frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fpls.2022.882441

Phylotranscriptomics Illuminates the Placement of Whole Genome Duplications and Gene Retention in Ferns

Jessie A. Pelosi[1], Emily H. Kim[2], W. Brad Barbazuk[1,3] and Emily B. Sessa[1]*

[1] Department of Biology, University of Florida, Gainesville, FL, United States, [2] Department of Microbiology and Cell Science, University of Florida, Gainesville, FL, United States, [3] Genetics Institute, University of Florida, Gainesville, FL, United States

Ferns are the second largest clade of vascular plants with over 10,000 species, yet the generation of genomic resources for the group has lagged behind other major clades of plants. Transcriptomic data have proven to be a powerful tool to assess phylogenetic relationships, using thousands of markers that are largely conserved across the genome, and without the need to sequence entire genomes. We assembled the largest nuclear phylogenetic dataset for ferns to date, including 2884 single-copy nuclear loci from 247 transcriptomes (242 ferns, five outgroups), and investigated phylogenetic relationships across the fern tree, the placement of whole genome duplications (WGDs), and gene retention patterns following WGDs. We generated a well-supported phylogeny of ferns and identified several regions of the fern phylogeny that demonstrate high levels of gene tree–species tree conflict, which largely correspond to areas of the phylogeny that have been difficult to resolve. Using a combination of approaches, we identified 27 WGDs across the phylogeny, including 18 large-scale events (involving more than one sampled taxon) and nine small-scale events (involving only one sampled taxon). Most inferred WGDs occur within single lineages (e.g., orders, families) rather than on the backbone of the phylogeny, although two inferred events are shared by leptosporangiate ferns (excluding Osmundales) and Polypodiales (excluding Lindsaeineae and Saccolomatineae), clades which correspond to the majority of fern diversity. We further examined how retained duplicates following WGDs compared across independent events and found that functions of retained genes were largely convergent, with processes involved in binding, responses to stimuli, and certain organelles over-represented in paralogs while processes involved in transport, organelles derived from endosymbiotic events, and signaling were under-represented. To date, our study is the most comprehensive investigation of the nuclear fern phylogeny, though several avenues for future research remain unexplored.

Keywords: fern, transcriptome, phylogenetics, polyploidy, whole genome duplication, biased gene retention

OPEN ACCESS

Edited by:
Li-Yaung Kuo,
National Tsing Hua University, Taiwan

Reviewed by:
James Clark,
University of Bristol, United Kingdom
David Wickell,
Cornell University, United States

****Correspondence:***
Jessie A. Pelosi
jessiepelosi@ufl.edu

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *23 February 2022*
Accepted: *16 June 2022*
Published: *14 July 2022*

Citation:
Pelosi JA, Kim EH, Barbazuk WB and Sessa EB (2022) Phylotranscriptomics Illuminates the Placement of Whole Genome Duplications and Gene Retention in Ferns. Front. Plant Sci. 13:882441. doi: 10.3389/fpls.2022.882441

INTRODUCTION

Ferns are the second largest group of vascular plants (with around 10,000 species, The Pteridophyte Phylogeny Group, 2016), the sister group to seed plants, and highly diverse (**Figure 1**). Molecular data have revolutionized our understanding of fern phylogenetics over the last two decades. Results from these studies have clarified the phylogenetic placement of ferns in the land plant phylogeny as sister to seed plants (Pryer et al., 2001; Wickett et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019; henceforth, 1KP), the placement of Equisetales (horsetails) and Psilotales (whisk ferns) as true ferns (Pryer et al., 2001; Knie et al., 2015), the recent origins of most extant fern diversity (Schneider et al., 2004; Schuettpelz and Pryer, 2009; Testo and Sundue, 2016), deep (Pryer et al., 2001, 2004; Schuettpelz and Pryer, 2007; Kuo et al., 2011, 2018a; Rothfels et al., 2015; Testo and Sundue, 2016) and shallow (e.g., Rothfels et al., 2008; Sessa et al., 2012a; Schuettpelz et al., 2016; Xu et al., 2019; Kinosian et al., 2020) relationships of many fern lineages, and the role of polyploidy in shaping fern evolution (Manton, 1950; Klekowski and Baker, 1966; Sessa et al., 2012b; Schneider et al., 2017). Despite this progress, however, several key divergences along the backbone of the fern phylogeny remain unresolved, including the relationships between the eusporangiate and leptosporangiate ferns (Pryer et al., 2001; Knie et al., 2015; One Thousand Plant Transcriptomes Initiative, 2019), Gleicheniales and Hymenophyllales (Rothfels et al., 2015; Kuo et al., 2018a), among the sister groups to the eupolypods (Rothfels et al., 2015; Testo and Sundue, 2016), and within the eupolypods II (Aspleniineae *sensu* Rothfels et al., 2012; Testo and Sundue, 2016; The Pteridophyte Phylogeny Group, 2016; Wei et al., 2017; Du et al., 2021). Many challenges make resolving these deep relationships difficult, including lineage-specific rate heterogeneity, nuclear-plastid incongruence, and polyploidy.

The vast majority of DNA-based studies of fern phylogeny and evolution to date have used primarily or exclusively plastid loci (e.g., Pryer et al., 2001, 2004; Schuettpelz and Pryer, 2007; Testo and Sundue, 2016; see Table 1 in Rothfels et al., 2015 for a summary of the main studies in deep fern phylogenetics), which act as a single linkage group (Lynch, 2007) and are maternally inherited in ferns (Gastony and Yatskievych, 1992; Vogel et al., 1998; Guillon and Raquin, 2000; Kuo et al., 2018b). Nuclear analyses of fern phylogenetics, in contrast, have lagged behind and have focused on just a few loci. Large-scale comparative studies at the genomic scale are also lacking within ferns (Marks et al., 2021; Szövényi et al., 2021), although several genome-sequencing projects have been recently completed (Li F. W. et al., 2018; Marchant et al., 2019; Huang et al., 2022) or are in progress (Cheng et al., 2018; Pelosi unpub. data). The average homosporous fern (including 99% of fern species, Haufler et al., 2016; The Pteridophyte Phylogeny Group, 2016) has a 1C genome size of 12.05 pg (Sessa and Der, 2016), and there is a staggering 282-fold difference in genome sizes across all ferns (both homosporous and heterosporous), from 0.26 Gb in the heterosporous water fern *Salvinia cucullata* to 73.19 Gb in *Tmesipteris elongata* (Hidalgo et al., 2017; Li F. W. et al., 2018, respectively) (note that heterosporous ferns, which make up ∼1% of fern diversity, have substantially smaller genomes, with an average 1C value of 2.43 pg, Sessa and Der, 2016). For these reasons, whole genome studies across the fern clade are generally unfeasible with current sequencing and assembly technology. The implementation of transcriptome sequencing for phylogenetic study has been applied throughout the green plants (e.g., Wickett et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019) and within ferns specifically in a small number of studies (Qi et al., 2018; Shen et al., 2018); these datasets, however, have not yet been combined and thoroughly interrogated. The enormous genomes of ferns, which may be a consequence of multiple rounds of whole genome duplication (WGD), and consequent difficulty with phasing alleles, parsing homeologs, and chimeric assemblies, have hampered the progress of nuclear-based phylogenetic studies in ferns.

Whole genome duplication, or polyploidy, is associated with nearly one-third of speciation events in ferns (Wood et al., 2009). Shifts in ecological niches (Marchant et al., 2016), phenotypes (Finigan et al., 2012), and environmental robustness (Van de Peer et al., 2017), or genetic changes such as biased gene retention (Van de Peer et al., 2017; Li Z. et al., 2018) and expression (Shan et al., 2020), and alternative splicing patterns (Zhou et al., 2011) may arise following WGDs. Genomes may also undergo large-scale post-polyploidy reorganizations (reviewed by Soltis et al., 2015); for example, following an allopolyploidy event (hybridization of two species accompanied by genome duplication), one subgenome often becomes "dominant" over the other (Alger and Edger, 2020; but see Krabbenhoft et al., 2021 for an example of extreme subgenome stability following an ancient duplication event). There are also several potential fates of individual duplicated genes (reviewed in Li et al., 2021). In most cases, one of the duplicate copies becomes non-functional, and will either be retained in the genome as a pseudogene or lost in the process of reorganization (for examples of gene silencing in *Tragopogon* see Buggs et al., 2011, 2010a,b). Alternatively, one of the duplicates may undergo neofunctionalization, where less effective purifying selection on one duplicate allows it to evolve a new function. A third possibility, sub-functionalization, posits that both duplicates retain complementary functions of the single pre-duplication gene. Alternatively, according to the Dosage Balance Hypothesis (DBH, Papp et al., 2003), genes with multiple interaction partners (such as transcription factors) are preferentially retained in duplicate following a WGD (Freeling, 2009; Defoort et al., 2019), implying that duplicates from small-scale events and WGDs should have different functions than those retained from polyploidy events. Several studies (e.g., Barker et al., 2008; Li et al., 2016; Li Z. et al., 2018) have found that duplicates from WGDs were enriched for different functions than the entire transcriptome or genome, and that these retained duplicates tended to converge on similar functions. Similar patterns and processes of genome evolution have yet to be explored in ferns.

There have been, on average, between two and four duplication events inferred in the ancestry of each extant fern species (One Thousand Plant Transcriptomes Initiative, 2019), with 19 (Huang et al., 2019) to 21 (One Thousand Plant Transcriptomes Initiative, 2019) putative events across the fern

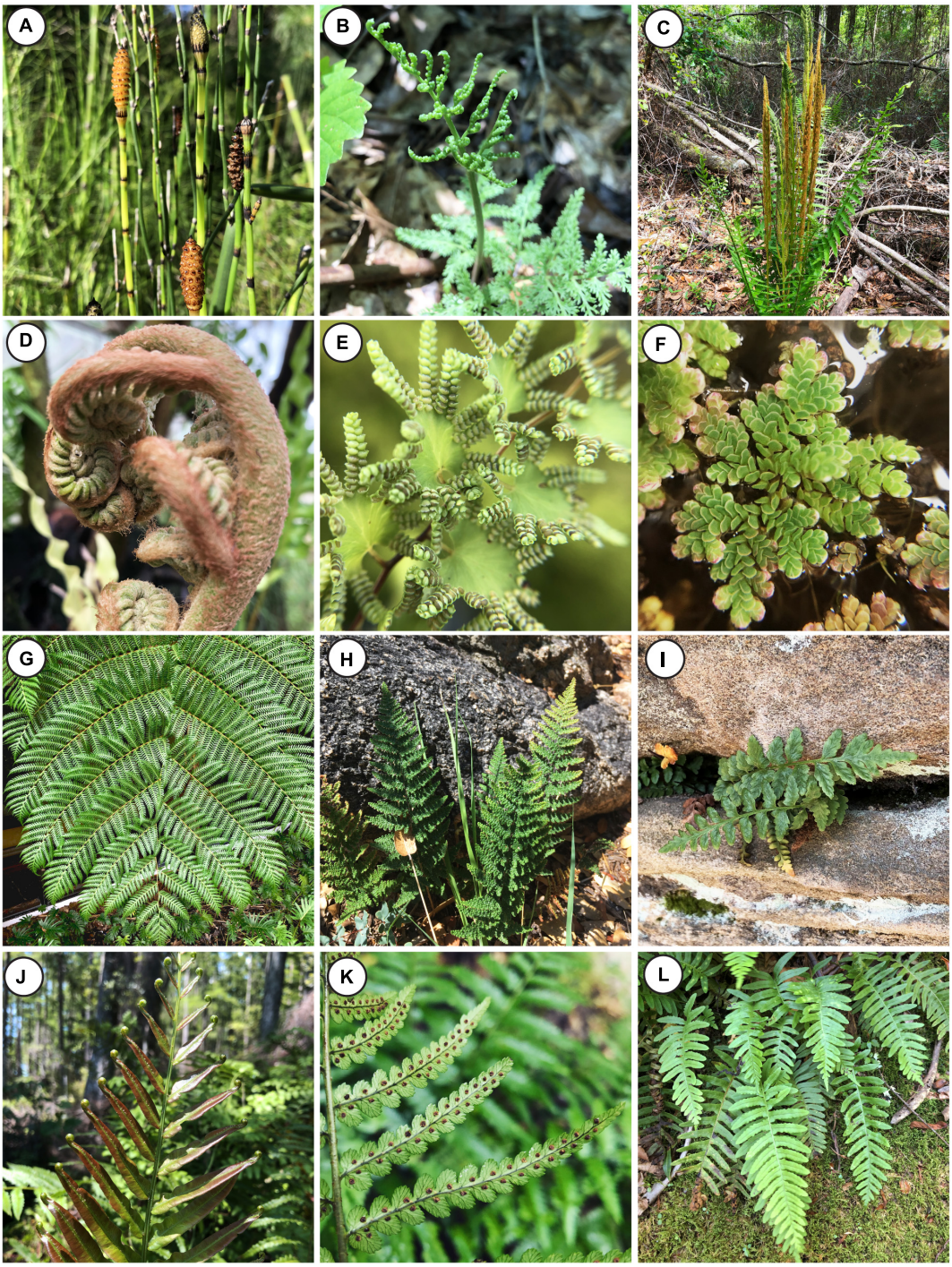

FIGURE 1 | The diversity of ferns. **(A)** *Equisetum hyemale* (Equisetaceae, Equisetales), **(B)** *Sceptridium dissectum* forma *dissectum* (Ophioglossaceae, Ophioglossales), **(C)** *Osmundastrum cinnamomeum* (Osmundaceae, Osmundales), **(D)** *Angiopteris evecta* (Marattiaceae, Marattiales), **(E)** *Lygodium microphyllum* (Lygodiaceae, Schizaeales), **(F)** *Azolla filiculoides* (Salviniaceae, Salviniales), **(G)** *Sphaeropteris cooperi* (Cyatheaceae, Cyatheales), **(H)** *Myriopteris wootonii* (Pteridaceae, Polypodiales), **(I)** *Asplenium* hybrid (Aspleniaceae, Polypodiales), **(J)** *Telmatoblechnum serrulatum* (Blechnaceae, Polypodiales), **(K)** *Dryopteris ludoviciana* (Dryopteridaceae, Polypodiales), **(L)** *Polypodium virginianum* (Polypodiaceae, Polypodiales). All images by JP.

phylogeny. A comprehensive understanding of the evolution and biology of ferns requires thorough study of the placement and aftermath of WGD events throughout the history of this clade. Here, we leverage publicly available data to reconstruct nuclear phylogenies for ferns. We use this phylogenetic background to investigate (1) the phylogenetic backbone of ferns, (2) the

placement of WGDs throughout the fern phylogeny, and (3) the fates and nature of duplicated genes following WGDs.

MATERIALS AND METHODS

Transcriptome Assembly and Annotation

Sequence data from Qi et al. (2018) and Shen et al. (2018) were downloaded from the NCBI SRA. The quality of raw sequences was assessed with FastQC[1] and reads were trimmed of adaptors and the first five bp with Trimmomatic ver. 0.39 (Bolger et al., 2014). Transcriptomes were assembled following the methods in One Thousand Plant Transcriptomes Initiative (2019); trimmed reads were assembled with SOAPdenovoTrans ver. 1.03 (Xie et al., 2014) with a kmer size of 25 bp. To remove any plastid sequence data that may have ended up in the nuclear assemblies, available fern plastome sequences were downloaded from NCBI (accessed July, 2020). BLASTN (Altschul et al., 1990) was used to compare assemblies against the plastome database and scaffolds or contigs with significant hits (*e*-value $< 1 \times 10^{-4}$, overlap $>$ 300 bp, bitscore $>$ 50) were removed from the assembly. Sequences with greater than or equal to 98% similarity were clustered with CD-HIT ver. 4.6.8 (Fu et al., 2012) to reduce transcript redundancy. Assemblies from 1KP were downloaded from the Cyverse repository for downstream analyses. Transcriptome completeness was assessed using BUSCO ver. 3.02 (Simão et al., 2015) with eukaryota odb9 lineage database (see **Supplementary Appendix 1**). Peptide and coding sequence (CDS) files were generated for each transcriptome using TransDecoder ver. 5.5.0 (Grabherr et al., 2011). We used Kruskal–Wallis tests by rank to determine whether assemblies from different studies were significantly different from one another for several metrics of interest, followed by pairwise comparisons using Wilcoxon rank sum tests with corrections for multiple testing using the Holm method (Holm, 1979).

Phylotranscriptomics

Peptide and corresponding CDS files for the outgroups *Amborella trichopoda, Arabidopsis thaliana, Physcomitrella patens,* and *Selaginella moellendorffii* were downloaded from Ensembl Plants 51 (Howe et al., 2020). To ensure that all major lineages of land plants were represented, peptide and CDS files for the gymnosperm *Ginkgo biloba* were also downloaded from Guan et al. (2019). Peptide files from outgroups and ferns with >55% BUSCO completeness (92.6%, 239 of 258) were passed to OrthoFinder ver. 2.3.11 (Emms and Kelly, 2015, 2019) to identify orthogroups (OGs). OGs identified with OrthoFinder were filtered using the custom bash script *get_orthogroups.sh* to generate datasets for single- and multi-copy datasets. OGs that were single-copy and contained at least 60, 75, and 85% of the input transcriptomes were used to generate the "SCO60," "SCO75," and "SCO85" datasets. Multi-copy OGs that contained all transcriptomes were used to generate the "MCO" dataset.

[1]www.bioinformatics.babraham.ac.uk/projects/fastqc/

The custom python script *extract_cds.py* (modified from Kasey K. Pham, pers. comm.) was used to obtain the corresponding coding sequences for each taxon in the filtered OGs. Sequences were aligned with the codon-aware alignment program MACSE ver. 2.0.4 (Ranwez et al., 2011) and gappy sites were removed with trimAl ver. 1.2 (Capella-Gutiérrez et al., 2009) by retaining sites that contained at least 50% of tips. Maximum likelihood gene trees were constructed from both the nucleotide and peptide alignments with IQTREE2 ver. 2.1.2 (Minh et al., 2020), with ModelFinder (Kalyaanamoorthy et al., 2017) and 1000 ultrafast bootstraps (Hoang et al., 2018) on 2 CPU threads where possible based on the recommendations from Shen et al. (2020). Note that trees constructed from the concatenated matrices required additional RAM and could not be run on only two CPU threads; these deviations are specified in our code. The optimal maximum likelihood gene trees for each locus were passed to Astral ver. 5.7.7 (Zhang et al., 2018) for the SCO datasets and Astral Pro ver. 1.1.2 (Zhang et al., 2020) for the MCO dataset to generate species trees under the multispecies coalescent (MSC). Concatenated alignments were used to generate a maximum likelihood species tree for the SCO datasets following the methods above, partitioned by locus (Chernomor et al., 2016). Discordance in the data was visualized using DiscoVista (Sayyari et al., 2018) and generalized Robinson-Foulds (GRF) distances between estimated species trees were calculated using the R package phangorn ver. 2.5.5 (Schliep, 2011).

Given heterogeneity among lineages (see Section "Discussion"), we compared the results from traditional models of sequence evolution to GHOST models (Crotty et al., 2020). GHOST is a mixture-model that takes a user-supplied number of components (k) and was developed to account for heterotachous evolution in datasets. For each locus in the SCO60 NT dataset, we ran IQTREE2 using the best-fitting model with classes k = 2–6 in both the linked and unlinked implementations of GHOST. The best number of classes was assessed using AIC. The optimal trees for the best-fitting number of classes under the GHOST model were passed to ASTRAL ver. 5.7.7 (Zhang et al., 2018) to construct the species tree as above. We also used a custom python script (*extract_codons.py*) written with the BioPython ver.1.79 module (Cock et al., 2009) to extract the first and second codon positions and third codon position from each untrimmed locus alignment in the SCO60 NT dataset. We then used trimAl on the first and second and third codon position alignments to remove sites with less than 50% tip occupancy. Gene trees were generated as above with IQTREE2 which were used to construct a species tree with ASTRAL.

Given that the phylogenetic position of two samples (*Onoclea sensibilis* ONSE and *Plagiogyria japonica* PLJA) suggested misidentifications, we extracted the longest plastid contig or scaffold from the initial assembly and used BLASTN (Altschul et al., 1990) against the nr database to determine their appropriate identification. The *Onoclea sensiblis* ONSE sample matched the *Matteuccia struthiopteris* chloroplast genome (KY427353.1) with a 98.7% identity compared to 91.26% identity to the *O. sensibilis* chloroplast genome (KY427354.1). The *P. japonica* PLJA sample matched the *P. subadnata* chloroplast genome (MN623362.1)

with 99.25% identity. These samples have been named according to their respective hits for downstream analyses.

In a supplemental analysis, we downloaded sequence data from Dong et al. (2019) for *Sphaeropteris lepifera* and assembled the transcriptome as above. Coding sequences of *Cystodium sorbifolium* and *Saccoloma campylurum* from Qi et al. (2018) were provided by the author as the raw sequence files were corrupted and not available on the NCBI SRA. For these three transcriptomes, we used BLASTP (Altschul et al., 1990) to identify significant hits to the 2884 OGs in the SCO60 dataset. The contig or scaffold with the highest bit-score was extracted and nucleotide alignments, gene trees, and species trees were constructed as above.

Divergence Time Estimation

We used SortaDate (Smith et al., 2018) to calculate the root-to-tip variance and bipartition support for each locus in the SCO60 NT dataset relative to the SCO60 NT MSC species tree. We then selected loci that were above the 90th percentile for bipartition support and below the 15th percentile for root-to-tip variance. This filtered dataset contained 99 single copy loci and was 136,137 bp in length. We generated a maximum likelihood tree with IQTREE2 as above using the SCO60 NT MSC species tree as a topological constraint to generate branch lengths relative to the number of substitutions per site. The resulting tree was then used to generate a dated phylogeny under a penalized likelihood method with treePL (Smith and O'Meara, 2012) with fifteen fossils (**Supplementary Appendix 2**) as calibration points along the phylogeny. To account for uncertainty in the dataset, we generated 100 bootstrap alignments from the 99 locus matrix and constructed trees from these bootstrap alignments using RAxML ver. 8.2.12 (Stamatakis, 2014). We then ran treePL on each of the bootstrap trees with three cross-validation runs and summarized them with treeAnnotator in BEAST ver. 2.5.0 (Suchard et al., 2018). We used the same methodology to date the supplemental tree with the inclusion of *C. sorbifolium, S. campylurum,* and *S. lepifera* with two additional fossil calibrations: we placed *Cyathocaulis* fossils (Tidwell and Nishida, 1993; Lantz et al., 1999) at 113–145 Ma for the stem Cyatheaceae as in Schuettpelz and Pryer (2009) and Du et al. (2021) and we used a stem age of 100.5–113 Ma for Lindsaeaceae based on the Regaladgo et al. (2017) as in Du et al. (2021). Although we did attempt to use MCMCTree (Yang, 2007) to estimate divergence times, due to the size of the dataset (e.g., number of loci, number of tips), Bayesian dating analyses were unfeasible.

Whole Genome Duplications

The program wgd ver. 1.1.1 (Zwaenepoel and Van de Peer, 2019) was used to generate paralog age distributions (K_S plots) for each transcriptome. Normal mixture models were fit to the K_S distributions using the R package mixtools ver. 1.2.0 (Benaglia et al., 2009); AIC values of models with one component were compared to those of models with more than one component to determine if fits were statistically different. The R package SiZer ver. 0.1–7 (Chaudhuri and Marron, 1999; Sonderegger, 2020) was used to detect significant changes in slope (at $\alpha = 0.05$). Briefly, SiZer identifies changes in slope based on changes in the first derivative in the curve (Chaudhuri and Marron, 1999). With respect to K_S plots, changes, particularly increases, would represent deviations from the background paralog distribution. Uncorrected interspecific K_S values were calculated for select species using wgd's one_v_one feature, which generates a K_S distribution of one-to-one orthologs for two taxa of interest. We compared the uncorrected K_S rates and subsequent WGD hypotheses with rates corrected using ksrates ver. 1.1.1 (Sensalari et al., 2022). Using a set of transcriptomes, ksrates accounts for differences in synonymous divergence attributable to lineage-specific rate heterogeneity using two focal taxa with an inferred WGD based on an initial K_S analysis and three to four outgroup taxa to correct K_S values. We selected taxa that represent major lineages and used BUSCO scores to inform our selection process.

Sets of transcriptomes were carefully selected based on hypotheses about the relative placement of WGDs from K_S plots and previous studies (Huang et al., 2019; One Thousand Plant Transcriptomes Initiative, 2019). Taxa were picked based on their phylogenetic position to represent major clades and we used transcriptome completeness (i.e., BUSCO scores) to aid in our decisions of which taxa to include. For taxa with multiple samples (e.g., *Psilotum nudum, Dicksonia antarctica*), we selected the assembly with the highest BUSCO score for our analyses. We extracted OGs where at least 45% of these transcriptomes were present at least once. OGs were aligned and gene trees were generated as above. Using subsets of the species tree and these gene trees, we performed several MAPS analyses (Li et al., 2015; Li Z. et al., 2018) to estimate the placement of WGDs in a phylogenomic context. Average gene birth (λ) and death (μ) rates for each dataset were estimated following Tiley et al. (2016). Briefly, we generated OGs using OrthoFinder for just the taxa of interest and filtered gene families to only include those with at least one copy in the outgroup and at least one copy in any other taxon in the dataset and removed any gene families with greater than 100 members in any taxon. We generated ten random subsets of 500 gene families to estimate λ and μ with WGDgc ver. 1.3 (Rabier et al., 2014) using the geometric mean of gene family size as the root prior and the "oneInBothClades" likelihood conditioning. To create a null distribution, we generated three sets of 1,000 simulated trees without a WGD using GuestTreeGen from PhyloGenData (Sjöstrand et al., 2013): 1,000 with the maximum likelihood values of λ and μ; 1,000 with λ and μ at half the ML values; and 1,000 with λ and μ at three times the ML values following (Li F. W. et al., 2018, Li Z. et al., 2018). We then generated 3,000 trees as above with WGDs in the midpoint of branches of interest (e.g., leading to nodes with higher subtree duplications relative to other nodes, nodes corresponding to K_S peaks) to create a positive distribution with a retention rate of 20%. We compared the MAPS analysis with the empirical data against these simulations using Fisher's exact tests.

We further compiled haploid (gametophytic) chromosome counts (n) and 1C genome sizes from Fujiwara et al. (2021), the Chromosome Count Database (Rice et al., 2015), and the Kew Plant *C*-Values Database (Leitch et al., 2019). For more recent inferred WGDs and where there was a direct sister clade without an inferred WGD, we compared chromosome number and genome size of the inferred polyploid taxa to their sister clade.

Gene Retention

Transcriptomes were blasted against the Araport11 *A. thaliana* protein dataset (Berardini et al., 2004). Gene ontology (GO) terms were sorted and tallied into GO Slim categories and visualized using custom perl scripts. Paralogs falling within ±1 SD of the mean of the putative WGD peak(s) and whose posterior probability was highest for the mixture component corresponding to the inferred WGD were also annotated using this pipeline. We used Chi-squared tests to determine if the GO composition of paralogs were significantly different from their respective full transcriptomes. Following Barker et al. (2008) and Shi et al. (2010), GO Slim categories with residuals >2 are considered to be overrepresented in paralog sequences compared to the full transcriptomes, while GO Slim categories with residuals <-2 are considered to be underrepresented. We further compiled GO annotations from all transcriptomes and paralogs for each WGD event identified along the phylogeny and compared the GO Slim composition of paralogs retained in each of these events.

RESULTS

Assemblies

We compiled a total of 261 fern transcriptomes that represent 230 species; after filtering for transcriptome completeness, there were a total of 242 transcriptomes representing 211 species from 43 families (89.6% of 48 total families) in all 11 orders. We found that BUSCO completeness was significantly different between all three studies (**Supplementary Figure 1A**, $\chi^2 = 74.565$, $df = 2$, $P < 0.001$), as well as the total number of scaffolds and contigs (**Supplementary Figure 1B**, $\chi^2 = 154.57$, $df = 2$, $P < 0.001$). There were additional significant differences in total transcriptome length (**Supplementary Figure 1C**, $\chi^2 = 75.388$, $df = 2$, $P < 0.001$) and number of predicted genes (**Supplementary Figure 1D**, $\chi^2 = 9.5959$, $df = 2$, $P < 0.01$) in which the 1KP assemblies were significantly shorter and had fewer predicted genes compared to the assemblies from the Qi et al. (2018) and Shen et al. (2018) studies, but the other assemblies were not different from one another. Assembly statistics for transcriptomes from each of the studies are given in **Table 1** and **Supplementary Appendix 1**. Note that these analyses and statistics do not include the assemblies for *C. sorbifolium* and *S. campylurum* since we only had access to the CDS files.

Clustering and Phylogenetics

OrthoFinder identified a total of 6,507,715 genes from 244 transcriptomes (239 ferns and 5 outgroups). Nearly all genes (94.9%) were assigned to one of 126,747 shared OGs. The mean and median OG size was 48.7 and 3.0 genes, respectively, with 1796 OGs that were represented at least once per transcriptome. After filtering and trimming gappy sites (see Section "Materials and Methods"), 2884, 1161, and 135 single-copy OGs were retained in the SCO60, SCO75, and SCO85 datasets, respectively (**Table 2**). A total of 1585 multi-copy OGs were retained in the MCO dataset (**Table 2**).

Tree topology was generally consistent across analyses with most nodes having full or high support values (**Figures 2**, **3** and **Supplementary Figure 2**). All orders, except Gleicheniales, were monophyletic; all families were monophyletic, except for Tectariaceae and Athyriaceae in some analyses (see Section "Discussion"). In nearly all cases, genera that contained multiple samples were monophyletic, though some were not (e.g., *Cheilanthes*). GRF distances between trees were relatively low, with generalized scores all equal to or less than 0.06 (GRF ≤ 0.06; **Supplementary Figure 3**). Analyses were clustered first by the type of analysis [maximum likelihood on concatenated data matrix (ML) vs. multispecies coalescent (MSC) in ASTRAL], then generally by data type (nucleotide vs. amino acid), and finally by the dataset used (single copy, multi-copy, and percent of missing taxa). The maximum likelihood trees were the most similar to each other, having GRF ≤ 0.03, with MSC analyses differing by up to GRF = 0.06. There were clusters of low GRF values within the MSC analyses, where SCO75, SCO60, and MCO MSC analyses were clustered with GRF ≤ 0.02, with clusters for nucleotide and amino acid data types. The tree that differed the most from the other analyses was the SCO85 MSC analysis on amino acids, which had GRF values from 0.04 to 0.06 compared to other trees. Interestingly, the topology of both linked and unliked GHOST species trees were identical to the SCO60 NT MSC tree (GRF = 0, **Supplementary Figures 2**, **3**). There were some topological differences when the first and second (CP12) and third (CP3) codon positions were analyzed separately in the SCO60 NT dataset. Tree topologies are compared in further detail for specific clades of interest in the Discussion.

Our penalized likelihood dating analysis resulted in family stem ages similar to those in previous studies (see **Supplementary Appendix 3** and **Supplementary Figures 4**–**6**). We estimated that ferns originated around 346.7 Ma (range 342.5–346.7 Ma), leptosporangiate ferns around 300.9 Ma (range 299.3–312 Ma), Polypodiales around 185.0 Ma (range 172.2–222.0 Ma), and eupolypods around 114.0 Ma (range 79.2–123.6 Ma). With the three additional taxa and 17 fossil calibrations (see Section "Materials and Methods") we estimated that ferns originated around 346.5 Ma (range 345.7–346.7 Ma), leptosporangiate ferns around 299.1 Ma (range 299–300.3 Ma), Polypodiales 166.9 Ma (range 160.6–181.9 Ma), and eupolypods around 95.3 Ma (range 86.2–111.6 Ma). Divergence times of specific families and clades are further compared in the Section "Discussion."

Whole Genome Duplications

Of the 239 fern transcriptomes analyzed (not including the three additional taxa in the supplemental analyses), 193 had evidence of at least one peak in their K_S plots (80.75%); of these, 163 transcriptomes had one peak, and 30 had two peaks (**Supplementary Appendix 4** and **Supplementary Figure 7**). Median peak K_S values ranged from 0.104 to 2.191, with nearly all inferred WGDs (25 out of 27) having median peak $K_S < 2$, suggesting that these analyses do not suffer from saturation effects. Briefly, at high synonymous substitution values (typically $K_S > 2$), a build-up of synonymous mutations not related to a duplication event may appear as a "saturation peak" in K_S plots (Vanneste et al., 2013), resulting in a false-positive

TABLE 1 | Assembly statistics for transcriptomes from three publications used in this study.

Study	No. transcriptomes	No. contigs + scaffolds	Assembly length (Mbp)	No. predicted genes	% Complete BUSCO genes
Shen et al., 2018	69	**147425** (53777, 660237)	**55.2** (28.3, 132.8)	**26514** (17527, 47079)	**85.07** (48.51, 95.71)
Qi et al., 2018	119	**121550** (55147, 483673)	**53.6** (24.1, 111.7)	**26883** (15893, 46757)	**87.74** (50.83, 96.04)
One Thousand Plant Transcriptomes Initiative, 2019	70	**13241** (61, 21403)	**36.7** (9.9, 67.7)	**23095** (648, 39285)	**66.44** (0, 94.39)
Total	258	**99084** (61, 660237)	**49.5** (9.9, 132.8)	**25756** (648, 47079)	**81.25** (0, 96.04)

Metric averages are given in bold, followed by minimum and maximum in parentheses: ***average*** *(minimum, maximum).*

TABLE 2 | Datasets constructed in this study and corresponding metrics.

Type of data	Dataset name	Percent of transcriptomes present	Number of orthogroups after filtering	Median orthogroup length (bp/AA)	Mean orthogroup length (bp/AA)	Total length (bp/AA)
Multi-copy	MCO	100%	1585	1,326.0	1,390.0	2,203,082
				462.3	441	732,267
Single-copy	SCO60	60%	2884	925.5	990.4	2,856,366
				307.5	329.1	949,237
	SCO75	75%	1161	768.0	795.7	923,894
				225.0	264.3	306,805
	SCO85	85%	135	690.0	712.3	96,165
				229.0	236.4	31,920

AA, amino acid; bp, base pair.

inference of a WGD. The two events with median $K_S > 2$ were supported using MAPS (see below). In general, SiZer identified significant increases in slope consistent with K_S plots (**Supplementary Figure 7**), although there were several cases where significant increases were detected when no discernable K_S peaks were found (**Supplementary Figure 7**). Furthermore, SiZer failed to detect any significant increases in slope of K_S distributions of transcriptomes that had more than one peak, such as *Alsophila* spp. (**Supplementary Figure 7**). We conducted a total of 18 MAPS analyses that utilized 147,273 empirical gene trees with an average of 8663 gene trees per analysis (range 8043–11334), and 102,000 simulated gene trees (**Supplementary Appendix 5**). While WGD inferences from corrected and uncorrected K_S values were generally consistent, there were some instances where they conflicted (**Figure 4**). Using a combination of evidences, we inferred 27 large- and small-scale WGDs throughout the phylogeny (**Figure 2**, **Supplementary Figures 8**, **9**, **Supplementary Appendix 5**); 18 large-scale WGDs (present in more than one sampled taxon) and 9 small-scale WGDs (present in only one sampled taxon). Sixteen of these WGDs were supported in both the MAPS and K_S analyses.

There were several independent, large-scale inferred WGD events in the extant eusporangiate ferns, with WGDs separately shared by Equisetales (EQUI), *Sceptridium* (OPHIO.1), *Ophioglossum* (OPHIO.2), Psilotales (PSIL.1), *Tmesipteris* (PSIL.2), and Marattiales (MARA) (**Figure 2**). A small-scale, unshared duplication in *Ophioderma pendula* (OPHIO.3) was also identified. In these lineages it was difficult to assess an additional lines of chromosome number or genome size evidence as there were no direct sister groups lacking an inferred WGD. For example, all *Equisetum* species sampled have $n = 108$ with a base chromosome number of $x = 108$, suggesting that they are all diploid (or ancient polyploids). Furthermore, repeated rounds of neopolyploidy or multiple cytotypes within a genus or species may obscure signals of paleopolyploidy events. For example, *Sceptridium, Ophioglossum*, and *Ophioderma* should have the same haploid chromosome number if they each underwent independent WGDs. However, the haploid chromosome number in *Sceptridium dissectum* is $n = 45$, while *Ophioglossum vulgatum* has $n = 240–1320$ and *O. pendula* has $n = 370$. In these instances, we could not use chromosome number as a line of evidence to support or reject our hypothesized WGD events. We do find chromosomal support for the PSIL.2 event, with the most common haploid chromosome number of *P. nudum* of $n = 52$ (although several counts have been reported ranging from $n = 46–210$), compared to $n = 104$ for *Tmesipteris tannensis*. Furthermore, the genome size estimated for *P. nudum* (1C = 32.8 pg) is just about half that reported for *T. tannensis* (1C = 74.84 pg), supporting our inference of the PSIL.2 WGD event.

While there was not a significant increase in duplicated gene trees relative to a null distribution in the putative WGD in Marattiales, there was a significant increase at the MRCA of *Angiopteris fokiensis* and *Ptisana pellucida* which is consistent with a WGD (**Supplementary Appendix 5**). However, both corrected and uncorrected K_S plots suggests a WGD shared by all Marattiales (**Figure 4**). Furthermore, the base chromosome numbers of genera in Marattiales do not show an increase at the node identified by the MAPS analysis alone (at the MRCA of *A. fokiensis* and *P. pellucida*). There are a range of haploid chromosome counts in Marattiales; *A. fokiensis* has $n = 40$, *Christensenia aesculifolia* has $n = 80$, and the base

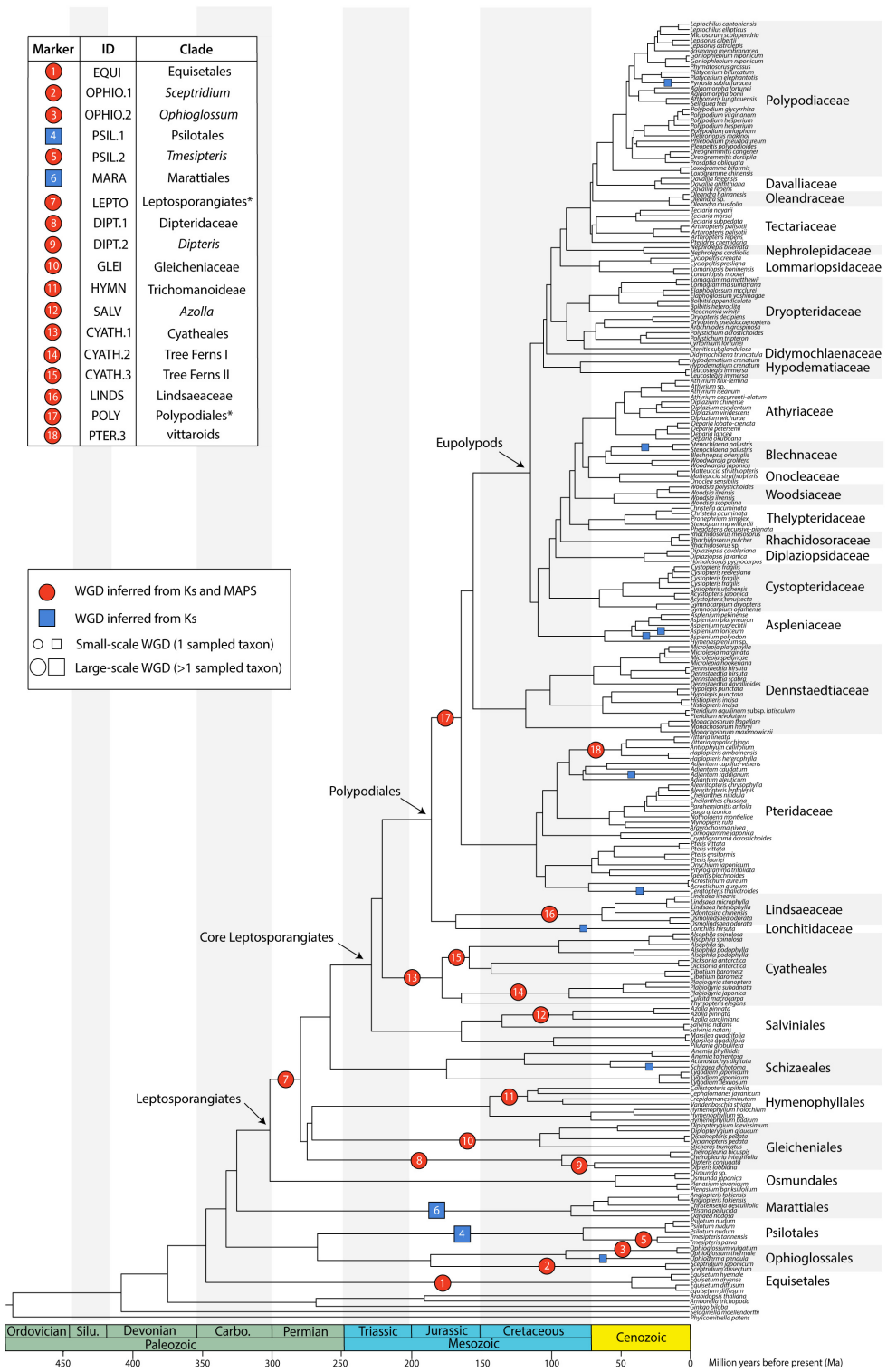

FIGURE 2 | Species tree generated from 2884 single copy nuclear loci (SCO60 dataset) under the multi-species coalescent. Divergence times are based on a penalized likelihood method in TreePL. Inferred whole genome duplications (WGDs) are placed on the tree (note that the age of the WGDs are not depicted, rather events are placed on the midpoint of the branches). Events inferred from K_S and MAPS analyses are shown as red circles, and events inferred from K_S evidence only are shown as blue squares. The size of the symbol reflects whether the event is large-scale (larger symbol, including more than one sampled taxon) or small-scale (smaller symbol, including only one sampled taxon). Asterisks for LEPTO and POLY events indicate there are no current names for the clades corresponding to the taxa affected by these events: LEPTO is shared by leptosporangiate ferns excluding Osmundales and POLY is shared by Polypodiales excluding Lindsaeineae and Saccolomatineae.

haploid chromosome number of *Ptisana* is n = 39, with no discernable pattern to confirm the within-Marattiales WGD event suggested by MAPS alone. While we cannot directly compare the karyotypes of Marattiales to a sister group that has not undergone a WGD, treating Osmundaceae as a diploid outgroup with x = 22 would further support the inference of a WGD at the base of Marattiales.

In the leptosporangiate ferns, we identified an event shared by all leptosporangiates excluding Osmundales (LEPTO, but see corrected K_S plots in **Figure 4** and **Supplementary Figure 9** for an alternative hypothesis). In the clade comprised of Hymenophyllales + Gleicheniales, we inferred four WGDs: events shared by Gleicheniaceae (GLEI), Dipteridaceae (DIPT.1), *Dipteris* (DIPT.2), and Trichomanoideae (HYMN). The placement of several of these events were ambiguous in both the corrected and uncorrected K_S analyses, while MAPS found strong support for the separate events. The DIPT.2 event requires further investigation, as the haploid chromosome number of *Dipteris conjugata* (n = 33) is nearly a quarter that of *Cheiropleuria dicuspis* (n = 116). Both *Dipteris* and *Cheiropleuria* have base chromosome numbers of x = 33, so it is possible the WGD event inferred here is a burst of gene duplications. HYMN is additionally supported by chromosome counts and genome sizes, although it is not identified in the corrected K_S analysis (**Figure 4** and **Supplementary Appendix 5**). The sampled *Hymenophyllum* all have n = 21 or 22 and *H. badium* has a 1C genome size of 16.15 pg, values which are nearly half that of those taxa hypothesized to have undergone the HYMN WGD event in Tichomanoideae, which have n = 36 and 1C genomes sizes from 25.04 to 25.24 pg (*Crepidomanes minutum* and *Cephalomanes javanicum*, respectively). If we compare the chromosome numbers of taxa that underwent the proposed DIPT.1 and GLEI WGD events to *Hymenophyllum* species which do not have a lineage-specific WGD event, the DIPT.1 and GLEI taxa have greater haploid chromosome numbers (n = 33–116 in Dipteridaceae and n = 39 in Gleicheniaceae) compared to *Hymenophyllum* (n = 22). These shifts in chromosome number may coincide with the inferred WGDs. In Schizaeales, a single unshared event was inferred in *Schizaea dichotoma* (SCHIZ).

In Salvinales, we found evidence for the WGD identified by Li F. W. et al. (2018) and One Thousand Plant Transcriptomes Initiative (2019), shared by *Azolla* species (SALV). There is greater than double the number of chromosomes in a haploid *Azolla pinnata* genome (n = 22) compared to *Salvinia natans* (n = 9). Interestingly, *S. natans* has a much greater genome size (1C = 1.82 pg) compared to *Azolla filiculoides* (syn. *A.* cf *caroliniana*), though this is not the case in other *Salvinia* species (e.g., *S. culcutta* has 1C = 0.25 pg). Three major WGD events were inferred in Cyatheales, one at the base shared by the order (CYATH.1), one shared by Culcitaceae and Plagiogyriaceae (CYATH.2), and one shared by Cibotiaceae, Cyatheaceae, and Dicksoniaceae (CYATH.3). In the MAPS analysis of CYATH.2, the empirical proportion of gene trees duplicated at the corresponding node of interest was significantly greater than the null distribution, but less than the positive distribution (**Supplementary Appendix 5**). While there are clear peaks in uncorrected K_S plots and evidence from MAPS, chromosome numbers are relatively uniform in Cyatheales. For example, the taxa that are proposed to have undergone the CYATH.3 WGD event have haploid chromosome numbers n = 65–69, while *Thrysopteris elegans* has n = 78. Furthermore, genome sizes of *Cibotium barometz* and *Alsophila spinulosa* (1C = 4.48 and 7.36 pg, respectively) are lower than that of *T. elegans* (1C = 10.23 pg). A similar pattern is observed in the CYATH.2, though there is greater variation in chromosome number n = 68–130 in *Plagiogyria* and *Culcita*. When we evaluated corrected K_S values, we found that the most recent peaks thought to represent CYATH.2 and CYATH.3 were found to be older than the inferred position by MAPS. The corrected K_S analysis instead suggests a single duplication at the base of Cyatheales, with an older event shared by Polypodiales + Salviniales + Cyatheales. It is possible that these inferred events within Cyatheales are bursts of gene duplication; genome sequencing projects focused on the tree ferns are in progress (or have been recently published, Huang et al., 2022) and should aim to tackle this question. We inferred two independent events in Lindsaeineae, one event in *Lonchitis hirsuta* (LONCH) and one shared by Lindsaeaceae (LINDS).

In Polypodiales, there was one inferred WGD shared by the order excluding Lindsaeineae and Saccolomatineae (POLY). One shared event was identified in the vittaroids within Pteridaceae (PTER.3), along with two independent events in *Adiantum raddianum* (PTER.2) and *Ceratopteris thalictroides* (PTER.1). Relative to *Adiantum,* the vittaroid ferns had much higher haploid chromosome numbers (n = 120) and genome sizes (1C = 32.81 and 33.17 pg for *Antrophym callifolium* and *Vittaria lineata*, respectively), which are nearly quadruple the chromosome numbers (n = 29, 30) and six times greater the genome sizes (1C = 5.17–5.58 pg for *A. caudatum* and *A. aleuticum*) of diploid *Adiantum.* Within the eupolypods, several small-scale events were also identified in *Asplenium loriceum* (ASPL.2), *Asplenium polyodon* (ASPL.1), *Stenochlaena palustris* (BLECH), and *Pyrrosia subfuracea* (PYRO). Compared to the base chromosome numbers in their respective genera (*Asplenium* x = 36, *Stenochlaena* and *Pyrrosia* x = 37), these taxa have twice the number of haploid chromosomes except for *Pyrrosia subfurfuracea* (n = 37).

Gene Retention Analyses

By comparing the number of putative paralogs within ±1SD of the K_S peak mean to the number of predicted genes from TransDecoder, we estimated that gene retention in ferns is low, with 11.97% of genes remaining in duplicate following an inferred WGD event (range 4.10–20.35%). Over- and under-represented GO Slim categories were similar throughout taxa and events (see Section "Discussion," **Supplementary Appendix 6**, Supplementary **Figures 11**, **12**) although lineage-specific differences are present. Generally, processes involved in binding (protein binding, nucleic acid binding, DNA binding, RNA binding), responses to stimuli (response to endogenous stimulus, response to abiotic stimulus, response to chemical), and certain organelles (nucleus, ribosome, endoplasmic reticulum, Golgi apparatus) were over-represented in paralogs. Processes involved in transport (transporter activity, transport, nuclear envelope), organelles derived from endosymbiotic events (mitochondrion,

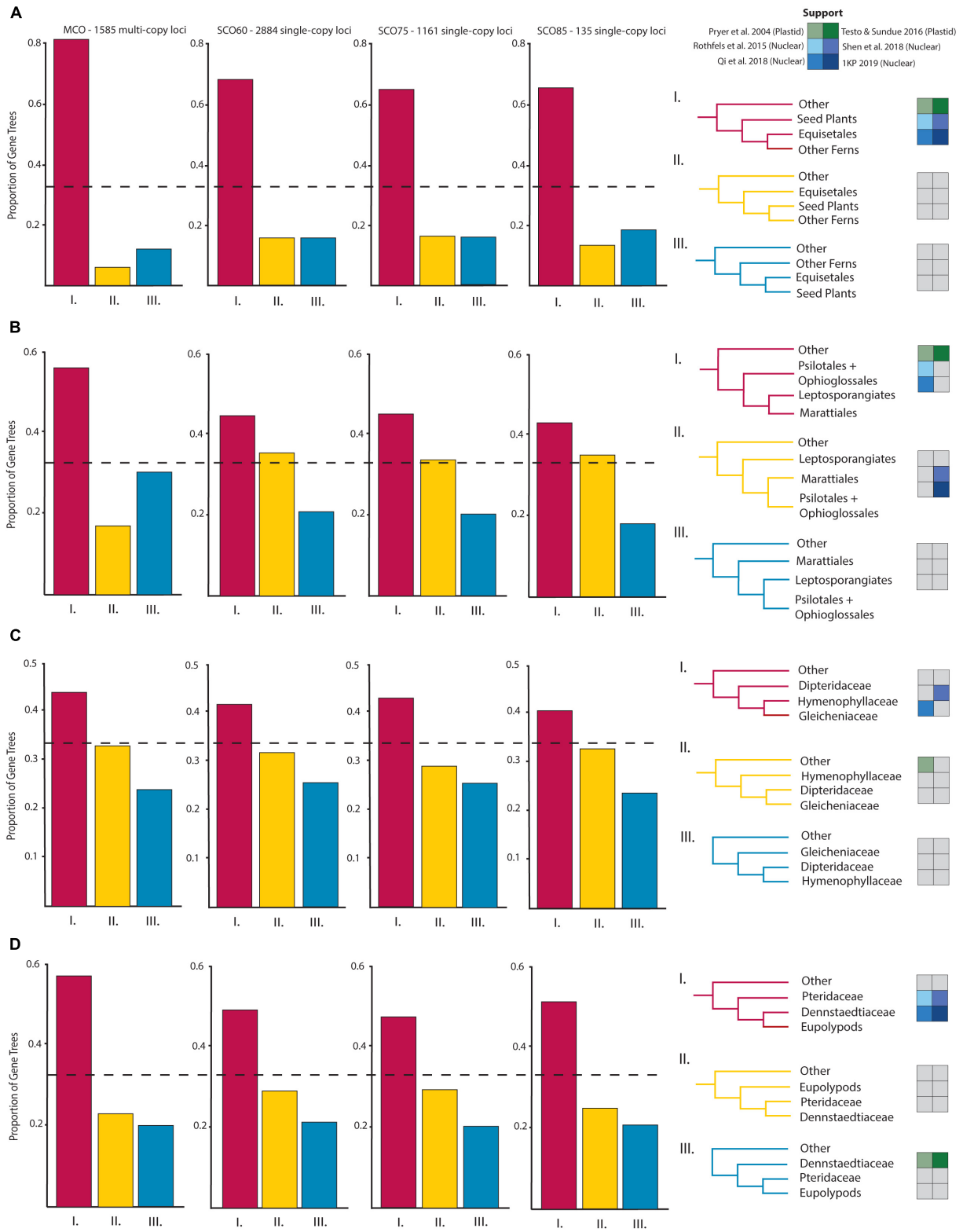

FIGURE 3 | Proportions of gene trees supporting the inferred species trees and alternative topologies for relationships of particular interest: **(A)** Horsetails (Equisetales) and other ferns, **(B)** Eusporangiate-leptosporangiate fern relationships, **(C)** Gleicheniales and Hymenophyllales, and **(D)** Pteridaceae, Dennstaedtiaceae, and the eupolypods. Topological support from select plastid (green, Pryer et al., 2004; Testo and Sundue, 2016) and nuclear (blue, Rothfels et al., 2015; Qi et al., 2018; Shen et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019) studies are given beside possible topologies.

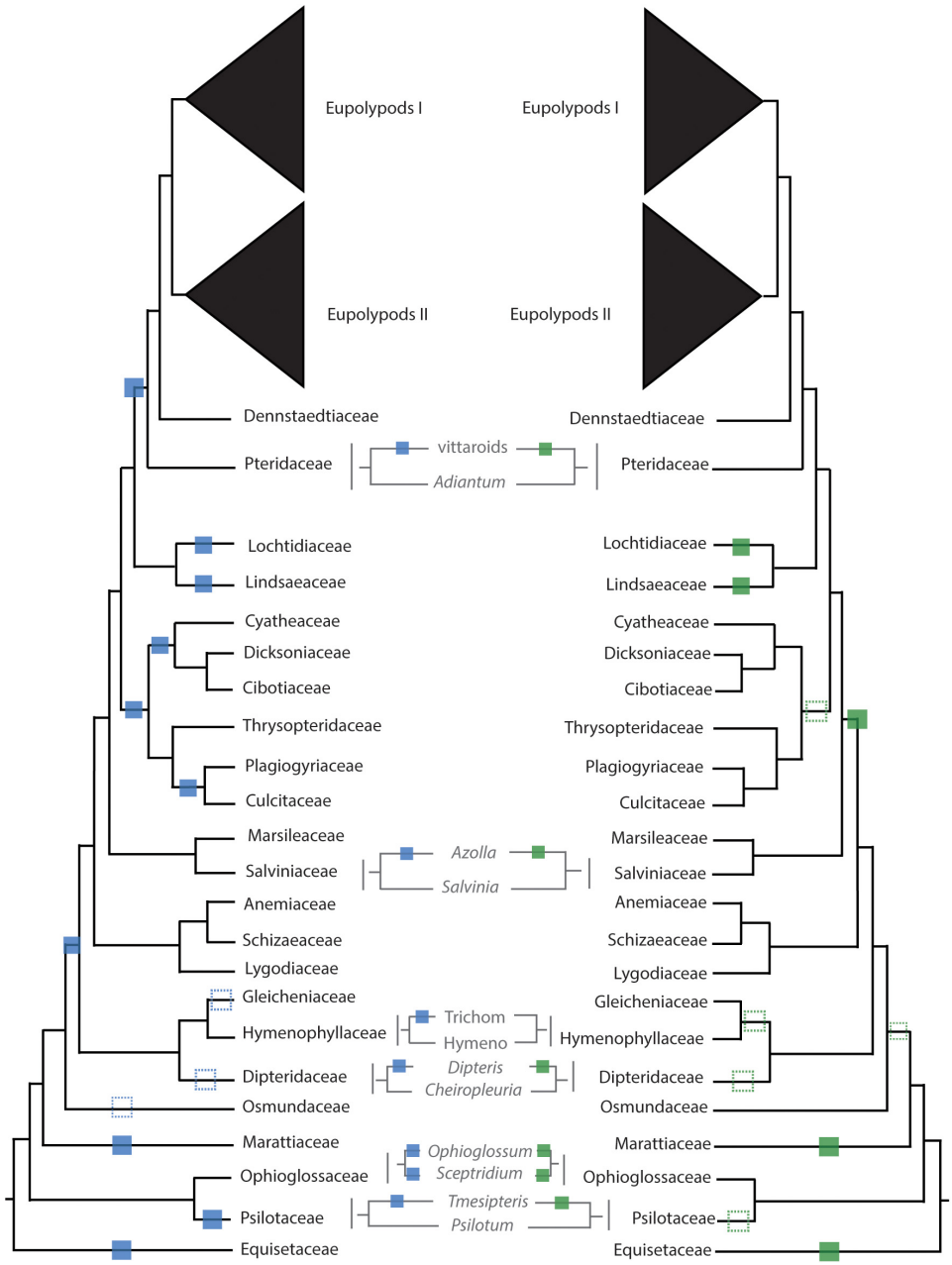

FIGURE 4 | Inferred whole genome duplication (WGD) events from uncorrected (left, blue) and corrected (right, green) K_S analyses. Unambiguous placements of inferred WGDs are depicted as solid squares; ambiguous placements are depicted as dotted squares. Trichom = Trichomanoideae, Hymeno = Hymenophylloideae.

chloroplast), and signaling (signaling receptor binding and activity) were under-represented in paralogs.

DISCUSSION

Resolving the Fern Phylogeny Backbone

Our understanding of the fern phylogeny has improved substantially over the last several decades, as our field has transitioned from morphological to molecular to genomic-based phylogenetic methods. The phylogenies we reconstructed here are largely consistent with the topologies from most recent plastid and nuclear analyses of ferns. Below we discuss several areas of the fern phylogeny that have been and remain difficult to resolve; for each clade we discuss relationships, gene tree-species tree conflict, and inferred ages, focusing primarily on the SCO60 NT MSC tree/dataset which contains the largest number of loci, and highlight points of discordance between this, our other datasets, and the literature.

Eusporangiate Ferns

The eusporangiate fern lineages form a successive grade to the leptosporangiate fern clade (**Figure 2**), with *Equisetum* sister to the rest of ferns, followed by a clade comprised of Ophioglossales and Psilotales, and Marattiales sister to the leptosporangiate ferns. We consistently found *Equisetum* to be sister to the rest of ferns (**Figure 2** and **Supplementary Figure 2**) with relatively low conflict among gene trees (**Figure 3A**), which agrees with most recent findings (Kuo et al., 2011; Knie et al., 2015; Rothfels et al., 2015; Testo and Sundue, 2016; Qi et al., 2018; Shen et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019), although some older studies have suggested *Equisetum* as sister to Marattiales (Pryer et al., 2001, 2004). Most previous studies have recovered a Devonian or Carboniferous origin of ferns, which is consistent with our finding a stem age of Equisetales of 346 Ma (range 341.5–347.0 Ma), although the age range in the literature spans 100 MY (**Supplementary Figures 4–6**), from 431 Ma (plastid data, Testo and Sundue, 2016) to 321 Ma (nuclear data, Shen et al., 2018). The placement of the clade consisting of Ophioglossales and Psilotales has differed across previous studies, with some suggesting that Ophioglossales and Psilotales together are sister to the rest of the ferns (Pryer et al., 2001, 2004), while most find them forming a grade leading up to leptosporangiate ferns (Knie et al., 2015; Rothfels et al., 2015; Testo and Sundue, 2016; Qi et al., 2018; Shen et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019; **Figure 2**). The placement of Psilotales and Ophioglossales relative to each other and to other ferns was congruent among analyses (**Supplementary Figure 2**) and loci. A similar range of ages exists for the stem of Psilotales and Ophioglossales (**Supplementary Figures 4–6**), which we recovered as 267 Ma (range 226.1–280.4 Ma), with ages from previous studies spanning from 368 Ma (plastid data, Testo and Sundue, 2016) to 173 Ma (nuclear, Shen et al., 2018).

In contrast with other eusporangiate fern relationships, there were relatively high levels of gene tree–species tree conflict in the placement of Marattiales, but the best-supported topology had Marattiales sister to the leptosporangiate ferns (**Figures 2**, **3B**). This topology is supported in most other large-scale fern phylogenies (Knie et al., 2015; Rothfels et al., 2015; Testo and Sundue, 2016; Kuo et al., 2018a; Qi et al., 2018), although some have found support for Psilotales and Ophioglossales sister to Marattiales (Shen et al., 2018; see also our third codon position analysis, **Supplementary Figure 2**) or Psilotales and Ophioglossales sister to the leptosporangiates (**Figure 3B**). The topology in One Thousand Plant Transcriptomes Initiative (2019) also shows substantial conflict in gene trees relative to the species tree here, with relatively equal proportions of gene trees supporting Marattiales sister to Ophioglossales and Psilotales, and leptosporangiates sister to Ophioglossales and Psilotales. We found a Marattialean stem age of 325 Ma (range 316.9–331.5 Ma), which is largely consistent with other findings (plastid: 365 Ma, Testo and Sundue, 2016; nuclear: 329–355 Ma, Rothfels et al., 2015; Qi et al., 2018). Although our sampling in this part of the phylogeny is sparse, the relationships we recovered within Marattiales are consistent with those found by Lehtonen et al. (2020) and May et al. (2020), with *Danaea* and *Ptisana* successively sister to a clade containing *Angiopteris* and *Christensenia*.

Gleicheniales and Hymenophyllales

The relationships between Gleicheniales and Hymenophyllales have not been previously resolved and the recovered topologies differed among studies with Hymenophyllales and Gleicheniales either forming a clade (Pryer et al., 2004), or not (Schuettpelz and Pryer, 2007; Testo and Sundue, 2016). Rothfels et al. (2015) found different topologies depending on the phylogenetic reconstruction approach used; using 25 low-copy nuclear loci, they found low support for a grade of Hymenophyllaceae, Gleichenaceae, and Dipteridaceae in their maximum likelihood tree, but recovered a clade of these families using a Bayesian approach. With entire plastome sequences, Kuo et al. (2018a) recovered two topologies with high support, one suggesting a grade of Hymenophyllales and Gleicheniales to the remaining leptosporangiate ferns, and the other suggesting a clade of Hymenophyllales and Gleicheniales.

We found that Hymenophyllaceae, Dipteridaceae, and Gleichenaceae formed a single clade, with Hymenophyllaceae sister to Gleicheniaceae, and Dipteridaceae sister to them (**Figures 2**, **3C**), suggesting that Gleicheniales may not be monophyletic. We identified high levels of gene tree–species tree conflict, with high proportions of gene trees supporting alternative topologies to the inferred species tree (**Figure 3C**). Short branches between critical nodes may represent a rapid divergence among these lineages (**Figure 2** and **Supplementary Figure 2**), suggesting a role for incomplete lineage sorting. In two trees, we recovered a clade of Gleicheniaceae and Hymenophyllales, with Dipteridaceae sister to the remaining ferns (**Supplementary Figure 2**, SCO85 AA MSC, MCO AA MSC). Within Hymenophyllaceae, we found support for the two subfamilies recognized by The Pteridophyte Phylogeny Group (2016): Trichomanoideae (*Callistopteris, Cephalomanes,* and *Vandenboschia*) and Hymenophylloideae (*Hymenophyllum*). The intra-familial relationships of these genera are similar to the topology recovered by Ebihara et al. (2006), except that we find *Callistopteris* and *Cephalomanes* are sister rather than a grade.

We found a stem age of Hymenophyllaceae of 271 Ma (range 270.6–273.1 Ma), giving this clade an origin in the Permian, which is consistent with other studies (Schuettpelz and Pryer, 2009; Qi et al., 2018; but see Testo and Sundue, 2016 for a possible Carboniferous origin 345 Ma). Our age for Gleicheniaceae is comparable to previous studies (287–263 Ma; Schuettpelz and Pryer, 2009; Qi et al., 2018, respectively) at around 271 Ma (range 270.6–273.1 Ma), while our age for Dipteridaceae is older than those previously recovered (196.1–239.7 Ma; Schuettpelz and Pryer, 2009; Rothfels et al., 2015; Testo and Sundue, 2016; Shen et al., 2018) at 274 Ma (range 272.9–279.0 Ma), and may reflect differences in topology in the dating analyses, although fossil Dipteridaceae suggest an origin in the early Triassic or late Paleozoic (Choo and Escapa, 2018).

Thus far, samples of Matoniaceae, the third family of Gleicheniales, have been lacking in large-scale nuclear fern phylogenies (Wickett et al., 2014; Rothfels et al., 2015; Qi et al., 2018; Shen et al., 2018;

One Thousand Plant Transcriptomes Initiative, 2019) and thus are not included here. Matoniaceae is a relatively small family with just four species in two genera (The Pteridophyte Phylogeny Group, 2016), but the omission of this family may alter the recovered topologies. Using plastid data, Pryer et al. (2004) and Schuettpelz and Pryer (2007) recovered Matoniaceae sister to Dipteridaceae; it is possible that additional sampling of this family in future work could help further resolve these relationships with nuclear data.

Cyatheales

Due in part to their low rates of molecular evolution (Korall et al., 2010), relationships among taxa in Cyatheales have been difficult to resolve (Korall et al., 2006). We recovered six families recognized by The Pteridophyte Phylogeny Group (2016) falling into two major clades, with one including Culcitaceae, Plagiogyriaceae, and Thyrsopteridaceae, and the other comprised of Cibotiaceae, Cyatheaeceae, and Dicksoniaceae (**Figure 2**). Our phylogeny agrees with the classification and phylogeny in Rothfels et al. (2015), Qi et al. (2018), and Shen et al. (2018), all of which used nuclear data. In contrast, our topology differs from the plastid-based phylogenies reconstructed by Schuettpelz and Pryer (2007) and Testo and Sundue (2016). Both plastid studies recovered a clade composed of Cyatheaceae and Cibotiaceae, with Dicksoniaceae sister to them, while the nuclear data recovered Dicksoniaceae and Cibotiaceae sister to one another, with Cyatheaceae sister to them. Our analysis of the first and second codon positions recovered Thrysopteridaceae sister to the clade composed of Cyatheaceae, Dicksonianceae, and Cibotiaceae, while our other analyses (including of the third codon position) found Thyrsopteridaceae sister to a clade consisting of Culcitaceae and Plagiogyriaceae (**Supplementary Figure 2**).

Stem ages of each tree fern family vary among studies (**Supplementary Figure 4**), likely a product of their abrupt shift to lower rates of molecular evolution compared to other ferns. We estimated that the split between the two major clades of tree ferns occurred around 164 Ma (range 110.8–192.5 Ma), although older (206–176 Ma, Schuettpelz and Pryer, 2009; Testo and Sundue, 2016, respectively) and younger (72–162 Ma, Rothfels et al., 2015; Qi et al., 2018, respectively) ages have been suggested. With a comprehensive sampling of 150 taxa in Cyatheales, Barrera-Redondo et al. (2018) estimated the stem age of Cyatheaceae, by far the largest family in the order, to be around 160 Ma, which is similar to both nuclear (157 Ma, Qi et al., 2018) and plastid (174–168 Ma, Schuettpelz and Pryer, 2009; Testo and Sundue, 2016, respectively) studies, but differs considerably from our estimate of 108.6 Ma (range 102.4–170.5 Ma). In our supplemental analysis including the *S. lepifera* transcriptome and additional fossil calibrations (see Section "Materials and Methods" and **Supplementary Figure 6**), we recovered a stem age of Cyatheaeceae at 121.1 Ma (range 119.8–124.6 Ma), which is still younger than previous estimates.

Polypodiales

Polypodiales is the largest of the fern orders and includes 80% of extant fern diversity (The Pteridophyte Phylogeny Group, 2016). Consistent with other studies (e.g., Testo and Sundue, 2016; Du et al., 2021), we find suborders Lindsaeineae and Saccolomatineae form a clade sister to the rest of Polypodiales. We found that Saccolomatineae was sister to a monophyletic Lindsaeineae (**Supplementary Figure 6**). Within Lindsaeineae, we recovered Cystodiaceae sister to a clade composed of Lindsaeaceae and Lonchitidaceae (**Figure 2** and **Supplementary Figure 6**). The monophyly and relationship of this clade to the rest of Polypodiales has been recovered by several recent studies (e.g., Schuettpelz and Pryer, 2007; Testo and Sundue, 2016; Qi et al., 2018; Du et al., 2021).

We consistently found Pteridaceae and Dennstaedtiaceae successively sister to the eupolypods, with Pteridaceae sister to Dennstaedtiaceae + the eupolypods (**Figure 2** and **Supplementary Figure 2**). This result was also seen in studies using nuclear loci (Rothfels et al., 2015; Shen et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019), whereas others (Pryer et al., 2004; Schuettpelz and Pryer, 2007; Testo and Sundue, 2016) recovered Dennstaedtiaceae and Pteridaceae successively sister to the eupolypods using plastid loci. Du et al. (2021), however, recovered Dennstaedtiaceae and Pteridaceae as a clade sister to the eupolypods. In Pteridaceae, the recovered major clades correspond to subfamilies as circumscribed in PPG I: Cheilanthoideae, Cryptogrammoideae, Parkerioideae, Pteridoideae, and Vittaroideae. Our estimated ages of Pteridaceae and Dennstaedtiaceae are similar to other studies (185–163 Ma, Schuettpelz and Pryer, 2009; Testo and Sundue, 2016, respectively) at 164 Ma (range 145.7–198.0 Ma) and 155 Ma (range 110.6–175.9 Ma; 325–146 Ma, Rothfels et al., 2015; Testo and Sundue, 2016, respectively), respectively. In agreement with Schuettpelz and Pryer (2009), we place the origin of the eupolypods in the early Cretaceous, at 113 Ma (range 79.2–123.6 Ma), although Du et al. (2021) found an older origin in the Jurassic at 160 Ma.

Eupolypods I (Polypodiineae)

In contrast to the relationships recovered in Kuo et al. (2011), Zhang and Zhang (2015), and Testo and Sundue (2016) we recovered Hypodematiaceae as sister to the rest of the eupolypods I (**Figure 2**), in agreement with Qi et al. (2018), Shen et al. (2018), and Du et al. (2021). Interestingly, Schuettpelz and Pryer (2007) and Rothfels et al. (2015) found that Hypodematiaceae and Didymochlaenaceae formed a clade, which was sister to the rest of the eupolypods I. Our analyses recovered Lomariopsidaceae and Nephrolepidaceae as successively sister to the remaining families in the eupolypods I (**Figure 2**), which is consistent with Testo and Sundue (2016), Qi et al. (2018), Shen et al. (2018), and Du et al. (2021), although Schuettpelz and Pryer (2007) found Lomariopsidaceae and Nephrolepidaceae form a clade rather than a grade.

The monophyly of Tectariaceae had varying levels of support throughout our analyses (**Supplementary Figure 2**). Some analyses (SCO85 AA ML, SCO60 NT ML), identified *Pteridrys cnemidaria* as sister to a clade consisting of Tectariaceae, Oleandraceae, Davalliaceae, and Polypodiaceae. Others (SCO75 AA MSC, SCO60 AA MSC, MCO NT MSC, MCO AA

MSC, codon positions one and two) recovered a clade consisting of *Pteridrys* and *Tectaria* sister to a clade of *Arthropteris*, Oleandraceae, Davalliaceae, and Polypodiaceae. Despite the different topologies recovered in our analyses, several recent studies have recovered Tectariaceae to be monophyletic (including all three genera sampled here) (Liu et al., 2014; Zhang et al., 2016; Zhou et al., 2018) although Zhou et al. (2018) suggest a new family (Pteridryaceae) be recognized. While Du et al. (2021) separate Tectariaceae from Pteridryaceae (Zhou et al., 2018) and Arthopteridaceae (suggested by Liu et al., 2013, but not recognized by The Pteridophyte Phylogeny Group, 2016), they found the three families form a clade sister to the remaining eupolypods I. Results for the remaining families within the eupolypods I (e.g., Dryopteridaceae and Polypodiaceae) were relatively consistent with those from previous studies and The Pteridophyte Phylogeny Group (2016).

Eupolypods II (Aspleniineae)

The relationships within the eupolypods II have been difficult to resolve due to heterogeneity in rates of molecular evolution among families and the rapid radiation of lineages in the group (e.g., Rothfels et al., 2012). In particular, the placement of Aspleniaceae varied amongst our analyses, with concatenated ML analyses finding Cystopteridaceae sister to the rest of the eupolypods II, with Aspleniaceae and Diplaziopsidaceae forming a clade nested within the eupolypods II (see **Supplementary Figure 2**), or a clade of Aspleniaceae and Diplaziopsidaceae sister to the rest of the eupolypods II (SCO85 NT MSC), although most MSC analyses resolved Aspleniaceae as sister to the rest of the eupolypods II with the remaining families forming a grade (**Supplementary Figure 2**). The latter relationship (depicted in **Figure 2**) reflects the result found in Testo and Sundue (2016), although Rothfels et al. (2012) resolved relationships of the eupolypods II by assessing and addressing model misspecifications in a Bayseian framework, resulting in Aspleniaceae nested within the eupolypods II rather than sister to the rest. They did, however, only use plastid data for these analyses; further work by Rothfels et al. (2015) using nuclear data support the topology of the eupolypods II in Rothfels et al. (2012). Using full plastome sequences, Wei et al. (2017) and Du et al. (2021) found Cystopteridaceae sister to the rest of the eupolypods II, with Aspleniaceae, Desmophlebiaceae (not sampled here), Hemidictyaceae (not sampled here), Diplaziopsidaceae, and Rhachidosoraceae forming a clade nested within eupolypods II (RHADD clade *sensu* Du et al., 2021). However, other nuclear datasets and particularly differing analyses, have contradicted this result (Qi et al., 2018; Shen et al., 2018), with concatenation-based ML analyses supporting the RHADD clade, while MSC analyses placed Aspleniaceae sister to the rest of the eupolypods II.

Challenges in Fern Phylogenetics

Despite recent advances in fern systematics and our ability to use thousands of markers throughout the genome to reconstruct phylogenetic relationships, challenges certainly remain in fully resolving the fern tree of life. In particular, we address the difficulties posed by lineage-specific rate heterogeneity, nuclear-plastid incongruence, and polyploidy.

Lineage-Specific Rate Heterogeneity

Differences in the rates of molecular evolution in different clades have made resolving relationships and estimating divergence times difficult across the entire tree of life (e.g., Beaulieu et al., 2015; Carruthers et al., 2020). In ferns, some clades of special interest regarding rate heterogeneity are Aspleniaceae (Rothfels et al., 2012; Wei et al., 2017), Cyatheales (Korall et al., 2010), Hymenophyllales (Schuettpelz and Pryer, 2006), Marattiales (Soltis et al., 2002), Osmundales (Rothfels et al., 2015), and vittaroid ferns (Rothfels and Schuettpelz, 2014; Grusz et al., 2016; see **Supplementary Figure 2**). Compared to the rest of the ferns, Cyatheales, Marattiales, and Osmundales have decelerated rates of molecular evolution (e.g., Soltis et al., 2002; Korall et al., 2010; Rothfels et al., 2015). Increased longevity and generation time has been posited as explanations for the rate heterogeneity seen in other plants (e.g., Gaut et al., 1992); for example, in angiosperms, annuals have been shown to have higher rates of molecular evolution than perennials and arborescent plants (Smith and Donoghue, 2008). A similar pattern has emerged in ferns; deceleration of molecular evolution has been linked to the evolution of arborescence in tree ferns (Korall et al., 2010). The cause for decreased rates in Osmundales and Marattiales may be linked to their long generation times, with individual clones of *Osmunda* suggested to live more than 1000 years (Wagner et al., 1978). In contrast, Aspleniaceae, Trichomanoideae, and vittaroid ferns have experienced accelerated rates of molecular evolution (Schuettpelz and Pryer, 2006; Rothfels et al., 2012; Rothfels and Schuettpelz, 2014; Grusz et al., 2016; Testo and Sundue, 2016). Shifts in life history, morphology, and/or ecological niche may explain changes in their rates of molecular evolution. For example, the biology of vittaroid ferns differs from their closest relatives, as they are tropical, simple-leaved mostly epiphytic plants (as opposed to arid-adapted, dissected-leaved, epipetric plants) and experience a 4.3 times faster rate of molecular evolution than cheilanthoid ferns (Rothfels and Schuettpelz, 2014). Given the range in rates across the fern phylogeny and the implications for divergence time estimation and phylogeny reconstruction, lineage-specific rate heterogeneity is one of the biggest challenges in modern fern phylogenetics.

Nuclear–Plastid Incongruence

The vast majority of plant (and specifically fern) systematics to date has relied on plastid loci (e.g., Chase et al., 1993; Pryer et al., 2001, 2004; Schuettpelz and Pryer, 2007; Testo and Sundue, 2016). However, the plastid is a maternally inherited, single linkage group, and therefore only captures part of evolutionary history (Lynch, 2007; but see Gonçalves et al., 2019 for a suggestion that all the loci in the plastid are not truly linked). One of the biggest challenges for nuclear phylogenetics in ferns has been working with single-copy (or low-copy) loci. Due to multiple rounds of polyploidy and tandem duplications, it is unlikely that any locus is truly single-copy in all fern genomes (although most gene families quickly revert back to

single copy, De Smet et al., 2013; Li et al., 2016). This has made it particularly challenging to identify and sequence single- or low-copy nuclear genes throughout ferns (Rothfels et al., 2013). Furthermore, resolving recalcitrant relationships among ferns requires several unlinked, bi-parentally inherited markers, which cannot be accomplished with just plastid markers. Recent advances in computational tools have made it possible to analyze multi-copy loci datasets such as ASTRAL-PRO (Zhang et al., 2020). Large-scale work with the nuclear genome has been relatively recent, but the results have been surprisingly congruent with plastid phylogenies (Wickett et al., 2014; Rothfels et al., 2015; Qi et al., 2018; Shen et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019). The same problematic nodes and relationships in plastid phylogenies have been replicated in nuclear phylogenies (**Figure 3**), though, as discussed at length above, some differences are consistent between plastid and nuclear phylogenies (e.g., the sister group to the eupolypods, **Figure 3D**).

Polyploidy

Manton's (1950) work revealed that ferns have high chromosome numbers (and later large genome sizes, e.g., Clark et al., 2016) and subsequent studies have confirmed that polyploidy is a prevalent phenomenon in extant ferns. The extent to which polyploidy has occurred throughout the phylogeny, particularly at deeper nodes, has only recently been explored (Huang et al., 2019; One Thousand Plant Transcriptomes Initiative, 2019; Li and Barker, 2020). Paralogs from WGDs make it difficult to assess orthology across multiple species using only a reciprocal homology search algorithm, especially when duplications are recurrent. Several clustering programs (e.g., Li et al., 2003; Yang and Smith, 2014; Emms and Kelly, 2015, 2019) and downstream phylogenetic methods (e.g., Zhang et al., 2020) have been developed to account for the presence of paralogs in phylogenomic datasets. While polyploidy is a prevalent process in ferns that can make it difficult to reconstruct evolutionary histories, new methods and the advent of high-throughput sequencing techniques have been instrumental for tackling this challenge.

Phylogenetic Placement of Whole Genome Duplications

Polyploidy has long been recognized as an important evolutionary mechanism in plants (Manton, 1950; Stebbins, 1950; Klekowski and Baker, 1966), although our view of the role of polyploidy in plant evolution has shifted dramatically throughout the past century (reviewed by Soltis et al., 2014b). One of the major debates in the polyploid literature is the role of genome duplications in diversification. Polyploid plants have been suggested to diversify at faster rates than diploids (Soltis et al., 2014a; Tank et al., 2015; Landis et al., 2018; Román-Palacios et al., 2020; but see Mayrose et al., 2011; Arrigo and Barker, 2012; Mayrose et al., 2015 for contrasting results). Although we do not address diversification here, it is of note that the two inferred WGDs shared by the largest number of taxa are at the base of the most diverse fern clades. By placing WGDs on the phylogeny constructed by Qi et al. (2018), Huang et al. (2019) found that clades of ferns that underwent two or more WGD events had higher diversification rates than other clades. They propose that the three rounds of fern radiations (Rothwell, 1987; Rothwell and Stockey, 2008; Schuettpelz and Pryer, 2009) correspond to WGDs at the base of the leptosporangiates, core leptosporangiates, and Polypodiales (Huang et al., 2019). While this is an enticing possibility, further testing will be required to explore these patterns, such as running these analyses with the varying hypotheses of the placement of WGDs in ferns posited by Huang et al. (2019), One Thousand Plant Transcriptomes Initiative (2019), and this study. Furthermore, generating a phylogeny with more comprehensive sampling or placing WGDs on an existing tree with extensive sampling may be more appropriate for these diversification analyses. Additional work will be required to determine if there is a relationship between polyploidy and diversification in ferns.

Paralog age distributions (K_S plots) have been the primary tool of choice for inferring WGDs; briefly, the synonymous distances between genes within a gene family (often defined by a clustering algorithm such as OrthoMCL, Li et al., 2003) are calculated and a histogram of the frequency or count of these distances is plotted. Peaks in the K_S distribution are interpreted as the result of large-scale duplication events (i.e., polyploidy; Lynch and Conery, 2000; Vanneste et al., 2013). By rigorously testing the applicability of K_S plots to infer ancient WGDs (paleopolyploidy events), Tiley et al. (2018) found that K_S plots should be used primarily as a hypothesis-building tool and should be supplemented with other lines of evidence. One of the biggest challenges in using paralog K_S values relative to speciation events (ortholog divergence) to infer WGDs is lineage-specific rate heterogeneity. Several methods have been developed to account for these rate differences (e.g., Barker et al., 2008; Sensalari et al., 2022) to yield corrected K_S values, which may reveal different phylogenetic placements of WGDs. Interestingly, when we compared WGD inferences from corrected and uncorrected K_S most of the WGD inferences had identical phylogenetic placements (12 identical phylogenetic placements out of 20, including LONCH; **Figure 4**). As expected, correcting K_S in lineages with shifts in molecular evolutionary rates such as the tree ferns and Osmundaceae resulted in different placements of the WGD; however, phylogenomic approaches generally supported the uncorrected K_S placements, rather than the corrected K_S placement. Even after correcting for rate heterogeneity, there were some cases where K_S between one-to-one orthologs were ambiguous (**Figure 4** and **Supplementary Figures 8**, **9**), likely due to rapid divergences (e.g., GLEI, DIPT.1) or differences in rates of molecular evolution (e.g., CYATH.1-3).

Recent advances in phylogenomic methods have been applied to placing WGDs on a species tree (e.g., Li et al., 2015; McKain et al., 2016; Li Z. et al., 2018). These methods examine the proportion of gene trees with a shared duplication event at the nodes of a species tree; many shared duplications support a large-scale duplication event. One of the limitations to MAPS is that the software requires a ladderized tree as input, which requires users to subsample their phylogeny. To test whether there was an effect of the sample choice, we ran MAPS using

each *P. nudum* sample in the PSIL analysis (see **Supplementary Appendix 5**), which resulted in nearly identical proportions of gene duplicated gene trees at each node in the species tree (e.g., <1% difference). Similarly, using different lineages may impact the inference of WGDs. We tested whether this may impact our results by using different *Equisetum* species in the EQUI analysis (see **Supplementary Appendix 5**) and again found minimal changes to the proportion of duplicated gene trees at each node and no change in the inference of the WGD event. While we did not find that taxon selection had an impact on the inference of the phylogenetic placement of WGDs, further work should be considered to determine if and how this process can affect inferences. Both K_S plots and gene tree-based methods are accessible for data generated from transcriptomes. Full genome assemblies may further be interrogated to assess duplication events using synteny (i.e., gene order; Li F. W. et al., 2018; Krabbenhoft et al., 2021). To infer WGDs in this study, we used a combination of synonymous distance between paralogs within species and one-to-one orthologs between species, a phylogenomic method implemented in MAPS and rigorous statistical testing to place WGDs on the fern phylogeny. In general, these lines of evidence were consistent.

Overall, the placement of WGDs identified here was similar to previous estimates (Huang et al., 2019; One Thousand Plant Transcriptomes Initiative, 2019). Of the 18 large-scale WGDs we inferred, 12 had identical placements in the phylogeny produced by Huang et al. (2019). Below we discuss events which are novel or whose placement differed between our study and previous inferences, starting at the base of phylogeny. We emphasize that the inferred events should be treated as hypotheses and require further study, including whole genome analyses.

In contrast with One Thousand Plant Transcriptomes Initiative (2019), we did not find a duplication shared by all ferns, although the placement of our duplications suggest that nearly all ferns sampled have a polyploid history (**Figure 2**). The placement of the duplication at the base of Ophioglossales found in One Thousand Plant Transcriptomes Initiative (2019) was not recovered in our study, although a WGD shared by *Ophioglossum* was identified (OPHIO.2). Unlike Huang et al. (2019), we did not find that this duplication was shared by *Ophioglossum* and *Ophioderma*; rather we found that *Ophioderma* underwent a separate duplication event (OPHIO.3). These inferences are supported by several lines of evidence including corrected and uncorrected K_S values and MAPS. In Psilotales, we identified a novel duplication in *Tmesipteris* relative to *Psilotum,* with support from both K_S analyses, MAPS, and karyotypes. All three studies agree on the placement of a WGD at the base of the leptosporangiates excluding Osmundales (LEPTO). While Huang et al. (2019) and One Thousand Plant Transcriptomes Initiative (2019) found that Osmundales underwent a separate duplication event our analyses suggested that the uncorrected K_S inference was ambiguous, the corrected K_S suggested an event shared by all leptosporangiate ferns, and MAPS failed to identify a significant difference in the observed and null distribution of duplicated gene trees. While One Thousand Plant Transcriptomes Initiative (2019) did not identify duplication events in Hymenophyllales and Gleicheniales (only two taxa were included from these groups in their phylogeny), our findings are similar to Huang et al. (2019), although we found novel WGDs on the branches leading to *Dipteris* (DIPT2) and Trichomanoideae (HYMN), both of which had taxa that were sampled in Huang et al. (2019).

Unlike One Thousand Plant Transcriptomes Initiative (2019), we did not infer a WGD at the base of Schizaeales, and the inferred event may instead be representative of LEPTO. The uncorrected K_S plots for *Azolla* spp. were ambiguous but the corrected K_S clearly identified the inferred WGD shared by *Azolla* based on MAPS; syntenic analysis further support this placement (Li F. W. et al., 2018). Although both Huang et al. (2019) and One Thousand Plant Transcriptomes Initiative (2019) identified events on the backbone of the phylogeny either shared by Cyatheales + Polypodiales or Salviniales, Cytheales, and Polypodiales, respectively, we did not identify a WGD event occurring at either of these locations. Within Cyatheales, Huang et al. (2019), One Thousand Plant Transcriptomes Initiative (2019), and Huang et al. (2022) identified a WGD shared by *Alsophila*, although our analyses suggests instead it is shared by Cibotiaceae, Cyatheaceae, and Dicksoniaceae (CYATH.3). Although Huang et al. (2022) generated a chromosome-scale genome for *A. spinulosa*, they used the same approaches used here (K_S analyses and MAPS), although further syntenic evidence will be required to verify the placement of CYATH.3, especially given the uniformly high chromosome numbers in Cyatheales (see Section "Results") and slower rates of molecular evolution in the tree ferns. Along the backbone of the phylogeny, we inferred a WGD shared by Polypodiales excluding Lindsaeineae and Saccolmatineae (POLY), the placement of which agrees with One Thousand Plant Transcriptomes Initiative (2019). Within Pteridaceae, we found one shared event at the base of the vittaroid ferns (PTER.3) that was not identified in One Thousand Plant Transcriptomes Initiative (2019). We did not find evidence of the WGD on the branch leading the eupolypods, which was recovered in Huang et al. (2019). Although we did identify a significantly greater proportion of duplicated gene trees than in the null distribution at the MRCA of the eupolypods, the observed distribution was not consistent with positive simulations of a WGD.

Gene Retention Following Whole Genome Duplications and Large-Scale Duplications

As with most other plants, we found that the rate of duplicate gene retention is low following paleopolyploidy or large-scale gene duplications, with an average of around 11.97% of genes remaining in duplicate. On average, duplicate genes have a retention rate around 10% in plants (Tiley et al., 2016), but vary depending on the age of the duplication (Li et al., 2016), with up to 76.3% of genes duplicated in *Glycine max* (Tiley et al., 2016) which underwent a recent duplication event. Genomes which have undergone more recent duplications

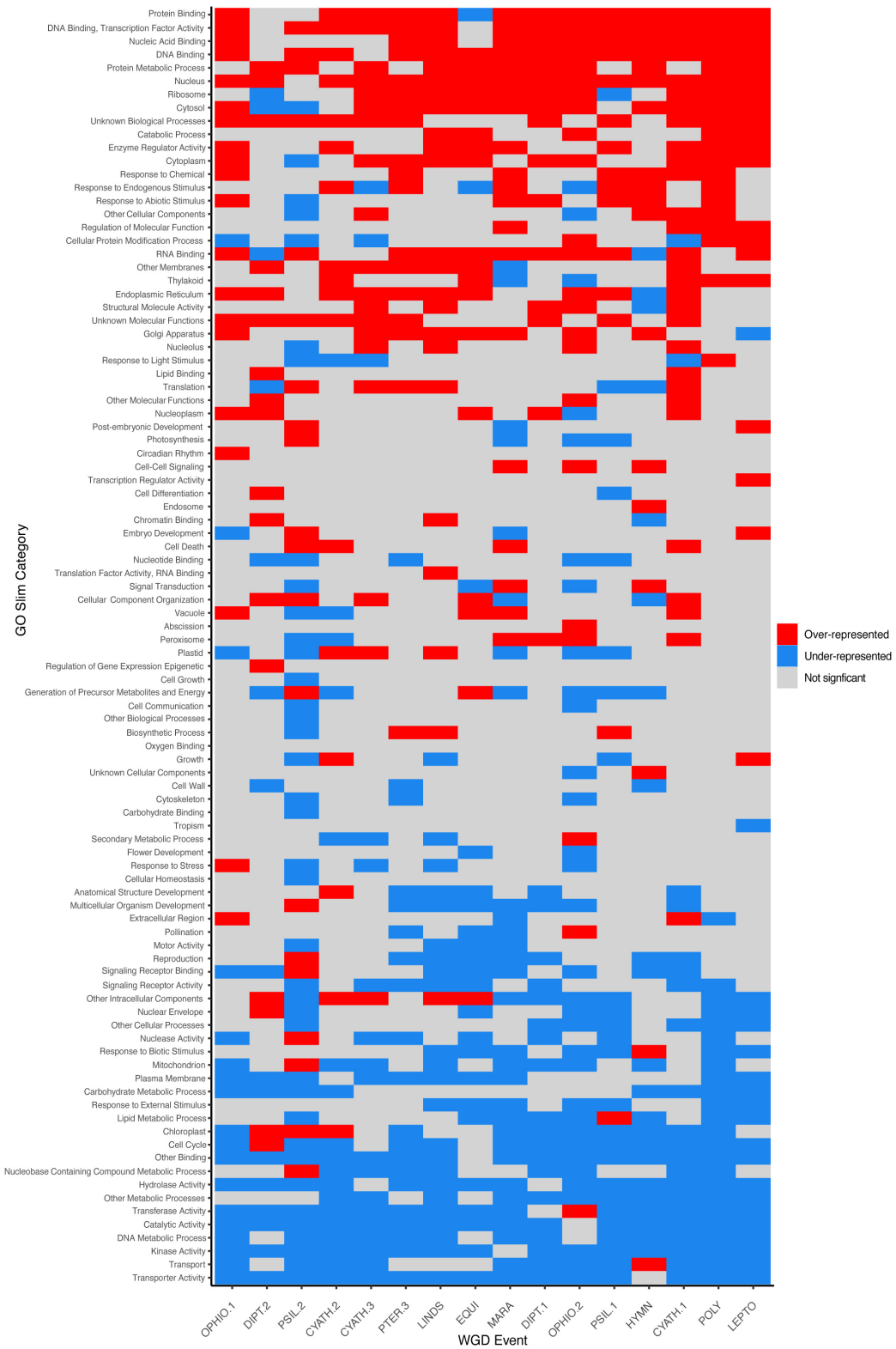

FIGURE 5 | Gene retention in ferns following large-scale duplications is biased. Gene ontology (GO) Slim composition of retained duplicates (paralogs) from 17 of the 18 large-scale whole genome duplications (WGDs) identified in this study are shown, where categories that are over-represented in paralogs are in red (Chi-squared residuals > 2) and under-presented as blue (Chi-squared residuals < -2). Terms with non-significant residuals are gray. Events are sorted by ascending median Ks values. Note that SALV is not shown here; while MAPS identified a significant proportion of duplicated gene trees, uncorrected K_S plots were ambiguous, although a WGD was identified with syntenic analyses in Li F. W. et al. (2018).

tend to show higher retention rates than older duplications, with the proportion of the gene families duplicated rapidly decaying with the age of events (Li et al., 2016). Given that the gene families we analyzed were not restricted to "core gene families" as in Li et al. (2016) and that the signal of WGDs becomes diluted as synonymous divergence increases, we were not able to replicate these findings in ferns. While our method of identifying paralogs from WGDs using K_S plots has been described and used before (e.g., Li Z. et al., 2018), it is important to note that other types of duplication events other than WGDs may contribute paralogs to the distribution. In some cases, particularly in older WGDs, peaks may be difficult to distinguish from the background processes affecting duplicated genes. Although we assigned paralogs to WGD events based on their probability of belonging to certain components in mixture models (see Section "Materials and Methods"), we may be including other duplicated genes that may not have arisen from WGDs.

While most duplicated genes are rapidly lost (Lynch and Conery, 2000), the functions and types of genes retained in duplicate following independent WGD events are similar (e.g., Barker et al., 2008; Freeling, 2009; Li et al., 2016; Li Z. et al., 2018; **Figure 5**). According to the DBH (Papp et al., 2003), the loss of duplicated copies of some but not all partners in an interaction would alter the stoichiometry of the product/network and therefore the ultimate function may be disrupted. Compared to full transcriptomes, we found that genes retained in duplicate have functions that are over-represented in binding processes (DNA binding, transcription factor activity, nucleic acid binding, protein binding, RNA binding, nucleotide binding), responses to stimuli (response to chemical, response to endogenous stimulus, response to abiotic stimulus), and certain organelles involved in gene and protein production and processing (nucleus, ribosome, endoplasmic reticulum, Golgi apparatus) (**Figure 5**). Binding processes (DNA binding, RNA binding, nucleic acid binding) were also found to be over-represented in paralogs retained in hexapods (Li Z. et al., 2018) and intermediate-age duplicates in angiosperms (Li et al., 2016). Processes involved in transport (transporter activity, transport, nuclear envelope), organelles derived from endosymbiotic events (mitochondrion, chloroplast), and signaling (signaling receptor binding and activity) were under-represented in paralogs (**Figure 5**). Similar patterns of gene loss are again seen in hexapods (transport, mitochondrion, nuclear envelope, Li Z. et al., 2018), and angiosperms (transport, transporter activity, mitochondrion, Li et al., 2016). While lineage-specific variation in patterns of gene retention is present (see below), some patterns appear to be conserved over deep evolutionary time among the kingdoms of life.

While the overall pattern of gene functions is similar across independent duplication events, there were several instances that suggest that other factors may also drive lineage-specific differences in retained duplicates. For example, retained duplicates following the inferred Psilotales (PSIL.1) and *Tmesipteris* (PSIL.2) events were over-representative of genes involved with endosymbiotic organelles (chloroplast, mitochondrion) and signal receptor binding, while duplicates from most other events were under-represented in those categories (**Figure 5**). Similarly, genes with functions related to the ribosome and cytosol were under-represented in retained duplicates following the *Dipteris* event (DIPT.2) (**Figure 5**), although this may be due to low sample size. In Asteraceae, Barker et al. (2008) found several GO Slim categories that were under-represented in duplicates following independent WGDs which we found to be over-represented in fern duplicates (e.g., DNA or RNA binding, nucleus) whereas others were over-represented (e.g., cytosol, protein metabolic process) or under-represented (e.g., chloroplast) in both analyses. Furthermore, the age of the inferred WGD may impact the function of gene sets retained, as in angiosperms (Li et al., 2016); for instance, genes with functions related to translation and metabolic processes were over-represented in older WGDs, but under-represented in more recent events and single-copy gene families. Gene function may therefore be an important factor in the long-term survival of duplicated genes.

While genic diploidization (i.e., the process of removal and loss of genes by molecular mechanisms, Li et al., 2021) is clearly occurring in ferns, perhaps through pseudogenization/gene silencing as hypothesized by Haufler and Soltis (1986), cytological diploidization may be slow to follow. Unlike other plants, chromosome number and genome size are positively correlated in ferns (Clark et al., 2016) and, taken with the relative stasis in genome size across ferns, may suggest that chromosomes are retained following WGDs rather than lost during genomic reorganization in angiosperms. A similar pattern has been observed in the catostomid fish *Myxocyprinus asiaticus*, which shows remarkable genome subgenome stability and retained synteny over 50 MY following a WGD (Krabbenhoft et al., 2021). In contrast, one subgenome tends to dominate in polyploid plants (Alger and Edger, 2020) and rearrangements drastically alter gene order and retention of synteny (Zhao and Schranz, 2018; but see VanBuren et al., 2020 for an example of subgenome stability in plants following a WGD 1 Ma). Whether similar patterns of genome reorganization or stability are present in ferns remains uncertain and an active area of research.

Importantly, transcriptomes are temporal and spatial "snapshots" of gene expression. Many of the transcriptomes used in this study were derived from young leaf material, although some are from other tissue types (e.g., fertile pinnae, gametophytes). Not all genes in the genome will be expressed in every tissue and therefore transcriptomes from single tissues will likely not represent the entirety of gene-space in a genome. The presence or absence of genes in a transcriptome may not necessarily correlate to the presence or absence of that gene in the genome but could rather be a product of differences in expression between tissues and time. The analysis presented here is one of the first to tackle functional gene retention following WGDs in ferns, but additional analyses will be needed to explore whether these patterns of gene retention are specific to transcriptomic study. As new genome assemblies are becoming available (e.g., *Adiantum, Alsophila, Ceratopteris*) patterns of gene retention should be further explored in more complete gene spaces.

CONCLUSION

Ferns are a ubiquitous part of global floras and occupy a pivotal evolutionary position sister to seed plants, yet genomic resources for this group are lacking. Using publicly available transcriptome data, we addressed fundamental questions about the evolution of ferns, particularly the nuclear phylogenetic backbone, the placement and number of WGDs along the phylogeny, and the fates of duplicated genes following WGDs. Despite using thousands of loci, areas of the fern phylogeny remain contentious, including the sister group to the leptosporangiate ferns, the relationships among Gleicheniales and Hymenophyllales, the sister group to the eupolypods, and the placement of Aspleniaceae within the eupolypods II. We recovered a number of paleopolyploidy events throughout the phylogeny and found that functions of genes retained in duplicate following polyploidy are largely convergent, with duplicate genes of similar function retained between events. Given the high number of polyploidy events in ferns, questions related to fern evolution must account for WGDs. As sequencing costs continue to decrease and genomics becomes more accessible, ferns will no longer remain one of the final frontiers in plant genomics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/jessiepelosi/ferntxms.

AUTHOR CONTRIBUTIONS

JP planned and designed the research. JP and EK analyzed the data. All authors wrote and revised the manuscript, read, and approved the final manuscript.

FUNDING

This work was funded by NSF DEB #1844930 to ES.

ACKNOWLEDGMENTS

We thank Nathan Catlin, Cody Howard, Qinyin Ling, Kasey Pham, Lindsey Riibe, and Weston Testo for their assistance and comments on this project; the Krabbenhoft Lab at the University at Buffalo for access to their computing resources; Amanda Grusz for her help to identify species in **Figure 1**; and the Sessa and Barbazuk labs for insightful discussions. We also thank the reviewers and editor for their constructive comments on this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.882441/full#supplementary-material

REFERENCES

Alger, E. I., and Edger, P. P. (2020). One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr. Opin. Plant Biol.* 54, 108–113. doi: 10.1016/j.pbi.2020.03.004

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Arrigo, N., and Barker, M. S. (2012). Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.* 15, 140–146. doi: 10.1016/j.pbi.2012.03.010

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., et al. (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. doi: 10.1093/molbev/msn187

Barrera-Redondo, J., Ramírez-Barahona, S., and Eguiarte, L. E. (2018). Rates of molecular evolution in tree ferns are associated with body size, environmental temperature, and biological productivity. *Evolution* 72, 1050–1062. doi: 10.1111/evo.13475

Beaulieu, J. M., O'Meara, B. C., Crane, P., and Donoghue, M. J. (2015). Heterogeneous rates of molecular evolution and diversification could explain the triassic age estimate for angiosperms. *Syst. Biol.* 64, 869–878. doi: 10.1093/sysbio/syv027

Benaglia, T., Cheauveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.* 32, 1–29.

Berardini, T. Z., Mundodi, S., Reiser, R., Huala, E., Garcia-Hernandez, M., Zhang, P., et al. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.* 135, 1–11.

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

Buggs, R. J. A., Chamala, S., Wu, W., Gao, L., May, G. D., Schnable, P. S., et al. (2010a). Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Mol. Ecol.* 19(Suppl. 1), 132–146. doi: 10.1111/j.1365-294X.2009.04469.x

Buggs, R. J. A., Elliott, N. M., Zhang, L., Koh, J., Viccini, L. F., Soltis, D. E., et al. (2010b). Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytol.* 186, 175–183. doi: 10.1111/j.1469-8137.2010.03205.x

Buggs, R. J. A., Zhang, L., Miles, N., Tate, J. A., Gao, L., Wei, W., et al. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr. Biol.* 21, 551–556. doi: 10.1016/j.cub.2011.02.016

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Carruthers, T., Sanderson, M. J., and Scotland, R. W. (2020). The implications of lineage-specific rates for divergence time estimation. *Syst. Biol.* 69, 660–670. doi: 10.1093/sysbio/syz080

Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H., Mishler, B. D., et al. (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. *Ann. Mol. Bot. Gard.* 80:528. doi: 10.2307/2399846

Chaudhuri, P., and Marron, J. S. (1999). SiZer for exploration of structures in curves. *J. Am. Stat. Assoc.* 94, 807–823. doi: 10.1080/01621459.1999.10474186

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., et al. (2018). 10KP: a phylodiverse genome sequencing plan. *Gigascience* 7, 1–9. doi: 10.1093/gigascience/giy013

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65, 997–1008. doi: 10.1093/sysbio/syw037

Choo, T. Y. S., and Escapa, I. H. (2018). Assessing the evolutionary history of the fern family Dipteridaceae (Gleicheniales) by incorporating both extant and

extinct members in a combined phylogenetic study. *Am. J. Bot.* 105, 1315–1328. doi: 10.1002/ajb2.1121

Clark, J., Hidalgo, O., Pellicer, J., Liu, H., Marquardt, J., Robert, Y., et al. (2016). Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytol.* 210, 1072–1082. doi: 10.1111/nph.13833

Cock, P. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermin, L. S., et al. (2020). GHOST: recovering historical signal from heterochously evolved sequence alignments. *Syst. Biol.* 69, 249–264. doi: 10.1093/sysbio/syz051

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2898–2903. doi: 10.1073/pnas.1300127110

Defoort, J., Van de Peer, Y., and Carretero-Paulet, L. (2019). The evolution of gene duplicates in angiosperms and the impact of protein-protein interactions and the mechanism of duplication. *Genome Biol. Evol.* 11, 2292–2305. doi: 10.1093/gbe/evz156

Dong, S., Xiao, Y., Kong, H., Feng, C., Harris, A. J., Yan, Y., et al. (2019). Nuclear loci developed from multiple transcriptomes yield high resolution in phylogeny of scaly tree ferns (Cyatheaceae) from China and Vietnam. *Mol. Phylogenet. Evol.* 139:106567. doi: 10.1016/j.ympev.2019.106567

Du, X.-Y., Lu, J.-M., Zhang, L.-B., Wen, J., Kuo, L.-Y., Mynssen, C. M., et al. (2021). Simultaneous diversification of Polypodiales and angiosperms in the Mesozoic. *Cladistics* 37, 518–539. doi: 10.1111/cla.12457

Ebihara, A., Dubuisson, J.-Y., Iwatsuki, K., Hennequin, S., and Ito, M. (2006). A taxonomic revision of Hymenophyllaceae. *Blum. J. Plant Tax Plant Geogr.* 51, 221–280. doi: 10.3767/000651906X622210

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

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

Finigan, P., Tanurdzic, M., and Martienssen, R. A. (2012). "Origins of novel phenotypic variation in polyploids," in *Polyploidy and Genome Evolution*, eds P. S. Soltis and D. E. Soltis (Cham: Springer), 57–76.

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. doi: 10.1146/annurev.arplant.043008.092122

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

Fujiwara, T., Liu, H., Meza-Torres, E. I., Morero, R., Vega, A. J., Liang, Z., et al. (2021). Evolution of genome space occupation in ferns: linking genome diversity and species richness. *Ann. Bot.* [Epub ahead of print]. doi: 10.1093/aob/mcab094

Gastony, G. J., and Yatskievych, G. (1992). Maternal inheritance of the chloroplast and mitochondrial genomes in cheilanthoid ferns. *Am. J. Bot.* 79, 716–722. doi: 10.1002/j.1537-2197.1992.tb14613.x

Gaut, B. S., Muse, S. V., Clark, W. D., and Clegg, M. T. (1992). Relative rates of nucleotide substitution at the rbcL locus of monocotyledonous plants. *J. Mol. Evol.* 35, 292–303. doi: 10.1007/BF00161167

Gonçalves, D. J. P., Simpson, B. B., Ortiz, E. M., Shimizu, G. H., and Jansen, R. K. (2019). Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Mol. Phylogenet. Evol.* 138, 219–232. doi: 10.1016/j.ympev.2019.05.022

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Grusz, A. L., Rothfels, C. J., and Schuettpelz, E. (2016). Transcriptome sequencing reveals genome-wide variation in molecular evolutionary rate among ferns. *BMC Genomics* 17:692. doi: 10.1186/s12864-016-3034-2

Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., et al. (2019). Updated genome assembly of Ginkgo biloba. *GigaScience Database* [Epub ahead of print]. doi: 10.5524/100613

Guillon, J. M., and Raquin, C. (2000). Maternal inheritance of chloroplasts in the horsetail *Equisetum variegatum* (Schleich.). *Curr. Genet.* 37, 53–56. doi: 10.1007/s002940050008

Haufler, C. H., Pryer, K. M., Schuettpelz, E., Sessa, E. B., Farrar, D. R., Moran, R., et al. (2016). Sex and the single gametophyte: revising the homosporous vascular plant life cycle in light of contemporary research. *Bioscience* 66, 928–937. doi: 10.1093/biosci/biw108

Haufler, C. H., and Soltis, D. E. (1986). Genetic evidence suggests that homosporous ferns with high chromosome numbers are diploid. *Proc. Natl. Acad. Sci. U.S.A.* 83, 4389–4393. doi: 10.1073/pnas.83.12.4389

Hidalgo, O., Pellicer, J., Christenhusz, M. J. M., Schneider, H., and Leitch, I. (2017). Genomic gigantism in the whisk-fern family (Psilotaceae): *Tmesipteris obliqua* challenges record holder *Paris japonica*. *Bot. J. Linn. Soc.* 183, 509–514. doi: 10.1093/botlinnean/box003

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). Ufboot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Howe, K. L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., et al. (2020). Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* 48, D689–D695. doi: 10.1093/nar/gkz890

Huang, C.-H., Qi, X., Chen, D., Qi, J., and Ma, H. (2019). Recurrent genome duplication events likely contributed to both the ancient and recent rise of ferns. *J. Integr. Plant Biol.* 62, 433–455. doi: 10.1111/jipb.12877

Huang, X., Wenling, W., Gong, T., Wickell, D., Kuo, L.-Y., Zhang, X., et al. (2022). The flying spider-monkey tree fern genome provides insights into fern evolution and arborescence. *Nat. Plants* 8, 500–512. doi: 10.1038/s41477-022-01146-6

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kinosian, S. P., Pearse, W. D., and Wolf, P. G. (2020). Cryptic diversity in the model fern genus *Ceratopteris* (Pteridaceae). *Mol. Phylogenet. Evol.* 152:106938. doi: 10.1016/j.ympev.2020.106938

Klekowski, E. J., and Baker, H. G. (1966). Evolutionary significance of polyploidy in the pteridophyta. *Science* 153, 305–307. doi: 10.1126/science.153.3733.305

Knie, N., Fischer, S., Grewe, F., Polsakiewicz, M., and Knoop, V. (2015). Horsetails are the sister group to all other monilophytes and Marattiales are sister to leptosporangiate ferns. *Mol. Phylogenet. Evol.* 90, 140–149. doi: 10.1016/j.ympev.2015.05.008

Korall, P., Pryer, K. M., Metzgar, J. S., Schneider, H., and Conant, D. S. (2006). Tree ferns: monophyletic groups and their relationships as revealed by four protein-coding plastid loci. *Mol. Phylogenet. Evol.* 39, 830–845. doi: 10.1016/j.ympev.2006.01.001

Korall, P., Schuettpelz, E., and Pryer, K. M. (2010). Abrupt deceleration of molecular evolution linked to the origin of arborescence in ferns. *Evolution* 64, 2786–2792. doi: 10.1111/j.1558-5646.2010.01000.x

Krabbenhoft, T. J., MacGuigan, D. J., Backenstose, N. J. C., Waterman, H., Lan, T., Pelosi, J. A., et al. (2021). Chromosome-level genome assembly of chinese sucker (*Myxocyprinus asiaticus*) reveals strongly-conserved synteny following a catostomid-specific whole genome duplication. *Genome Biol. Evol.* 13:evab190. doi: 10.1093/gbe/evab190

Kuo, L.-Y., Li, F.-W., Chiou, W.-L., and Wang, C.-N. (2011). First insights into fern matK phylogeny. *Mol. Phylogenet. Evol.* 59, 556–566. doi: 10.1016/j.ympev.2011.03.010

Kuo, L.-Y., Qi, X., Ma, H., and Li, F.-W. (2018a). Order-level fern plastome phylogenomics: new insights from Hymenophyllales. *Am. J. Bot.* 105, 1545–1555. doi: 10.1002/ajb2.1152

Kuo, L.-Y., Tang, T.-Y., Li, F.-W., Su, H.-J., Chiou, W.-L., Huang, Y.-M., et al. (2018b). Organelle genome inheritance in *Deparia ferns* (Athyriaceae, Aspleniineae, Polypodiales). *Front. Plant Sci.* 9:486. doi: 10.3389/fpls.2018.00486

Landis, J. B., Soltis, D. E., Li, Z., Marx, H. E., Barker, M. S., Tank, D. C., et al. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* 105, 348–363. doi: 10.1002/ajb2.1060

Lantz, T. C., Rothwell, G. W., and Stockey, R. A. (1999). *Conantiopteris schuchmanii*, gen. et sp. nov., and the role of fossils in resolving the phylogeny of Cyatheaceae s.l. *J. Plant Res.* 112, 361–381. doi: 10.1007/PL00013890

Lehtonen, S., Poczai, P., Sablok, G., Hyvönen, J., Karger, D. N., and Flores, J. (2020). Exploring the phylogeny of the marattialean ferns. *Cladistics* 36, 569–593. doi: 10.1111/cla.12419

Leitch, I. J., Johnston, E., Pellicer, J., Hidalgo, O., and Bennett, M. D. (2019). *Pteridophyte DNA C-Values Database Release 6.0*. Available online at: https://cvalues.science.kew.org/ (accessed April 2022).

Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* 4, 460–472. doi: 10.1038/s41477-018-0188-8

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

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

Li, Z., and Barker, M. S. (2020). Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *Gigascience* 9:giaa004. doi: 10.1093/gigascience/giaa004

Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., and De Smet, R. (2016). Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28, 326–344. doi: 10.1105/tpc.15.00877

Li, Z., McKibben, M. T. W., Finch, G. S., Blischak, P. D., Sutherland, B. L., and Barker, M. S. (2021). Patterns and processes of diploidization in land plants. *Annu. Rev. Plant Biol.* 72, 387–410. doi: 10.1146/annurev-arplant-050718-100344

Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., et al. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4713–4718. doi: 10.1073/pnas.1710791115

Liu, H.-M., He, L.-J., and Schneider, H. (2014). Towards the natural classification of tectarioid ferns: confirming the phylogenetic relationships of Pleocnemia and Pteridrys (eupolypods I). *J. Syst. Evol.* 52, 161–174. doi: 10.1111/jse.12073

Liu, H.-M., Jiang, R.-H., Guo, J., Hovenkamp, P., Perrie, L. R., Shepherd, L., et al. (2013). Towards a phylogenetic classification of the climbing fern genus Arthopteris. *Taxon* 62, 688–700. doi: 10.12705/624.26

Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates, Inc.

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Manton, I. (1950). *Problems of Cytology and Evolution in the Pteridophyta*. Cambrdige, MA: Oxford University Press.

Marchant, D. B., Sessa, E. B., Wolf, P. G., Heo, K., Barbazuk, W. B., Soltis, P. S., et al. (2019). The C-Fern (*Ceratopteris richardii*) genome: insights into plant genome evolution with the first partial homosporous fern genome assembly. *Sci. Rep.* 9:18181. doi: 10.1038/s41598-019-53968-8

Marchant, D. B., Soltis, D. E., and Soltis, P. S. (2016). Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytol.* 212, 708–718. doi: 10.1111/nph.14069

Marks, R. A., Hotaling, S., Frandsen, P. B., and VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* 7, 1571–1578. doi: 10.1038/s41477-021-01031-8

May, M. R., Contreras, D. L., Sundue, M. A., Nagalingum, N. S., Looy, C. V., and Rothfels, C. J. (2020). Inferring the total-evidence timescale of marattialean fern evolution in the face of model sensitivity. *BioRxiv* [Preprint]. doi: 10.1101/2020.09.25.313643

Mayrose, I., Zhan, S. H., Rothfels, C. J., Arrigo, N., Barker, M. S., Rieseberg, L. H., et al. (2015). Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014). *New Phytol.* 206, 27–35. doi: 10.1111/nph.13192

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., et al. (2011). Recently formed polyploid plants diversify at lower rates. *Science* 333:1257. doi: 10.1126/science.1207205

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., dePamphilis, C. W., et al. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol. Evol.* 8, 1150–1164. doi: 10.1093/gbe/evw060

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Papp, B., Pál, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197. doi: 10.1038/nature01771

Pryer, K. M., Schneider, H., Smith, A. R., Cranfill, R., Wolf, P. G., Hunt, J. S., et al. (2001). Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* 409, 618–622. doi: 10.1038/35054555

Pryer, K. M., Schuettpelz, E., Wolf, P. G., Schneider, H., Smith, A. R., and Cranfill, R. (2004). Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. *Am. J. Bot.* 91, 1582–1598. doi: 10.3732/ajb.91.10.1582

Qi, X., Kuo, L.-Y., Guo, C., Li, H., Li, Z., Qi, J., et al. (2018). A well-resolved fern nuclear phylogeny reveals the evolution history of numerous transcription factor families. *Mol. Phylogenet. Evol.* 127, 961–977. doi: 10.1016/j.ympev.2018.06.043

Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol. Biol. Evol.* 31, 750–762. doi: 10.1093/molbev/mst263

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* 6:e22594. doi: 10.1371/journal.pone.0022594

Regaladgo, L., Schmidt, A. R., Müler, P., Kobbert, M. J., Schneider, H., and Heinrich, J. (2017). The first fossil of Lindsaeaceae (Polypodiales) from the Cretaceous amber forest of Myanmar. *Cretaceous Res.* 72, 8–12. doi: 10.1016/j.cretres.2016.12.003

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N. A., Salman-Minkov, A., et al. (2015). The Chromosome Count Database (CCDB) – a community resource of plant chromosome numbers. *New Phytol.* 206, 19–26. doi: 10.1111/nph.13191

Román-Palacios, C., Molina-Henao, Y. F., and Barker, M. S. (2020). Polyploids increase overall diversity despite higher turnover than diploids in the Brassicaceae. *Proc. Biol. Sci.* 287:20200962. doi: 10.1098/rspb.2020.0962

Rothfels, C. J., Larsson, A., Kuo, L.-Y., Korall, P., Chiou, W.-L., and Pryer, K. M. (2012). Overcoming deep roots, fast rates, and short internodes to resolve the ancient rapid radiation of eupolypod II ferns. *Syst. Biol.* 61, 490–509. doi: 10.1093/sysbio/sys001

Rothfels, C. J., Larsson, A., Li, F.-W., Sigel, E. M., Huiet, L., Burge, D. O., et al. (2013). Transcriptome-mining for single-copy nuclear markers in ferns. *PLoS One* 8:e76957. doi: 10.1371/journal.pone.0076957

Rothfels, C. J., Li, F.-W., Sigel, E. M., Huiet, L., Larsson, A., Burge, D. O., et al. (2015). The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *Am. J. Bot.* 102, 1089–1107. doi: 10.3732/ajb.1500089

Rothfels, C. J., and Schuettpelz, E. (2014). Accelerated rate of molecular evolution for vittarioid ferns is strong and not driven by selection. *Syst. Biol.* 63, 31–54. doi: 10.1093/sysbio/syt058

Rothfels, C. J., Windham, M. D., Grusz, A. L., Gastony, G. J., and Pryer, K. M. (2008). Toward a monophyletic Notholaena (Pteridaceae): resolving patterns of evolutionary convergence in xeric-adapted ferns. *Taxon* 57, 712–724. doi: 10.1002/tax.573005

Rothwell, G. W. (1987). Complex paleozoic filicales in the evolutionary radiation of ferns. *Am. J. Bot.* 74, 458–461. doi: 10.1002/j.1537-2197.1987.tb08628.x

Rothwell, G. W., and Stockey, R. A. (2008). "Phylogeny and evolution of ferns: a paleontological perspective," in *Biology and Evolution of Ferns and Lycophytes*, eds T. A. Ranker and C. H. Haufler (Cambridge, MA: Cambridge University Press), 332–366. doi: 10.1093/aob/mcs017

Sayyari, E., Whitfield, J. B., and Mirarab, S. (2018). DiscoVista: interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* 122, 110–115. doi: 10.1016/j.ympev.2018.01.019

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593. doi: 10.1093/bioinformatics/btq706

Schneider, H., Liu, H.-M., Chang, Y.-F., Ohlsen, D., Perrie, L. R., Shepherd, L., et al. (2017). Neo- and Paleopolyploidy contribute to the species diversity of Asplenium - the most species-rich genus of ferns. *J. Syst. Evol.* 55, 353–364. doi: 10.1111/jse.12271

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

Schuettpelz, E., Chen, C.-W., Kessler, M., Pinson, J. B., Johnson, G., Davila, A., et al. (2016). A revised generic classification of vittarioid ferns (Pteridaceae) based on molecular, micromorphological, and geographic data. *Taxon* 65, 708–722. doi: 10.12705/654.2

Schuettpelz, E., and Pryer, K. M. (2006). Reconciling extreme branch length differences: decoupling time and rate through the evolutionary history of filmy ferns. *Syst. Biol.* 55, 485–502. doi: 10.1080/10635150600755438

Schuettpelz, E., and Pryer, K. M. (2007). Fern phylogeny inferred from 400 leptosporangiate species and three plastid genes. *Taxon* 56:1037. doi: 10.2307/25065903

Schuettpelz, E., and Pryer, K. M. (2009). Evidence for a Cenozoic radiation of ferns in an angiosperm-dominated canopy. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11200–11205. doi: 10.1073/pnas.0811136106

Sensalari, C., Maere, S., and Lohaus, R. (2022). ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* 38, 530–532. doi: 10.1093/bioinformatics/btab602

Sessa, E. B., and Der, J. (2016). "Evolutionary genomics of ferns and lycophytes," in *Genomes and Evolution of Charophytes, Bryophytes, and Ferns Advances in Botanical Research*, ed. S. A. Rensing (Amsterdam: Elsevier), 215–254.

Sessa, E. B., Zimmer, E. A., and Givnish, T. J. (2012a). Phylogeny, divergence times, and historical biogeography of New World *Dryopteris* (Dryopteridaceae). *Am. J. Bot.* 99, 730–750. doi: 10.3732/ajb.1100294

Sessa, E. B., Zimmer, E. A., and Givnish, T. J. (2012b). Reticulate evolution on a global scale: a nuclear phylogeny for New World *Dryopteris* (Dryopteridaceae). *Mol. Phylogenet. Evol.* 64, 563–581. doi: 10.1016/j.ympev.2012.05.009

Shan, S., Boatwright, L., Liu, X., Ch, A., Erbali, Fu, C., et al. (2020). Transcriptome dynamics of the inflorescence in reciprocally formed allopolyploid *Tragopogon miscellus* (Asteraceae). *Front. Genet.* 11:888. doi: 10.3389/fgene.2020.00888

Shen, H., Jin, D., Shu, J.-P., Zhou, X.-L., Lei, M., Wei, R., et al. (2018). Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. *Gigascience* 7, 1–11. doi: 10.1093/gigascience/gix116

Shen, X.-X., Li, Y., Hittinger, C. T., Chen, X.-X., and Rokas, A. (2020). An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nat. Commun.* 11:6096. doi: 10.1038/s41467-020-20005-6

Shi, T., Huang, H., and Barker, M. S. (2010). Ancient genome duplications during the evolution of kiwifruit (Actinidia) and related Ericales. *Ann. Bot.* 106, 497–504. doi: 10.1093/aob/mcq129

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Sjöstrand, J., Arvestad, L., Lagergren, J., and Sennblad, B. (2013). GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* 14:209. doi: 10.1186/1471-2105-14-209

Smith, S. A., Brown, J. W., and Walker, J. F. (2018). So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One* 13:e0197433. doi: 10.1371/journal.pone.0197433

Smith, S. A., and Donoghue, M. J. (2008). Rates of molecular evolution are linked to life history in flowering plants. *Science* 322, 86–89. doi: 10.1126/science.1163197

Smith, S. A., and O'Meara, B. C. (2012). treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28, 2689–2690. doi: 10.1093/bioinformatics/bts492

Soltis, D. E., Segovia-Salcedo, M. C., Jordon-Thaden, I. E., Majure, L. C., Miles, N. M., Mavrodiev, E. V., et al. (2014a). Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytol.* 202, 1105–1117. doi: 10.1111/nph.12756

Soltis, D. E., Visger, C. J., and Soltis, P. S. (2014b). The polyploidy revolution then…and now: stebbins revisited. *Am. J. Bot.* 101, 1057–1078. doi: 10.3732/ajb.1400178

Soltis, P. S., Marchant, D. B., Van de Peer, Y., and Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* 35, 119–125. doi: 10.1016/j.gde.2015.11.003

Soltis, P. S., Soltis, D. E., Savolainen, V., Crane, P. R., and Barraclough, T. G. (2002). Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4430–4435. doi: 10.1073/pnas.032087199

Sonderegger, D. (2020). *SiZer R package version 0.1–4*.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stebbins, G. L. (1950). *Variation and Evolution in Plants*. New York, NY: Cambridge University Press.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

Szövényi, P., Gunadi, A., and Li, F.-W. (2021). Charting the genomic landscape of seed-free plants. *Nat. Plants* 7, 554–565. doi: 10.1038/s41477-021-00888-z

Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., et al. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* 207, 454–467. doi: 10.1111/nph.13491

Testo, W., and Sundue, M. (2016). A 4000-species dataset provides new insight into the evolution of ferns. *Mol. Phylogenet. Evol.* 105, 200–211. doi: 10.1016/j.ympev.2016.09.003

The Pteridophyte Phylogeny Group (2016). A community-derived classification for extant lycophytes and ferns. *J. Syst. Evol.* 54, 563–603. doi: 10.1111/jse.12229

Tidwell, W. D., and Nishida, H. (1993). A new fossilized tree fern stem, Nishidacaulis burgii gen. et sp. nov., from Nebraska-South Dakota, USA. *Rev. Paleobot. Palynol.* 78, 55–67. doi: 10.1016/0034-6667(93)90017-O

Tiley, G. P., Ané, C., and Burleigh, J. G. (2016). Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biol. Evol.* 8, 1023–1037. doi: 10.1093/gbe/evw058

Tiley, G. P., Barker, M. S., and Burleigh, J. G. (2018). Assessing the performance of ks plots for detecting ancient whole genome duplications. *Genome Biol. Evol.* 10, 2882–2898. doi: 10.1093/gbe/evy200

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A. E., Wang, H., et al. (2020). Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat. Commun.* 11:884. doi: 10.1038/s41467-020-14724-z

Vanneste, K., Van de Peer, Y., and Maere, S. (2013). Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* 30, 177–190. doi: 10.1093/molbev/mss214

Vogel, J. C., Russell, S. J., Rumsey, F. J., Barrett, J. A., and Gibby, M. (1998). Evidence for maternal transmission of chloroplast DNA in the genus Asplenium (Aspleniaceae, Pteridophyta). *Bot. Acta* 111, 247–249. doi: 10.1111/j.1438-8677.1998.tb00704.x

Wagner, W. H. Jr., Wagner, F. S., Miller, C. N. Jr., and Wagner, D. H. (1978). New observations on the royal fern hybrid *Osmunda X ruggii*. *Rhodora* 80, 92–106.

Wei, R., Yan, Y.-H., Harris, A. J., Kang, J.-S., Shen, H., Xiang, Q.-P., et al. (2017). Plastid phylogenomics resolve deep relationships among eupolypod II ferns with rapid radiation and rate heterogeneity. *Genome Biol. Evol.* 9, 1646–1657. doi: 10.1093/gbe/evx107

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4859–E4868. doi: 10.1073/pnas.1323926111

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13875–13879. doi: 10.1073/pnas.0811575106

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30, 1660–1666. doi: 10.1093/bioinformatics/btu077

Xu, K., Zhang, L., Rothfels, C. J., Smith, A. R., Viane, R., Lorence, D., et al. (2019). A global plastid phylogeny of the fern genus *Asplenium* (Aspleniaceae). *Cladistics* 36, 22–71. doi: 10.1111/cla.12384

Yang, Y., and 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. doi: 10.1093/molbev/msu245

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* 19:153. doi: 10.1186/s12859-018-2129-y

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S. (2020). ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* 37, 3292–3307. doi: 10.1093/molbev/msaa139

Zhang, L., Schuettpelz, E., Rothfels, C. J., Zhou, X.-M., Gao, X.-F., and Zhang, L.-B. (2016). Circumscription and phylogeny of the fern family Tectariaceae based on plastid and nuclear markers, with the description of two new genera: *Draconopteris* and *Malaifilix* (Tectariaceae). *Taxon* 65, 723–738. doi: 10.12705/654.3

Zhang, L.-B., and Zhang, L. (2015). Didymochlaenaceae: a new fern family of eupolypods I (Polypodiales). *Taxon* 64, 27–38. doi: 10.12705/641.4

Zhao, T., and Schranz, M. E. (2018). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Harvard Dataverse* 116, 2165–2174. doi: 10.7910/dvn/bdma7a

Zhou, R., Moshgabadi, N., and Adams, K. L. (2011). Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. *Proc. Natl. Acad. Sci. U.S.A.* 108, 16122–16127. doi: 10.1073/pnas.1109551108

Zhou, X.-M., Zhang, L., Lu, N. T., Gao, X.-F., and Zhang, L.-B. (2018). Pteridryaceae: a new fern family of Polypodiineae (*Polypodiales*) including taxonomic treatments. *J. Syst. Evol.* 56, 148–173. doi: 10.1111/jse.12305

Zwaenepoel, A., and Van de Peer, Y. (2019). wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* 35, 2153–2155. doi: 10.1093/bioinformatics/bty915

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.