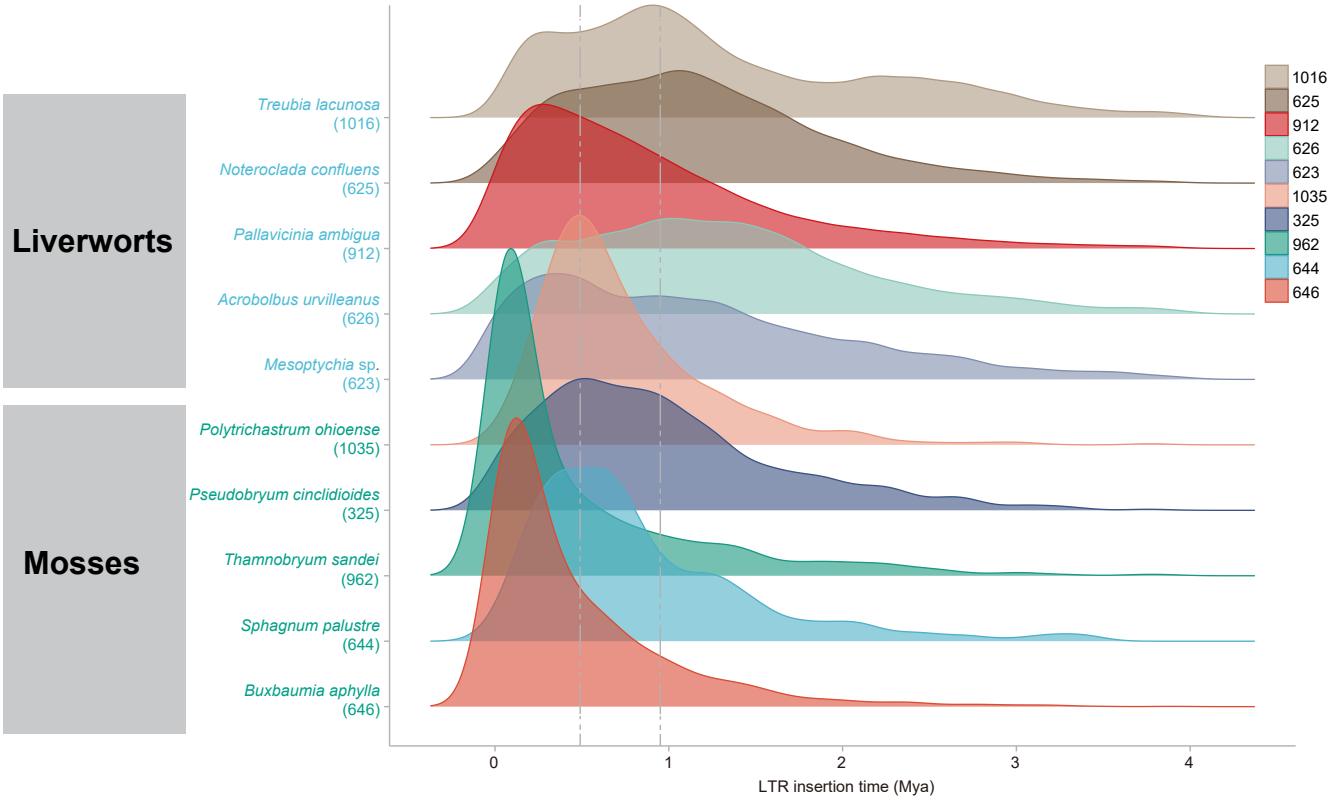




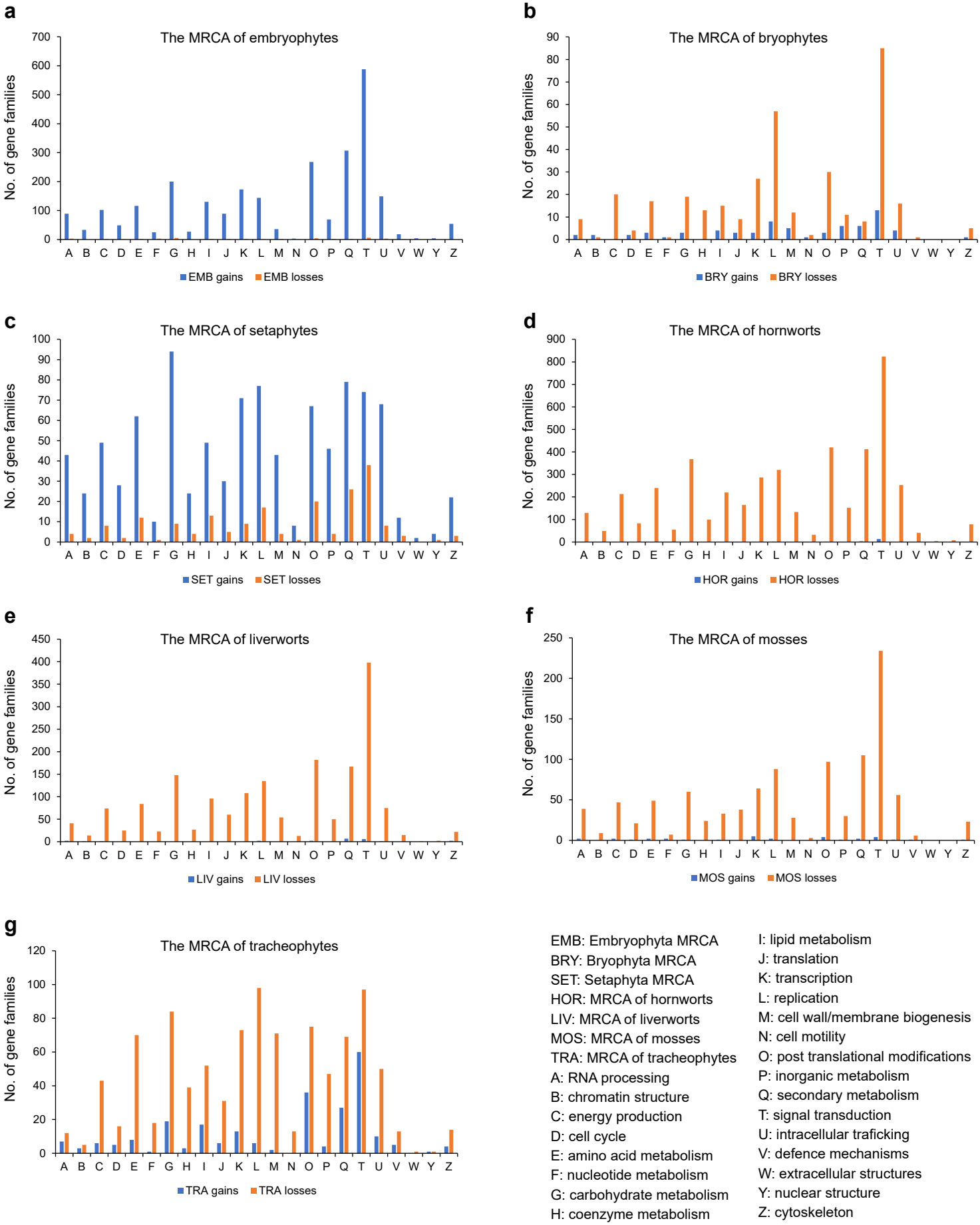




Supplementary Figure 38. Genome features of ten representative bryophytes. Only those chromosomes > 5 Mb were plotted. The outer ring shows individual chromosomes. Inner fromrings outside to inside: location of RLC5 elements, the distribution of the mCHH level, the distribution of the mCHG level, the distribution of the mCG level, the distribution of the Gypsy, the distribution of the Copia, the Gene density, the distribution of the GC content.

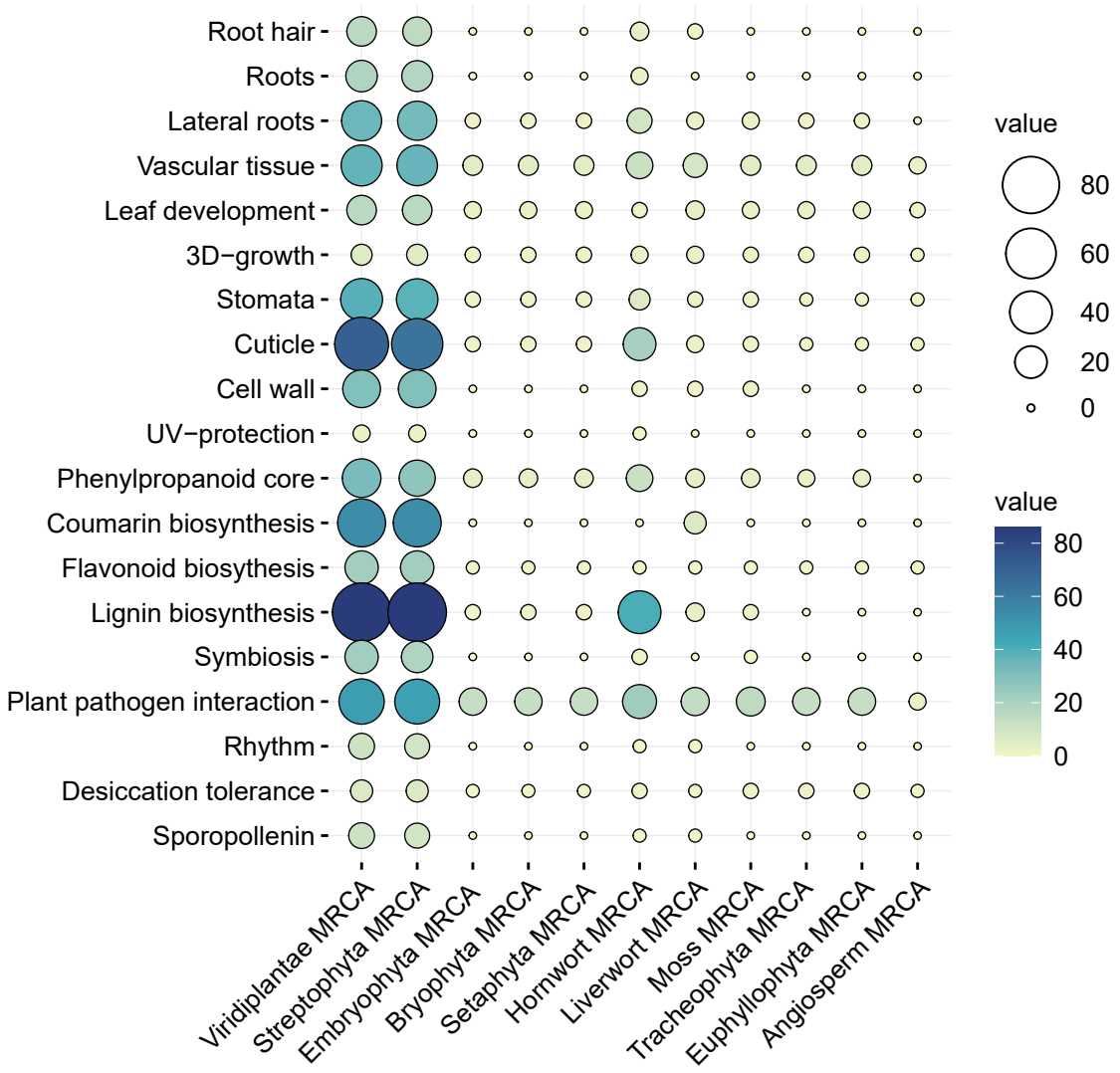


Supplementary Figure 39. Comparison of putative insertion time of LTR-RTs among liverworts and mosses with relatively large genomes.

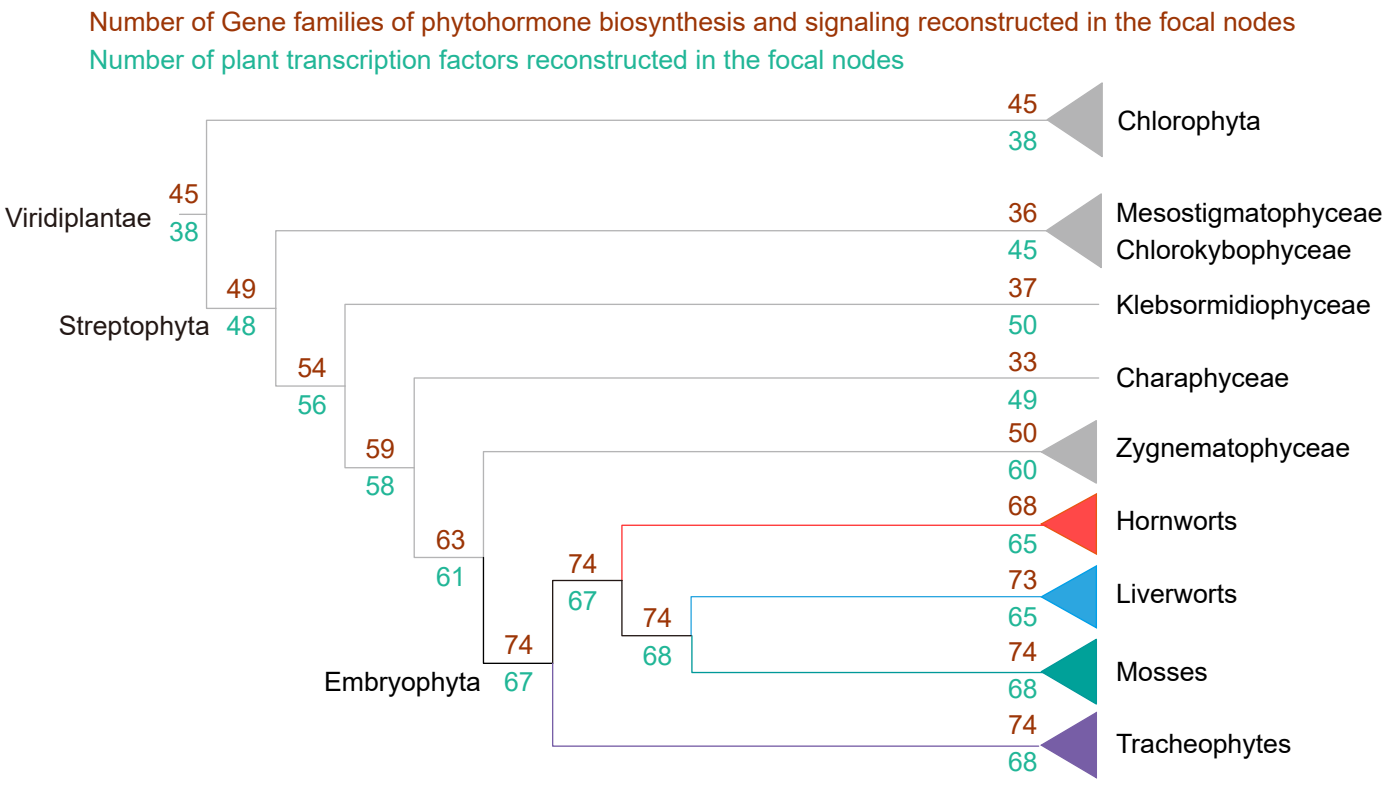


EMB: Embryophyta MRCA
BRY: Bryophyta MRCA
SET: Setaphyta MRCA
HOR: MRCA of hornworts
LIV: MRCA of liverworts
MOS: MRCA of mosses
TRA: MRCA of tracheophytes
A: RNA processing
B: chromatin structure
C: energy production
D: cell cycle
E: amino acid metabolism
F: nucleotide metabolism
G: carbohydrate metabolism
H: coenzyme metabolism
I: lipid metabolism
J: translation
K: transcription
L: replication
M: cell wall/membrane biogenesis
N: cell motility
O: post translational modifications
P: inorganic metabolism
Q: secondary metabolism
T: signal transduction
U: intracellular trafficking
V: defence mechanisms
W: extracellular structures
Y: nuclear structure
Z: cytoskeleton

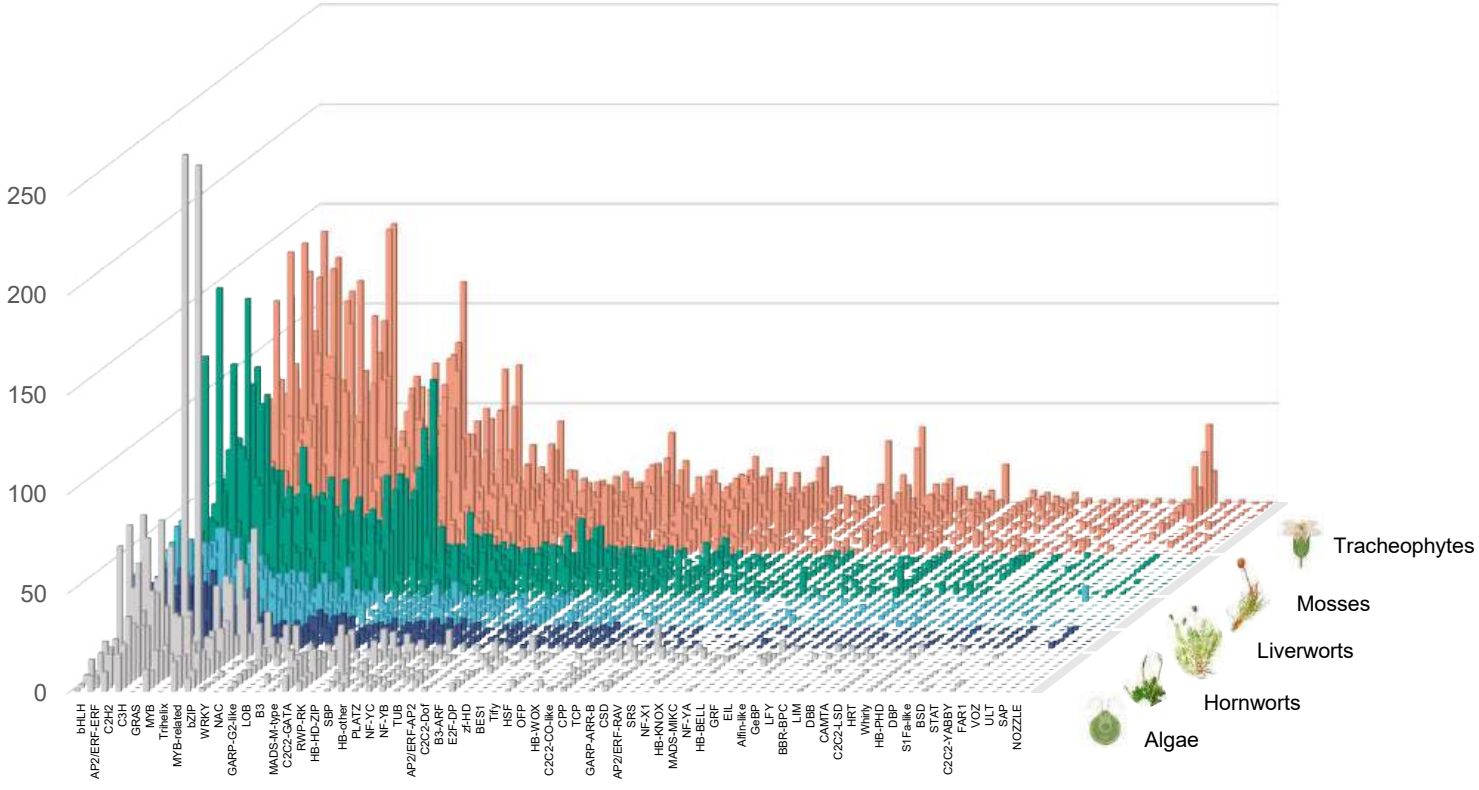
Supplementary Figure 40. Bar charts showing the number of gene family gains and losses representated in each COG (Clusters of Orthologous Groups of proteins) functional categories for the seven focal nodes, the MRCA of embryophytes (a), bryophytes (b), setaphytes (c), hornworts (d), liverworts (e), mosses (f), and tracheophytes (g).



Supplementary Figure 41. Bubble plot showing the number of genes absent (compared with the MRCA of angiosperms) in focal nodes inferred with Dollo parsimony algorithm in COUNT, for selected key pathways related to developmental and physiological/biochemical networks that may have facilitated adaptation of plants to the terrestrial environment.

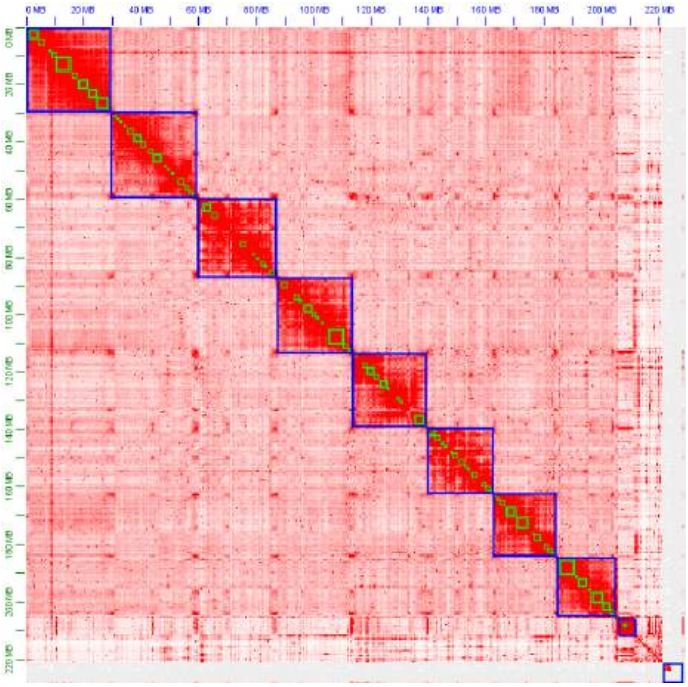


Supplementary Figure 42. Ancestral reconstrution of gene families for TFs and Phytohormone related genes across Viridiplantae tree with COUNT using Dollo parsimony principle. The phylogeny is simplified from the 190-taxa phylogenetic tree as shown in Figure S2.

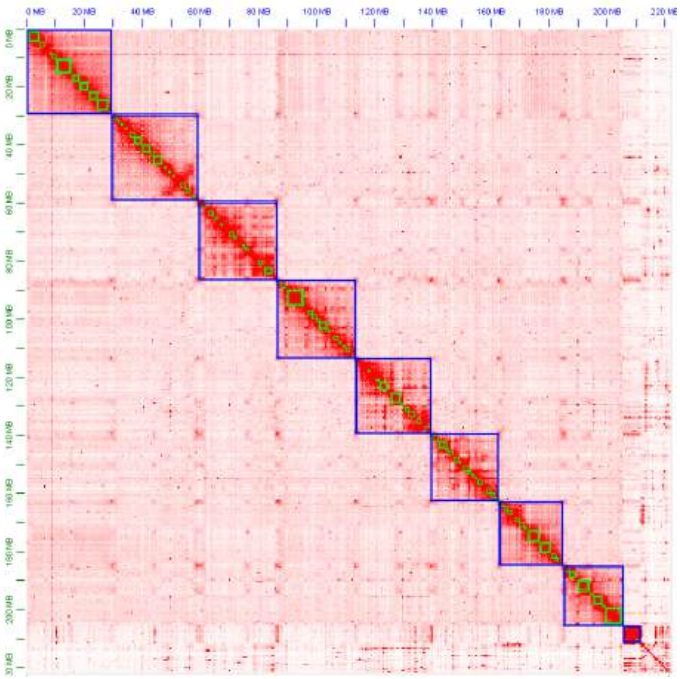


Supplementary Figure 43. Abundance of transcription factors across 54 representative Viridiplantae species (Fig. 3). Five plant groups are colored accordingly. Transcription factors (Supplementary Table 30) are sorted by their total abundances across the 54 investigated plants from left to right. The images of exemplar algae and tracheophyte are from BioRender (https://biorender.com/).

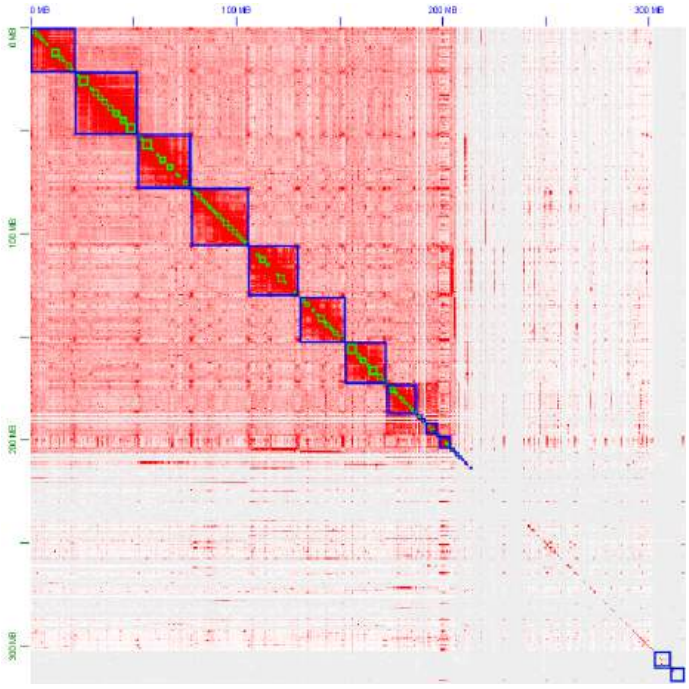
Raw sssembly with authentic *Marchantia* ONT reads



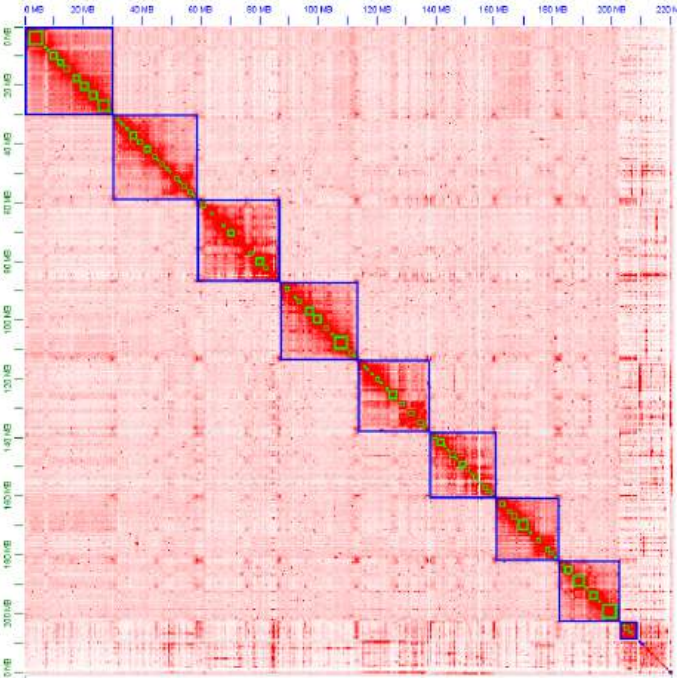
Clean assembly with authentic *Marchantia* ONT reads (decontaminated)



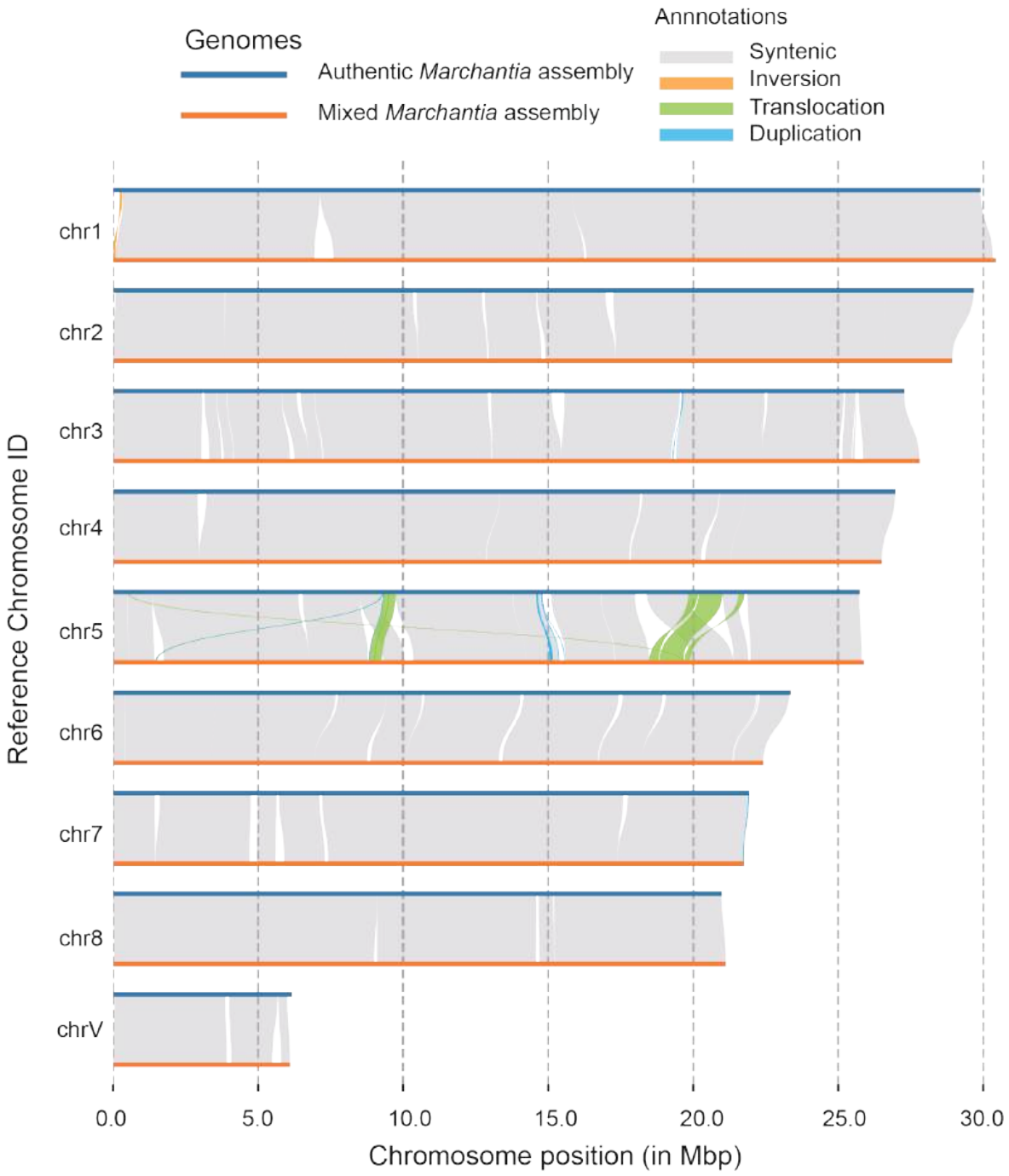
Mixed assembly of *Marchantia* and exogenous ONT reads (raw)



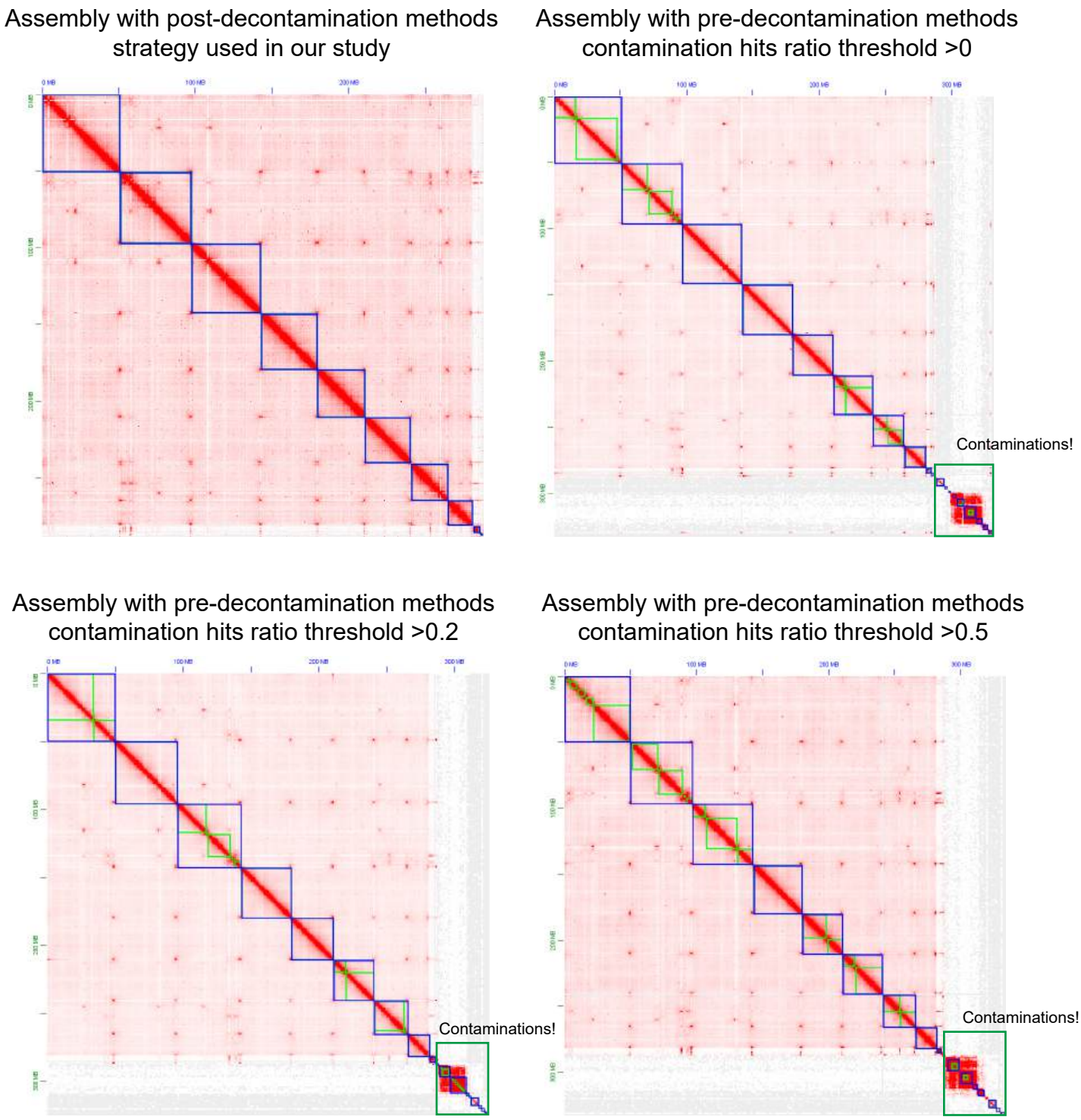
Mixed assembly of *Marchantia* and exogenous ONT reads (decontaminated)



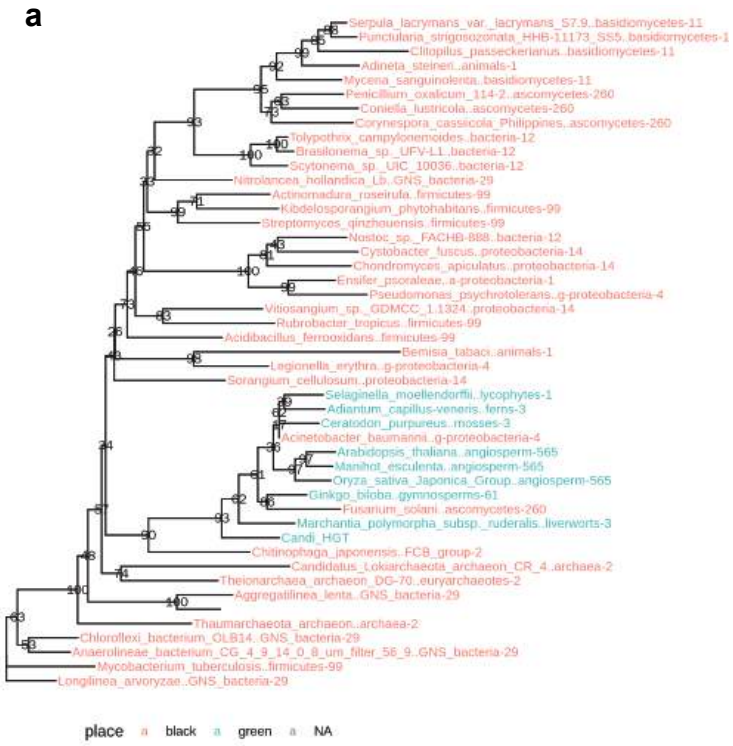
Supplementary Figure 44. Hi-C heatmap generated by Juicebox for different assemblies of *Marchantia polymorpha*. Due to the limited sequencing depth of *Marchantia* reads (~20x), both the authentic and mixed assemblies have some fragmented sequences that cannot be anchored with Hi-C scaffolding methods.



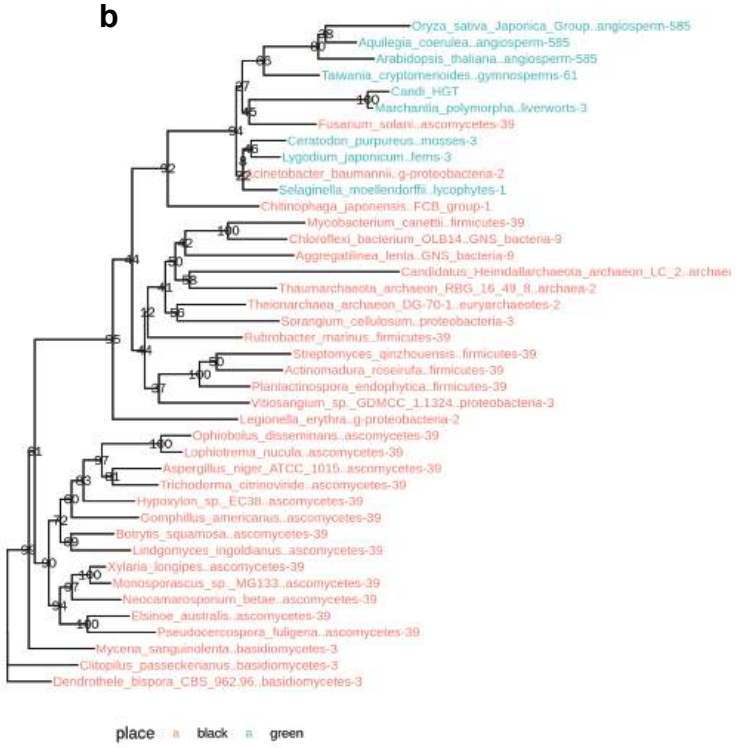
Supplementary Figure 45. Synteny plot between the authentic *Marchantia* assembly (decontaminated) and the mixed *Marchantia* assembly decontaminated following our pipeline. The synteny relationships are generally good, except for some translocations, which might be caused by some misplaced short fragmented contigs due to the limited sequencing depth (~20x) of ONT data and insufficient Hi-C data.



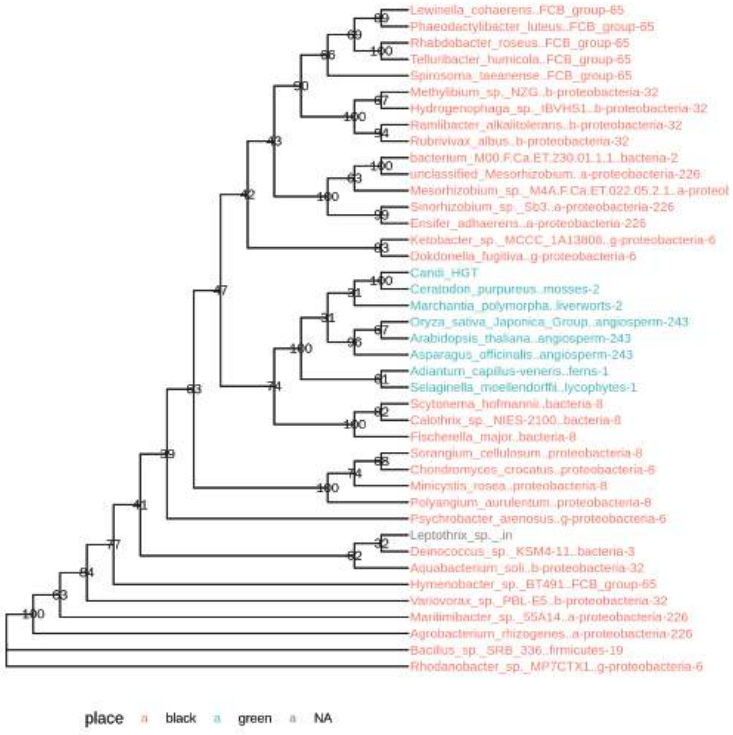
Supplementary Figure 46. Hi-C heatmap generated by Juicebox for different assemblies for *Conocephalum conicum*. The assemblies adopting pre-decontamination methods usually contain varying degrees of contaminated sequences as indicated in the green boxes. These contaminations are isolated and have no interaction with the main genome.



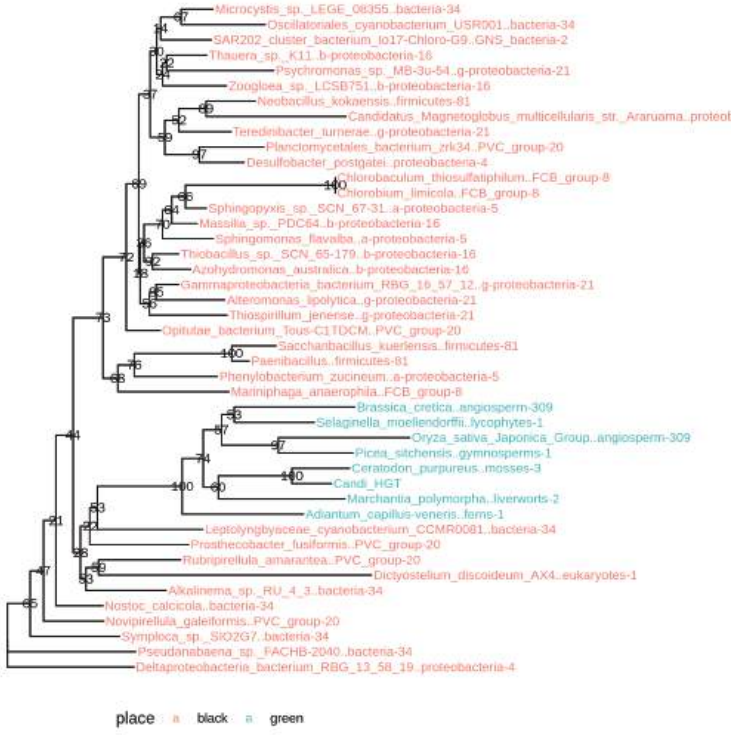
OG0049234-955-*jg15969*.t1.tree
strong support



OG0013970-664-*jg14390*.t1.tree
strong support



OG0000409-1010-*jg10155.t1*.tree
moderate support



OG0000021-1011-*jg4238.t1*.tree
weak support

Supplementary Figure 47. Examples of HGT trees showing different bootstrap support values for the candidate HGT genes and their donor lineages. The id information of the Candi_HGT was indicated below each figure.