Ancestral gene duplications in mosses characterized by integrated phylogenomic analyses

Bei Gao1,2,3, Mo-Xian Chen4, Xiao-Shuang Li5, Yu-Qing Liang1, Dao-Yuan Zhang1, Andrew J. Wood5, Melvin J. Oliver6*, and Jian-Hua Zhang3,7*

1State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China
2School of Life Sciences, The Chinese University of Hong Kong, Hong Kong, China
3State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, China
4Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
5Department of Plant Biology, Southern Illinois University-Carbondale, Carbondale, IL 62901-6509, USA
6Division of Plant Sciences, Interdisciplinary Plant Group, University of Missouri, Columbia, MO 65211, USA
7Department of Biology, Faculty of Science, Hong Kong Baptist University, Hong Kong, China

*Authors for correspondence. Melvin J. Oliver. E-mail: olivermj@missouri.edu; Jian-Hua Zhang. E-mail: jhzhang@hkbu.edu.hk

Received 30 March 2020; Accepted 6 September 2020; Article first published online 10 September 2020

Abstract Mosses (Bryophyta) are a key group occupying an important phylogenetic position in land plant (embryophyte) evolution. The class Bryopsida represents the most diversified lineage, containing more than 95% of modern mosses, whereas other classes are species-poor. Two branches with large numbers of gene duplications were elucidated by phylogenomic analyses, one in the ancestry of all mosses and another before the separation of the Bryopsida, Polytrichopsida, and Tetraphidopsida. The analysis of the phylogenetic progression of duplicated paralogs retained on genomic syntenic regions in the Physcomitrella patens genome confirmed that the whole-genome duplication events WGD1 and WGD2 were re-recognized as the ψ event and the Funariidae duplication event, respectively. The ψ polyploidy event was tightly associated with the early diversification of Bryopsida, in the ancestor of Bryidae, Dicranidae, Timmidae, and Funariidae. Together, four branches with large numbers of gene duplications were unveiled in the evolutionary past of P. patens. Gene retention patterns following the four large-scale duplications in different moss lineages were analyzed and discussed. Recurrent significant retention of stress-related genes may have contributed to their adaption to distinct ecological environments and the evolutionary success of this early-diverging land plant lineage.

Key words: gene duplication, moss, phylogenomic, Physcomitrella, polyploidy, WGD.

1 Introduction

Gene duplications provided the genetic raw materials for evolutionary innovation and are considered important driving forces in diversification and evolution (Ohno, 1970; Van de Peer et al., 2017), and the evolutionary past of land plants is characterized by recurrent ancestral polyploidy events (Blanc & Wolfe, 2004; Cui et al., 2006; Van de Peer et al., 2017). The duplicated genes in Arabidopsis can be traced back to five rounds of polyploidy, the α–β–γ–ε–ζ series, where the ε event occurred in the ancestor of angiosperms and ζ event in the ancestor for all seed plants (Blanc et al., 2003; Bowers et al., 2003; Jiao et al., 2011); however, this is still controversial (Ruprecht et al., 2017). Echoing the α–β–γ series, whole-genome duplication (WGD) series, termed ρ–σ–τ, was reported in grasses (Jiao et al., 2014). The gamma event appears associated with the species radiation of more than 75% of angiosperms (Friedman, 2009; Jiao et al., 2012; Vekemans et al., 2012). Large-scale transcriptomic analyses also revealed that hundreds of WGDs may have occurred in the evolutionary past of different green plant lineages (Leebens-Mack et al., 2019). However, both the phylogenetic relationships and the evolutionary history of mosses, an important clade in early-diverging land plant lineages (Mack et al., 2019), are less investigated.

A polyploidy event, characterized by a Ks peak around 0.5–0.9, was identified in the evolutionary past of the moss Physcomitrella patens (Rensing et al., 2008). Recently,
the chromosome-scale assembly of the *P. patens* genome deciphered the event as two recurrent WGDs: the younger WGD2 with a $K_s$ peak around 0.5–0.65 and the older WGD1 with a $K_s$ peak around 0.75–0.9 (Lang et al., 2018). Paralogous $K_s$ frequency distributions in a number of moss transcriptomes indicated widespread ancestral genomic duplications in mosses (Lang et al., 2018). Multiple genome duplications in peat mosses (class Sphagnopsida) and one ancestral polyploidy event that occurred in the ancestor of mosses were also reported (Devos et al., 2016).

High-throughput sequencing technologies generated adequate transcriptome data for discovery of ancestral large-scale genome duplications by analyzing $K_s$-based age distributions (Blanc & Wolfe, 2004; Cui et al., 2006) and/or large-scale gene tree-based phylogenomic approaches (Jiao et al., 2011; Jiao et al., 2012; Yang et al., 2018). For mosses, transcriptomes for 11 moss species were generated in the pilot study for the 1000 Green Plant Transcriptome Project (1KP; https://www.onekp.com; Wickett et al., 2014), and the transcriptomes of *Syntrichia caninervis* and *Bryum argenteum* have been reported in our previous studies (Gao et al., 2014; Gao et al., 2017). Recently, a phylotranscriptomic study in pleurocarpous mosses (Hypnales) generated a large set of transcriptomic data (Johnson et al., 2016). Despite the rapid growing body of sequencing data in mosses, the precise phylogenetic positioning of large-scale genome duplication events remains unresolved in Bryopsida, the largest class of mosses (Newton et al., 2006). The detection of these ancestral duplication events has been hampered by the unresolved phylogenetic relationships of the Bryopsida with other classes of mosses; however, phylogenetic relationships have been proposed (Shaw et al., 2010a, 2010b, 2011).

The *P. patens* chromosome structure and intragenomic synteny comparison provided unequivocal evidence for two recurrent ancestral WGDs (Lang et al., 2018), but the precise phylogenetic position for the two events was unresolved. Whether or not the WGDs were shared with other mosses or simply species-specific polyploidy events remained an open question. Although intergenome synteny comparison remains unavailable at present due to the lack of a closely related moss genome, resolving the phylogenetic placement of the two *P. patens* WGDs could be achieved by investigating the phylogenetic timing of the duplicated syntelog pairs. For example, the precise phylogenetic timing of the core-eudicot gamma triplication event was successfully resolved by tracing and summarizing the phylogenetic timing of *Vitis* syntelogs (Jiao et al., 2012).

To resolve the phylogeny of the mosses and phylogenetic positioning of ancestral genomic duplication events, we used phylotranscriptomic analyses of augmented transcriptomic data to generate reliable species phylogenies (Wickett et al., 2014; Ran et al., 2018; Unruh et al., 2018). This was coupled with a large-scale phylogenomic analysis investigating duplication signals from thousands of gene trees to determine the phylogenetic positioning of ancestral genomic duplication events to avoid the negative effects of “L shaped” background duplications that impact paralogue-based $K_s$ frequency analyses (Jiao et al., 2011; Devos et al., 2016; Johnson et al., 2016).

## 2 Material and Methods

### 2.1 Transcriptome data collection

Transcriptomes or genomes from 27 Bryophyta species, representing a wide range of lineages, along with the lycophyte (*Selaginella moellendorffii*) and the liverwort (*Marchantia polymorpha*), were collected (Table S1). We limited the analysis to species that occupy significant phylogenetic positions in the moss phylogeny (Wickett et al., 2014; Johnson et al., 2016; Lang et al., 2018) for computational efficiency. Only five of the available transcriptomes in Hypnales, representing different phylogenetic clades (Johnson et al., 2016), were used. For the Dicranidae, transcriptomes were included for five species from four studies (Gao et al., 2014; Wickett et al., 2014; Johnson et al., 2016; Lang et al., 2018). When assembled transcriptomes were unavailable, Illumina reads deposited in the NCBI-SRA were obtained, cleaned using Trimmomatic v0.36 (Bolger et al., 2014), and de novo assembled using Trinity v2.6.6 (Haas et al., 2013). Coding regions (CDS) were predicted for all transcripts and translated into peptide sequences via the TransDecoder v5.3.0 pipeline (https://github.com/TransDecoder/). The cd-hit v4.6 algorithm (Li & Godzik, 2006) was employed to remove redundantly assembled sequences with a sequence identity threshold of 0.98 ($\leq 0.98$) and minimum coverage of 0.9 ($\geq 0.9$). Only peptide sequences $\geq 50$ amino acids were retained. Genomic CDS for *Physcomitrella patens, M. polymorpha, and S. moellendorffii* were downloaded from the Phytozone v12.1.6 database (https://phytozone.jgi.doe.gov/).

### 2.2 Clustering transcripts into homolog clusters

To cluster transcripts into homolog clusters, all-by-all pair-wise bidirectional BlastP (Buchfink et al., 2015) comparisons (e-value $\leq 1e$-5) were conducted among the 29 transcriptomes. Only alignments with more than 50% coverage were introduced into the OrthoFinder v2.2.6 pipeline (Emms & Kelly, 2015). Homolog clusters were required to have at least one sequence from each of the three genomes (*P. patens, M. polymorpha, and S. moellendorffii*); homolog clusters with less than 10 taxa were discarded and 5684 homolog clusters were retained.

Protein sequences in each homolog cluster were aligned using MAFFT v7.310 (Katoh & Standley, 2013), employing the L-INS-i strategy and 5000 iterations. Coding regions were forced on the protein sequence alignment using Phyutility v2.2.6 and trimmed by a minimal column occupancy of 0.1 (Smith & Dunn, 2008). Gene trees for each homolog cluster were constructed using FastTree v2.1.10 (Price et al., 2010) with 1000 bootstrap replicates. At this stage (Fig. S1), gene clusters may include alternatively spliced isoforms and/or recent paralogs (high sequence identity). The Yang and Smith pipeline (Yang & Smith, 2014) was employed to inspect monophyletic or paraphyletic clades in each tree consisting of genes from a single taxon, and these clades were reduced to retain the longest unambiguous alignment, thus masking the transcript isoforms and recently duplicated paralogs in each transcriptome. The original protein sequences in each gene cluster (before trimming) were realigned using MAFFT v7.310 (Katoh & Standley, 2013). The alignments were trimmed and back-translated using TRIMAL v1.4 (Capella-Gutierrez et al., 2009). The generated CDS alignments for homolog clusters were divided into two subsets: one containing the putatively one-to-one
orthogroups (one gene from each species) for the species tree estimation and the other containing multicopy homolog clusters, to identify ancestral gene duplications.

2.3 Species tree construction
Homolog clusters were masked to maximize the number of one-to-one orthogroups (Table S2) and utilized for species tree estimation using both concatenated sequence supermatrix and coalescence-based summary methods (Johnson et al., 2016; Unruh et al., 2018). However, we recognize that some ancient paralogs may be misidentified as orthologs due to lineage or species-specific nonorthologous gene losses (Robertson et al., 2017). Reduced coding sequence alignments for the 649 one-to-one orthogroups were concatenated and gaps were inserted where taxa were missing. Three partition strategies were employed: (i) original concatenated alignment unpartitioned; (ii) one partition for the first and second codon positions, and a separate partition for the third codon position; and (iii) a recommended partition scheme (PartitionFinder v2.1.1; Lanfear et al., 2017) that treated each codon position as an individual partition. The best-fit partitioning scheme divided concatenated alignment into 611 partitions. The concatenated species tree was estimated using RAxML v8.2.12 (Stamatakis, 2014) under the GTRGAMMA evolutionary model and 500 bootstrap replicates with S. moellendorfﬁi as the outgroup. To validate the species tree, we constructed trees using the first and second codon positions and the concatenated protein alignments under the PROTGAMMAUTO model.

For the coalescence-based summary approach, individual maximum-likelihood (ML) unrooted gene trees for CDS alignments were estimated for each of the 649 single-copy orthogroups using RAxML v8.2.12 (Stamatakis, 2014) under the GTRGAMMA model with 500 bootstrap replicates. The ASTRAL-III v5.6.2 algorithm (Zhang et al., 2018) was used to estimate the coalescence-based species tree from individual gene trees, and support values were calculated from 100 replicates of multi-locus bootstrapping.

2.4 Phylogenetic placement of ancestral gene duplication events
The 5035 multicopy homolog clusters were employed for the identification of ancestral gene duplications. Maximum-likelihood gene trees were constructed for each homolog clusters using RAxML v8.2.12 under the GTRGAMMA model with 100 rapid bootstraps and with the gene sequences from Selaginella or Marchantia as outgroups.

The Phylogenetic Placement of Polyploidy Using the Genomes (PUG) algorithm (McKain et al., 2016; Unruh et al., 2018) was employed to identify paleopolyploidy events using the species tree generated from the concatenated sequence supermatrix as a guide. The PUG algorithm, based on species-tree to gene-tree reconciliation, was run for each of the 5035 gene trees using the “estimate_paralogs” option to count unique duplication nodes. For each paralogous pair, the coalescence node (potential duplication node) was identified from the gene family phylogeny, and species composition for both the duplication clades (subtrees) and the sister lineage was queried and scrutinized to count the duplication events for the internal nodes in the species tree. The placement was considered acceptable for the duplication of that paralogous pair only if the combined species composition of the subtree and the sister lineage resolved to the same node in the species tree (McKain et al., 2016; Unruh et al., 2018). Results from PUG were filtered for a minimal bootstrap value of 50 and 80 for each paralogous pair coalescence node.

Paralogous pairs retained on syntenic blocks (syntelogs) in the P. patens genome were identified using the McScanX algorithm (Tang et al., 2008; Wang et al., 2012) and their phylogenetic timing was traced from corresponding gene family phylogenies, similar to the strategy employed to analyze the core-eudicot GAMMA event in eudicots (Jiao et al., 2012). Specifically, the P. patens syntelog pairs were input as paralogous pairs in PUG to obtain their coalescence nodes in corresponding phylogenetic trees.

2.5 Synonymous divergence estimation of duplicated genes
Divergence patterns for duplicated genes were explored by calculating the synonymous substitutions per synonymous site (Ks, also known as Ds) for each paralogous pair. To capture the “pure” Ks peak signal of ancestral duplications, the paralogous pairs with a coalescence node at focal duplication nodes were extracted from the gene trees with a minimum bootstrap support value (BSV) of 50%. Protein sequence alignments were generated (Edgar, 2004) and back-translated (Suyama et al., 2006) for each duplicated sequence pair, and then alignment gaps and stop codons were removed. Ks values for each paralogous gene pair were calculated using Ks_Calculator v2.0 software (Wang et al., 2010) with the “GY” model (Goldman & Yang, 1994). The derived paralogous Ks values in each species were further analyzed using EMMIX (McLachlan et al., 1999) with one to ten components, using 100 random and 10 k-means starting values. The Bayesian Information Criterion was used to select the number of components that best fit the data.

2.6 Functional enrichment analyses for homolog groups with ancient duplications
Physcomitrella patens genes were annotated using InterProScan v5.34-73.0 (Jones et al., 2014) and Blast2GO v5.3.5 (Jones et al., 2014) to obtain gene ontology (GO) labels. Assuming that all genes clustered into the same homolog cluster share similar functions (Jiao et al., 2011; Ren et al., 2018), the GO labels of each P. patens gene were assigned to each of the corresponding homolog clusters. The functional enrichment analyses were subsequently performed for homolog clusters containing ancient duplications at the focal nodes. To assess the overrepresented GO categories for each homolog cluster, the Cytoscape v3.7.1 (Shannon et al., 2003) plugin BiNGO v3.0.3 (Maere et al., 2005b) was utilized to conduct a hypergeometric analysis and Benjamini and Hochberg False Discovery Rate (FDR) multiple test correction (Benjamini & Hochberg, 1995). The goSlim_plant ontology vocabulary was selected as the namespace. GO terms with statistical FDR ≤ 0.05 were considered as significantly overrepresented. The statistical results were depicted with heatmaps using the heatmap package in the R statistical environment.
3 Results and Discussion

3.1 Transcriptomes and homolog clusters

A total of 29 species (Fig. 1) were included for the phylogenomic analyses, including 12 de novo assembled transcriptomes (Table S1). The nonredundant protein sequences for each species (listed in Table S1) ranged from 11 140 (Rhodobryum ontariense) to 45 329 (Eosphagnum inretortum). A total of 643 020 nonredundant peptide sequences from the three genomes (Physcomitrella patens, Marchantia polymorpha, and Selaginella moellendorffii) and 26 de novo transcriptomes were sorted into 31 779 homolog clusters, which reduced to 5684 by retaining only homolog clusters that included a minimum of 10 taxa and at least one sequence in each of the three reference genomes. The 5684 homolog clusters were collapsed to retain only one representative sequence for mono- and paraphylogenetic clades that contained sequences from single taxa, and terminal branches with abnormal lengths were eliminated (Fig. S1). The 649 one-to-one orthogroups were used for species tree estimation and the remaining 5035 multicopy homolog groups for phylogenetic placement of WGD events. The masking of transcript isoforms and recent paralogs did not alter the detection of WGDs (Figs. S2 and S3), and the trimmed gene family trees removed spurious terminal branches, eliminating unreliable members (Yang & Smith, 2014).

3.2 Moss phylogeny reconstructed from single-copy orthogroups

Alignments from the 649 putative single-copy gene families were concatenated to generate a supermatrix that contained 833 313 sites. The contributions of each species to the single-copy orthogroups are presented in Table S2. With the exception of the placement of Hypnum imponens and Anomodon rostratus, relative to Thuidium delicatulum within Hypnales and the clade of Polytrichopsida plus Tetraphidopsida (Fig. S4),
the ML species tree produced a well-resolved moss phylogeny with BSVs of 100% for all other nodes. Although the BSVs for the three nodes in Hypnales were not 100%, all were larger than 90% in the three partition schemes (Fig. 1). The partitioning scheme recommended by PartitionFinder produced a moss phylogeny with better BSV support values for the three nodes in Hypnales nodes (BSV > 95%), supporting the phylogeny presented in Fig. 1. The Polytrichopsida plus Tetraphidopsida clade was supported when a species tree was constructed using partitioning scheme II (Fig. 5s), and the concatenated protein alignment (Fig. S6) also generated the identical species tree topologies with 100% BSVs.

A coalescence-based species tree was estimated from the 649 individual gene family trees using the ASTRAL-III v5.6.2 algorithm (Zhang et al., 2018). With the exception of the detailed deep phylogeny in Hypnales, the ASTRAL-III species tree topologies were consistent with the phylogeny reconstructed from concatenated supermatrix (Fig. S4). The Polytrichopsida plus Tetraphidopsida monophyletic clade was supported in the ASTRAL-III analysis (BSV of 100%).

Overall, the species tree topologies were consistent with the accepted moss taxonomy (Shaw et al., 2011) and with the recently resolved ordinal phylogeny of mosses (Liu et al., 2019). Takakia lepidozioides (Takakiotropsida) was confidently placed as sister to all other mosses, followed by the Sphagnopsida, as concluded in the study of Liu et al. (2019) and suggested by others (Nickrent et al., 2000; Qiu et al., 2006; Shaw et al., 2010a; Fig. 1). Our analyses resolved Takakiotropsida as a sister to all the other mosses, with 100% bootstrap support in all ML and ASTRAL-III analyses.

Within the species-rich class Bryopsida (true mosses), the subclasses Funariidae and Timmiidae (Timmia austriae) were recognized as graded sister groups to the clade of Dicranidinae plus Bryidae. These four subclasses together constituted a large proportion (>95%) of modern mosses (Newton et al., 2006) and demonstrated a relatively short ancestral branch length, indicative of a rapid species diversification process (Fig. 1). For expediency, we refer to them as the BDTF clade (Bryidae, Dicranidae, Timmiidae, and Funariidae). The other species-poor Bryopsida subclasses Diphysciidae (Diphuscium foliosum) and Buxbaumidinae (Buxbaumia aphylla) were resolved as laddered sister groups to the BDTF clade. The high-confidence supermatrix-based species tree topology generated here (Fig. 1) was employed in subsequent phylogenomic analyses for the detection of ancestral gene duplication events.

3.3 Phylogenetic placement of ancestral gene duplications

Ks frequency plots have been used to detect signals for ancestral WGDs for mosses (Rensing et al., 2007; Szovenyi et al., 2015; Lang et al., 2018). Lang et al. (2018) detected Ks peak signals in most Bryopsida and Sphagnopsida species and circumscribed two rounds of WGDs based on chromosomal synteny in the P. patens genome. However, detection of WGD events using paralome-based Ks distributions is difficult (Vekemans et al., 2012), and the saturation effect of Ks and rate heterogeneity among gene families and lineages can blur Ks-based age peaks (Jiao et al., 2011). A large-scale gene tree-based phylogenomic method (Jiao et al., 2011) avoids such issues, and it was used successfully in WGD studies (Jiao et al., 2010; Devos et al., 2016; McKain et al., 2016; Unruh et al., 2018).

We examined gene trees constructed from the 5035 multipholytom homolog groups using the PUG algorithm (McKain et al., 2016; Unruh et al., 2018), based on the constructed moss phylogeny, to detect six focal nodes with large numbers of gene duplications, as highlighted by the PUG algorithm (Fig. 2). The ancestral node with the largest number of gene duplications was identified at the ancestor of the BDTF clade (N20 in Fig. 2, 988 duplications BSV ≥ 80%). This supported a genome duplication event shared by all species in BDTF, designated as the ψ event or BDTF duplication. Large-scale duplication events were also indicated in the ancestor of the Funarioideae (N4 in Fig. 2, 794 duplications BSV ≥ 80%), which were not shared with Encalypta streptocarpus (Encalyptaceae) and in the ancestor of the Bryopsida, Polytrichopsida, and Tetraphidopsida (BPT) clade (N23 in Fig. 2, 265 duplications BSV ≥ 80%). The analyses captured the Sphagnopsida-wide genome duplication event (N1, 401 duplications BSV ≥ 80%) and the ancestral moss-wide duplication (N26, 357 duplications BSV ≥ 80%) reported earlier (Devos et al., 2016). Our analyses suggested that T. lepidozioides (Takakiotropsida) was included in the moss-wide genome duplication event. The analyses using the BSV ≥ 50% were also labeled and highlighted at each internal node and yielded results consistent with BSV ≥ 80% (Fig. 2). It should be noted that we cannot exclude the possibility that the large numbers of gene duplications detected on N26 and N23 are derived from accumulations of individual gene duplication, as genomic structural evidence is not available. The large number of gene family duplications occurring before the rapid speciation in Hypnales was also captured; however, the number of detected gene duplications is low (N14, 149 duplications with BSV ≥ 80%) and does not strongly support a recent WGD event. A comprehensive analysis of gene family expansions in the Hypnales was reported in the study of Johnson et al., 2016; our work only corresponds to ancestral gene duplications (as a part of the expansions) that are supported by the generated gene family phylogenies.

3.4 Two WGDs in P. patens

Phylogenetic timing of duplicated paralogs retained on genomic synteny regions in the P. patens genome was summarized in the moss phylogeny (Fig. 3 and Table 1). More than 95% of the syntellog pairs were phylogenetically identified as products of the N20 (BDTF) and N4 (Funarioideae) duplications. Using P. patens syntellogs as anchors, 93 and 114 duplication nodes (BSVs of 80% and 50%) were detected in the ancestor to the BDTF clade. Similarly, 287 and 339 duplications (BSVs of 80% and 50%) were detected in the ancestor of Funarioideae (shared by Physcomitrella and Funaria). All other ancestral nodes contained few (less than 1.5%) of the duplicated syntellogs (Table 1).

Syntellogs derived from N20 and N4 duplications were connected using CIRCOS v0.69-6 (Krzywinski et al., 2009; Figs. 3A, 3B) and compared to the P. patens chromosomal synteny karyotype (Lang et al., 2018). The spatial distribution of the duplicated syntellogs provided strong evidence that the older WGD1 is congruent to the BDTF clade (N20) duplication, whereas the younger WGD2 is consistent with the Funarioideae (N4) duplication. For example, the Chr5 vs Chr6, Chr7 vs Chr11, and Chr9 vs Chr15 are chromosome pairs representing products
of the younger WGD2 events (Lang et al., 2018), and they are heavily populated with connected duplicated syntelogs (blue lines Fig. 3B), indicative of Funarioideae (N4) duplications. Consistent phylogenetic timing evidence for the older WGD1 patterns could also be observed in the analyses; chromosomes 18/21 and 19/22 in \( P.\ patens \) were derived from the two WGD events from a single ancestral chromosome, largely congruent with \(([\text{chr18,chr21}]\text{WGD2},[\text{chr19,chr22}]\text{WGD2}]\text{WGD1}; \text{Lang et al., 2018}).

The phylogenomic analyses provided robust evidence that the ancestral \( \psi \) genome duplication event occurred before diversification within the BDTF clade and the data are consistent with the reported WGD1 event in \( P.\ patens \) (Lang et al., 2018) and the WGD2 event in the ancestor of the Funarioideae.

3.5 Synonymous divergence of \( P.\ patens \) syntelogs from the Funarioideae and \( \psi \) duplications

To evaluate the sequence divergence of the syntelogs after paleopolyploidy, we generated and analyzed the \( K_s \) frequency distributions for the syntelog pairs that traced the Funarioideae and \( \psi \) duplication events in corresponding gene family phylogenies (Fig. 3). The observed \( K_s \) frequency distribution peaks were consistent with the phylogenetic placement results, as described above, and the Funarioideae and \( \psi \) duplications are in parallel with major \( K_s \) peaks around 0.6 (Figs. 3C, 3D) and 0.9 (Figs. 3E, 3F), respectively.

The synonymous divergence curves for syntelogs derived from both Funarioideae and \( \psi \) duplications were obviously right-skewed and the \( K_s \) frequency distributions could be modeled with two major components containing one “younger” peak component and one “older” right-skewed component (blue and green for Funarioideae duplications in Figs. 3C, 3D, red and purple for \( \psi \) duplications in Figs. 3E, 3F). The analysis using the BSV threshold at 50% for each component was described by “color/mean/proportion,” in agreement with previous analyses (Jiao et al., 2011, 2012). The \( K_s \) frequency distribution of syntelog pairs derived from the Funarioideae duplications was modeled with three components: blue/0.59/0.67, green/0.88/0.27, and black/1.36/0.06 (Fig. 3C), and the \( K_s \) frequency of \( \psi \) event syntelogs contained two statistically significant components: red/0.9/0.775 and purple/1.5/0.225 (Fig. 3E).

Thus, the green component (Figs. 3C, 3D) in the Funarioideae duplications could severely blur the major component of the \( \psi \) duplications, as shown in Fig. 3G. This synonymous distance distribution pattern would make it difficult to infer ancestral genome duplications directly from the \( K_s \) distribution without the phylogenetic timing information. We also analyzed the \( K_s \) frequency distribution for all syntelogs in
P. patens; the four major $K_S$ components were recaptured (Fig. 3H) and were similar to that reported in the study of Lang et al., 2018. We described each component using “component no./mean/proportion”: 1/0.56/0.21, 2/0.76/0.33, 3/1.09/0.41, and 4/1.8/0.05, and according to the $K_S$ distribution analyzed above. The youngest component (1/0.56/0.21) is largely in parallel to the Funarioideae duplication event. However, the other two major components (2/0.76/0.33 and 3/1.09/0.41) could be a mixture of both Funarioideae and $\psi$ duplications.

3.6 Nature of the $\psi$ duplication event
To validate the “two-component” divergence pattern of the $\psi$ genes, we extracted all paralogous pairs with a coalescence node at N20 (BSV $\geq$ 50%) tracing the $\psi$ events in gene family trees. $K_S$ values were estimated for these “low background” sets of $\psi$ duplicated paralogous pairs. In parallel with the age estimations (Fig. S7), $K_S$ frequency distributions for the BDTF clade species were modeled with two significant components (Fig. S7, A–R); for example, the two major components “red/0.9/0.51” and “purple/1.26/0.4” seen in the P. patens syntelog.
The syntelog pairs were identified in the K5 distribution for the ψ duplicated genes inferred from gene trees (Fig. S7L). The inferred *P. patens* paralogous pairs with a coalescence node N4 (BSV ≥ 50%) tracing the Funarioideae duplication event, two major components, “blue/0.63/0.66” and “green/0.99/0.3,” were identified (Fig. S7T), which was consistent with the syntelog-based analyses. Synonymous substitution rate frequency distributions of orthologs between *P. patens* and other BDTF mosses demonstrated slightly younger peaks around 0.54–0.86 (Fig. S8), further confirming that the ψ event occurred before the diversification of the BDTF clade.

Heterogeneity in evolutionary rates between different gene families and K5 saturation effects may partially explain the distinctive “two components” in the age distribution of the ψ-derived duplicates. Several studies suggested that the right-skewed K5 distribution might be better modeled using a log-normal distribution (Cui et al., 2006; Zwaenepoel & Van de Peer, 2018), but log-transformed K5 could not perfectly resolve the two recurrent WGDs in *P. patens* (Lang et al., 2018), because the two peaks are close to each and mixed together. The two components are not the result of two recurrent genome duplication events or a triplication event as that would have been detected in *P. patens* (Lang et al., 2018). It is possible that genome segments were duplicated before the tetraploidization of the two subgenomes that constitute the ψ polyploidy event; however, specific genomic segments that were significantly older than others were not observed.

Moreover, we cannot fully rule out the possibility that the older component of the ψ duplications might be the result of errors introduced in the large-scale phylogenomic procedure, especially when a number of transcriptomes were included. It is possible that some “older” paralogous pairs might be misidentified on a “younger” node when a sequence from the sister taxa was absent due to gene losses and the incompleteness of transcriptomes. For example, when the genes from Encalypta were not expressed in the transcriptome, the paralogs duplicated in the ancestor of *P. patens*, Funaria, and Encalypta would be placed in the ancestor of *P. patens* and Funaria. Nevertheless, the phylogenomic analyses, together with the convincing genome synteny structural evidence, probed the accurate phylogenetic position of the two WGD events.

It is possible that the ψ duplication was an allopolyploidy event resulting from dioecious female and male gametophyte genomes, as suggested by Rensing et al. (2007), where the divergences existed before the WGD event. This then indicates that the dated nodal ages traced the divergence time of the separation of the two subgenomes and not the polyploidy event (Doyle & Egan, 2010). We favor this hypothesis, because the allopolyploidization of dioecious gametophytes could yield a monoecious plant that would enhance the dispersal of breeding populations (Rensing et al., 2007) and provide extra ecological advantages for polyploids.

### 3.7 Evolutionary rate variations among mosses

Variations in nucleotide substitutional rates in the evolution of angiosperm genomes is common and linked to distinctive life histories (Smith & Donoghue, 2008).

---

**Table 1** Phylogenetic timing of the *Physcomitrella patens* syntelogs inferred from gene family phylogenies

| Ancestral Nodes | N4a | N5 | N20b | N21 | N22 | N23 | N24 | N25 | N26 |
|-----------------|-----|----|------|-----|-----|-----|-----|-----|-----|
| BSV ≥ 80        | 287 | 2  | 93   | 0   | 2   | 2   | 0   | 0   | 3   |
| Percent         | 73.8% | 0.5% | 23.9% | 0%  | 0.5% | 0.5% | 0%  | 0%  | 0.8% |
| BSV ≥ 50        | 339 | 8  | 114  | 2   | 2   | 2   | 0   | 0   | 4   |
| Percent         | 71.8% | 1.5% | 24.2% | 0.4% | 0.4% | 0.4% | 0.4% | 0%  | 0.8% |

Abbreviations: BDTF, Bryidae, Dicranidae, Timmildae, and Funariidae; BSV, bootstrap support value; MRCA, Most Recent Common Ancestor.

aThe node of N4 corresponds to the MRCA of Funarioideae in the moss phylogeny.
bThe node of N20 corresponds to the MRCA of BDTF clade in the moss phylogeny.
There were distinctive \( K_S \) peaks for the shared \( \psi \) event, from 0.45 (T. austriaca) to 1.32 (Bryum argenteum; Fig. S7, A–R), suggesting that there are substantial evolutionary rate variations among the different lineages after the \( \psi \) event. Setting the \( \psi \) event as the ancestral calibration point, we compared the nucleotide substitutional rates among the different lineages using \( K_S \) estimations of \( \psi \) paralogs (Fig. 4).

There were no conspicuous lineage-wide rate shifts, but there were significant within-lineage differences in evolutionary rates (Fig. 4). The species within the Hypnales clade exhibited similar substitutional rates as the \( \psi \) event paralogs, but for B. argenteum, the rate was significantly elevated. Evolutionary rate elevations were also seen in the Dicranidae (Syntrichia caninervis) and the Funariidae (Funaria hygrometrica). Each of the three species, B. argenteum, S. caninervis, and F. hygrometrica, exhibited major \( K_S \) peaks (Fig. S7) that exceeded 1.0 \( K_S \) for the \( \psi \) event. In contrast, significantly slower nucleotide substitutional rates could be observed for T. austriaca, E. streptocarpa, and Scouleria aquatica, with major \( K_S \) peaks (Fig. S7) smaller than 0.6. Additionally, duplication signals were also observed in all the \( K_S \) frequency distributions for each of the whole paranome in BDTF moss species (Fig. S9).

3.8 Gene retention after ancient large-scale duplications

The four ancestral branches with large numbers of gene duplications documented here could have contributed to not only to gene family expansions but also functional innovations in the evolutionary past of mosses. The potential functional bias among the duplicated paralogs that “survived” the ancient duplications was evaluated using GO enrichment analyses for the homolog clusters that contained the ancient duplications. Recurrent gene duplications within a specific gene family were rare, as only 23 homolog clusters were detected to contain duplication signals for all four ancient duplication events (Fig. 5A), and in P. patens genome, we detected only seven such gene families (Fig. 5B). If paralogs generated from all four duplications were retained, they could have generated as many as 24-fold gene member expansions; thus, gene losses would be common after the duplication events. The Venn diagram analyses confirmed that the majority of the homolog clusters (1516 out of 2180 homolog clusters, 69.5%) contained only one of the ancient duplication events. Homolog clusters that contained two (503 clusters, 23%) and three (138 clusters, 6.3%) duplication events were much reduced. Similar patterns were observed in the homolog clusters containing duplication events for P. patens paralogs; 1288 (81%), 248 (15.6%), and 45 (2.8%) homolog clusters that contained gene duplications were derived from single, two, and three duplication events, respectively. For gene duplications supported by both moss and P. patens paralogs, a conspicuously large proportion of the overlapping gene families survived both the \( \psi \) and the Funarioideae duplication events, indicating that duplication signals for these two younger WGD events in Funarioideae were persistent in gene family phylogenies, such as that depicted in Fig 3A.

Duplicated genes originated from the \( \psi \) and Funarioideae events remain pervasive in the P. patens genome (Lang et al., 2018); however, the paralogous pairs that survived the older duplication events would be rarer. We observed paralogous pairs in several transcription factor families that survived the moss-wide and BPT duplications. The genes clustered in orthogroup_243 encode bZIP transcription factors and P. patens genes were found in both of the child clades that constituted a moss-wide duplication. The more recent duplications (the \( \psi \) and Funarioideae events) were observed in the gene family phylogeny (Fig. 5a). Orthogroup_362 encodes the Nin-like transcription factors where moss-wide
and BPT duplication nodes were well-supported by the tree phylogeny (Fig. S11). Similarly, the WRKY (Fig. S12) and SBP (Fig. S13) transcription factors in *P. patens* also survived the ancestral moss-wide duplication. The genes in orthogroup_462 encode homologs of vascular associated death 1, a regulator of cell death and defense vascular responses (Lorrain et al., 2004), and also retain duplicated genes derived from the moss-wide and BPT duplications (Fig. S14). In all cases (Figs. S10–S14), the more recent ψ and Funarioideae duplication events contributed extra family members for specific clades in the gene family phylogeny. The preservation of the paralogs derived from ancient duplication events and the recurrent acquisition of family members exert evolutionary and ecological significance for functional renovations and the species radiation process (Rensing, 2014).

To determine if genes of specific functional categories were preferentially retained after the ancestral large-scale duplication events, we performed a GO term enrichment analyses for the homolog clusters containing gene duplications (Figs. S15–S18). The retained gene duplications were identified by the duplicated paralog pairs that mapped to the four focal nodes (N26, N23, N20, and N4) in the ML gene trees. We observed both conserved and distinct GO slim plant functional categories associated with the four ancient duplication events (Figs. 5C and 6). Despite the conclusion that gene families containing all four recurrent duplications are rare, several GO categories such as “response to stress,” “carbohydrate metabolic process,” “anatomical structure morphogenesis,” and “post-embryonic development” were significantly overrepresented for all four duplication events.

Fig. 5. Analyses of the homolog clusters containing ancestral gene duplications. A, The Venn diagram showing the overlap of four sets of homolog clusters containing ancestral duplications with nodal bootstrap values not less than 50%. B, The Venn diagram showing the overlap of four sets of homolog clusters containing ancestral duplications supported by *Physcomitrella patens* paralogs with nodal bootstrap values not less than 50%. C, The gene ontology functional enrichment analysis of duplicated gene families for each of the four ancestral large-scale duplication events. The statistical FDR corrected P values were depicted using simplified color schemes: gray colors represented not significant GO terms, light blue for statistically overrepresented GO terms with FDR ≤ 0.05, and dark blue for terms with FDR ≤ 0.01. FDR, False Discovery Rate; GO, gene ontology.
Duplicated gene families involved in “cell communication” and “signal transduction” were only enriched after the latest three duplication events. Duplications in gene families participating in “cell differentiation” and “transporter activity” were especially enriched for the moss-wide duplication event, indicating that these fundamental activities might have been well established in the ancestor of all mosses. A number of GO terms such as “DNA-binding transcription factor activity” and “generation of precursor metabolites and energy” were enriched after the most recent duplication event shared by *Physcomitrella* and *Funaria*, which is consistent with the determination that the *P. patens* genomic duplication contributed a large number of metabolic genes to its genome (Rensing et al., 2007). The enrichment of genes related to transcriptional regulatory functions may further suggest a conserved gene retention pattern across land plants, as observed in peat mosses (Devos et al., 2016), ferns (Li et al., 2018), and angiosperms.
Gene duplications in mosses

3.9 Implications of the ancestral large-scale duplications in mosses

Paleopolyploidy events provide substantial raw genetic materials for evolution and drive species radiation (Schranz et al., 2012; Vanneste et al., 2014; Van de Peer et al., 2017); however, it was suggested that polyploidy could be an evolutionary “dead end” (Mayrose et al., 2011; Arrigo & Barker, 2012). The data we generated indicated that the ψ event may have provided an impetus for a major species radiation in the Bryopsida. This is strengthened by the observation that the early-diverging Bryopsida lineages (Diphyisciidae and Buxbaumiidae) that are outside of the ψ event are species-poor, and the BDTF clade containing more than 95% of current moss species (Newton et al., 2006) experienced the ψ event.

However, unlike the GAMMA event in the core eudicots where triplicated syntenic regions were found in the Vitis genome (Jaillon et al., 2007), the ψ event in the mosses only generated duplicated syntenic blocks (as seen in the P. patens genome). The ψ event in the P. patens lineage likely resulted in a duplication of the seven ancestral moss chromosomes into the current 13 (one chromosome loss) intermediate chromosomes (Lang et al., 2018).

Our phylogenomic analyses circumscribed the P. patens WGD2 as a Funarioideae duplication event that was shared by both P. patens and F. hygrometrica. However, some F. hygrometrica accessions have been reported to contain 14 or 56 chromosomes, and 51% of the accessions contained 28 chromosomes (Fritsch, 1991). The P. patens accession (Gransden 2004) has 27 chromosomes (Lang et al., 2018) and some isolates of P. patens have 14 or 28 chromosomes (Engel, 1968). This suggests that some species within the Funarioideae may have escaped the duplication event or that independent polyploidization events are common in the mosses. It is also possible that some BDTF species might have escaped the ψ event, which needs to be tested. This is especially true when the transcriptomic/genomic data become available for F. hygrometrica accessions with low (e.g., n = 7) chromosome numbers, even though low chromosome numbers alone could be unequivocal evidence for escaping the ancestral WGD; for example, the five Arabidopsis thaliana chromosomes contained three WGDs, whereas the ancestral eudicot genome was reconstructed as seven chromosomes (Salse et al., 2009).

4 Conclusions

A robust moss phylogeny was constructed using an extensive matrix of putative one-to-one orthologs derived from transcriptomic data. Consistent evidence that supported the occurrence of a whole-genome duplication event (termed ψ) before the species radiation within the BDTF clade was generated. We circumscribed the WGD1 and WGD2 events in the ancestor of BDTF clade and Funarioideae, respectively. We identified another two branches with large numbers of gene duplications: one in the ancestor of all mosses and another before the separation of Polytrichopsida, Tetraphidopsida, and Bryopsida. Functional assessment of homolog clusters containing gene duplications offers a possible insight into the evolutionary forces that have impacted bryophyte diversification.

Acknowledgements

This study was supported by the NSFC-Xinjiang Key Project (Grant No. U1703233 to DZ) and National Science Foundation (Collaborative Research: Dimensions: Grant No. 1638972 to MJO). The School of Life Sciences of The Chinese University of Hong Kong provided the computational resources for this project. BG conceived the study, performed the bioinformatic analyses, and wrote the manuscript. MC, DZ, XL, YL, and AJW contributed to the result discussion. MJO and JZ contributed to the data interpretation and critically revised and improved the manuscript. All authors read and approved the manuscript.

References

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. Current Opinion in Plant Biology 15: 140–146.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Statistical Methodology 57: 289–300.
Gene duplications in mosses

Stevenson DW, Ullrich KK, Wickett NJ, DeGironimo L, Edger PP, Jordan-Thaden IE, Joya S, Liu T, Melkonian B, Miles NW, Pokorny L, Quigley C, Thomas P, Villareal JC, Augustin MM, Barrett MD, Ba continuously and rapidly. The EMMIX software for the fitting of mixtures of normal and t-components. Journal of Statistical Software 4: 1–14.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. 2018. The timescale of early land plant evolution. Proceedings of the National Academy of Sciences USA 115: E2274–E2283.

Newton AE, Wikström N, Bell N, Lowe Forrest L, Ignatov MS. 2006. Dating the diversification of pteridophytic mosses. In: Newton AE, Tangney RS eds. Pteridophytic mosses: Systematics and evolution. Boca Raton: CRC Press. 337–366.

Nichol DL, Parkinson CL, Palmer JD, Duff RJ. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Molecular Biology and Evolution 17: 1885–1895.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer.

Pan Z, Pitt WG, Zhang Y, Wu N, Tao Y, Truscott TT. 2016. The upside-down water collection system of Syntrichia caninervis. Nat Plants 2: 16076.

Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5: e9490.

Qiu YL, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrowska O, Lee J, Kent L, Rees J, Estabrook GF, Hendry TA, Taylor DW, Testa CM, Ambros M, Crandall-Stotter BJ, Duff RJ, Stech M, Frey W, Quandt D, Davis CC. 2006. The deepest divergences in land plants inferred from phylogenomic evidence. Proceedings of the National Academy of Sciences USA 103: 15511–15516.

Ran JH, Shn TT, Wang MM, Wang XQ. 2018. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between gnetales and angiosperms. Proceedings of the Royal Society B: Biological Sciences 285: 20181012.

Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, Ma H, Qi J. 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Molecular Plant 11: 414–428.

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. Current Opinion in Plant Biology 17: 43–48.

Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R. 2007. An ancient gene duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. BMC Evolutionary Biology 7: 130.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schwein G, Shiu SH, Stueber K, Theodoulou FL, Tu H, van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Misher BD, Reski R, Grigoriev IV, Quatrano RS, Taylor DW, Testa CM, Ambros M, Crandall-Stotter BJ, Duff RJ, Stech M, Rees J, Quandt D, Davis CC. 2016. The upside-down water collection system of Syntrichia caninervis. Nat Plants 2: 16076.
reveals evolutionary insights into the conquest of land by plants. Science 319: 64–69.

Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. 2017. Lineage-specific reduplication is a mechanism to explain time-lags between genome duplication and evolutionary diversification. Genome Biology 18: 111.

Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, Persson S. 2017. Revisiting ancestral polyploidy in plants. Science Advances 3: e1603195.

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C. 2009. Reconstruction of monocotelydaneous proto-chromosomes reveals faster evolution in plants than in animals. Proceedings of the National Academy of Sciences USA 106: 14908–14913.

Scharzen ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: The WGD radiation lag-time model. Current Opinion in Plant Biology 15: 147–153.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, RAMAGE D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 13: 2498–2504.

Shaw AJ, Cox CJ, Buck WR, Devos N, Buchanan AM, Cave L, Seppelt R, Shaw B, Larrain J, Andrus R, Greilhuber J, Temsch EM. 2010a. Newly resolved relationships in an early land plant lineage: Bryophyta class sphagnopsida (peat mosses). American Journal of Botany 97: 1511–1531.

Shaw AJ, Devos N, Cox CJ, Boles SB, Shaw B, Buchanan AM, Cave L, Seppelt R. 2010b. Peatmoss (sphagnum) diversification associated with miocene northern hemisphere climatic cooling? Molecular Phylogenetics and Evolution 55: 1139–1145.

Shaw AJ, Szovenyi P, Shaw B. 2011. Bryophyte diversity and evolution: Windows into the early evolution of land plants. American Journal of Botany 98: 352–369.

Smith SA, Dunn CW. 2008. Phyloutility: A phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24: 715–716.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322: 86–89.

Stamatakis A. 2014. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Suyama M, Torreres D, Bork P. 2006. Pal2nal: Robust conversion of protein sequence alignments into the corresponding codon sequences. Nucleic Acids Research 34: W609–612.

Szovenyi P, Perroud PF, Symeonidi A, Stevenson S, Quattrano RS, Rensing SA, Cumming AC, McDaniel SF. 2015. De novo assembly and comparative analysis of the ceratodon purpureus transcriptome. Molecular Ecology Resources 15: 203–215.

Tang HB, Wang XY, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Research 18: 1944–1954.

Unruh SA, McKain MR, Lee YI, Yakawa T, McCormick MK, Shefferson RP, Smithson A, Leebens-Mack JH, Pires JC. 2018. Phylogenomic analysis and genome evolution of the cypripedioideae (orchidaceae). American Journal of Botany 105: 631–640.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nature Reviews Genetics 18: 411–424.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the cretaceous-paleogene boundary. Genome Research 24: 1334–1347.

Vekemans D, Proost S, Vanneste K, Coenen H, Vlaene T, Ruelens P, Maere S, Van de Peer Y, Geuten K. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: Significance for mads-box gene and species diversification. Molecular Biology and Evolution 29: 3793–3806.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Lin H, Marler B, Guo H, Kissinger JC, Paterson AH. 2012. Mcscanx: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research 40: 449.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. Kaks_calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 8: 77–80.

Wickett NJ, Mirara S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafala E, Der JP, Graham SW, Mathews S, Melkonian M, Solits DE, Solits PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchian TM, Augustin MM, Wang J, Zhang Y, Tian Z, Yan Z, Wu X, Sun X, Wong CK, Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences USA 111: E4859–4868.

Wood AJ, Jenks MA. 2007. Plant desiccation tolerance: Diversity, distribution, and new-world evolution. In: Jenks MA, Wood AJ eds. Plant desiccation tolerance. Hoboken: Blackwell Publishing Ltd. 1–10.

Yang Y, Moore MJ, Brockington SF, Mikenas J, Olivieri J, Walker JF, Smith SA. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in caryophyllales, including two allopolyploidy events. New Phytologist 217: 855–870.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. Molecular Biology and Evolution 31: 3081–3092.

Zhang C, Rabbee M, Sayyari E, Mirarab S. 2018. Astral-iii: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19: 153.

Zwaenepoel A, Van de Peer Y. 2018. Wgd—Simple command line tools for the analysis of ancient whole genome duplications. Bioinformatics 35: 2155–2155.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12683/suppinfo:

Fig. S1. Phylogenomic workflow utilized in this study for species tree construction and identification of ancestral gene duplications in mosses.

Fig. S2. Exemplary gene tree generated in FastTree before masking the single-taxa monophyletic and paraphyletic clades. Clades that were constituted with genes from a single species were highlighted in red and the two ancestral duplications in Funariaeideae and BDTF clade were labeled with stars.

Fig. S3. Exemplary gene tree generated in RAxML after masking the single-taxa monophyletic and paraphyletic clades. Clades constituted with genes from a single species

J. Syst. Evol. 60(1): 144–159, 2022 www.jse.ac.cn
in Fig. S2 were masked and one representative sequence retained using the Yang & Smith (2014) pipeline. The detection of the two ancestral duplications in Funarioidae and BDTF (labeled with stars) were not affected by this masking procedure.

**Fig. S4.** Comparison of species trees constructed using supermatrix-based and coalescence-based approaches. Left panel depicts the maximum-likelihood nucleotide supermatrix tree topology from RAxML. The right panel depicts the coalescence-based species tree inferred from ASTRAL-III, generated from the 649 maximum-likelihood gene trees of each single-copy orthogroup. Nodal bootstrap support values indicated by asterisks are 100%. The supermatrix-based and coalescence-based approaches generated consistent species tree topology at the subclass level in mosses, the only variation was discovered in the clade of Hypnales.

**Fig. S5.** The maximum-likelihood species tree generated using the 1st and 2nd codon positions of the concatenated alignment supermatrix. Branch lengths are scaled to substitutions per site. The MRCA node for BDTF clade was indicated, and the relatively short branch length might be congruent with the rapid speciation following the WGD event.

**Fig. S6.** The maximum-likelihood species tree generated from the concatenated protein alignment using RAxML under the PROTGAMMAAUTO model. Branch lengths are scaled to substitutions per site. The MRCA node for BDTF clade was indicated and highlighted.

**Fig. S7.** \( K_5 \) distribution of all inferred duplicated paralogs in different species tracing the \( \Psi \) and Funarioidae duplications. (A-S) Paralogous pairs with coalescence node at \( N_20 \) (BSV \( \geq 50 \)) were identified in each BDTF clade species and corresponding \( K_5 \) distributions were analyzed. (S-T) \( K_5 \) distribution analyses of paralogous pairs with coalescence node at \( N_4 \) (BSV \( \geq 50 \)) in Funaria hygrometrica and Physcomitrella patens.

**Fig. S8.** Distribution of synonymous substitutions (\( K_S \)) in orthologous pairs between species of Bryidae, Dicranidae, Timmiidae and Physcomitrella patens. The modal \( K_S \) values were calculated from Kernal Density Estimation function in R statistical environment and labeled in the orthologous \( K_S \) distribution, providing supporting evidence for the relatively young age of the BDTF clade.

**Fig. S9.** Paralogous \( K_S \) frequency distribution of whole paralogomes in the mosses. Nodal \( K_S \) values were estimated and redundancies were eliminated according to Maere et al. (2005). The R package ‘mclust’ was employed to conduct the finite mixture model analyses and gaussian components with mean values in [0.3, 2] and proportions were depicted in the figures. Putative gaussian components corresponding to the \( \Psi \) were highlighted in red fonts.

**Fig. S10.** Exemplar maximum-likelihood gene family (OG-243) phylogeny contains Physcomitrella patens paralogs that survived ancestral duplication. Genes in orthogroup 243 encode the bZIP transcription factors. \( P. \) patens genes were highlighted with arrows and the moss-wide duplication node was labeled as brown star and the BPT duplication was labeled as a yellow star. More recent duplications supported by tree topology were labeled as green circles.

**Fig. S11.** Exemplar maximum-likelihood gene family (OG-362) phylogeny contains Physcomitrella patens paralogs that survived ancestral duplications. Genes in orthogroup 362 encode the Nin-like transcription factors. \( P. \) patens genes were highlighted with arrows and the moss-wide duplication node was labeled as brown star and the BPT duplication was labeled as a yellow star. More recent duplications supported by tree topology were labeled as green circles.

**Fig. S12.** Exemplar maximum-likelihood gene family (OG-504) phylogeny contains Physcomitrella patens paralogs that survived ancestral duplications. Genes in orthogroup 504 encode the WRKY transcription factors. \( P. \) patens genes were highlighted with arrows and the moss-wide duplication node was labeled as brown star. More recent duplications supported by tree topology were labeled as green circles.

**Fig. S13.** Exemplar maximum-likelihood gene family (OG-1137) phylogeny contains Physcomitrella patens paralogs that survived ancestral duplications. Genes in orthogroup 1137 encode the SBP transcription factors. \( P. \) patens genes were highlighted with arrows and the moss-wide duplication node was labeled as brown star. More recent Funarioidae duplication supported by tree topology was highlighted with green circle.

**Fig. S14.** An exemplar maximum likelihood gene tree from OrthoGroup 462 contains Physcomitrella patens paralogs that survived ancestral duplications. Genes in orthogroup 462 encode homologs of VAD1 (Vascular Associated Death), a regulator of cell death and defense responses in vascular tissues. \( P. \) patens genes were highlighted with arrows and the moss-wide duplication node was labeled as brown star and the BPT duplication was labeled as a yellow star. More recent \( P. \) patens gene duplications supported by tree topology were labeled as green circles.

**Fig. S15.** GO enrichment network depicting the overrepresented categories for the orthogroups containing the moss-wide gene duplications. The node sizes were in proportion to the number of orthogroups in each of the GO categories and the adjusted p-values were depicted by the node color.

**Fig. S16.** GO enrichment network depicting the overrepresented categories for the orthogroups containing gene duplications in the ancestor of BPT (Bryopsida, Polytrichopsida and Tetraphidopsida) clade. The node sizes were in proportion to the number of orthogroups in each of the GO categories and the adjusted p-values were depicted by the node color.

**Fig. S17.** GO enrichment network depicting the overrepresented categories for the orthogroups containing gene duplications in the ancestor of Funaria and Physcomitrella. The node sizes were in proportion to the number of orthogroups in each of the GO categories and the adjusted p-values were depicted by the node color.

**Fig. S18.** GO enrichment network depicting the overrepresented categories for the orthogroups containing the Funarioidae gene duplications in the ancestor of Funaria and Physcomitrella. The node sizes were in proportion to the number of orthogroups in each of the GO categories and the adjusted p-values were depicted by the node color.

**Table S1.** Summary of plant genomes and transcriptomes included in the phylogenomic analyses.

**Table S2.** Species contributions to single-copy orthologous groups.