Molecular Phylogenetics and Evolution 98 (2016) 29–40

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta) ☆

Matthew G. Johnson [a,*], Claire Malley [b], Bernard Goffinet [c], A. Jonathan Shaw [d], Norman J. Wickett [a,b,*]

[a] Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, United States
[b] Program in Biological Sciences, Northwestern University, 2205 Tech Drive, O.T. Hogan Hall, Room 2-144, Evanston, IL 60208, United States
[c] Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Rd., Storrs, CT 06269, United States
[d] Department of Biology, Duke University, Box 90338, Durham, NC 27708, United States

ARTICLE INFO

Article history:
Received 8 October 2015
Revised 7 January 2016
Accepted 11 January 2016
Available online 23 January 2016

Keywords:
Bryophyte
Phylogenomics
Rapid radiation
Orthology

ABSTRACT

The pleurocarpous mosses (i.e., Hypnanae) are a species-rich group of land plants comprising about 6,000 species that share the development of female sex organs on short lateral branches, a derived trait within mosses. Many of the families within Hypnales, the largest order of pleurocarpous mosses, trace their origin to a rapid radiation less than 100 million years ago, just after the rise of the angiosperms. As a result, the phylogenetic resolution among families of Hypnales, necessary to test evolutionary hypotheses, has proven difficult using one or few loci. We present the first phylogenetic inference from high-throughput sequence data (transcriptome sequences) for pleurocarpous mosses. To test hypotheses of gene family evolution, we built a species tree of 21 pleurocarpous and six acrocarpous mosses using over one million sites from 659 orthologous genes. We used the species tree to investigate the genomic consequences of the shift to pleurocarpy and to identify whether patterns common to other plant radiations (gene family expansion, whole genome duplication, or changes in the molecular signatures of selection) could be observed. We found that roughly six percent of all gene families have expanded in the pleurocarpous mosses, relative to acrocarpous mosses. These gene families are enriched for several gene ontology (GO) terms, including interaction with other organisms. The increase in copy number coincident with the radiation of Hypnales suggests that a process such as whole genome duplication or a burst of small-scale duplications occurred during the diversification. In over 500 gene families we found evidence of a reduction in purifying selection. These gene families are enriched for several terms in the GO hierarchy related to "tRNA metabolic process." Our results reveal candidate genes and pathways that may be associated with the transition to pleurocarpy, illustrating the utility of phylotranscriptomics for the study of molecular evolution in non-model species.

1. Introduction

The Bryophyta (mosses) are one of three phyla of non-vascular land plants, and comprise more than 13,000 species (Magill, 2014). Although it is one of the oldest groups of land plants, with fossils dating to at least the lower Permian (Smoot and Taylor, 1986), a significant amount of genus-level diversity has been generated in bursts that are coincident with the diversification of extant ferns and angiosperms in the Mesozoic (Laenen et al., 2014). Approximately 42% of moss species diversity (Crosby et al., 1999) belong to the pleurocarpous mosses or Hypnanae (Buck et al., 2005) a monophyletic "crown group" of mosses typically defined by the development of female gametangia on short, lateral branches lacking differentiated vegetative leaves (La Farge-England, 1996). This is in contrast to the ancestral growth form of mosses, acrocarpy and cladocarpy, where gametangia (and thus sporophytes) are usually terminal on upright shoots, or branches bearing well developed leaves. Recent fossil discoveries have pushed the origin of the pleurocarpous growth form into the late Permian (de Souza et al., 2012), but the major diversification of extant pleurocarpous moss lineages occurred much later; in fossil-calibrated phylogenies, the most speciose pleurocarpous moss families (i.e., Amblystegiaceae, Hypnaceae, or Brachytheciaceae) are estimated to have diversified within the last 100 million years (Laenen et al., 2014; Newton et al., 2007; Shaw et al., 2003).

☆ This paper was edited by the Associate Editor Stefanie M. Ickert-Bond.

* Corresponding authors at: Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, IL 60022, United States (N.J. Wickett).
E-mail addresses: mjohnson@chicagobotanic.org (M.G. Johnson), nwickett@chicagobotanic.org (N.J. Wickett).

http://dx.doi.org/10.1016/j.ympev.2016.01.008
1055-7903/

The superorder Hypnanae comprises four orders: Hypnodendrales, Ptychomniales, Hookeriales, and Hypnales, with the latter including most of the diversity, namely ±4000 species in about 430 genera and 42 families (Goffinet et al., 2009; Huttunen et al., 2013). Just eight families hold the majority of genera, 5% of the genera hold the majority of species and nearly 200 genera are monospecific. Suprageneric taxa within Hypnanae were traditionally circumscribed using morphological traits and habitat type. Given the uneven distribution of species among genera and families and the rapid tempo of diversification in Hypnales, it is not surprising that most of these morphologically-defined taxa are not monophyletic (Cox et al., 2010). Lineage-through-time plots revealed that Hypnales, unlike its sister group (Hookeriales), underwent a rapid, explosive diversification rather than a gradual diversification early in its history (Shaw et al., 2003). Many of the families within Hypnales likely diversified after the radiation of angiosperm forests (Laenen et al., 2014; Newton et al., 2007), which is hypothesized to have been a major driver of diversification in other epiphytic and understory-dwelling plants, such as leptosporangiate ferns and liverworts (Feldberg et al., 2014; Schneider et al., 2004). The diversification of leptosporangiate ferns parallels that of major families in Hypnales in both timing (late Cretaceous) and extant diversity in similar habitats (e.g., low-light, epiphytic; (Li et al., 2014)), which raises the possibility that a signature of the radiation can be found in the genomes of pleurocarpous mosses.

Recent genomic evidence has revealed that whole genome duplication (WGD) events are associated with many gene family expansions in land plants (Barker et al., 2008; Blanc and Wolfe, 2004a; Jiao and Paterson, 2014; Jiao et al., 2011; Li et al., 2015; Maere et al., 2005). Although most of the gene copies generated by WGD events are lost due to fractionation and subsequent "red iploidization" or nonfunctionalization (Jiao et al., 2011), in *Arabidopsis* some of the duplications led to neo- and/or sub-functionalization (Moore and Purugganan, 2005), resulting in the evolution of parallel gene networks (Blanc and Wolfe, 2004a) with copy-specific gene regulation (Spangler et al., 2012). Paleopolyploidy events in plants may provide opportunities for positive selection (or, at the very least reduced purifying selection) on retained duplicated copies of genes. Gene duplications in some gene families have been associated with key innovations in angiosperms; for example, the expansion of the CYCLOIDEA gene family is linked to the development of zygomorphic flowers. Members of the CYCLOIDEA gene family that have been implicated in symmetry contain conserved domains with sites under positive selection following duplication (e.g. Chapman et al., 2008).

The identification of genomic signatures, such as WGD, associated with radiations in non-model organisms has benefited from the emergence of sequencing techniques that reduce genomic complexity, such as transcriptome sequencing (Cannon et al., 2015; Yang et al., 2015). These methods allow researchers to generate large, comparative data sets without having to invest considerable resources in sequencing entire genomes, which can vary enormously in size. By using only the coding portion of the genome in a phylotranscriptomic approach, it is possible to provide an evolutionary context to inferred paleopolyploidy, gene family expansion, and shifts in selective regimes of duplicated genes, without having to sequence the extensive non-coding portion of a genome. To date, all phylogenetic analyses involving pleurocarpous mosses have sampled few discrete nuclear coding loci and focused on dense taxon sampling (Cox et al., 2010; Huttunen et al., 2012; Shaw et al., 2003). In contrast, transcriptome sequencing generates enough data to leverage methods that have the potential to identify genomic signatures, such as evidence of WGD, associated with the radiation of Hypnales, as has been observed in other land plant radiations.

Here, we identify homologous and orthologous genes from sequenced transcriptomes using a phylogeny-based approach (Yang and Smith, 2014). We construct a species tree from over 650 nuclear, protein-coding loci for 21 pleurocarpous mosses, five acrocarpous mosses, and the well-annotated proteome of the model moss *Physcomitrella patens*. Our sampling includes *Aulacomnium palustre*, an exemplar of the Aulacomniales resolved by Bell et al. (2007) as the sister-group to the Hypnanae. We use the species tree to infer the history of gene duplication events, gene family expansions, signatures of whole genome duplications, and shifts in rates of selection. Functional analysis of gene families that underwent expansion during the diversification of Hypnales will inform future studies on the genetic basis of the radiation of this most speciose lineage of pleurocarpous mosses.

2. Materials and methods

2.1. RNA extraction and sequencing

We generated 25 transcriptomes for this study from 21 pleurocarpous and four acrocarpous moss species (Table 1). Two of our samples are from the same species, *Aulacomnium palustre*, which was sequenced twice due to its critical phylogenetic position. All samples were wild-collected and then placed in clear plastic containers within a growth chamber for at least one week prior to RNA extraction. Tissue was sampled from young, green shoots and flash-frozen in liquid nitrogen, then ground to a powder with a mortar and pestle. Total RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma–Aldrich, St. Louis, MO, USA) with no modification to the standard protocol. RNA was quantified using a Qubit Fluorometer (Life Technologies, Grand Island, NY, USA). RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc, Santa Clara, CA, USA) at the Northwestern University Center for Genetic Medicine (Chicago, IL, USA). RNA-Seq libraries for Illumina sequencing were prepared at BGI (Shenzhen, China) using the TruSeq RNA Sample Preparation Kit v2 (Illumina Inc., San Diego CA, USA). Three of the libraries (NW_1, NW_2, and NW_3, see Table 1) were multiplexed and sequenced on one lane of Illumina HiSeq2000 (2×100 Paired End) in February, 2014. A second batch of ten samples (accession numbers NW_45 through NW_62, see Table 1) was multiplexed and sequenced across two lanes in July, 2014, and the remaining fourteen samples were multiplexed and sequenced across two lanes in November, 2014. All sequencing was carried out at BGI (Shenzhen, China). All unedited sequence reads were deposited in the NCBI Sequence Reads Archive (BioProject PRJNA296787).

2.2. Transcriptome assembly and filtering

Raw reads were demultiplexed and adapters were removed by BGI (Shenzhen, China) prior to delivering the data. We further trimmed the sequences with Trimmomatic (Bolger et al., 2014) using the following parameters: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN: 36. We assembled each transcriptome from the filtered reads with the Trinity pipeline (Grabherr et al., 2011; Haas et al., 2013) using default parameters. We applied a hierarchal filtering approach in order to reduce the complexity of downstream analyses, and to reduce the possibility of contamination of our datasets by transcripts from associated organisms. First, the transcripts were translated using the Transdecoder (version 20140704, http://transdecoder.github.io) tool included with Trinity. Transdecoder chooses the most valid open reading frames from each transcript using a likelihood approach that incorporates domain similarity matches to the Pfam database (Finn et al., 2014) using the hmmscan function in HMMER (Johnson et al.,

Table 1
Transcriptome assembly statistics and herbarium voucher information. Buck–NY, Goffinet, Quandt–CONN, Shaw–DUKE.

ID	Species	Voucher collection	Paired reads (Millions)	Trinity transcripts (thousands)	Percent of transcripts passing filter (BLAST and Transdecoder) (%)	Masked transcripts kept	Homologous gene family trees	BUSCOs
NW-1	*Climacium americanum*	Goffinet 11684	58.86	129.4	33.7	20,018	12,295	390
NW-2	*Thuidium delicatulum*	Goffinet 11686	67.17	137.6	46.1	20,289	12,501	391
NW-3	*Hypnum cupressiforme*	Goffinet 11687	72.38	235.6	31.7	22,035	12,735	392
NW-45	*Pleurozium schreberi*	Goffinet 11700	23.11	108.1	37.3	18,098	11,965	388
NW-51	*Aulacomnium palustre*	Goffinet 11701	38.23	135.4	40.9	19,894	12,313	390
NW-53	*Bryoandersonia illecebra*	Buck 63016	34.75	111.8	41.8	20,891	12,702	392
NW-55	*Rhytidiadelphus subpinnatus.*	Goffinet 11708	34.76	82.4	48.0	18,490	12,068	391
NW-56	*Kindbergia praelonga*	Goffinet 11707	31.73	132.8	34.5	21,146	12,528	387
NW-57	*Callicladium haldanianum*	Goffinet 11688	25.38	169.4	33.6	19,332	11,991	381
NW-59	*Hylocomium brevirostre*	Goffinet 11699	30.56	211.8	31.1	25,976	13,205	386
NW-60	*Hylocomium splendens*	Goffinet 11696	26.69	147.6	35.3	20,902	12,375	391
NW-61	*Calliergon cordifolium*	Goffinet 11693	29.07	300.3	27.0	21,494	12,417	389
NW-62	*Aulacomnium palustre*	Goffinet 11702	23.79	147.0	39.3	20,489	12,335	390
NW-65	*Anomodon rostratus*	Goffinet 11711	12.88	115.3	35.8	19,720	12,253	390
NW-66	*Thelia asprella*	Goffinet 11712	22.01	135.6	37.6	19,994	12,421	391
NW-69	*Pilotrichella flexillis*	Quandt DR158/ WP281	22.91	128.0	37.5	20,181	12,535	391
NW-71	*Antitrichia curtipendula*	Shaw 17555	21.87	101.5	44.2	20,736	12,481	385
NW-72	*Meteoridium remotifolium*	Quandt DR152/ WP281	22.78	102.5	42.0	18,649	12,066	387
NW-74	*Rhytidiopsis robusta*	Shaw 17554	19.18	113.1	37.8	20,312	12,453	386
NW-76	*Forrstroemia trichomitria*	Shaw 17557	21.27	113.8	40.9	19,161	12,365	389
NW-77	*Rhodobryum ontariense*	Goffinet 11803	13.34	53.3	60.6	11,296	9470	382
NW-79	*Dicranum scoparium*	Goffinet 11775	20.08	120.7	37.1	13,577	9843	383
NW-84	*Platyhypnidium riparioides*	Goffinet 11802	19.28	100.4	39.1	18,969	12,011	385
NW-85	*Plagiothecium laetum*	Goffinet 11851	17.19	144.7	29.5	20,620	12,560	382
NW-86	*Leucobryum glaucum*	Goffinet 11773	20.64	171.5	30.3	12,566	9450	386

2010). All transcripts with a valid translation were searched against a custom BLAST database containing protein sequences from 22 land plant nuclear genomes, including the moss *Physcomitrella patens*, downloaded from Phytozome (phytozome.jgi.doe.gov). We accepted protein matches in blastp (version 2.2.29) with an e-value below 10^{-10}. For the next stages of analysis, we included all transcripts that had a significant hit to the proteome database, as well as all Trinity-annotated isoforms of that same transcript.

2.3. Clustering transcripts into homologous gene families

In order to cluster transcripts from all species, remove redundant isoforms, and construct a species tree from low-copy genes, we employed the phylogenetic clustering method described by Yang and Smith (2014). In this pipeline (hereafter, the Yang/Smith Pipeline), all transcripts that passed the above filtering procedures were first grouped using an all-vs-all BLAST search of nucleotide sequences from every species. Significant hits (e-value < 10^{-5}) were clustered, using the software MCL (Enright et al., 2002), into gene families using a hit-fraction of 0.3 and an inflation parameter of 2.0. All clusters containing sequences from at least four species (of 26 total, including the *Physcomitrella* proteome) were aligned using MAFFT (Katoh and Standley, 2013) and nucleotide gene trees were reconstructed from peptide sequences with RAxML (Stamatakis, 2014) under the CAT model.

At this stage in the pipeline, the gene clusters may include isoforms as inferred by Trinity. To account for the possibility that some of these may actually be paralogs, rather than alternative splice forms, the Yang/Smith Pipeline extracts monophyletic or paraphyletic groupings of transcripts from a single taxon. These clades are reduced to contain only the sequence with the longest

unambiguous alignment. We also trimmed the gene family trees to remove terminal branches whose length exceeded an absolute (0.3 substitutions/site) or relative (more than 10× longer than its sister branch) cutoff. These terminal branches represent transcripts with potentially spurious homology to the other transcripts in the gene family tree, and were removed from further analysis. This approach retains multiple isoforms from the same Trinity component if they are not part of the same clade on the gene tree; however, recent lineage-specific paralogs will be lost. Since the goal of this project is to identify genomic changes prior to, or coincident with the diversification of Hypnales, rather than species-specific changes, the masking of lineage-specific duplications does not impact our conclusions. We refer to the subset of transcripts that remain following this phylogenetic transcript clustering as the **masked** dataset.

When using transcriptomes for gene discovery, rather than quantifying relative expression, it is important to assess whether or not it is likely that sequences for all possible transcripts were recovered. One method to approximate whether or not we sequenced all possible transcripts that were present in the tissues we sequenced is to determine how well we have recovered a core set of genes for our taxonomic group. Although no such set is known for mosses (or even land plants), a set of core genes is defined for all eukaryotes. The Basic Universal Single Copy Orthologs (BUSCOs) are curated from all metazoan and fungal genomes, and maintained as a set of profile Hidden Markov Models (HMMs), and an inferred ancestral amino acid sequence for each orthogroup is provided (Simão et al., 2015). To determine whether our masked dataset contained BUSCOs, we first searched the translated amino acids from our masked dataset against the ancestral sequences using BLASTP. Sequences with hits were then searched against the 429 BUSCO profile HMMs using HMMER. In order to accept a match, the hmmsearch score had to exceed a minimum score threshold defined for each BUSCO.

2.4. Species tree reconstruction

Many methods of species tree reconstruction rely on the identification of orthologous sequences, that is, sequences that arose by speciation rather than duplication. The Yang/Smith Pipeline identifies sets of orthologous genes (orthogroups) by decomposing the unrooted homologous gene family trees into subtrees where a monophyletic outgroup (here, acrocarpous mosses) is sister to a monophyletic ingroup (pleurocarpous mosses). The extracted subtrees represent inferred orthologous gene families, i.e. gene families for which the most recent common ancestor (the ancestral node) underwent a speciation event and not a duplication event. More than one of these orthologous subtrees may appear within a homologous gene tree if, for example, a gene duplication occurred prior to the divergence of the ingroup and outgroup.

We further filtered the orthologous gene trees by requiring that each species be represented by exactly one transcript, and refer to this subset as the **one-to-one orthologs.** The final matrix for species tree reconstruction represented 659 orthogroups where all 26 species are represented, with a total alignment length of 361,745 amino acid residues. Transdecoder produces an amino acid file and a coding domain sequence (CDS; nucleotides) for each putative protein. We aligned the proteins from each orthogroup with MAFFT, and back-translated the sequences using the corresponding nucleotide sequences using TrimAl (Capella-Gutierrez et al., 2009). We concatenated all of the coding regions into a supermatrix using phyutility (Smith and Dunn, 2008). This matrix comprised 659 genes and 1,062,897 nucleotides, with all 26 species represented for each gene. We reconstructed the species tree in RAxML using the GTRGAMMA model using two partitions per gene (one partition for the first and second codon positions, and another partition for the third). We evaluated nodal support using 200 bootstrap replicates.

We also reconstructed the Maximum Quartet Support Species Tree (MQSST) using ASTRAL (Mirarab et al., 2014). Individual gene trees were reconstructed using RAxML with the GTRGAMMA model, including a single maximum likelihood tree as well as 200 "fast bootstrap" trees. We evaluated support on the ASTRAL trees with a "gene-wise jackknife method." We generated 200 pseudoreplicates of the dataset by sampling 10% of the maximum likelihood gene trees without replacement and calculated a MQSST tree using ASTRAL on each subset.

We also repeated both the supermatrix and MQSST reconstruction methods using the corresponding amino acid alignment. The PROTGAMMA models were used for the supermatrix and individual gene tree reconstructions in RAxML.

2.5. Gene family expansion in Hypnales

We generated a table of gene family occupancy by counting the number of transcripts present for each species in each of the 27,299 homolog gene family trees generated by the Yang/Smith Pipeline. Unlike the one-to-one ortholog set of gene trees used for phylogenetic reconstruction, each gene homolog gene family can contain many transcripts from each species. To track the phylogenetic history of these gene families and identify expansions, we used the program Count (Csurös, 2010) to reconstruct ancestral states. Count uses Wagner parsimony (with a 20% penalty for gains) to identify nodes where: (a) a gene family has more than one member and (b) the gene family has exactly one member at the immediately ancestral node.

To identify gene family expansions associated with the diversification of Hypnales (the largest order of pleurocarpous mosses), we were interested in gene family expansions reconstructed at the following nodes (Fig. 1): (A) the common ancestor of all Hypnales minus *Plagiothecium* (a genus revealed to be sister to the rest of Hypnales), (B) the common ancestor of all Hypnales (including *Plagiothecium*), (C) the common ancestor of Hypnales and *Aulacomnium*, and (D) the common ancestor of the Bryidae (includes Hypnales, *Aulacomnium*, and *Rhodobryum*). In order to assign GO annotations to transcripts of non-model organisms, we used the annotation pipeline Trinotate (trinotate.github.io). Briefly, the pipeline searches transcripts (and their protein translations) against curated functional annotation databases, including Pfam and SwissProt. We assigned GO annotations to each homologous gene family cluster by recording all GO annotations from Trinotate made to each transcript in the cluster.

We performed a gene ontology enrichment analysis using the orthogroups with inferred expansions on one of the four nodes of interest, using the Python package goatools (version 0.5.4, github.com/tanghaibao/goatools). All GO categories annotated for all 15,459 homologous gene family clusters was used as the baseline, and we controlled for multiple testing using the False Discovery Rate method (Benjamini and Hochberg, 1995).

2.6. Evidence of paleopolyploidy

Whole genome duplication (WGD) events can be detected from transcriptome data by finding pairs of paralogous sequences within the transcriptome (Barker et al., 2009; Blanc and Wolfe, 2004b; Yang et al., 2015). The synonymous substitution rate (Ks) is calculated from each pair of paralogs. In principle, the distribution of Ks values should approximate an exponential distribution, reflecting the age distribution of gene duplication events—many pairs of genes with low Ks values, and fewer pairs with larger Ks values. This distribution could arise from many, ongoing small-scale duplications occurring throughout the history of the lineage,

followed by the nonfunctionalization of one duplicate. However, a WGD event would result in a very large number of paralogous pairs all having the same age. If a histogram of Ks values among pairs of parologs in a transcriptome has multiple peaks at intermediate values of Ks, it may be evidence of a paleopolyploidy event.

We could not analyze the transcriptome data for the presence of recent paralogs using the masked dataset, which we used for phylogenetic analysis and ortholog detection. Masking removes recent paralogs, which would bias the estimation of Ks from paralog pairs. We also could not use the raw output from Trinity, which produces transcripts with names such as "c1250_g1_i1." The first field refers to a "component" of the DeBruijn graph, the second field to a "gene" identifier, and the third field to a putative "isoforms" for the transcript. However, these isoforms may not correspond to real alternative splice variants, but may also contain parologous gene sequences. Therefore, rather than keeping only the longest isoform, or the isoform with the highest coverage for each gene component, we used CD-Hit-EST (Fu et al., 2012) to cluster the transcripts with a high percent identity threshold (−c option, 98%) and high alignment overlap threshold (−aS option, 90%) to reduce the isoforms to a set of non-redundant transcripts for each species.

To detect ancient parology, we began with the protein sequences that matched the non-redundant transcript set from CD-Hit-EST cluster for each species separately. We clustered these protein sequences for each species again with CD-Hit, but with much lower thresholds for percent identity (−c 0.4) and alignment overlap (−aS 0.75) to maximize cluster inclusiveness. For each cluster that was not a singleton, we constructed pairwise amino acid alignments among all proteins in the cluster using MAFFT. The corresponding nucleotide transcript sequences were forced into the amino acid alignments using pal2nal (Suyama et al., 2006), and all gap regions and internal stop codons were removed.

For each pair of paralogous nucleotide sequences, we calculated the synonymous substitution rate (Ks) with KaKs_Calculator (Zhang et al., 2006), using the "GY" method (Goldman and Yang, 1994), also known as F3x4. We investigated the presence of multiple normal distributions of Ks values using mixture models, implemented in the R package mclust (Fraley et al., 2012). We evaluated mixture models with between one and ten components, and the best fit model was chosen using the Bayesian Information Criterion (BIC).

2.7. Detecting changes in selection

We investigated the effect of the Hypnales radiation on signatures of molecular selection using the codeml package implemented in PAML (version 4.7; (Yang, 2007). For this analysis we used a subset of orthogroups where (1) all six acrocarpous mosses were present and (2) at least ten pleurocarpous mosses were present. We aligned protein sequences with MAFFT and back-translated using the corresponding CDS sequences using trimAL (version 1.4.rev15 Capella-Gutierrez et al., 2009), which also removed codons in the sequence matrix if they were present in fewer than five sequences. We calculated a tree for each orthogroup using FastTree (Price et al., 2010). The common ancestor of pleurocarpous moss sequences in each orthogroup was determined using the Python package ETE2 (Huerta-Cepas et al., 2010), which also assisted in running codeml. All branches descending from the common ancestor of pleurocarp sequences were marked as "foreground" for branch-model analysis. We estimated the ratio of non-synonymous to synonymous nucleotide substitution rate (dN/dS or omega) under two models: in the "M0" model, all branches are assumed to have the same omega, but in the "bfree" model, separate omegas are estimated for the "foreground" and "background" branches. Because the M0 model is nested within the bfree model, the significance of the bfree model can be determined with a Likelihood Ratio Test (LRT) with one degree of freedom. We tested the significance of the LRT against a chi-squared distribution, and accounted for multiple tests by accepting p-values less than 0.0001.

For genes with evidence of different rates of evolution in pleurocarps and acrocarps, we conducted a GO Enrichment analysis using the same procedure as above. We summarized the GO Enrichment results using ReviGO (Supek et al., 2011), which produces a visualization of the semantic similarity among GO categories.

A conceptual diagram illustrating our entire analysis can be found in Supplemental Fig. 1.

3. Results and discussion

3.1. Transcriptome assembly and orthology detection

Across our 25 assembled transcriptomes, we recovered between 53,000 and 301,000 transcripts (Table 1). After applying our filtering steps, assuring that the transcript contained a valid protein using Transdecoder and that the protein had BLAST hits against known land plant proteomes, we retained between 32,349 and 81,018 protein coding sequences per species. The Yang/Smith Pipeline then clustered these transcripts from all 25 transcriptomes and the *Physcomitrella patens* proteome into 27,299 homologous gene families that contained transcripts from at least four different species. We then produced the "masked" dataset by removing sequences if transcripts from the same species form monophyletic or paraphyletic groups on the multispecies gene family trees. On average, 19,393 translated proteins were retained (Table 1), and each transcriptome had at least one sequence in an average of 12,054 homologous gene families. Following the filtering and clustering steps, we retained between 12,566 and 25,976 transcripts within 27,299 homolog family trees (Table 1).

To approximate whether we sampled the pool of all possible transcripts as deeply as possible, we searched for the 429 BUSCOs (universal orthologous genes) defined for all eukaryotes (Simão et al., 2015). We were able to detect between 382 and 391 BUSCOs in our 25 transcriptomes. For comparison, we could find 391 BUSCOs in the *Physcomitrella* proteome, suggesting that this is the maximum number for mosses, and that the remaining 38 orthogroups are not universal to all eukaryotes if plants are included. The authors of the BUSCOs pipeline have begun development of a more plant-specific set of universal orthologs (buscos.ezlab.org). We therefore accept 391 as the maximum number of eukaryote BUSCOs that can be found in mosses.

Because we used different multiplexing schemes, we also tested whether the number of reads correlated with our assessment metrics, but the number of reads per sample did not correlate with the number of Trinity transcripts ($F_{23} = 3.2$, $r^2 = 0.08$, $p = 0.09$) or the number of proteins retained in the masked dataset ($F_{23} = 3.1$, $r^2 = 0.08$, $p = 0.09$). The correlation between the number of reads and the number of BUSCOs recovered was significant ($F_{23} = 8.0$, $r^2 = 0.23$, $p = 0.01$). On average, we recovered three additional BUSCOs from the eight samples with more than 30 million reads, compared to the seventeen samples with fewer than 30 million reads. Overall, this suggests that additional sequencing of each species would not change the inferences made here.

We also assessed the completeness of our transcriptomes by comparison to the well-curated proteome of *Physcomitrella patens*. Although 21,312 proteins from *Physcomitrella* clustered with transcripts from at least four of our transcriptomes, only 7778 gene families in the masked data set contain a protein from the model moss proteome. This is likely due to the masking step of the Yang/Smith pipeline, which removes sequences if they form a

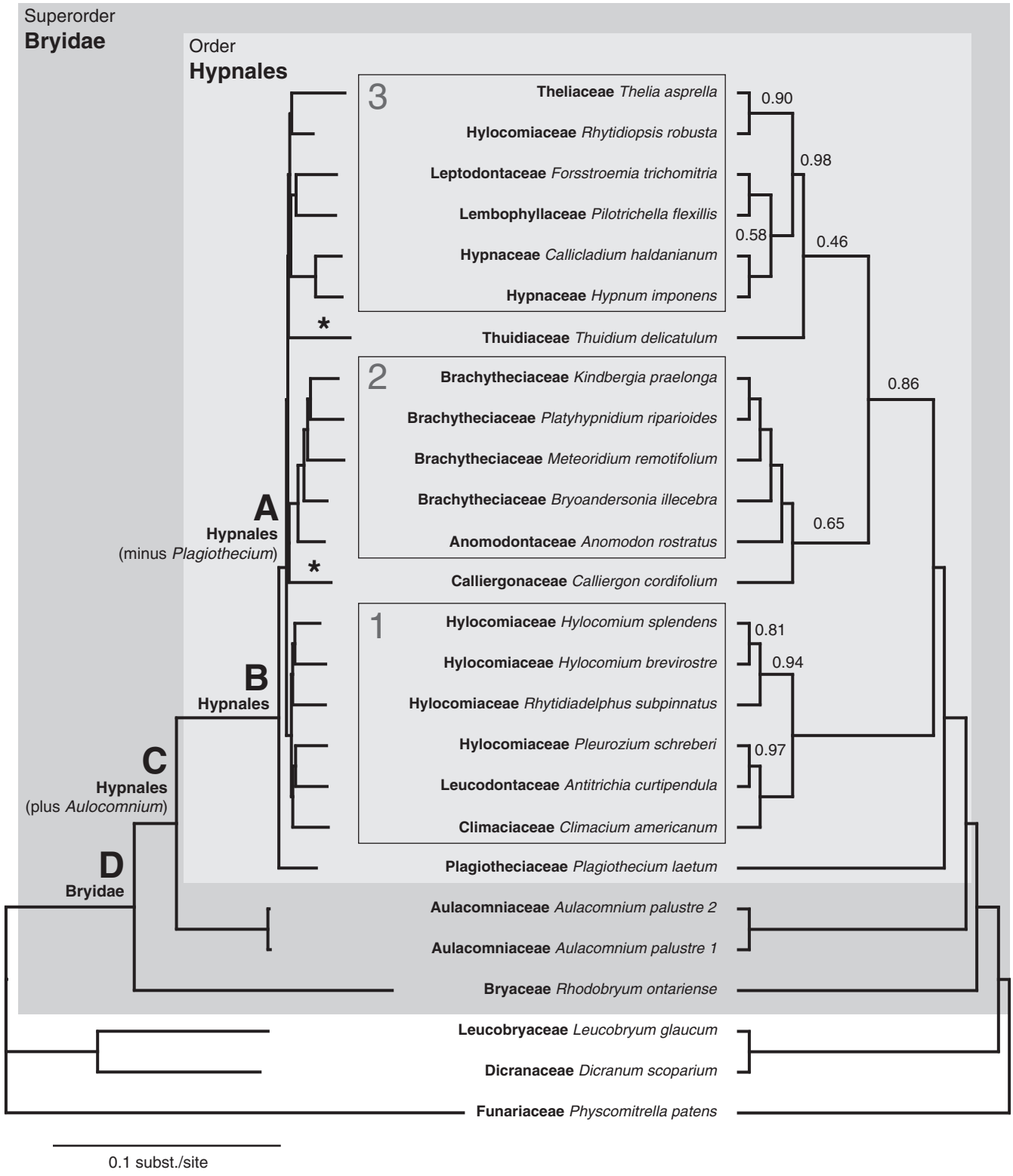

Fig. 1. Species trees constructed from 659 orthogroups in 21 pleurocarpous and five acrocarpous mosses, and the number of homologous gene family expansions at key nodes on the tree. Left: Maximum-likelihood nucleotide tree from RAxML. All nodes are supported at 100% bootstrap support (200 pseudoreplicates) except for the nodes indicated by stars. The letters indicate four key nodes at which gene family expansions were calculated: A: Hypnales minus *Plagiothecium*, B: Hypnales, C: Hypnales mosses plus *Aulacomnium*, D: Superorder Bryidae. Right: Maximum Quartet Support Species Tree from ASTRAL, reconstructed from nucleotide gene trees estimated in RAxML. Nodal support is 100% except where indicated by genewise-jackknife.

monophyletic group that includes only sequences from the same taxon. As a result of the whole genome triplication event that occurred during the diversification of Funariaceae (Rensing et al., 2007), many *Physcomitrella* paralogs clustered together in our gene family trees at a 3–1 ratio relative to other mosses. On the initial gene family trees, the *Physcomitrella* paralogs would share a common ancestor within *Physcomitrella*, and the masking step would retain only one of these paralogs per gene family.

There are also a large percentage of homolog gene families (over 19,000) that do not contain a representative from *Physcomitrella*, but do contain transcripts from at least four different transcriptomes. The transcriptome sequences could only progress to this point in our pipeline if they had a significant (*e*-value < 10^{-5}) BLAST hit to a land plant proteome. Therefore, some of these gene families may be ancestral land plant families that have since been lost in the lineage leading to *Physcomitrella*. Alternatively, because Yang/Smith Pipeline is intended primarily as a way of identifying low-copy orthologs for phylogenetic inference, gene families may be circumscribed narrowly. Higher-order clustering may reveal homology between gene families. Additional genome sequences from other bryophytes would assist in this effort.

3.2. Species tree reconstruction

We constructed species trees from 659 orthogroups that had a one-to-one orthology between acrocarpous and pleurocarpous mosses, as well as a representative sequence from all 26 taxa. Both the Maximum Likelihood (ML) Tree and the (Maximum Quartet Support Species Tree (MQSST) have strong backbone support, reflecting relationships among acrocarpous mosses, and their relationships to Hypnales (Fig. 1). When rooted using *Physcomitrella patens*, *Aulacomnium palustre* (both accessions) is sister to the pleurocarpous mosses (Hypnales), as expected (Bell et al., 2007; Cox et al., 2010). Sister to this clade is *Rhodobryum*; together with *Aulacomnium* and Hypnales these species represent the subclass Bryidae (Goffinet et al., 2009; Stech and Frey, 2008). Previous phylogenetic evidence supports Bryidae as sister to Dicranidae (Chang and Graham, 2011; Cox et al., 2010), which in our dataset is represented by *Leucobryum* and *Dicranum*.

Within the pleurocarpous mosses, *Plagiothecium* is resolved as sister to the rest of Hypnales with maximum support in both tree reconstruction approaches (Fig. 1), consistent with previous efforts using one or a few genes (Cox et al., 2010; Huttunen et al., 2012; Merget and Wolf, 2010). Two other multi-species relationships are maximally supported using all methods. Clade 1 contains four of our five samples from the Hylocomiaceae plus *Climacium* (Climaciaceae) and *Antitrichia* (Leucodontaceae), and is resolved as sister to the remaining Hypnales (minus *Plagiothecium*) with maximal support in both the RAXML and ASTRAL trees. Clade 2 contains four Brachytheciaceae, which compose a monophyletic sister lineage to *Anomodon* (Anomodontaceae), and relationships within this clade are fully resolved in both trees. Clade 3 contains the remaining species of Hylocomiaceae (Rhytidiopsis) along with two species of Hypnaceae, Forsstroemia (Leptodontaceae), Pilotrichella (Lembophyllaceae), and Thelia (Theliaceae). Rhytidiopsis has been accommodated in the Hylocomiaceae (Buck and Vitt, 1986) but affinities to Thelia had first been proposed by Chiang and Schaal (2000), and then by Huttunen et al. (2012). These three well-resolved clades are consistent with those recovered from inferences from discrete loci from all genomic compartments (Huttunen et al., 2012).

Two branches on the maximum likelihood tree did not receive 100% support. Inspection of bootstrap trees reveals that full phylogenetic resolution within Hypnales is impeded by the positions of two species: *Thuidium delicatulum* and *Calliergon cordifolium* (Supplemental Fig. 2). Lineage movement analysis of the bootstrap replicates revealed the two species are as likely to be sister-species (25%) as they are in their maximum likelihood arrangement (26%). Likewise, the placement of *Thuidium* sister to Clade 3 (46%) and *Calliergon* as sister to Clade 2 (65%) have the lowest gene-wise jackknife values of any relationship on the ASTRAL tree. We expect that denser taxon sampling of transcriptomes within Hypnales and the less species-rich orders of pleurocarpous mosses would allow us to more confidently reconstruct the affinities of *Thuidium* and *Calliergon*. However, while we are unable to completely resolve relationships in these clades, the focus of this study is to reconstruct the history of gene family evolution at higher phylogenetic levels.

The backbone of the phylogeny was reconstructed with equal confidence using amino acid characters (Supplemental Fig. 3). The resolution of clades within Hypnales was less certain; Clades 1 and 2 were fully supported using both RAxML and ASTRAL, but the support for Clade 3 was similarly reduced by the placement of *Thuidium* and *Calliergon*. The relationships among the three major clades within Hypnales were unresolved with both methods using the amino acid matrix.

Our results suggest that intra-genomic phylogenetic conflict complicates the resolution of family-level relationships within Hypnales. Earlier studies, which focused effort on taxon sampling, with few genes, had similar difficulty resolving the same relationships (Cox et al., 2010; Huttunen et al., 2012). Although we have employed a phylotranscriptomic approach, it is likely that the relationships within Hypnales may not be resolved without a broader taxonomic sampling. Specifically, several major families of Hypnales were not sampled in our phylogeny, and the addition of transcriptomes from the other orders of pleurocarpous mosses (e.g. Hookeriales) would root the Hypnales phylogeny more accurately. Though not suitable for taxonomic revision, our data do present a large step forward in the genomic sampling effort, increasing the number of genes sequenced in mosses by two orders of magnitude over previous studies. We anticipate that the phylogenetic framework presented here, particularly with respect to the identification of orthologous gene families, will provide a foundation for more taxon-dense phylogenetic studies in the future.

3.3. Gene family expansion analysis

Despite some instances of low phylogenetic resolution within Hypnales, the strong support among backbone clades of the mosses enables us to infer patterns of gene family evolution. Because we are using transcriptomes, gene family membership may be reduced due to incomplete expression of the proteome. However, we can treat the number of distinct transcripts per species (an approximation of gene copy number, not relative expression) as a discrete trait that evolves along the species tree. By reconstructing the "gene family occupancy" for each gene family at each node on the species tree, we minimize the noise associated with gene loss and under-expression in terminal taxa. Specifically, we identified expansions in homologous gene families at four robustly supported (100% in all methods), nested nodes, marked in Fig. 1: (A) Hypnales minus *Plagiothecium*, (B) Hypnales, (C) Hypnales plus *Aulacomnium*, and (D) Bryidae.

We reconstructed gene family occupancy using a Wagner parsimony approach in the software Count (Csurös, 2010). Expansions were identified by two criteria at each of the four target nodes: the gene family had to be reconstructed as (1) containing multiple paralogs at the target node and (2) exactly one paralog at the immediately ancestral node. Gene family contractions were identified by inverse criteria. Our analysis revealed that homologous gene family expansions at the four target nodes were extremely enhanced compared to homologous gene family contractions (Fig. 2A, Supplemental Table 2). The highest discrepancy between expansions and contractions occurred at the Hypnales + *Aulacomnium* (C) node, with 799 expansions and only 11 contractions.

Across all four target nodes, 1712 homologous gene families exhibited low ancestral occupancy and high occupancy in Hypnales. Sixty-one GO categories (Fig. 2B, Supplemental Table 1) were enriched among the homologous gene families that had expanded in Hypnales. The enriched categories included "post-embryonic morphogenesis" (GO:0009986) and "developmental process

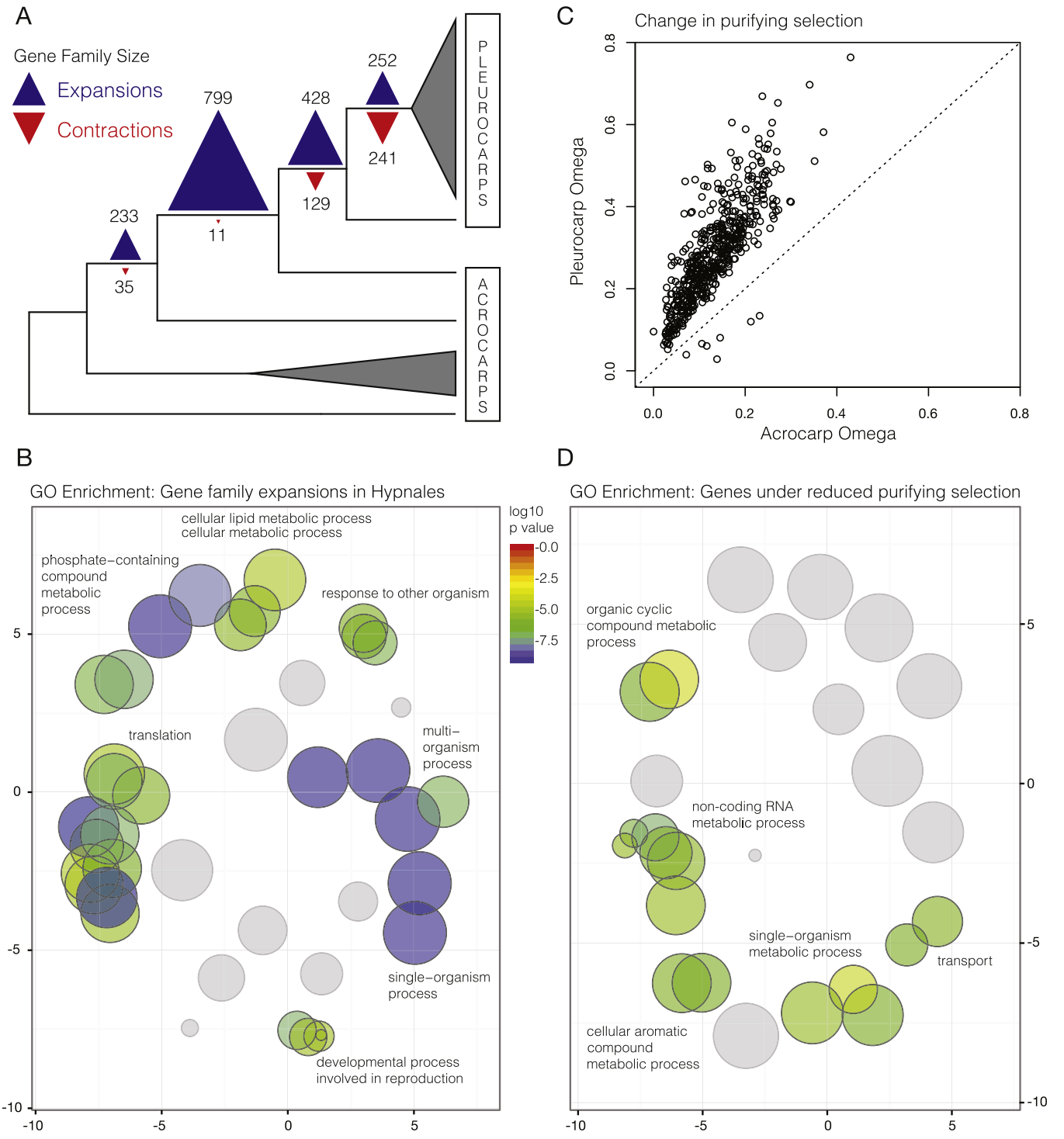

Fig. 2. Gene family expansion in pleurocarpous mosses is associated with enriched gene ontology (GO) categories and reduced purifying selection. (A) Expansions and contractions in homologous gene families reconstructed by Wagner parsimony at four nested nodes that represent common ancestors of the pleurocarpous mosses. (B) Multidimensional scaling of semantic similarity among GO categories for gene families that have expanded in pleurocarpous mosses. Two GO terms (circles) have similar semantic similarity if they are siblings in the GO heirarchy or are related by inheritance. Overlapping circles of terms share similar characteristics and are labeled in the figure with one representative GO term. (C) Comparison of omega values (ratio of non-synonymous to synonymous substitution rate) in 526 orthogroups where a two-omega model was preferred to a single omega for the tree. The remaining 2220 gene family trees for which the two-omega model was not preferred are not shown. Each point represents a gene tree, where the horizontal axis is the omega inferred from the "background" branches (acrocarpous mosses), and the vertical axis is the omega inferred from the "foreground" branches (pleurocarpous mosses). The dotted line represents equivalent omegas in the two sets of branches; only seven genes have a lower inferred omega in pleurocarpous mosses, compared to acrocarpous mosses. An omega over 1.0 represents evidence of positive selection, but an increased omega in the foreground branches may represent reduced purifying selection. (D) Semantic similarity plot of GO terms that are enriched in the set of gene families shown to have reduced purifying selection. Overlapping sets of GO categories are represented by one GO term. For a full list of all enriched categories, see the supplemental information.

involved in reproduction" (GO:0003006). Of particular interest were several categories involved in interactions among organisms, such as "response to external biotic stimulus" (GO:0043207) and "defense response to other organism" (GO:0098542).

These functional annotations are intriguing because the shared derived growth form of pleurocarpous mosses (a shift in reproductive structure from terminal to lateral, and a generally more prostrate and branching growth form) may be associated with traits that later facilitated the rapid radiation of Hypnales. Because most of the major families in Hypnales diversified concurrent with the diversification of other plants (Laenen et al., 2014), the opportunity for new biotic interactions may have arisen. For example, the diversification of angiosperms may have presented mosses with a novel substrate, driving neofunctionalization of genes responsible for external stimuli and defense responses. It is clear from our results that many gene families have expanded coincident with the radiation of the pleurocarpous mosses. However, it is unclear when the gene families expanded during the evolution of pleurocarpous mosses, because our sampling was limited to Hypnales. Future studies could identify whether many of the gene families instead expanded in the shared ancestor of Hypnales and Hookeriales (the other major order of pleurocarpous mosses), or perhaps prior to the divergence of earlier pleurocarpous lineages, namely the Hypnodendrales and the Ptychomniales (Bell et al., 2007). Detailed studies involving gene knockouts are also needed to confirm that these gene families have significant impact on life history traits in pleurocarpous mosses.

3.4. Evaluation of paleopolyploidy in pleurocarpous mosses

The increase in gene family occupancy that we observe can be explain either by (1) a whole genome duplication (WGD) event, or (2) many small-scale duplications (SSDs). If the large increase in gene family occupancy in the pleurocarpous mosses indicates a whole genome duplication, it does not appear to have been accompanied by an increase in the base chromosome number. Most of the species in our transcriptome dataset have a base chromosome count of 10, 11, or 12 (Fritsch, 1991), below a threshold that has been used previously for inferring polyploidy in mosses (Crawford et al., 2009).

We observe a consistent pattern of gene family occupancy ratios (2:1 or 3:1 pleurocarp:acrocarp) in about 6% of the gene family trees in the masked dataset. Many gene family duplications localize to specific branches on our species tree (Figs. 1 and 2A), particularly the common ancestor of Hypnales and *Aulacomnium*, suggesting that the gene family expansions share a common age, indicative of WGD. However, Hypnales shares a more recent common ancestor with other groups of pleurocarpous mosses (such as Hookeriales) that were not sampled as part of this study. The apparent increase in gene family occupancy at the common ancestor of Hypnales and *Aulacomnium* may be the result of SSDs that occurred along the long branch that separates these groups (Fig. 1), but could not be reconstructed to more specific nodes due to our taxon sampling. Data from additional groups of mosses, including the other orders of pleurocarpous and protopleurocarpous mosses, would be necessary to pinpoint the age of gene duplications and better distinguish WGD from SSD using the gene family occupancy reconstruction method.

We were also unable to detect a clear signal of WGD in our 25 transcriptomes using a Ks-based method. Although our pipeline reliably recovered the WGD previously described from *Physcomitrella* (Rensing et al., 2007), none of our transcriptomes showed an obvious intermediate "peak" of Ks values between 0.5 and 2.0 (Fig. 3, Supplemental Fig. 1). Using the mclust method to fit Gaussian distributions to the Ks values, the best fit (evaluated by BIC score) was typically between 6 and 9 components (Supplemental Table 3), which would suggest evidence of several WGD events. However, in most cases the BIC values for several values of "g" (the number of components) were very similar. When only the model with the best BIC value was considered for each transcriptome, the means of distributions for each species did not show consistent overlap (Supplemental Fig. 4). We therefore consider any signal of WGD with this method to be weak. If an ancient WGD event occurred, the intermediate peak of Ks values may be obscured due to the age of the duplication event, as seen in the *Amborella* proteome (Amborella Genome Project et al., 2013). Additionally, we may not be able to observe consistent peaks across Hypnales if the substitution rates are too variable (Barker et al., 2009).

3.5. Molecular signatures of selection

When a gene is duplicated, one or both copies may take on a new function (neofunctionalization), and these innovations result from positive, or reduced purifying selection acting on one or both copies (Blanc and Wolfe, 2004a; Freeling, 2009). We investigated the signature of selection in all orthologous gene trees found by the Yang/Smith pipeline that contained sequences from all acrocarpous mosses and at least ten pleurocarpous mosses (2746 orthogroups). We reconstructed gene trees for each orthogroup from back-translated nucleotide sequences and estimated branch-wise models of molecular evolution using CodeML. In the "M0" model, a single ratio of synonymous to nonsynonymous substitution rates (dN/dS, or omega) was inferred for the entire tree. In the "two-omega" model, separate omegas are estimated for the "pleurocarpous" and "acrocarpous" branches, and we determined whether this was a significantly better fit to the data using a Likelihood Ratio Test ($p < 0.0001$ to correct for multiple tests). Of the 2746 orthogroups tested, the two-omega model was preferred in 526 orthogroups, and for 519 of these, omega was greater in the pleurocarpous moss lineages (Fig. 2C). None of the inferred omega values were greater than one (which would suggest positive selection), but the large number of genes with elevated omegas in pleurocarpous lineages relative to acrocarpous lineages supports a hypothesis of reduced purifying selection.

We performed a GO enrichment analysis of the functionally annotated *Physcomitrella* proteins in the 519 orthogroups with reduced purifying selection, and revealed 19 enriched GO categories (Fig. 2D, Supplemental Table 4). Many of these GO categories belong to a single "family" of GO categories (Supplemental Fig. 5), the most specific of which is "metabolic process" (GO: 006399). To determine whether this was related to codon usage bias, we analyzed all transcriptomes and the *Physcomitrella* coding domain sequences using four metrics of codon usage calculated by the program codonW (http://codonw.sourceforge.net/). However, none of the metrics showed significant differences in codon usage between acrocarps and pleurocarps (Supplemental Table 5).

For seven orthogroups for which the two-omega model was preferred, omega was greater in the background (acrocarp) branches (Table 2). The *Physcomitrella* proteins in these orthogroups have been studied in controlled differential expression experiment, and several pairs of these seven genes are known to have correlated co-expression (see phytozome.jgi.gov). *Physcomitrella* genes Phpat.011G004300.1 (Membrane-associated hemotopoietic protein) and Phpat.011G069500.1 (Ankyrin repeat and protein kinase domain-containing protein) are known to be significantly co-expressed (Pearson's coefficient 0.82). In contrast, Phpat.001G030000.1 (Putative RNA Polymerase II regulator) and Phpat.024G039100.1 (histone H-3) are known to be significantly inversely expressed (Pearson's coefficient −0.33). If the co-expression of these genes is maintained in Hypnales, it would suggest that entire gene networks have shifted regimes of molecular

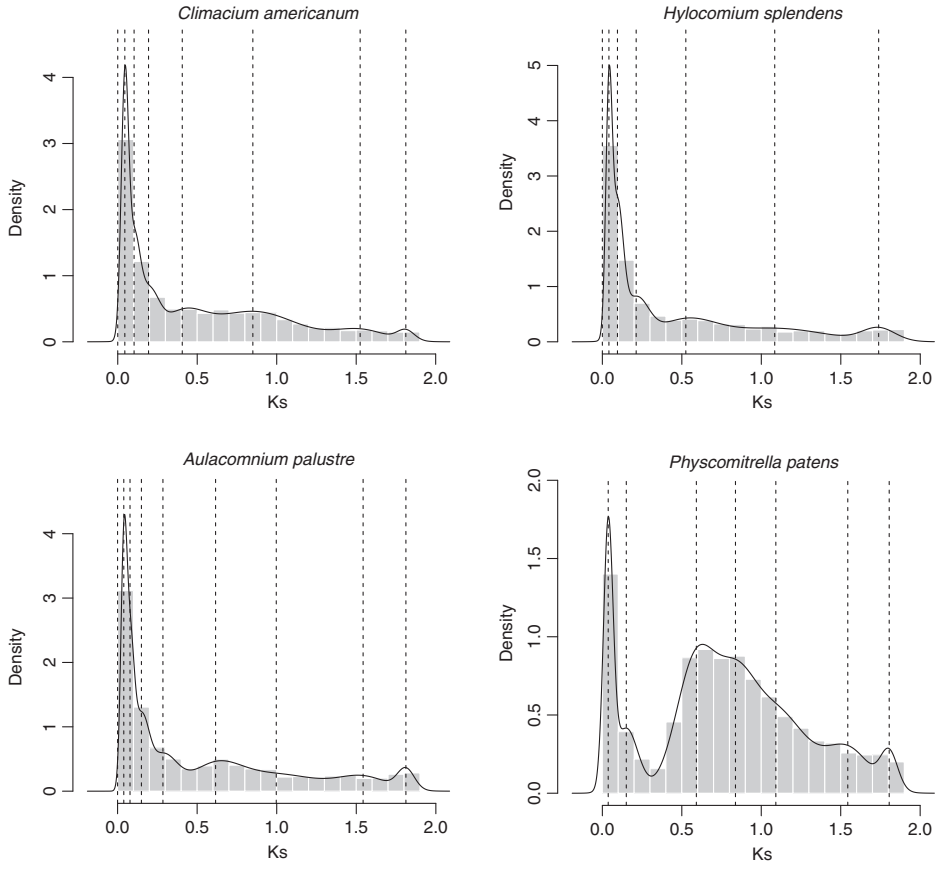

Fig. 3. Distribution of synonymous substitution rates (Ks) among pairs of paralogous genes within selected species. The dotted vertical lines represent the mean values of component distributions inferred by mclust under the model with the highest BIC score. The solid curved line is the inferred density distribution for the mixture model by mclust. For plots of all species, see Supplemental Fig. 2. *Climacium americanum* and *Hylocomium splendens* are pleurocarpous mosses, *Aulacomnium palustre* is a "protopleurocarpous" moss generally considered to be acrocarpous, and *Physcomitrella patens* is acrocarpous.

Table 2
Description of gene families with increased purifying selection in Hypnales, relative to the acrocarpous mosses.

Cluster ID	Acrocarp omega	Pleurocarp omega	*Physcomitrella* Gene ID	*Physcomitrella* gene annotation
cluster7933	0.145	0.051	Phpat.001G030000.1	Putative RNA Polymerase II regulator
cluster4410	0.212	0.12	Phpat.011G069500.1	Ankyrin repeat and protein kinase domain-containing protein
cluster5502	0.071	0.039	Phpat.011G004300.1	Membrane-associated hemotopoietic protein
cluster6993	0.106	0.066	Phpat.024G039100.1	Histone-H3
cluster10685	0.139	0.028	Phpat.007G045900.1	U2 small nuclear ribonucleoprotein B
cluster8238	0.116	0.061	Phpat.001G146700.1	ATP Binding/DNA Binding/Helicase
cluster3968	0.232	0.134	Phpat.012G062300.1	Hypothetical protein F4|10.140

selection coincident with the radiation of Hypnales. A controlled-condition differential expression study is needed to determine if the genes that have shifted regimes of selection have retained correlated expression in Hypnales pleurocarpous mosses.

It is possible that selection may act differently upon gene copies that result from small-scale gene duplications (SSDs), compared to genes that result from WGD. For example, a WGD generates gene copies in roughly equal proportions throughout enzymatic pathways, while SSDs may cause dosage imbalances and result in poor pathway flux (Lynch and Conery, 2000). As such, an alternative explanation to WGD is that the excess of gene family occupancy in Bryidae (and in Hypnales) is the result of several SSDs. In an investigation of the fate of gene copies resulting from small-scale duplications in land plants, Carretero-Paulet and Fares (2012) found that gene copies resulting from SSDs had reduced purifying selection in three angiosperm species, unlike gene copies resulting from WGD events. However, they did not observe this effect in *Physcomitrella patens*. Because we see a similar pattern (reduced purifying selection) in gene families that have expanded in Hypnales, small-scale duplications may be more plausible than a WGD. Additional taxon sampling and genome-wide analyses of synteny, particularly of other lineages within Bryidae and among other orders of the pleurocarpous mosses, are required to further distinguish between whole genome and small-scale duplications as sources of expanded gene families in pleurocarpous mosses.

4. Conclusions

This study is the first to investigate the genomic signatures associated with a rapid radiation in the largest order of pleurocarpous mosses, the lineage that accounts for the largest proportion of extant moss diversity. We describe here a set of 659 orthologous gene families and demonstrate their utility for phylogenetic

reconstruction in pleurocarpous mosses. These genes and analyses will likely form the foundation for future analyses of pleurocarp diversity, and our phylogenetic hypothesis provides a starting point to ask whether genomic features common to other rapid radiations in land plants occurred in pleurocarps. Our results suggest that both gene family expansion and a relaxation of purifying selection on many genes are significant features of the radiation of Hypnales and provide a set of candidate genes that, with further refinement, may be used in functional studies of pleurocarp development. The utility of transcriptome data for the phylogenetic analysis of molecular evolution depends on careful curation of datasets, including removal of contaminants, detection of homologous and orthologous sequences, and data analysis that allows the presence of missing data. Future phylotranscriptomic work, in this group and others, will rely on the continued development of bioinformatics pipelines to handle the challenges of working with transcriptome data. However, we have shown here that the use of transcriptomes to discover fundamental evolutionary processes that underlay the radiation of pleurocarpous mosses yields significant results and shows great promise for testing evolutionary hypotheses more broadly in mosses.

Acknowledgments

We would like to thank the Genomics Core Facility at the Northwestern University Center for Genetic Medicine Center and BGI for quality assurance and sequencing. We also thank D. Quandt (University of Bonn, Germany) for supplying two of our moss specimens. We thank B. Shaw for permission to use the photos in the graphical abstract. The masked transcriptome dataset, orthogroup assignment, individual gene alignments, and GO categories can be found in the Dryad depository: http://dx.doi.org/10.5061/dryad.475g7. This research was funded by National Science Foundation grants to AJS (DEB-1239980), BG (DEB-1240045), and NJW (DEB-1239992).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.01.008.

References

Amborella Genome Project, Albert, V.A., Barbazuk, W.B., dePamphilis, C.W., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., Soltis, D.E., Soltis, P.S., Wessler, S.R., Wing, R.A., Ammiraju, J.S.S., Chamala, S., Chanderbali, A.S., Determann, R., Ralph, P., Talag, J., Tomsho, L., Walts, B., Wanke, S., Chang, T.H., Lan, T., Arikit, S., Axtell, M.J., Ayyampalayam, S., Burnette, J.M., De Paoli, E., Estill, J.C., Farrell, N.P., Harkess, A., Jiao, Y., Liu, K., Mei, W., Meyers, B.C., Shahid, S., Wafula, E., Zhai, J., Zhang, X., Carretero-Paulet, L., Lyons, E., Tang, H., Zheng, C., Altman, N.S., Chen, F., Chen, J.Q., Chiang, V., Fogliani, B., Guo, C., Harholt, J., Job, C., Job, D., Kim, S., Kong, H., Li, G., Li, L., Liu, J., Park, J., Qi, X., Rajjou, L., Burtet-Sarramegna, V., Sederoff, R., Sun, Y.H., Ulvskov, P., Villegente, M., Xue, J.Y., Yeh, T.F., Yu, X., Acosta, J.J., Bruenn, R.A., de Kochko, A., Herrera-Estrella, L.R., Ibarra-Laclette, E., Kirst, M., Pissis, S.P., Poncet, V., 2013. The amborella genome and the evolution of flowering plants. Science 342, 1241089. http://dx.doi.org/10.1126/science.1241089.

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J., Rieseberg, L.H., 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25, 2445–2455. http://dx.doi.org/10.1093/molbev/msn187.

Barker, M.S., Vogel, H., Schranz, M.E., 2009. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biol. Evol. 1, 391–399. http://dx.doi.org/10.1093/gbe/evp040.

Bell, N.E., Quandt, D., O'Brien, T.J., Newton, A.E., 2007. Taxonomy and phylogeny in the earliest diverging pleurocarps: square holes and bifurcating pegs. The Bryologist. 110, 533–560. http://dx.doi.org/10.1639/0007-2745(2007)110[533:TAPITE]2.0.CO;2.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. Ser. B (Methodological) 57, 289–300. http://dx.doi.org/10.2307/2346101?ref=no-x-route:68fa25040cda789006fe1143e3a461e1.

Blanc, G., Wolfe, K.H., 2004a. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16, 1679–1691. http://dx.doi.org/10.1105/tpc.021410.

Blanc, G., Wolfe, K.H., 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16, 1667–1678. http://dx.doi.org/10.1105/tpc.021345.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

Buck, W.R., Cox, C.J., Shaw, A.J., Goffinet, B., 2005. Ordinal relationships of pleurocarpous mosses, with special emphasis on the Hookeriales. System. Biodiv. 2, 121–145. http://dx.doi.org/10.1017/S1477200004001410.

Buck, W.R., Vitt, D.H., 1986. Suggestions for a new familial classification of pleurocarpous mosses. Taxon 35, 21. http://dx.doi.org/10.2307/1221034.

Cannon, S.B., McKain, M.R., Harkess, A., Nelson, M.N., Dash, S., Deyholos, M.K., Peng, Y., Joyce, B., Stewart, C.N., Rolf, M., Kutchan, T., Tan, X., Chen, C., Zhang, Y., Carpenter, E., Wong, G.K.-S., Doyle, J.J., Leebens-Mack, J., 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol. Biol. Evol. 32, 193–210. http://dx.doi.org/10.1093/molbev/msu296.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. http://dx.doi.org/10.1093/bioinformatics/btp348.

Carretero-Paulet, L., Fares, M.A., 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol. Biol. Evol. 29, 3541–3551. http://dx.doi.org/10.1093/molbev/mss162.

Chang, Y., Graham, S.W., 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. Am. J. Bot. 98, 839–849. http://dx.doi.org/10.3732/ajb.0900384.

Chapman, M.A., Leebens-Mack, J.H., Burke, J.M., 2008. Positive selection and expression divergence following gene duplication in the sunflower CYCLOIDEA gene family. Mol. Biol. Evol. 25, 1260–1273. http://dx.doi.org/10.1093/molbev/msn001.

Chiang, T.-Y., Schaal, B.A., 2000. The internal transcribed spacer 2 region of the nuclear ribosomal DNA and the phylogeny of the moss family Hylocomiaceae. Plant System. Evol. 224, 127–137.

Cox, C.J., Goffinet, B., Wickett, N.J., Boles, S.B., Shaw, A.J., 2010. Moss diversity: a molecular phylogenetic analysis of genera. Phytotaxa 9, 175–195.

Crawford, M., Jesson, L.K., Garnock-Jones, P., 2009. Correlated evolution of sexual system and life history traits in mosses. Evolution 63, 1129–1142. http://dx.doi.org/10.1111/j.1558-5646.2009.00615.x.

Crosby, M.R., Magill, R.E., Allen, B., He, S., 1999. A checklist of the Mosses [WWW Document]. <http://www.mobot.org/MOBOT/tropicos/most/checklist.shtml> (accessed 8.28.15).

Csurös, M., 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26, 1910–1912. http://dx.doi.org/10.1093/bioinformatics/btq315.

de Souza, I.C.C., Recardi Branco, F.S., Vargas, Y.L., 2012. Permian bryophytes of Western Gondwanaland from the Paraná Basin in Brazil. Palaeontology 55, 229–241. http://dx.doi.org/10.1111/j.1475-4983.2011.01111.x.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res. 30, 1575–1584. http://dx.doi.org/10.1093/nar/30.7.1575.

Feldberg, K., Schneider, H., Stadler, T., Schäfer-Verwimp, A., Schmidt, A.R., Heinrichs, J., 2014. Epiphytic leafy liverworts diversified in angiosperm-dominated forests. Sci. Rep. 4. http://dx.doi.org/10.1038/srep05974.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucl. Acids Res. 42, D222–D230. http://dx.doi.org/10.1093/nar/gkt1223.

Fraley, C., Raftery, A.E., Murphy, T.B., Scrucca, L., 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.

Freeling, M., 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60, 433–453. http://dx.doi.org/10.1146/annurev.arplant.043008.092122.

Fritsch, R., 1991. Index to Bryophyte Chromosome Counts, Bryophytorum Biblioteca, Bryophytorum Bibliotheca. Science Publishers, Stuttgart, DE.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. http://dx.doi.org/10.1093/bioinformatics/bts565.

Goffinet, B., Buck, W.R., Shaw, A.J., 2009. Morphology, anatomy, and classification of the Bryophyta. In: Goffinet, B., Shaw, A.J. (Eds.), Bryophyte Biology. Cambridge, pp. 55–138.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman,

N., Regev, A., . Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. http://dx.doi.org/10.1038/nbt.1883.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protocols 8, 1494–1512. http://dx.doi.org/10.1038/nprot.2013.084.

Huerta-Cepas, J., Dopazo, J., Gabaldón, T., 2010. ETE: a python environment for tree exploration. BMC Bioinform. http://dx.doi.org/10.1186/1471-2105-11-24.

Huttunen, S., Bell, N., Bobrova, V.K., Buchbender, V., Buck, W.R., Cox, C.J., Goffinet, B., Hedenäs, L., Ho, B.-C., Ignatov, M.S., Krug, M., Kuznetsova, O., Milyutina, I.A., Newton, A., Olsson, S., Pokorny, L., Shaw, J.A., Stech, M., Troitsky, A., Vanderpoorten, A., Quandt, D., 2012. Disentangling knots of rapid evolution: origin and diversification of the moss order Hypnales. J. Bryol. 34, 187–211. http://dx.doi.org/10.1179/1743282012Y.0000000013.

Huttunen, S., Ignatov, M.S., Quandt, D., Hedenäs, L., 2013. Phylogenetic position and delimitation of the moss family Plagiotheciaceae in the order Hypnales. Bot. J. Lin. Soc. 171, 330–353. http://dx.doi.org/10.1111/j.1095-8339.2012.01322.x.

Jiao, Y., Paterson, A.H., 2014. Polyploidy-associated genome modifications during land plant evolution. Philos. Trans. Roy. Soc. B 369, 20130355. http://dx.doi.org/10.1098/rstb.2013.0355.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S. E., Schuster, S.C., Ma, H., Leebens-Mack, J., dePamphilis, C.W., 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100. http://dx.doi.org/10.1038/nature09916.

Johnson, L.S., Eddy, S.R., Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinform. 11, 431. http://dx.doi.org/10.1186/1471-2105-11-431.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. http://dx.doi.org/10.1093/molbev/mst010.

La Farge-England, C., 1996. Growth form, branching pattern, and perichaetial position in mosses: cladocarpy and pleurocarpy redefined. The Bryologist 99, 170. http://dx.doi.org/10.2307/3244546.

Laenen, B., Shaw, B., Schneider, H., Goffinet, B., Paradis, E., Désamoré, A., Heinrichs, J., Villarreal, J.C., Gradstein, S.R., McDaniel, S.F., Long, D.G., Forrest, L.L., Hollingsworth, M.L., Crandall-Stotler, B., Davis, E.C., Engel, J., von Konrat, M., Cooper, E.D., Patiño, J., Cox, C.J., Vanderpoorten, A., Shaw, A.J., 2014. Extant diversity of bryophytes emerged from successive post-Mesozoic diversification bursts. Nat. Commun. 5, 5134. http://dx.doi.org/10.1038/ncomms6134.

Li, F.-W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J., Burge, D.O., Pokorny, L., Larsson, A., Chen, T., Weststrand, S., Thomas, P., Carpenter, E., Zhang, Y., Tian, Z., Chen, L., Yan, Z., Zhu, Y., Sun, X., Wang, J., Stevenson, D.W., Crandall-Stotler, B.J., Shaw, A. J., Deyholos, M.K., Soltis, D.E., Graham, S.W., Windham, M.D., Langdale, J.A., Wong, G.K.-S., Mathews, S., Pryer, K.M., 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. Proc. Natl. Acad. Sci. USA 111, 6672–6677. http://dx.doi.org/10.1073/pnas.1319929111.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., Barker, M.S., 2015. Early genome duplications in conifers and other seed plants. Sci. Adv. 1, e1501084. http://dx.doi.org/10.1126/sciadv.1501084.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., Van de Peer, Y., 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA 102, 5454–5459. http://dx.doi.org/10.1073/pnas.0501102102.

Magill, R.E., 2014. Moss diversity: new look at old numbers. Phytotaxa.

Merget, B., Wolf, M., 2010. A molecular phylogeny of Hypnales (Bryophyta) inferred from ITS2 sequence-structure data. BMC Res. Notes 3, 320. http://dx.doi.org/10.1186/1756-0500-3-320.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548. http://dx.doi.org/10.1093/bioinformatics/btu462.

Moore, R.C., Purugganan, M.D., 2005. The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol. 8, 122–128. http://dx.doi.org/10.1016/j.pbi.2004.12.001.

Newton, A., Wikstrom, N., Bell, N., Lowe Forrest, L., Ignatov, M., 2007. Dating the diversification of the pleurocarpous mosses. In: Newton, A.E., Tangney, R.S. (Eds.), Pleurocarpous Mosses: Systematics and Evolution, Systematics and Evolution. CRC Press, pp. 337–366. http://dx.doi.org/10.1201/9781420005592.ch17.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490. http://dx.doi.org/10.1371/journal.pone.0009490.

Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y., Reski, R., 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. BMC Evol. Biol. 7, 130. http://dx.doi.org/10.1186/1471-2148-7-130.

Schneider, H., Schuettpelz, E., Pryer, K.M., Cranfill, R., Magallón, S., Lupia, R., 2004. Ferns diversified in the shadow of angiosperms. Nature 428, 553–557. http://dx.doi.org/10.1038/nature02361.

Shaw, A.J., Cox, C.J., Goffinet, B., Buck, W.R., Boles, S.B., 2003. Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (Bryophyta). Evolution 57, 2226–2241. http://dx.doi.org/10.2307/3448774.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btv351.

Smith, S.A., Dunn, C.W., 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24, 715–716. http://dx.doi.org/10.1093/bioinformatics/btm619.

Smoot, E.L., Taylor, T.N., 1986. Structurally preserved fossil plants from Antarctica: II. A permian moss from the transantarctic mountains. Am. J. Bot. 73, 1683. http://dx.doi.org/10.2307/2444234.

Spangler, J.B., Subramaniam, S., Freeling, M., Feltus, F.A., 2012. Evidence of function for conserved noncoding sequences in *Arabidopsis thaliana*. New Phytol. 193, 241–252. http://dx.doi.org/10.1111/j.1469-8137.2011.03916.x.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btu033.

Stech, M., Frey, W., 2008. A morpho-molecular classification of the mosses (Bryophyta). Nova Hedw. 86, 1–21. http://dx.doi.org/10.1127/0029-5035/2008/0086-0001.

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE 6, e21800. http://dx.doi.org/10.1371/journal.pone.0021800.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucl. Acids Res. 34, W609–W612. http://dx.doi.org/10.1093/nar/gkl315.

Yang, Y., Moore, M.J., Brockington, S.F., Soltis, D.E., Wong, G.K.-S., Carpenter, E.J., Zhang, Y., Chen, L., Yan, Z., Xie, Y., Sage, R.F., Covshoff, S., Hibberd, J.M., Nelson, M.N., Smith, S.A., 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. Mol. Biol. Evol. http://dx.doi.org/10.1093/molbev/msv081.

Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31, 3081–3092. http://dx.doi.org/10.1093/molbev/msu245.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591. http://dx.doi.org/10.1093/molbev/msm088.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.K.-S., Yu, J., 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genom. Proteom. Bioinform. 4, 259–263. http://dx.doi.org/10.1016/S1672-0229(07)60007-2.