Phylogenetic Analysis of ferns

Ferns, the sister clade to Spermatophyta (seed plants), consists of Equisetidae (horsetails), Ophioglossidae which is composed of Psilotales (whisk ferns) and Ophioglossales, Marattiidae, and the species diversity-richest subclass Polypodiidae. The former 3 subclasses constitute a paraphyly group, namely eusporangiate ferns, while the Polypodiidae itself is a monophyly group, namely leptosporangiate ferns, comprised of 7 orders. A major disputation concerning the phylogenetic relationship of Ferns remains (Christenhusz & Chase, 2014; Rothfels *et al.*, 2015; Shen *et al.*, 2017). Here, inspired by the freshly assembled 22 high-quality fern transcriptomes and the availability of recently released genome and transcriptome resources, as summarized in Table S1, we reconstructed a phylogenetic tree of 107 fern species covering the whole backbone of fern group, as shown in Fig. SM1-4, whereof three methods, including ASTRAL-Pro2 (Zhang & Mirarab, 2022), STAG (Emms & Kelly, 2018), and concatenation-based method, and four categories of analysis, using horsetails, seed plants, seed plants plus Lycopods, seed plants plus Lycopods and Bryophytes as outgroup species, respectively, were conducted to explore a consensus fern phylogeny. Another round of phylogenetic analysis was done after the publication of *Marsilea vestita* genome to further resolve the occurred uncertainty from the 107 fern dataset as shown in Fig. SM5-6.

The results of phylogenetic inference of the 107 fern dataset varied across different methods mainly above the relationship between Hymenophyllales, Gleicheniales and Dipteridaces (see methods for circumscription), as shown in Fig. SM1-4. Considering the method ASTRAL-Pro2, applied upon different datasets varied in outgroups, for instance, horsetails, the Hymenophyllales and Gleicheniales formed a monophyly group (local posterior probability, localPP = 0.65) with Gleicheniales closer to Hymenophyllales than Dipteridaces implicating the paraphyly nature of Gleicheniales, which was also evidenced in dataset with Lycopods as outgroup (localPP = 0.86). Nonetheless, the datasets with seed plants and Bryophytes as outgroup presented another scenario where the Gleicheniales was still closer to Hymenophyllales while Dipteridaces was sister to the last common ancestor of core leptosporangiate ferns and Schizaeales with localPP as 0.96 and 0.71, respectively. Apparent conflicts occurred at the phylogenetic location of Dipteridaces. In regard to concatenation-based method, the results were harmonious across datasets with varied outgroups, that Hymenophyllales, Gleicheniales, Dipteridaces and the last common ancestor of core leptosporangiate ferns and Schizaeales branched off from each other sequentially, all with bootstrap support higher than 0.72. With respect to the STAG method, although it brought generally lower branch support, concordant results were deduced regarding the phylogenetic location of the above four clades, wherein the Hymenophyllales and Gleicheniales formed a monophyly group while the Gleicheniales is a paraphyly group. Since the support value of internal bipartitions from STAG method denotes the realistic occurrence of that bipartition in the input trees, it's less biased than the concatenation-based method which generally exaggerates branch support upon the long concatenated alignment and thus renders the uncertain results look "highly reliable". The accuracy of quartet-based method ASTRAL-Pro2, which calculates the support values as localPP representing the highest posterior probability among all the three possible topologies and infers the phylogeny which optimizes the overall quartet similarity, is dependent on the errors of estimated gene trees. The displayed

phylogenetic uncertainty of ASTRAL-Pro2 method across datasets hinted that the estimation of inputted gene tree topologies might not be robust to the introduced impact of disparate branch lengths of varied outgroups.

To further resolve such phylogenetic uncertainty, we added the newly released *Marsilea vestita* genome and reimplemented the phylogenetic analysis using the ASTRAL-Pro2 method upon both nucleotide and peptide alignment. The results unequivocally supported a monophyly group comprised of Hymenophyllales and Gleicheniales independent of outgroups or aligned molecules as shown in Fig. SM5-6. Gleicheniales was supported to form a paraphyly group where the Gleicheniales was closer to Hymenophyllales than Dipteridaces in all datasets except the one with Bryophytes as outgroup, whose conflicting branching pattern was only supported by localPP as 0.46. Compared to the preceding phylogenetic results which might be blurred by the error associated with the gene tree estimation, the addition of another high-quality genome sequence might not only amplify the reliability of delineated orthogroups but also diminish the uncertainty brought by sequence errors in both tree searching and model comparison processes. Altogether, the updated topology derived from the 108 ferns (107 ferns plus *M. vestita*) dataset conferred more consistency and credibility in the regard of both data quality and methodological efficacy. Provided the minor collisions of other branches across datasets with varied outgroups, we accepted the phylogeny which shared the most consistency with the remaining as the consensus phylogenetic tree. The results upon peptide alignments exhibited generally more uncertainties than nucleotide alignments in the regard of more indefinite branches with lower localPP as a consequence of less informative sites.

To properly test competing tree topology hypotheses, we conducted the AU test, KH test and SH test on alternative topologies pertaining to Marattiales, Polypodiidae and Ophioglossidae (consisting of Psilotales and Ophioglossales) on the 136 single-copy gene family (SOG) dataset. As shown in Figure R1b, the topology (topology 1) in which Marattiales is sister to Polypodiidae has distinctly higher mean and median p-values in all tests than the alternative topology in which Marattiales is sister to Ophioglossidae. Mann-Whitney U test also shows that topology 1 has significantly higher p-value than topology 2 (p-value < 0.001 in all tree topology tests). There are also more SOGs that significantly reject topology 2 (i.e., p-value < 0.05 in the tree topology test) across all tree topology tests compared to topology 1, as shown in Figure R1c. Likewise, we also conduct the AU test, KH test and SH test on alternative topologies pertaining to Hymenophyllales and Gleicheniales (consisting of Dipteridaceae and Gleicheniaceae). As shown in Figure R2b, the topology (topology 1) in which Gleicheniales is a paraphyletic clade and Hymenophyllales is closer to Gleicheniaceae than Dipteridaceae obtained consistently higher mean and median p-values in all tests compared to the other two alternative topologies. Mann-Whitney U test further shows that topology 1 has significantly higher p-value than topology 3 (p-value < 0.05 in the SH test). The number of SOGs that significantly reject alternative topologies (i.e., p-value < 0.05 in the tree topology test) is smallest for topology 1 across all tree topology tests, as shown in Figure R2c. Altogether, the tree topology test analyses provide additional support for our inferred consensus species tree.

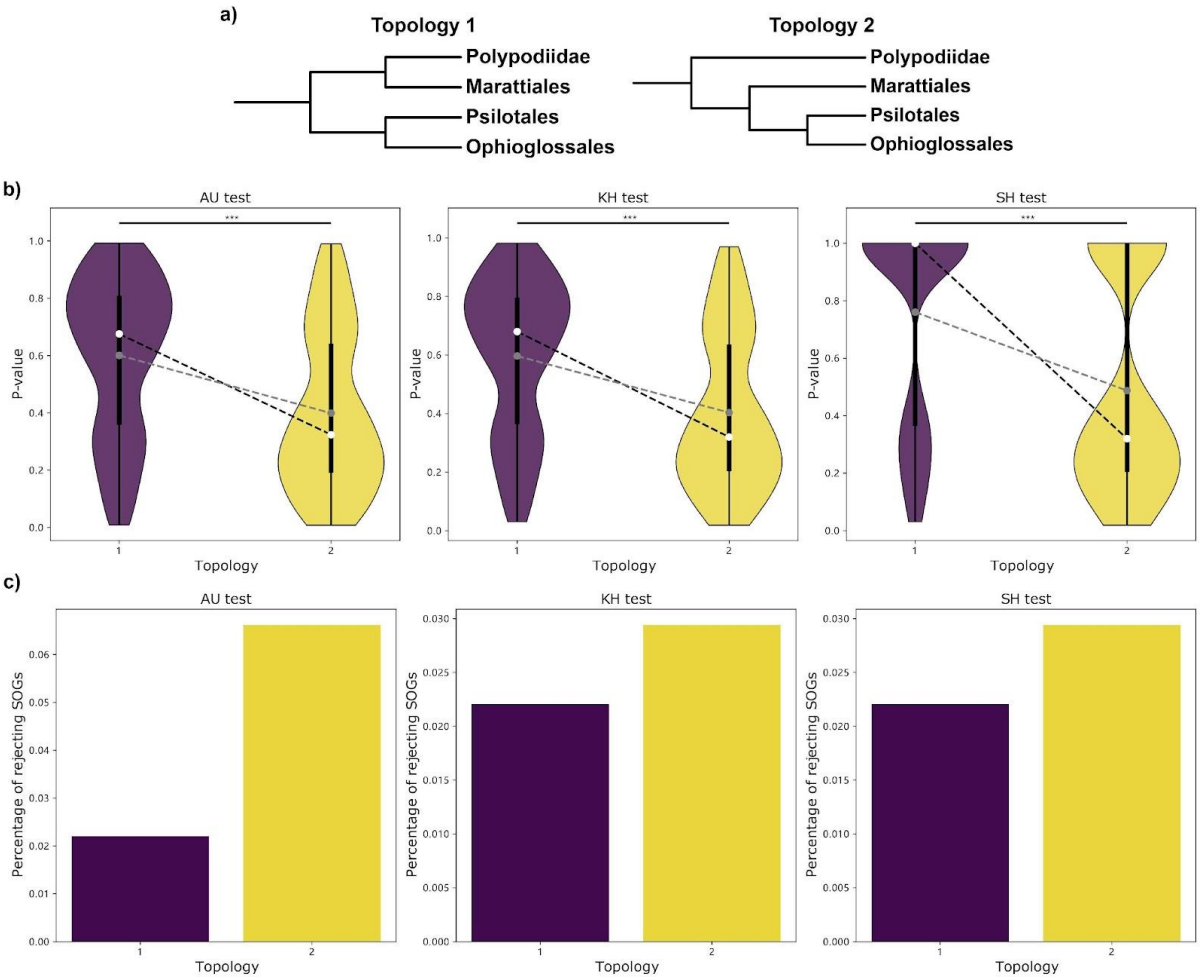


Figure R1 | AU test, KH test and SH test on alternative topologies pertaining to Marattiales, Polypodiidae and Ophioglossidae (consisting of Psilotales and Ophioglossales). a) Two alternative topologies pertaining to Marattiales, Polypodiidae and Ophioglossidae are presented. b) Violin plots of p-value obtained from AU test, KH test and SH test are presented in which white dots denote the median values connected by black dashed lines, gray dots denote the mean values connected by gray dashed lines, the first and third quartiles are represented as black bars, the minimum and maximum are represented as black solid lines. The significance levels calculated from the Mann-Whitney U test are denoted as either ns (0.05 <= p-value), '*' (0.01 <= p-value < 0.05), '**' (0.001 <= p-value < 0.01), or '***' (p-value < 0.001). c) Bar plots showing the number of SOGs that significantly reject alternative topologies (p-value < 0.05) are presented.

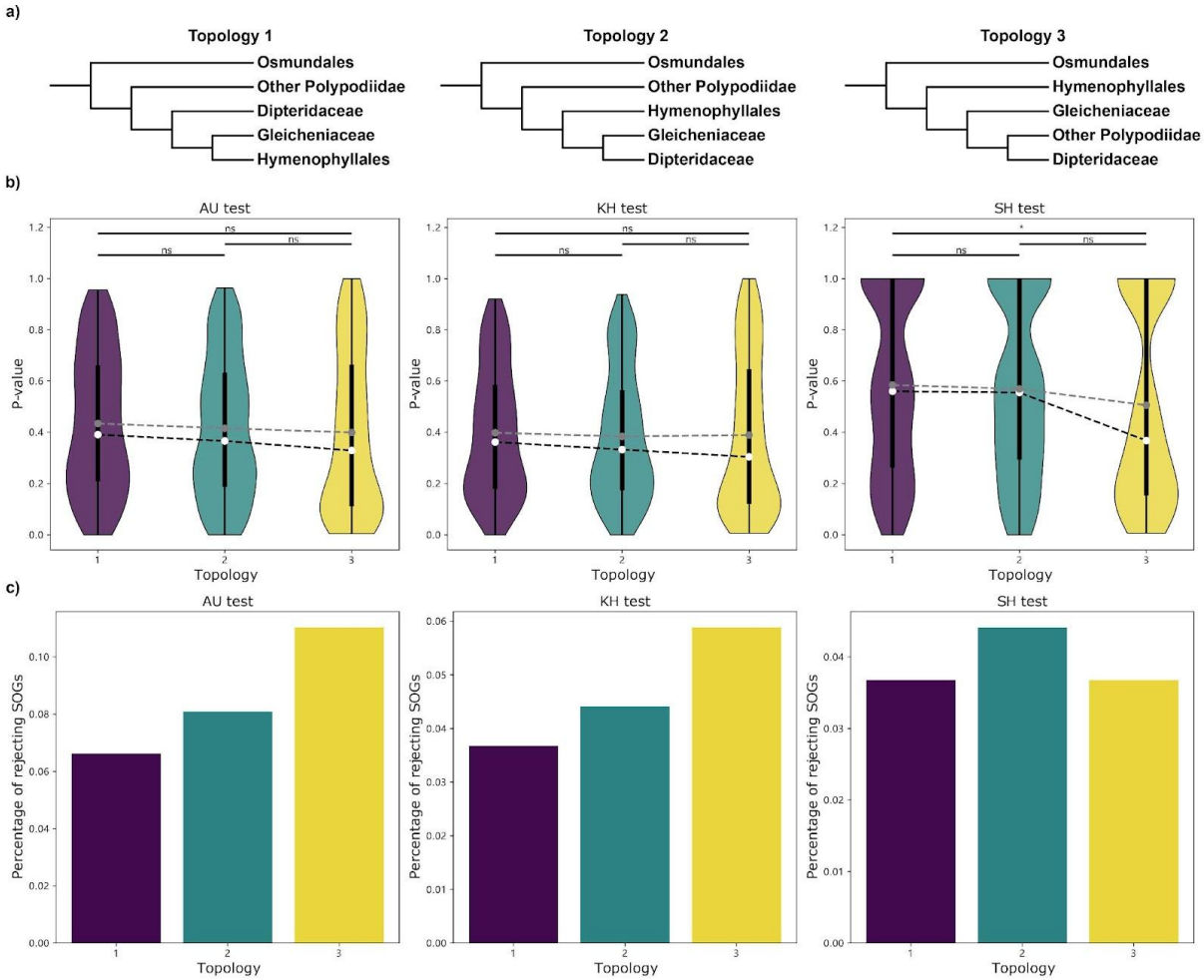


Figure R2 | AU test, KH test and SH test on alternative topologies pertaining to Hymenophyllales and Gleicheniales (consisting of Dipteridaceae and Gleicheniaceae). a) Three alternative topologies pertaining to Hymenophyllales and Gleicheniales are presented. b) Violin plots of p-value obtained from AU test, KH test and SH test are presented in which white dots denote the median values connected by black dashed lines, gray dots denote the mean values connected by gray dashed lines, the first and third quartiles are represented as black bars, the minimum and maximum are represented as black solid lines. The significance levels calculated from the Mann-Whitney U test are denoted as either ns (0.05 <= p-value), '*' (0.01 <= p-value < 0.05), '**' (0.001 <= p-value < 0.01), or '***' (p-value < 0.001). c) Bar plots showing the number of SOGs that significantly reject alternative topologies (p-value < 0.05) are presented.

The phylogeny derived from nucleotide dataset with seed plants as outgroup was eventually accepted as the consensus tree and used in subsequent analysis as shown in Fig 1.

Absolute Divergence Time Analysis of ferns

Phylogenetic branch lengths in metric of substitutions per site represent the evolutionary distance of molecules, as a consequence of varied substitution rates and divergence time. With the external fossil calibration information, phylogenetic branch lengths can be

transformed into absolute geological time given a model of molecular clock and prior distribution of node age (dos Reis *et al.*, 2016). Here we estimated the absolute divergence time of the 108 ferns phylogenetic tree as shown in Fig 1, by means of Bayesian molecular dating under the independent rate and LG general amino acid substitution model (Le & Gascuel, 2008) with 18 soft fossil constraints (as summarized in Table S2). The 95% Highest Posterior Density (HPD) and posterior mean absolute divergence time estimates for the origin of each major clade was summarized in Table 1.

Whole-Genome Duplication (WGD) Analysis of Monilophyta

Here we reappraised the WGD episode throughout the backbone of ferns based on the consensus phylogeny inferred above. K_S-age distribution and phylogenomic method was implemented to probe the remnants of WGDs. The construction of K_S-age distribution of whole paranome, defined as the collection of all gene families in a single genome, and collinear gene pairs (anchor pairs), defined as paralogues residing in collinear blocks, alludes the plausibility of WGDs by deducing the corresponding components from mixture modeling and pinpoints putative WGDs in a phylogenetic context by rescaling the peak K_S value of orthologous K_S-age distribution into the timescale of focal species. Phylogenomic analysis given the model of gene tree - species tree reconciliation and gene family evolution assuming small-scale gene duplication and whole genome duplication can estimate the retention rate of the provided WGD models and interrogate competing hypothesis (Chen *et al.*, 2023). As shown in Fig 1, there were in total 18 hypothetical WGDs supported within the backbone of ferns, 5 among which attained collinear support. Below a circumspect discussion follows.

The ASPL WGD event of *Asplenium*

The paralogous K_S-age distributions of *Asplenium nidus* and *Asplenium cf. x lucrosum* were deduced of a lognormal component of whole paranome with modes 0.38 and 0.34, respectively, as shown in Fig. SM7, which was explicitly older than the split within *Asplenium* while younger than the split between *Asplenium* and remaining ferns, implying a shared WGD event of *Asplenium*, proposed as ASPL. Phylogenomic results revealed the retention rate of this WGD as posterior mean 0.0688 with 95% equal-tail confidence interval (CI) as 0.0095 - 0.1141 in the relaxed branch-specific model, as shown in Fig. SM9. The dramatic gene loss of *A. platyneuron* around K_S 0.4 compared to *A. nidus* and *A. cf. x lucrosum* might have eroded the signal on *Asplenium* node in terms of less reconciled duplications, as shown in Fig. SM7.

The ADNE WGD event of *Adiantum nelumboides*

Adiantum nelumboides displayed a lognormal component of whole paranome with mode 0.06, younger than the split with remaining ferns, as such the lognormal cluster a of anchor K_S with mode 0.09, as shown in Fig. SM7 and S8, which unambiguously supported a lineage-specific WGD of *A. nelumboides*, proposed as ADNE. The phylogenomic results conferred unequivocal support for this WGD with high retention rate as posterior mean 0.8883 with 95% equal-tail CI as 0.7886 - 0.9838 in the relaxed branch-specific model, as shown in Fig. SM9, in accordance with its tetraploidy (Zhong *et al.*, 2022).

The ADRA WGD event of *Adiantum raddianum*

A lognormal component of whole paranome for *Adiantum raddianum* with mode 0.55 was inferred, as shown in Fig. SM7, overlapping the split with other *Adiantum* ferns, invoking the uncertainty of the phylogenetic placement of this hypothetical WGD, either shared by *Adiantum* or lineage-specific to *A. raddianum*. The phylogenomic results explicitly rejected the alternative scenario of a *Adiantum* shared WGD but favored a lineage-specific WGD of its own in terms of posterior mean retention rate 0.0529 compared to negligible 0.0104 in the relaxed branch-specific model, with 95% equal-tail CI as 0.0046 - 0.1121 compared to 0.0007 - 0.0241, as shown in Fig. SM9, in conformance with the lack of recent WGD of *A. capillus veneris* genome (Fang *et al.*, 2022). This hypothetical WGD of *A. raddianum* which might occur shortly after the divergence with other *Adiantum* ferns, was referred to as ADRA.

The VITT WGD event of *Vittaria*

The paralogous K_S-age distributions of whole paranome for *Vittaria appalachiana* and *Vittaria lineata* exhibited a lognormal component with modes as 0.61 and 0.56, respectively, as shown in Fig. SM7, older than the bipartition of one another while younger than the split with remaining ferns, betokening a shared WGD event of *Vittaria* ,proposed as VITT. The retention rate of this hypothetical WGD was inferred as posterior mean 0.1429 with 95% equal-tail CI as 0.0587 - 0.2158 in the relaxed branch-specific model, as shown in Fig. SM9.

The CERA WGD event of *Ceratopteris*

Ceratopteri thalictroides and *Ceratopteris richardii* presented a lognormal component of whole paranome K_S-age distributions with modes as 1.1 and 1.29, respectively, which were explicitly older the departure of one another while younger than the divergence with remaining ferns, as such the lognormal cluster c of anchor K_S for *C. richardii* with mode 1.4, as shown in Fig. SM7 and S8, indicating a shared WGD event of *Ceratopteris*, proposed as CERA. The retention rate was inferred as posterior mean 0.13 with 95% equal-tail CI as 0.0338 - 0.2045 in the relaxed branch-specific model, as shown in Fig. SM9.

The LIND WGD event of *Lindsaea*

The paralogous K_S-age distributions of whole paranome for *Lindsaea linearis* and *Lindsaea microphylla* manifested a lognormal component with modes as 0.74 and 0.77, respectively, as shown in Fig. SM7, which was older than the split within *Lindsaea* while younger than the separation with remaining ferns, evincing a shared WGD event of *Lindsaea*, proposed as LIND. The retention rate was inferred as posterior mean 0.1757 with 95% equal-tail CI as 0.1282 - 0.2214 in the relaxed branch-specific model, as shown in Fig. SM9.

The LOHI WGD event of *Lonchitis hirsuta*

A lognormal component with mode 0.47 emerged in the paralogous K_S-age distributions of whole paranome for *Lonchitis hirsuta*, as shown in Fig. SM7, which was apparently younger than the divergence with remaining ferns and endorsed a lineage-specific WGD of *L. hirsuta*, proposed as LOHI. The retention rate was inferred as posterior mean 0.0587 with 95% equal-tail CI as 0.0058 - 0.1148 in the relaxed branch-specific model, as shown in Fig. SM9.

The CYAT WGD event of Cyatheales

Alsophila spinulosa, *Alsophila latebrosa* and *Cibotium barometz* showed a lognormal component of whole paranome K_S-age distributions with modes as 0.31, 0.25 and 0.19, respectively, which overlapped the split between members of Cyatheales while younger than the separation with remaining ferns, as such the lognormal cluster b of anchor K_S for *A. spinulosa* with mode 0.32, as shown in Fig. SM7 and S8,implying a shared WGD event of Cyatheales, proposed as CYAT. The phylogenomic results granted unequivocal support for this WGD with retention rate as posterior mean 0.3900 with 95% equal-tail CI as 0.3252 - 0.4540 in the relaxed branch-specific model, as shown in Fig. SM10. The slower evolving nature of other Cyathealean ferns might have mingled the WGD with the bulk of recent small-scale gene duplications and obfuscate the signal informed by mixture modeling.

The AZOL WGD event of *Azolla*

Azolla filiculoides was deduced a lognormal cluster b of anchor K_S with mode as 1.01, as shown in Fig. SM8, which was older than the split with *A. cf. caroliniana* while overlapped with the divergence with *Salvinia*, implying a likely shared WGD event of *Azolla*, proposed as AZOL. The alternative scenario of sharing with *Salvinia* has been already rejected in previous study (Chen *et al.*, 2023). The retention rate was inferred as posterior mean 0.132 with 95% equal-tail CI as 0.0291 - 0.2140 in the relaxed branch-specific model, as shown in Fig. SM10.

The COLE WGD event of core leptosporangiate

Numerous species including *Pilularia globulifera*, *Pteris vittata*, *Cryptogramma acrostichoides*, *Gaga arizonica*, *Adiantum capillus-veneris*, *Dennstaedtia davallioides*, *Leucostegia immersa*, *Polystichum acrostichoides*, *Polypodium amorphum*, *Polypodium glycyrrhiza*, *Asplenium platyneuron*, *Homalosorus pycnocarpos*, *Gymnocarpium dryopteris*, *Cystopteris protrusa*, *Cystopteris reevesiana*, *Cystopteris fragilis*, *Woodsia scopulina*, *Woodsia ilvensis*, *Deparia lobato-crenata*, *Athyrium filix-femina*, *Onoclea sensibilis* and *Blechnum spicant* all furnished support for a shared ancient WGD event of core leptosporangiate (Polypodiales + Cyatheales + Salviniales), in terms of the deduced lognormal component of whole paranome K_S-age distributions earlier than the split of each other, as well as the lognormal clusters of anchor K_S for *A. capillus-veneris* and *Alsophila spinulosa* with modes as 2.15 and 0.78, respectively, as shown in Fig. SM7 and S8, which were undisputedly older than the separation with other core leptosporangiate ferns but younger than the split with remaining ferns, proposed as COLE. Overwhelming evidence rooted the authenticity of this ancient WGD event.

The HYGL WGD event of Hymenophyllales + Gleicheniales

Members of the clade Hymenophyllales + Gleicheniales, including *Diplopterygium laevissimum*, *Dicranopteris pedata*, *Dicranopteris curranii*, *Abrodictyum obscurum* and *Hymenophyllum bivalve* and *Dipteris conjugata* all exhibited a lognormal component of whole paranome overlapped the split with each other while younger than the divergence with remaining ferns, as shown in Fig. SM7, implying a shared WGD event with each other, proposed as HYGL. The phylogenomic results granted indisputable support for this WGD with retention rate as posterior mean 0.38 and 95% equal-tail CI as 0.3283 - 0.4404 in the relaxed branch-specific model, as shown in Fig. SM11.

The POLY WGD event of Polypodiidae

An ancient WGD event shared by Polypodiidae obtained support from the whole paranome of *Osmunda javanica* and *Osmundastrum cinnamomeum* in terms of deduced lognormal component with modes as 0.93 and 0.83, respectively, as shown in Fig. SM7, which were apparently older than the split within Polypodiidae while younger than the divergence with remaining ferns, proposed as POLY. The retention rate was inferred as posterior mean 0.0516 with 95% equal-tail CI as 0.0078 - 0.0963 in the critical branch-specific model, as shown in Fig. SM11. This likely ancient WGD might be the oldest among all the hypothetical WGDs, which invited less retained duplicates and the wreck of phylogenomic signal. Conclusive verification requires the collinear evidence from further genome assembly.

The MARA WGD event of Marattiales

Members of Marattiales, including *Angiopteris evecta*, *Marattia attenuata*, *Christensenia aesculifolia*, *Angiopteris fokiensis*, *Ptisana pellucida* and *Danaea nodosa*, all presented an inferred lognormal component older than the split within Marattiales while younger than the divergence with remaining ferns, as shown in Fig. SM7, implying a shared WGD event of Marattiales, proposed as MARA. The phylogenomic results contributed explicit support for this WGD with retention rate as posterior mean 0.2199 and 95% equal-tail CI as 0.1056 - 0.3201 in the relaxed branch-specific model, as shown in Fig. SM11.

The OPHI WGD event of *Ophioglossum*

The paralogous K_S-age distributions of whole paranome for species from *Ophioglossum* all manifested a very recent lognormal component with modes less than K_S 0.1, overlapping the split within *Ophioglossum* while younger than the split with remaining ferns, as shown in Fig. SM7, implying a shared WGD event of *Ophioglossum*, proposed as OPHI. The phylogenomic results conferred unequivocal support for this WGD with retention rate as posterior mean 0.2769 and 95% equal-tail CI as 0.2220 - 0.3278 in the relaxed branch-specific model, as shown in Fig. SM12.

The BOTR WGD event of Botrychioideae

Sceptridium dissectum exhibited a lognormal component of whole paranome K_S-age distributions with mode as 0.29, which was older than the bipartition with *Botrypus virginianus* while younger than the split with remaining ferns, as shown in Fig. SM7, suggesting a shared WGD event of Botrychioideae, proposed as BOTR. The retention rate was inferred as posterior mean 0.0735 with 95% equal-tail CI as 0.0197 - 0.1117 in the relaxed branch-specific model, as shown in Fig. SM12. The slower substitution rate of *B. virginianus* might conceal the K_S signal in recent small-scale gene duplications from mixture modeling.

The PSIL WGD event of Psilotales

Psilotum nudum showed a lognormal component of whole paranome K_S-age distributions with mode as 0.35, as shown in Fig. SM7, older than the bipartition with *Tmesipteris* while younger than the split with remaining ferns, hinting a shared WGD event of Psilotales, proposed as PSIL. The retention rate was inferred as posterior mean 0.0570 with 95%HPD as 0.0052 - 0.1140 in the relaxed branch-specific model, while the alternative scenario of a lineage-

specific WGD of *P. nudum* was rejected in terms of apparently smaller retention rate with posterior mean as 0.0250 and 95% equal-tail CI as 0.0011 - 0.0675, as shown in Fig. SM12. Analogously, the slower substitution rate of *Tmesipteris* might cache the K_S signal from mixture modeling.

The OPPS WGD event of Ophioglossales and Psilotales

The paralogous K_S-age distributions of whole paranome for *Sceptridium dissectum* and *Botrypus virginianus* were deduced a lognormal component with modes as 1.0 and 1.06, respectively, as shown in Fig. SM7, which were older than the bipartition of Ophioglossales and Psilotales while younger than the divergence with remaining ferns, implicating a shared WGD event of Ophioglossales and Psilotales, proposed as OPPS. The retention rate was inferred as posterior mean 0.0563 with 95% equal-tail CI as 0.0102 - 0.0999 in the critical branch-specific model, as shown in Fig. SM12. Considering that this likely WGD might be the oldest amid our analysis, collinear evidence enabled by future genome assembly is imperative to draw a dispositive conclusion.

The EQUI WGD event of Equisetales

All members of Equisetales, including *Equisetum arvense*, *Equisetum diffusum*, *Equisetum giganteum* and *Equisetum hyemale*, exhibited a lognormal component of paralogous K_S-age distributions of whole paranome older than the separation with each other while younger than the divergence with remaining ferns, as shown in Fig. SM7, promising a shared ancient WGD event of Equisetales, proposed as EQUI. The phylogenomic results issued prominent support for this WGD with retention rate as posterior mean 0.1918 and 95% equal-tail CI as 0.0444 - 0.3691 in the relaxed branch-specific model, as shown in Fig. SM12.

Absolute dating of WGD for genome-available ferns

The available genome assemblies of *Azolla filiculoides* and *Ceratopteris richardii* enabled the absolute dating of the recent WGD events experienced in their evolutionary past. The ages of AZOL and CERA WGD were dated with posterior mean 115.45 mya and 93.50 mya, respectively, with 95% HPD as 83.20 - 146.68 mya and 63.73 - 121.41 mya respectively, as shown in Fig. SM13. See methods for the detailed procedure of absolute dating.

Materials and Methods

Construction of K_S-based age distributions

K_S-age distributions for all paralogous genes (paranome) of genomes and transcriptomes were constructed by ksrates (v1.1.1) (Sensalari *et al.*, 2021). In brief, the ksrates pipeline entails firstly translating the coding nucleotide sequences into peptide sequences assuming standard genetic code, filtering out sequences whose sequence length is not divisible by 3, containing invalid codons or in-frame stop codon, after which an all-versus-all blastp was implemented with *E*-value set as 1×10^{-10} in BLASTP (v2.11.0+) (Camacho *et al.*, 2009) and the resultant subject-query hit table was fed into MCL (v14-137) (Van Dongen, 2000) with clustering inflation factor set as 3.0 to delineate paralogous gene families while filtering out gene families whose size is larger than 200, secondly calling the aligner MUSCLE (v3.8.1551) (Edgar, 2004) under default parameter to obtain a multiple sequence alignment (MSA) at the

protein level for each paralogous gene family while filtering out sequence pairs whose gap-stripped alignment length was shorter than 100, which was then back-translated into a codon alignment and subsequently fed into the CODEML function within PAML (v4.9j) (Yang, 2007) to acquire the maximum likelihood estimate (MLE) of K_S values under non-pairwise mode using the default control file defined by wgd (v1.1.1) (Zwaenepoel & Van de Peer, 2018) and then calling FastTree (v2.1.11) (Price *et al.*, 2010) upon the peptide MSA under default parameter to attain a midpoint-rooted phylogenetic tree of each paralogous gene family for retrieving the weight of each paralogous gene pair with or without outliers, and eventually building the K_S-age distribution with de-redundancy achieved by node-weighted method after excluding outliers. The collinear gene pairs (anchor pairs) were identified by i-ADHoRe (v3.0.01) (Proost *et al.*, 2011) under the default control file defined by wgd and the weight values for anchor pairs whose corresponding K_S values were between 0.05 and 20 were recalculated and reassigned while the weight of remaining pairs was set as zero. For orthologous K_S-age distributions, the process of MCL clustering was supersede as reciprocal best hits (RBH) searching to identify orthologous gene pairs while the weighting process was revoked on that only one-versus-one orthologues were inferred. CD-HIT (v4.8.1) (Fu *et al.*, 2012) was applied for the de-redundancy of transcriptome assemblies with the clustering threshold set as 0.99 before K_S-age analysis.

Correction of differences pertaining to synonymous substitution rates

The correction of synonymous substitution rates was realized in ksrates. The principle leans on a number of trios of species, including a focal species, a sister species and an outgroup species. The disparate synonymous substitution rate between focal species and sister species since the divergence is represented indeed by the branch-specific contribution of accumulated synonymous substitutions per synonymous site in respective branches. The mode of orthologous K_S-age distributions inferred from the kernel density estimate (KDE) using Gaussian kernels within python package scipy was designated as the proxy of peak K_S value of each orthologous K_S-age distribution. 200 iterations of bootstrap with replacements was implemented for each orthologous K_S-age distribution and the mean along with standard deviations (STD) of mode across the replicates was determined as the final peak K_S value representing divergence distance and its associated STD. The original accumulated synonymous substitutions per synonymous site of focal species-sister species pair consisting of branch-specific contribution of both species since diversification was transformed into two times the branch-specific contribution of focal species with the prop of outgroup species to resemble the timescale of focal species. The mean of rescaled peak K_S values of focal species-sister species pair against various outgroup species was taken as consensus adjusted peak K_S value. The maximum number of trios was set as 20. The species tree inferred by ASTRAL-Pro2 using seed plants as outgroup species was adopted in ksrates.

Construction of orthologous families and single-copy gene trees

Orthofinder (v2.5.4) (Emms & Kelly, 2019) was performed upon the protein sequences of 107 ferns and outgroup species with inflation factor set as 3 to delineate the orthologous families. No single-copy gene families were identified by Orthofinder, probably due to the universal and unique gene duplication and loss scenario across species and gene isoforms (Steenwyk *et*

al., 2022). To recover reliable and adequate single-copy gene families, we constructed mostly single-copy gene families (Li *et al.*, 2020), wherein most species were in single-copy while the remaining species had no more than 4 copies which were assumed to be transcript variants of the same gene, by retaining the longest copy, if applied, of each species. We referred the mostly single-copy gene families as single-copy gene families thereafter. In total, 140, 112, 107 and 57 single-copy gene families were constructed from dataset 107 ferns, 107 ferns plus seed plants, 107 ferns plus seed plants and Lycopods, 107 ferns plus seed plants, Lycopods and Bryophytes, respectively. MAFFT (v7.475) (Katoh & Standley, 2013) was performed to obtain a peptide multiple sequence alignment (MSA) for each single-copy gene family with the parameter "–auto". Trimal (v1.4.1) (Capella-Gutiérrez *et al.*, 2009) was then performed to trim the MSA and back-translate it into a codon-level nucleotide MSA with parameter "-automated1". IQ-TREE (v1.6.12) (Nguyen *et al.*, 2014) was implemented on each codon-level nucleotide MSA wherein ModelFinder (Kalyaanamoorthy *et al.*, 2017) was called to find the best-fit codon substitution model in terms of Bayesian Information Criterion (BIC) upon which a maximum likelihood (ML) gene tree was inferred and assigned with bootstrap support values from 1000 ultrafast (Hoang *et al.*, 2017) bootstrap replicates with parameter "-bnni" to further optimize each bootstrap iteration through a hill-climbing nearest neighbor interchange (NNI) search based directly on the corresponding bootstrap alignment to avoid severe model violations. The same process was further applied upon the 108 ferns dataset including *Marsilea vestita*, in which totally 136, 108, 103 and 55 single-copy gene families were reconstructed from dataset varied in outgroups with both nucleotide and peptide molecules.

Species tree inference

Three methods, ASTRAL-Pro2 (Zhang & Mirarab, 2022), STAG (Emms & Kelly, 2018), and concatenated-based method were implemented to infer the species tree. The acquired individual ML single-copy gene trees and gene name-species name map files were imported into ASTRAL-Pro2 and STAG under default parameters to estimate a consensus species tree with support values for each bipartition denoting local posterior probabilities (localPP) and the proportion of individual estimates of the species tree that contain that bipartition, respectively.

For concatenated-based method, the individual codon-level nucleotide MSA of single-copy gene families were concatenated and then fed into IQ-TREE to infer a ML super-gene tree as above. *Sticherus truncatus*, *Dicranopteris curranii*, *Dicranopteris pedata*, *Diplopterygium laevissimum* and *Diplopterygium glaucum* were grouped and named as Gleicheniales−II clade while *Cheiropleuria integrifolia*, *Dipteris lobbiana* and *Dipteris conjugata* were grouped and named as Gleicheniales−I clade due to their distinct phylogenetic relationship.

Estimation of absolute divergence time

Mcmctree (v4.9j) (Yang, 2007) was implemented upon the concatenated peptide MSA of single-copy gene families of 108 ferns dataset with seed plants as outgroup to infer the absolute divergence time for each bipartition. The independent rates model which assumes a log-normal distribution of evolutionary rates across branches was selected and 18 fossil calibrations of soft constraint from Nitta et al. (2022) were adopted to refine the divergence time of internal nodes, as summarized in Table S2. Fossils that calibrate clades within

Gleicheniaceae or Hymenophyllaceae were avoided for their indefinite phylogenetic location. LG amino acid substitution matrix was selected and a gamma model of rate variation was assumed with alpha as 0.5 and 5 categories in discrete gamma. Parameters controlling the birth-death process was set as 1 1 0.1 to generate uniform age priors on nodes that didn't have a fossil calibration. Gamma priors for the transition/transversion rate ratio and shape parameter for variable rates among sites were set as 6 2 and 1 1. A Dirichlet-gamma prior was set upon the mean rate across loci and the variance in logarithm as 2 20 1 and 1 10 1. The first 2000 iterations was discarded as burn-in and then 20,000,000 iterations were performed with sampling per 1000 iterations. The effective sample size (ESS) of all parameters was larger than 200, suggesting adequate sampling and convergence.

Phylogenomic analysis of gene tree - species tree reconciliation

To estimate the retention rate and interrogate hypothetical WGDs over competing scenarios in different clades, we implemented 4 categories of statistical gene tree - species tree reconciliation analysis using Whale (v.2.0.3) (Zwaenepoel & Van de Peer, 2019), as shown in Fig. SM9-12. Firstly, Orthogroups of each category of species were inferred by Orthofinder (v2.5.4) (Emms & Kelly, 2019) with inflation factor as 3. Gene families were filtered to assure at least one gene from both descendants present at the root and to avoid large gene family size which contains much noises and causes computational downshift via "orthofilter.py" (https://github.com/arzwa/Whale.jl), from which 1000 gene families were randomly selected as subsequent inputs. PRANK (v.150803) (Löytynoja, 2014) was utilized to obtain a MSA for each gene family and MrBayes (v.3.2.6) (Ronquist *et al.*, 2012) was then applied to infer the posterior distributions of gene trees under the LG + GAMMA model, with iterations set as 110,000 and sample frequency as 10 to get in total 11,000 posterior samples. ALEobserve (Szöllősi *et al.*, 2013) was subsequently performed upon the tree samples to construct the conditional clade distribution with a burn-in of 1000. Two gene family evolution models, namely relaxed branch-specific model and critical branch-specific model, were applied as previous study (Chen *et al.*, 2023) to estimate the retention rates of hypothetical WGDs for each category, as shown in Fig. SM9-12. Hypothetical WGDs with retention rates higher than 0.05 were regarded as supported WGDs considering the incompleteness of transcriptome assemblies and the stochasticity of sampled gene families.

Absolute dating of WGDs

Phylogenetic dating of AZOL and CERA WGD was proceeded as follows. Firstly, an orthogroup comprising orthologues from 8 other species and a pair of anchor which was assumed to be retained from the corresponding WGD was constructed per anchor pair by searching the reciprocal best hits (RBH) between anchor pair and the transcriptomes or genomes of other species by Diamond (v2.0.5.143) (Buchfink *et al.*, 2021) under default parameter, as shown in Fig. SM14. K_S range 0.36 - 2.00 was confined for the age of anchor pairs to be adopted in terms of densest aggregation of duplicates and avoidance towards saturation for AZOL WGD and K_S range 0.41 - 2.0 was bounded for CERA WGD. Secondly, the peptide sequences of each individual orthogroup was aligned by MAFFT (v7.475) (Katoh & Standley, 2013) under default parameter and then concatenated as a single peptide MSA. The numbers of concatenated orthogroups were 45 and 14 for AZOL and CERA WGD,

respectively. The adopted fossil calibrations followed Table S2 at corresponding phylogenetic location while the boundaries of root for AZOL WGD were set as minimum bound 168 mya based on the minimum bound of fossil calibration “Stem Lygodiaceae” and safe maximum bound 345 mya as the fossil calibration “Stem Osmundaceae” as shown in Fig. SM14. Mcmctree (v4.9j) (Yang, 2007) was implemented for the Bayesian molecular dating for each WGD with the parameters same as above. The ESS of all parameters was larger than 200, indicating adequate sampling and convergence. The posterior distribution of time estimate for the node joining the anchor pair was retrieved and the 95% HPD, posterior mean, median and mode were adopted to characterize the age of WGD, as shown in Fig. SM13.

Data availability

The fern genome and transcriptome assemblies involved in this paper were summarized in Table S1. The *Cycas* genome assembly was downloaded from CNGB data center at entry PwRftGHfPs5qG3gE. The *Selaginella*, *Physcomitrella* and *Marchantia* genome assembly were all downloaded from Phytozome v13 at genome ID 91, 318 and 320, respectively.

Reference

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**(4): 366-368.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**(1): 421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**(15): 1972-1973.

Chen H, Fang Y, Zwaenepoel A, Huang S, Van de Peer Y, Li Z. 2023. Revisiting ancient polyploidy in leptosporangiate ferns. *New Phytologist* **237**(4): 1405-1417.

Christenhusz MJM, Chase MW. 2014. Trends and concepts in fern classification. *Annals of Botany* **113**(4): 571-594.

Clark J, Hidalgo O, Pellicer J, Liu H, Marquardt J, Robert Y, Christenhusz M, Zhang S, Gibby M, Leitch IJ, et al. 2016. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytologist* **210**(3): 1072-1082.

Condamine FL, Silvestro D, Koppelhus EB, Antonelli A. 2020. The rise of angiosperms pushed conifers to decline during global cooling. *Proceedings of the National Academy of Sciences* **117**(46): 28867-28875.

dos Reis M, Donoghue PCJ, Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics* **17**(2): 71-80.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**(1): 113.

Emms DM, Kelly S. 2018. STAG: Species Tree Inference from All Genes. *bioRxiv*: 267914.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**(1): 238.

Fang Y, Qin X, Liao Q, Du R, Luo X, Zhou Q, Li Z, Chen H, Jin W, Yuan Y, et al. 2022. The genome of homosporous maidenhair fern sheds light on the euphyllophyte evolution and defences. *Nature Plants*.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23): 3150-3152.

Herendeen PS, Friis EM, Pedersen KR, Crane PR. 2017. Palaeobotanical redux: revisiting the age of the angiosperms. *Nature Plants* **3**(3): 17015.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**(2): 518-522.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**(6): 587-589.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**(4): 772-780.

Kinosian SP, Rowe CA, Wolf PG. 2022. Why Do Heterosporous Plants Have So Few Chromosomes? *Frontiers in Plant Science* **13**.

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**(7): 1307-1320.

Lehtonen S, Silvestro D, Karger DN, Scotese C, Tuomisto H, Kessler M, Peña C, Wahlberg N, Antonelli A. 2017. Environmentally driven extinction and opportunistic origination explain fern diversification patterns. *Scientific Reports* **7**(1): 4831.

Li F-W, Nishiyama T, Waller M, Frangedakis E, Keller J, Li Z, Fernandez-Pozo N, Barker MS, Bennett T, Blázquez MA, et al. 2020. Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature Plants* **6**(3): 259-272.

Löytynoja A 2014. Phylogeny-aware alignment with PRANK. In: Russell DJ ed. *Multiple Sequence Alignment Methods*. Totowa, NJ: Humana Press, 155-170.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**(1): 268-274.

Pelosi JA, Kim EH, Barbazuk WB, Sessa EB. 2022. Phylotranscriptomics Illuminates the Placement of Whole Genome Duplications and Gene Retention in Ferns. *Frontiers in Plant Science* **13**.

PPG. 2016. A community-derived classification for extant lycophytes and ferns. *Journal of Systematics and Evolution* **54**(6): 563-603.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**(3): e9490.

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

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* **61**(3): 539-542.

Rothfels CJ, Li F-W, Sigel EM, Huiet L, Larsson A, Burge DO, Ruhsam M, Deyholos M, Soltis DE, Stewart Jr. CN, et al. 2015. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *American Journal of Botany* **102**(7): 1089-1107.

Sensalari C, Maere S, Lohaus R. 2021. ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* **38**(2): 530-532.

Shen H, Jin D, Shu J-P, Zhou X-L, Lei M, Wei R, Shang H, Wei H-J, Zhang R, Liu L, et al. 2017. Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. *GigaScience* **7**(2).

Steenwyk JL, Goltz DC, Buida TJ, III, Li Y, Shen X-X, Rokas A. 2022. OrthoSNAP: A tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. *PLOS Biology* **20**(10): e3001827.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology* **62**(6): 901-912.

Van Dongen SM. 2000. *Graph clustering by flow simulation.*

Wang F-G, Wang A-H, Bai C-K, Jin D-M, Nie L-Y, Harris AJ, Che L, Wang J-J, Li S-Y, Xu L, et al. 2022. Genome size evolution of the extant lycophytes and ferns. *Plant Diversity* **44**(2): 141-152.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**(8): 1586-1591.

Zhang C, Mirarab S. 2022. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**(21): 4949-4950.

Zhong Y, Liu Y, Wu W, Chen J, Sun C, Liu H, Shu J, Ebihara A, Yan Y, Zhou R, et al. 2022. Genomic Insights into Genetic Diploidization in the Homosporous Fern Adiantum nelumboides. *Genome Biology and Evolution* **14**(8).

Zwaenepoel A, Van de Peer Y. 2018. wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**(12): 2153-2155.

Zwaenepoel A, Van de Peer Y. 2019. Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates. *Molecular Biology and Evolution* **36**(7): 1384-1404.

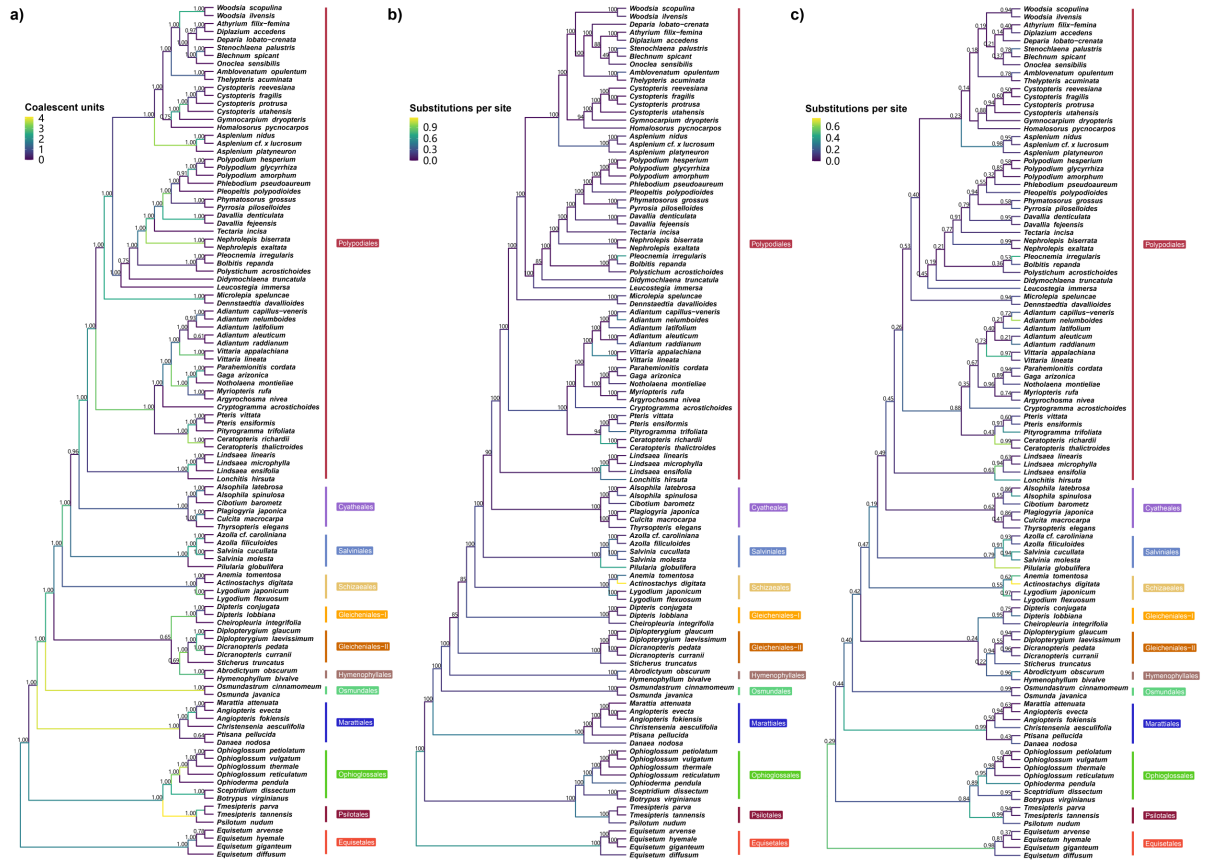


Fig. SM1 | Fern cladogram inferred from 107 ferns dataset of horsetails outgroup. a) fern cladogram inferred by ASTRAL-Pro2, wherein branches were colored in terms of the coalescent units and support values for bipartitions representing local posterior probability. b) fern cladogram inferred by concatenation-based method, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing ultrafast bootstrap approximation. c) fern cladogram inferred by STAG, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing the proportion of gene trees manifesting that bipartition. The clade strips indicating affiliations at order level were exhibited as vertical bars with distinct colors.

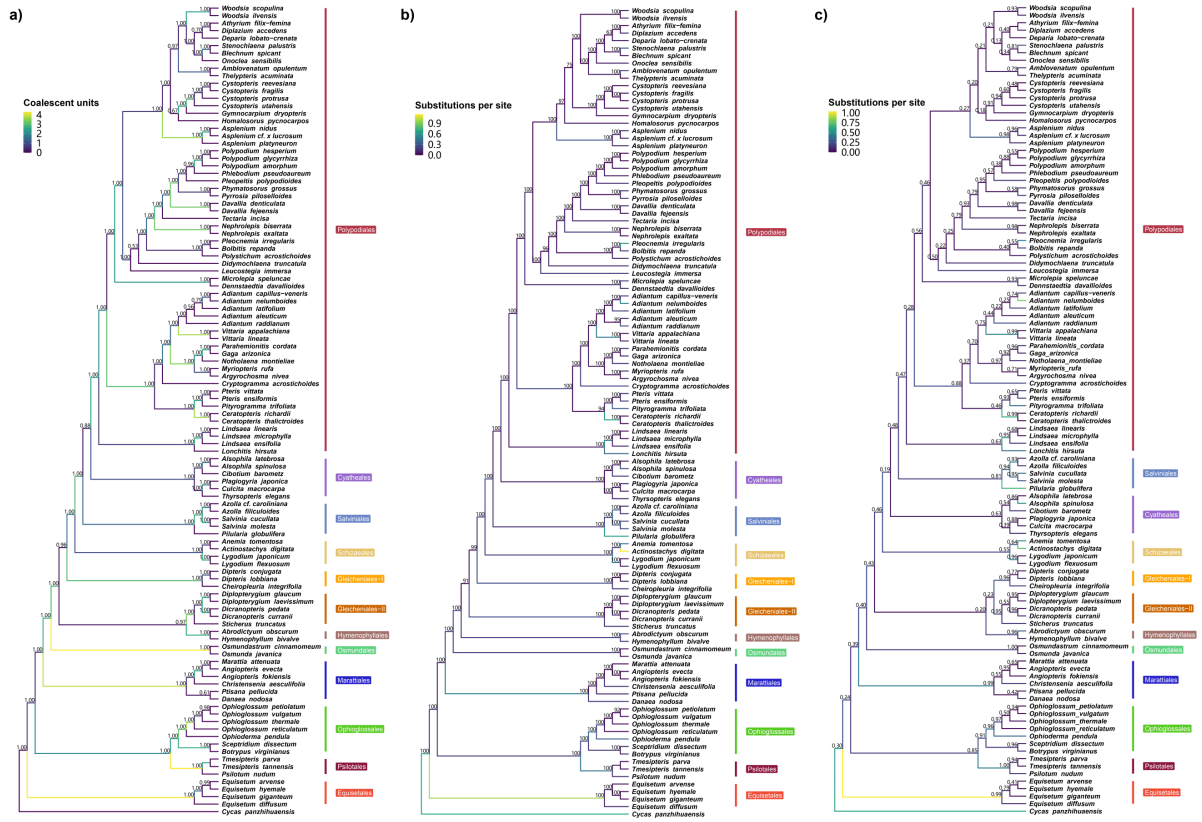


Fig. SM2 | Fern cladogram inferred from 107 ferns dataset of seed plants outgroup. a) fern cladogram inferred by ASTRAL-Pro2, wherein branches were colored in terms of the coalescent units and support values for bipartitions representing local posterior probability. b) fern cladogram inferred by concatenation-based method, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing ultrafast bootstrap approximation. c) fern cladogram inferred by STAG, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing the proportion of gene trees manifesting that bipartition. The clade strips indicating affiliations at order level were exhibited as vertical bars with distinct colors.

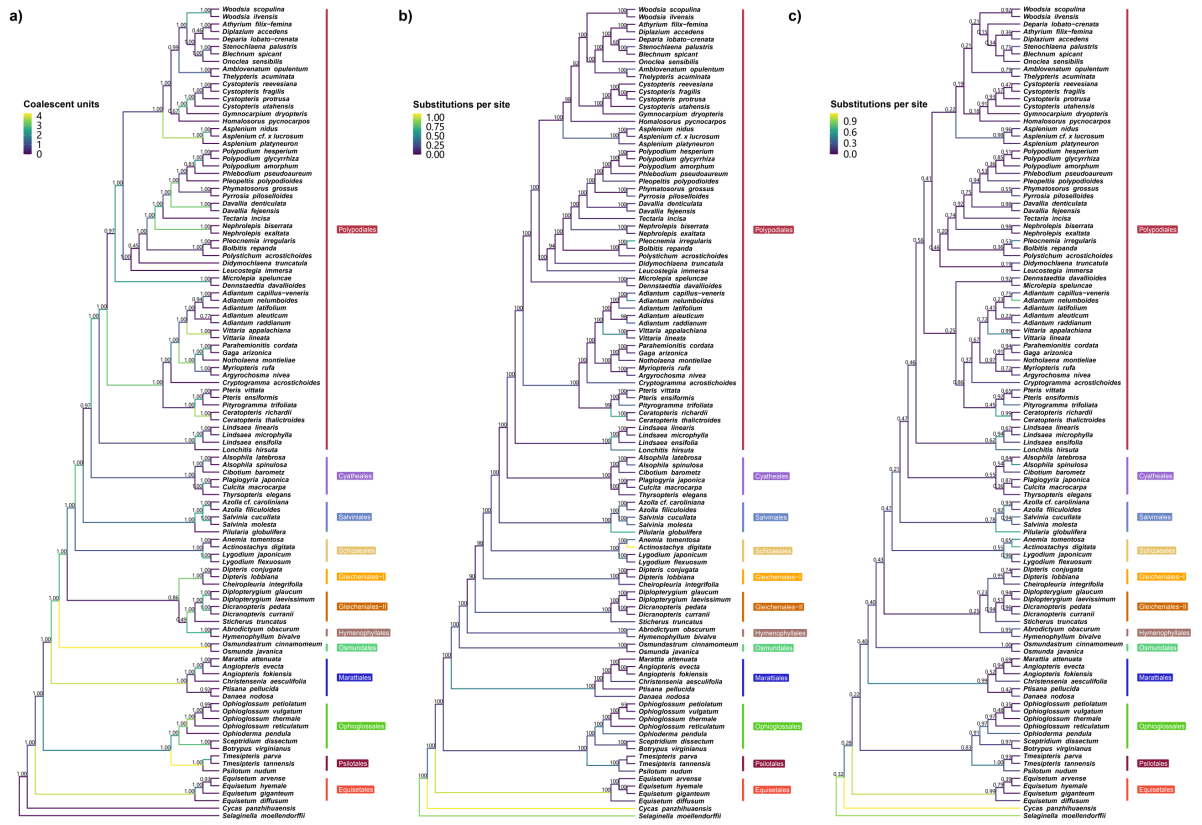


Fig. SM3 | Fern cladogram inferred from 107 ferns dataset of seed plants plus Lycopods outgroup. a) fern cladogram inferred by ASTRAL-Pro2, wherein branches were colored in terms of the coalescent units and support values for bipartitions representing local posterior probability. b) fern cladogram inferred by concatenation-based method, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing ultrafast bootstrap approximation. c) fern cladogram inferred by STAG, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing the proportion of gene trees manifesting that bipartition. The clade strips indicating affiliations at order level were exhibited as vertical bars with distinct colors.

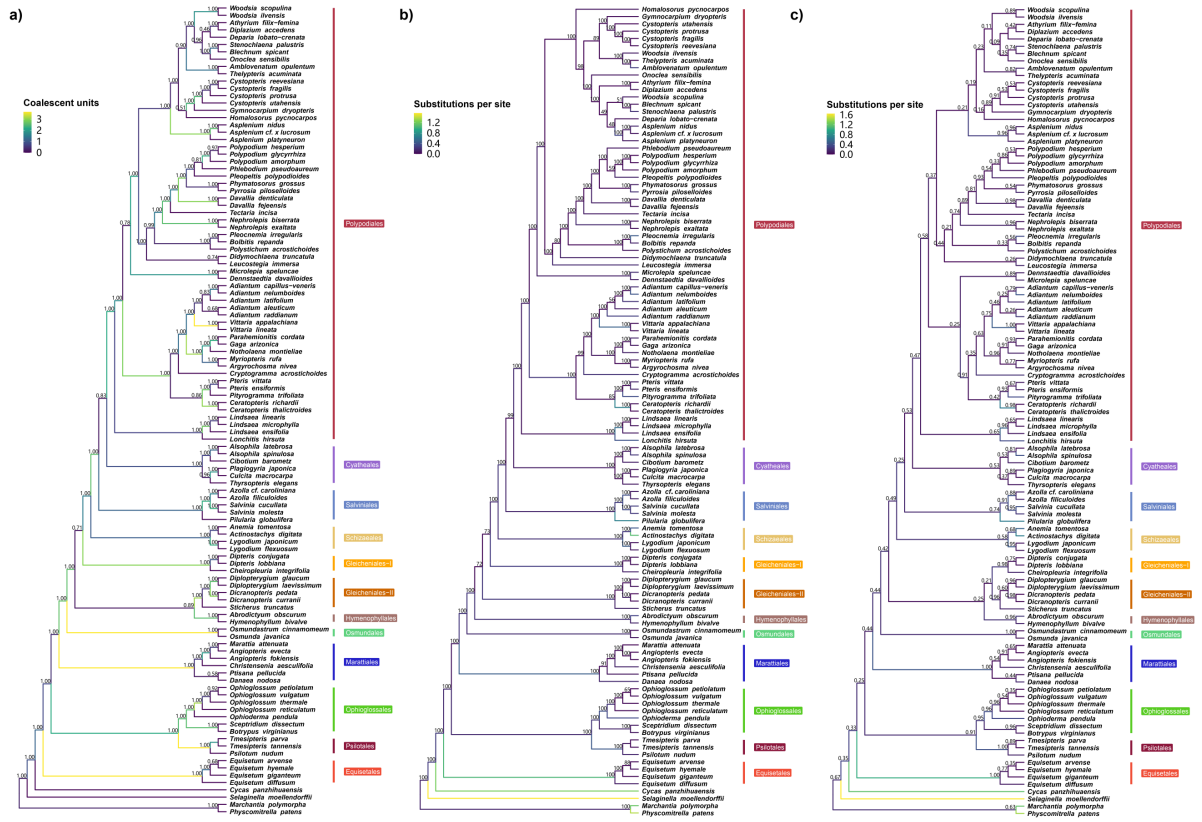


Fig. SM4 | Fern cladogram inferred from 107 ferns dataset of seed plants plus Lycopods and Bryophytes outgroup. a) fern cladogram inferred by ASTRAL-Pro2, wherein branches were colored in terms of the coalescent units and support values for bipartitions representing local posterior probability. b) fern cladogram inferred by concatenation-based method, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing ultrafast bootstrap approximation. c) fern cladogram inferred by STAG, wherein branches were colored in terms of the substitutions per site and support values for bipartitions representing the proportion of gene trees manifesting that bipartition. The clade strips indicating affiliations at order level were exhibited as vertical bars with distinct colors.

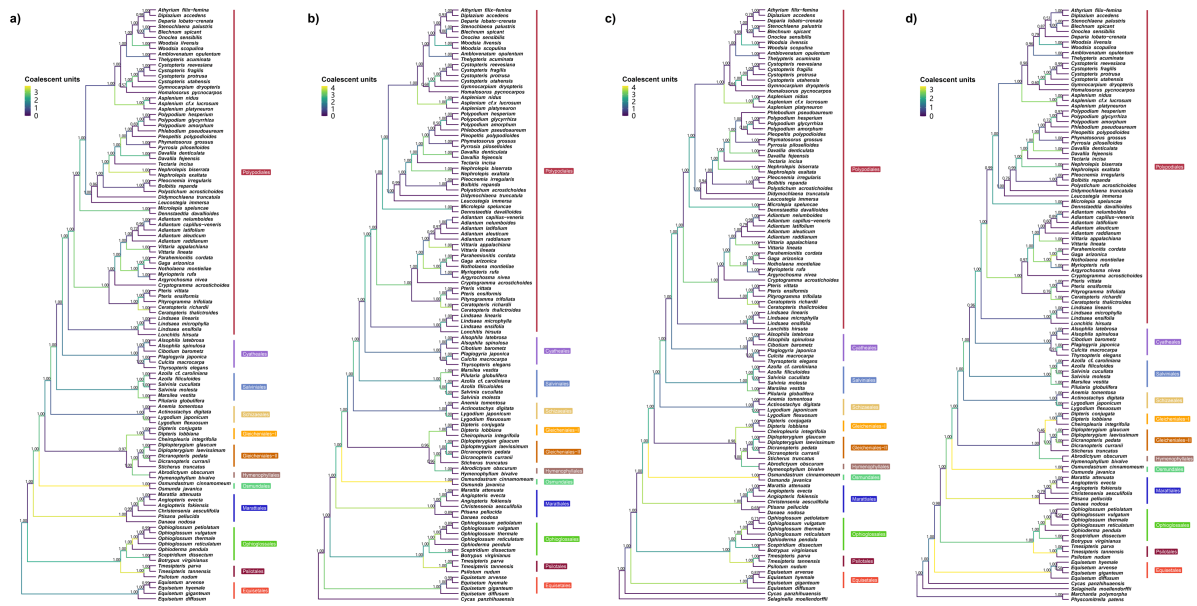


Fig. SM5 | Fern cladogram inferred from nucleotide alignment of 108 fern dataset using ASTRAL-Pro2. a) fern cladogram inferred from horsetails outgroup dataset. b) fern cladogram inferred from seed plants outgroup dataset. c) fern cladogram inferred from seed plants plus Lycopods outgroup dataset. d) fern cladogram inferred from seed plants plus Lycopods and Bryophytes outgroup dataset. Branches were colored in terms of the coalescent units and support values for bipartitions representing local posterior probability. The clade strips indicating affiliations at order level were exhibited as vertical bars with distinct colors.

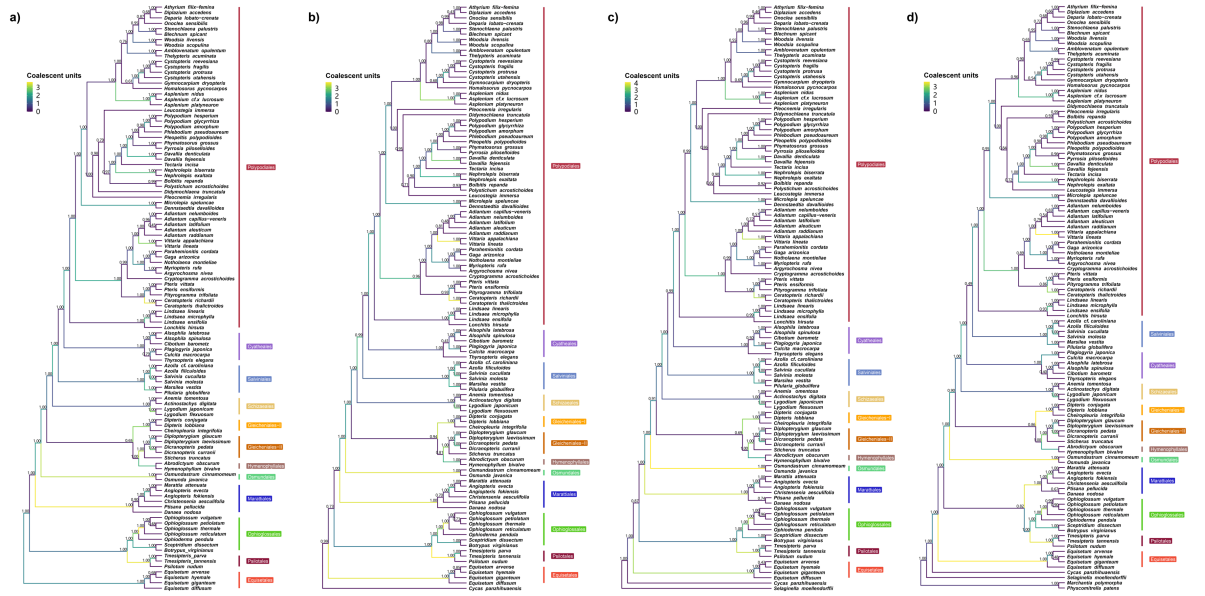


Fig. SM6 | Fern cladogram inferred from peptide alignment of 108 fern dataset using ASTRAL-Pro2. a) fern cladogram inferred from horsetails outgroup dataset. b) fern cladogram inferred from seed plants outgroup dataset. c) fern cladogram inferred from seed plants plus Lycopods outgroup dataset. d) fern cladogram inferred from seed plants plus Lycopods and Bryophytes outgroup dataset. Branches were colored in terms of the coalescent units and support values for bipartitions representing local posterior probability. The clade strips indicating affiliations at order level were exhibited as vertical bars with distinct colors.

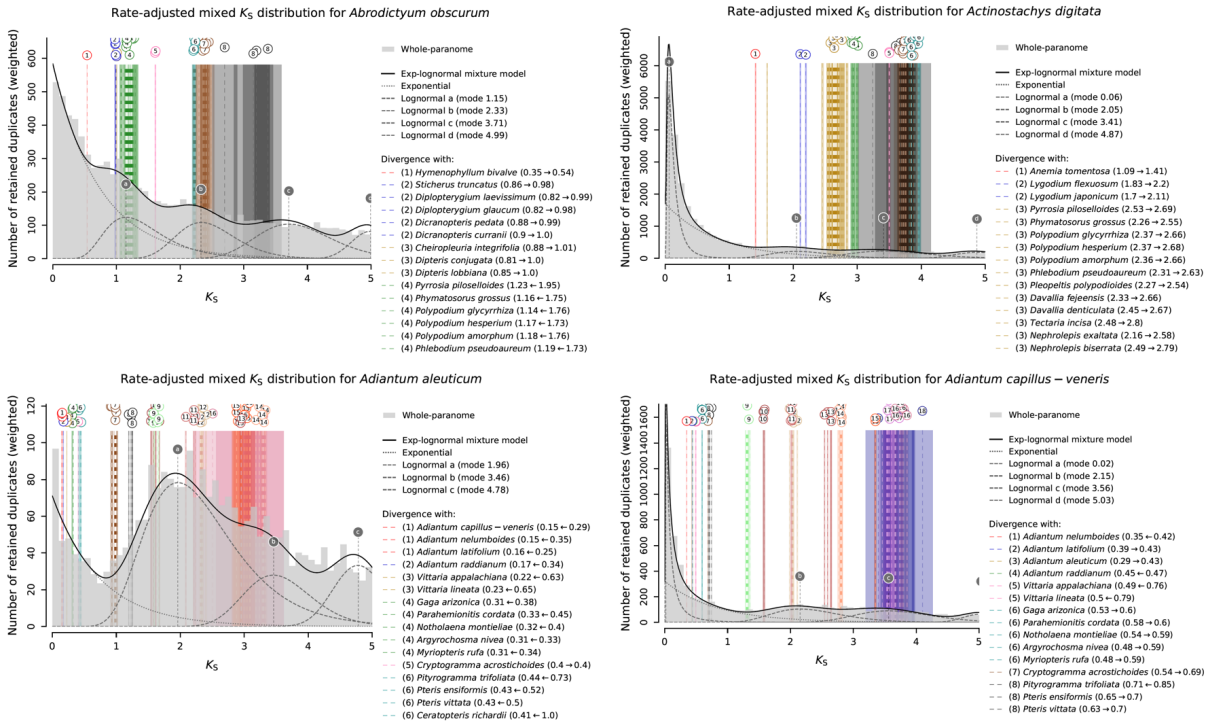


Fig. SM7 | The rate-adjusted mixed K_S-age distributions of whole paranome for 108 ferns. K_S-age distributions for the whole paranome were overlaid with rate-corrected speciation events in colored vertical lines. The overall mixture model in the dark solid line of each paralogous K_S-age distribution consists of an exponential component in dotted gray curve and optimized log-normal components in dashed gray curves. Each log-normal component is labeled with a letter, shown as vertical dashed gray lines with circular labels. Rate-corrected modes of orthologous K_S-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the K_S shifts resulted from the substitution rate correction. Speciation events were truncated for presenting to fit the space while complete representation of speciation events is available in supplementary documents.

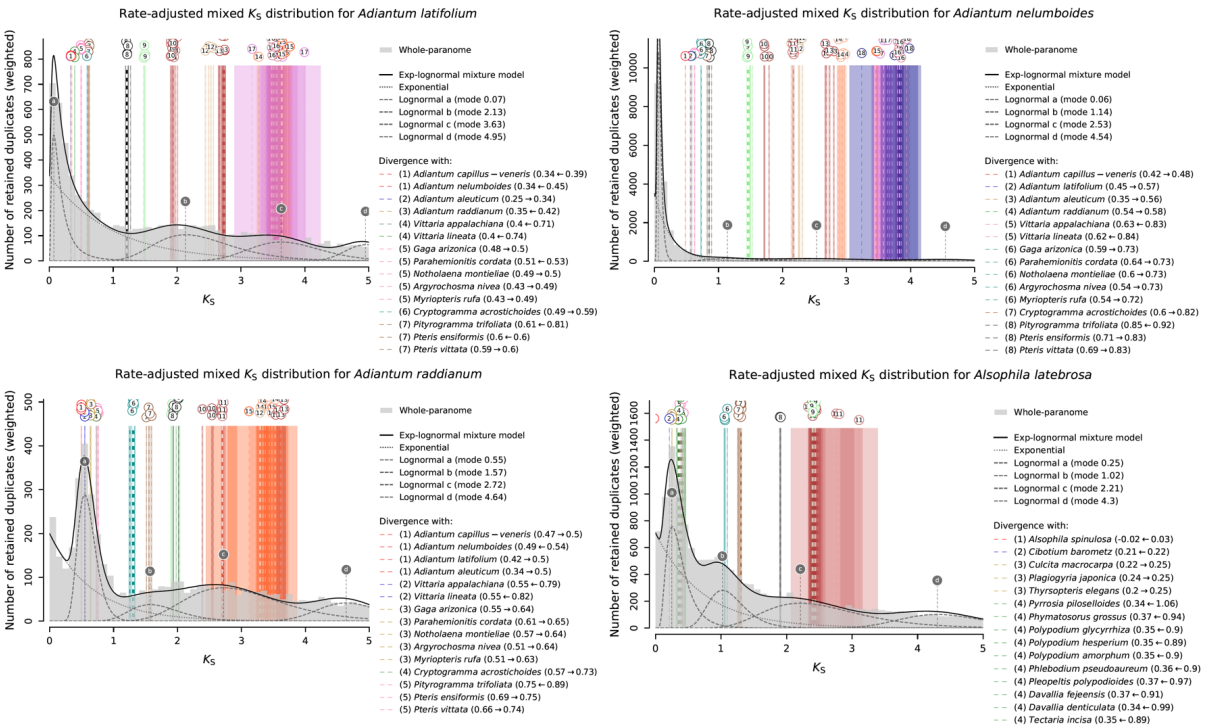


Fig. SM7 (continued)

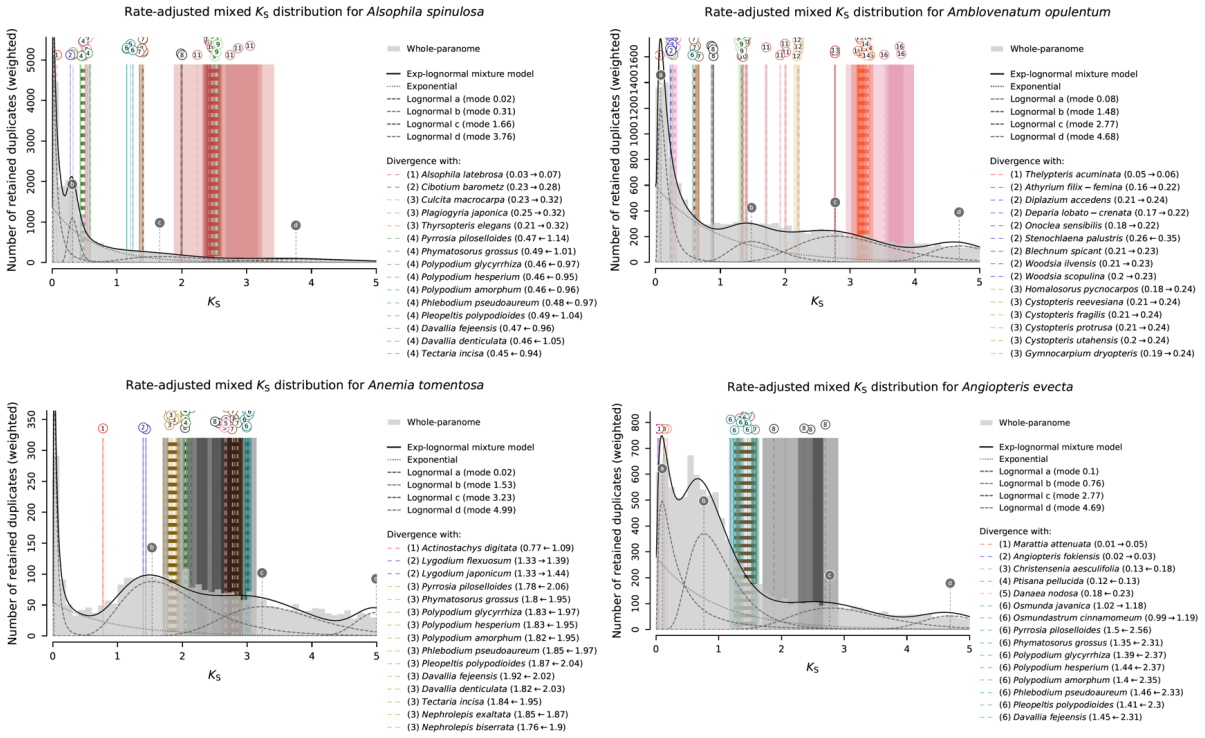


Fig. SM7 (continued)

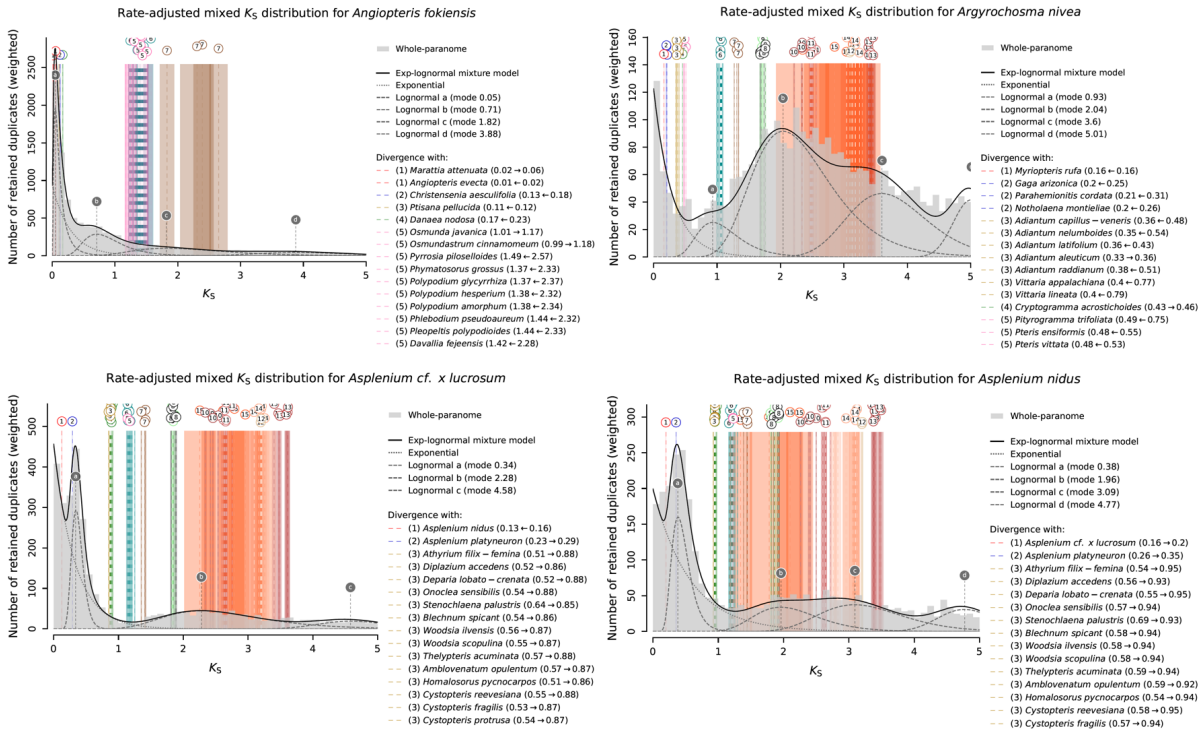


Fig. SM7 (continued)

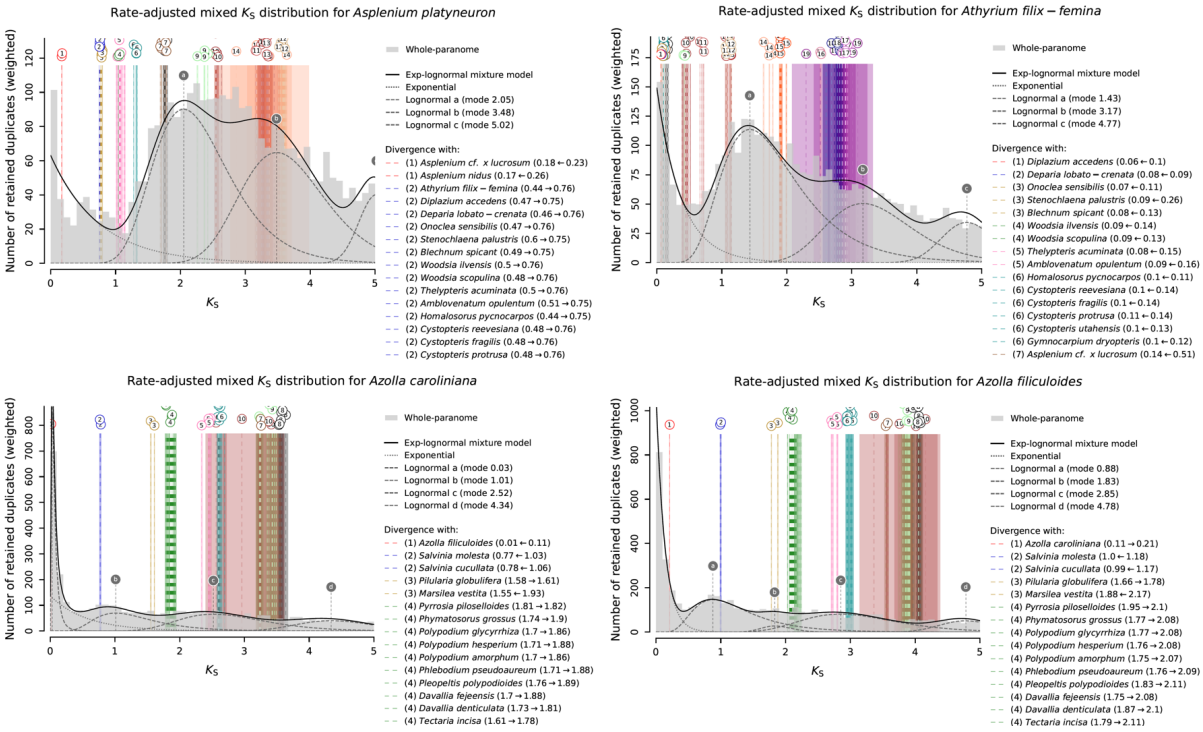


Fig. SM7 (continued)

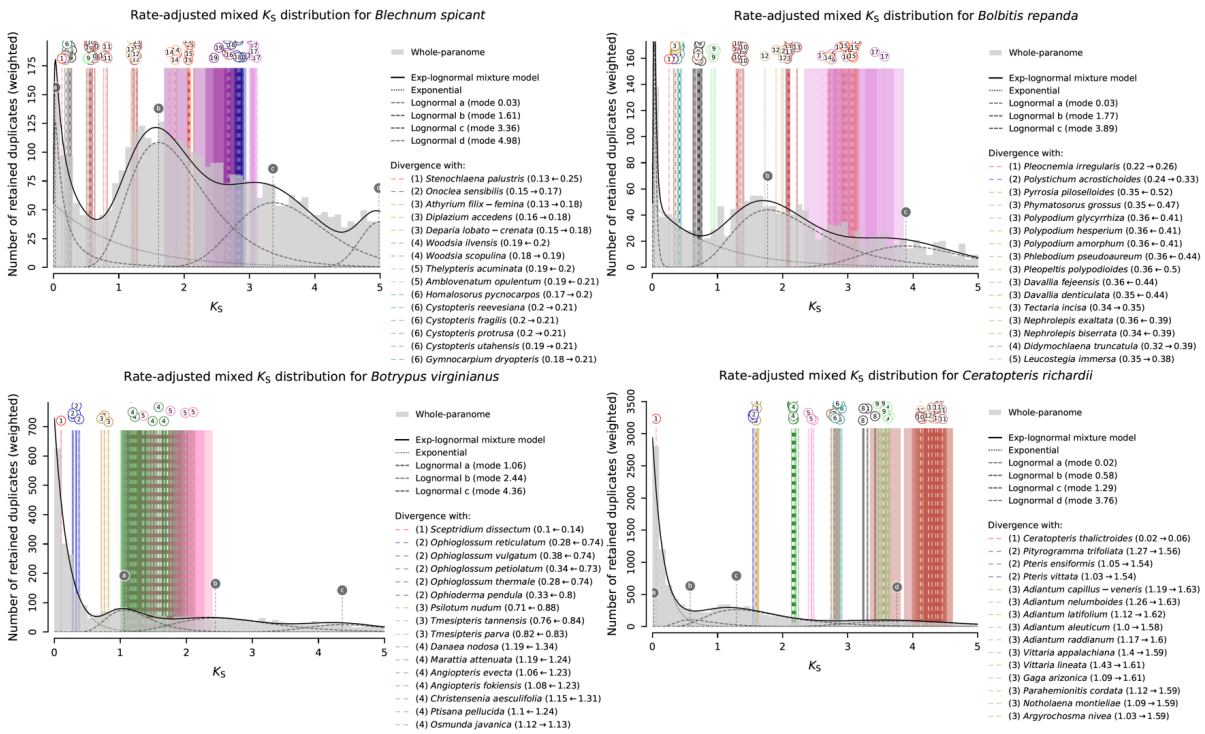


Fig. SM7 (continued)

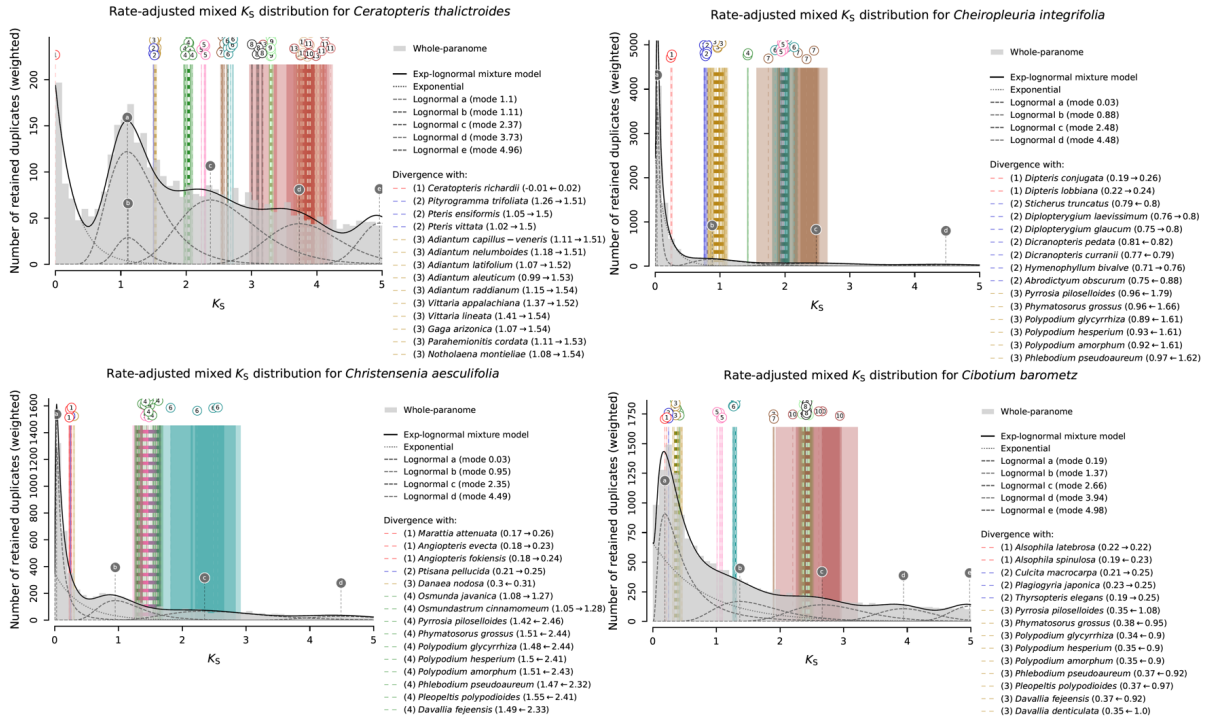


Fig. SM7 (continued)

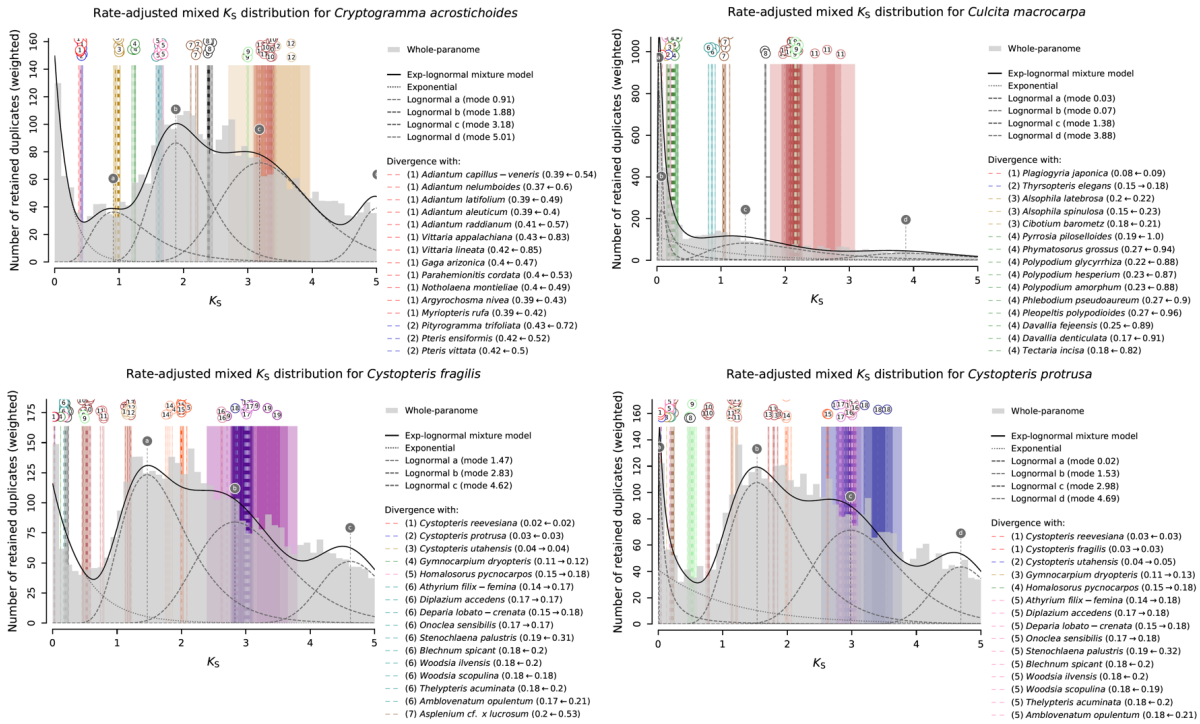


Fig. SM7 (continued)

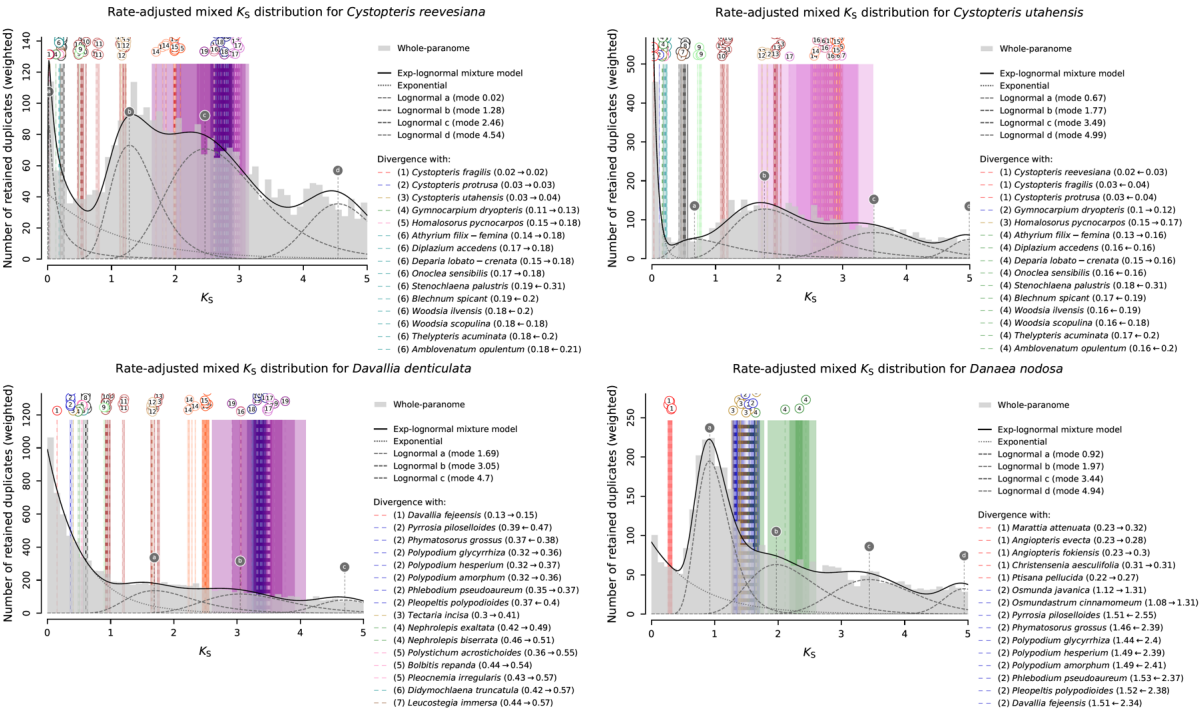


Fig. SM7 (continued)

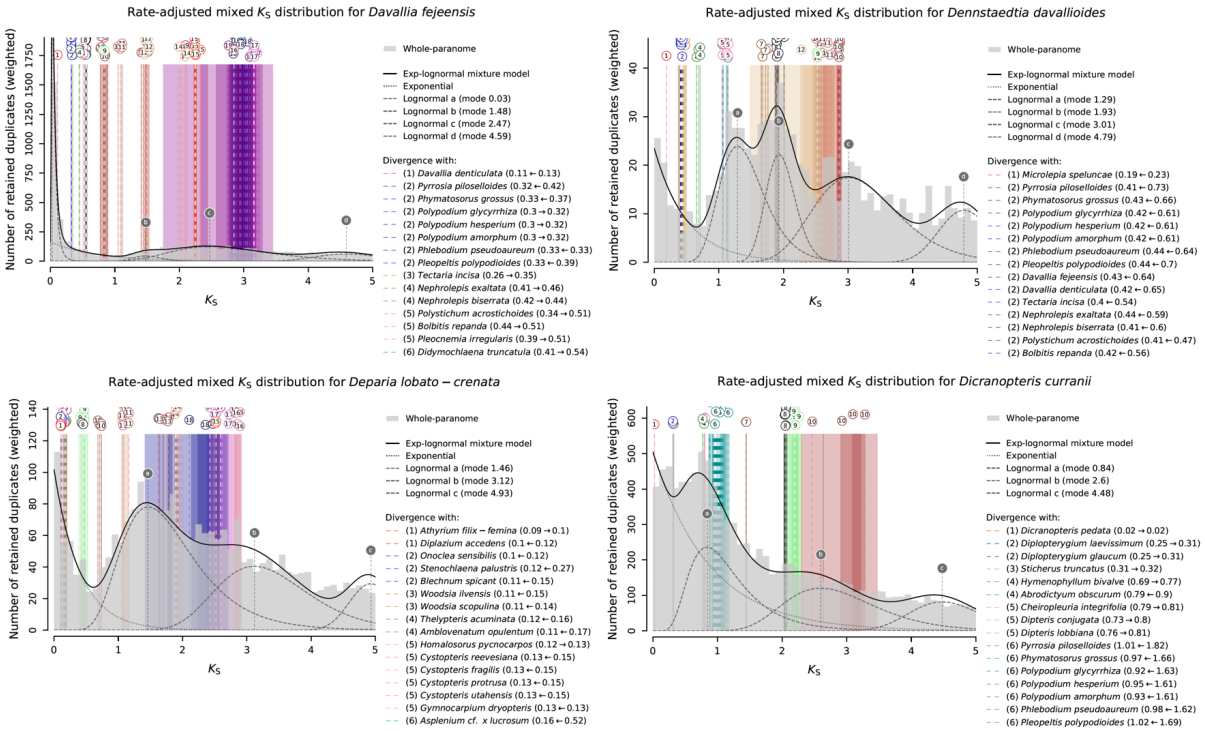


Fig. SM7 (continued)

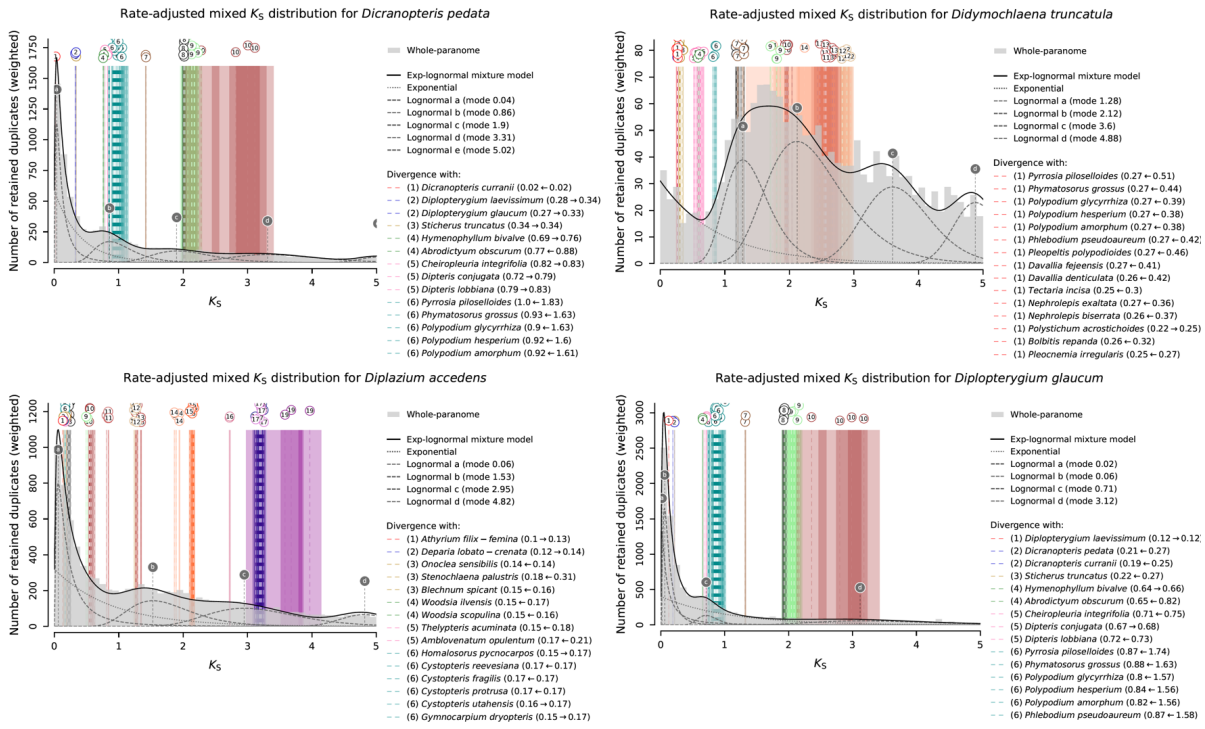


Fig. SM7 (continued)

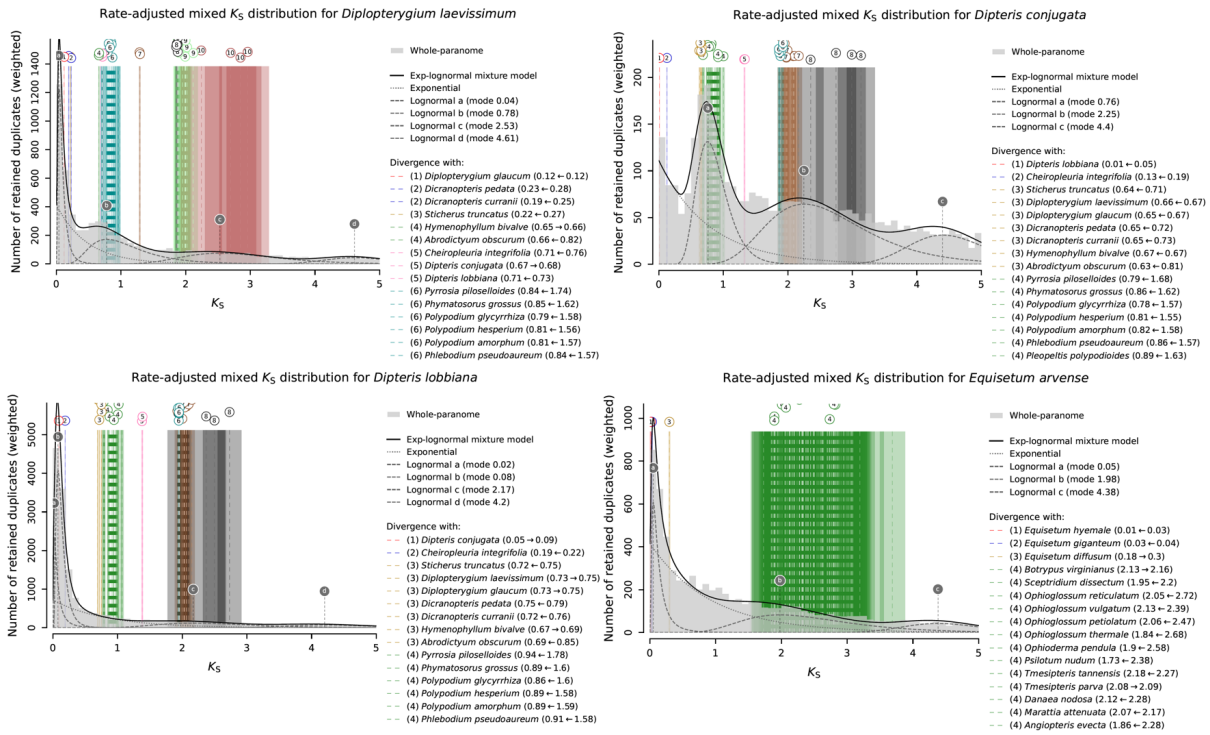


Fig. SM7 (continued)

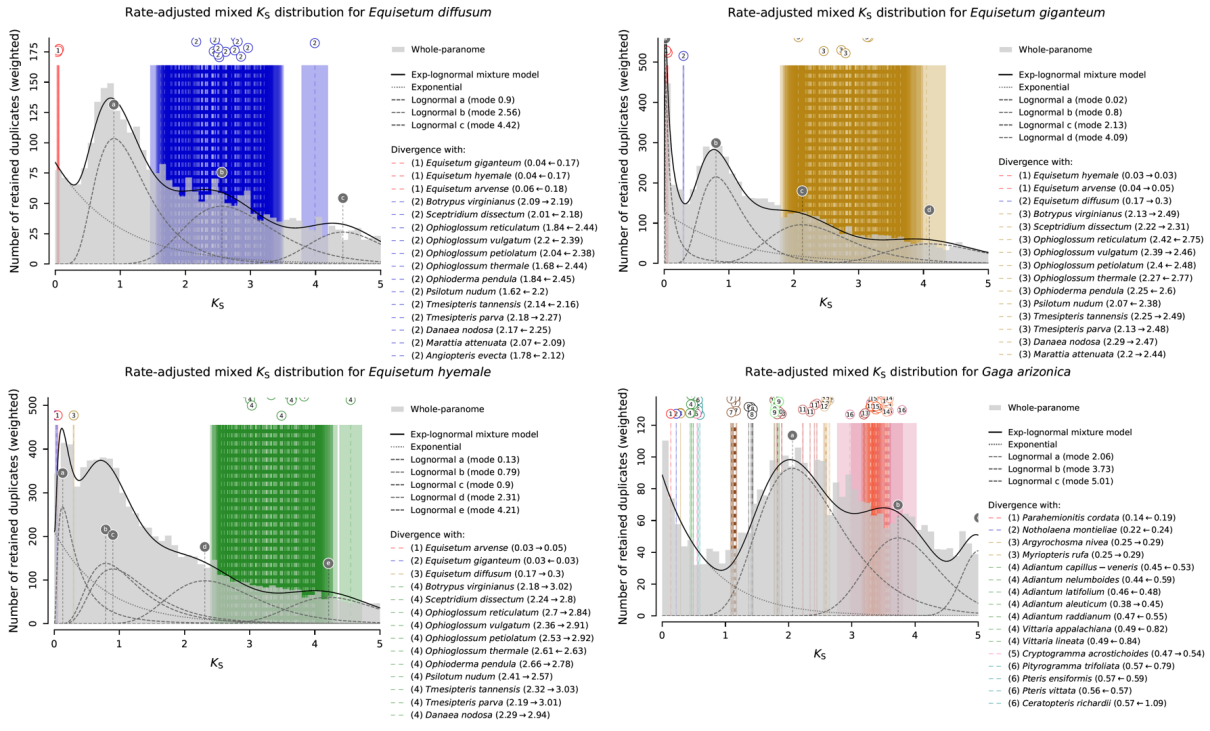


Fig. SM7 (continued)

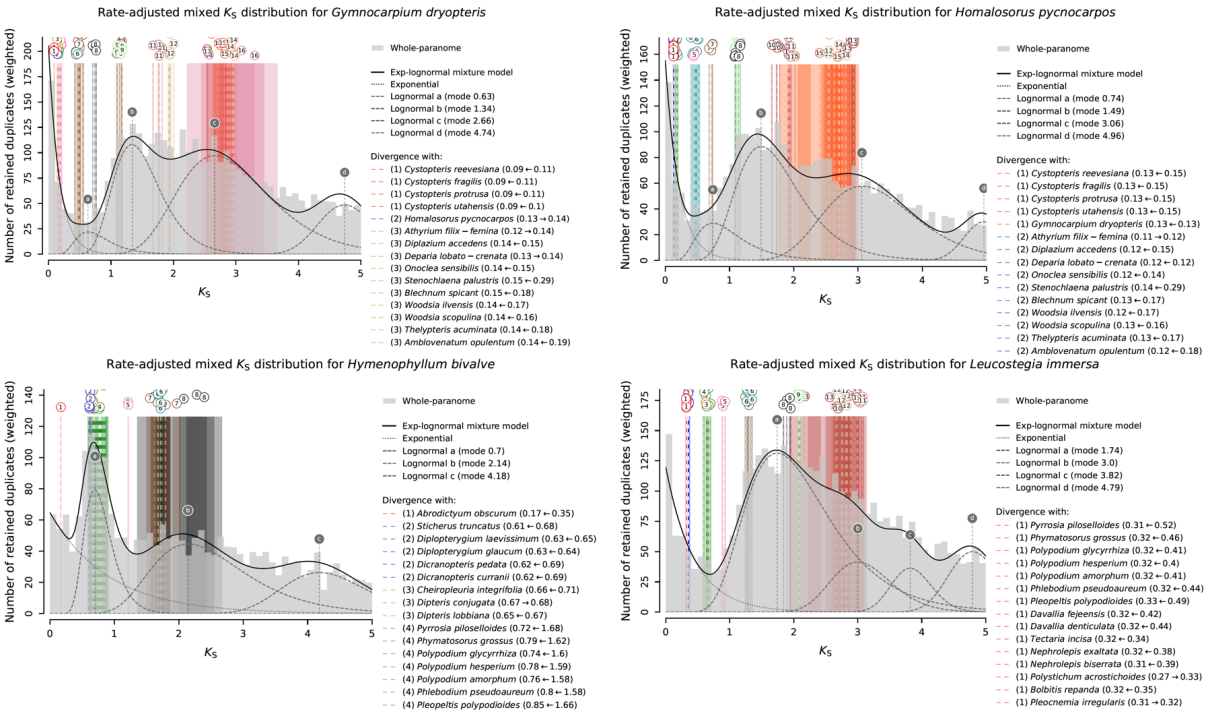


Fig. SM7 (continued)

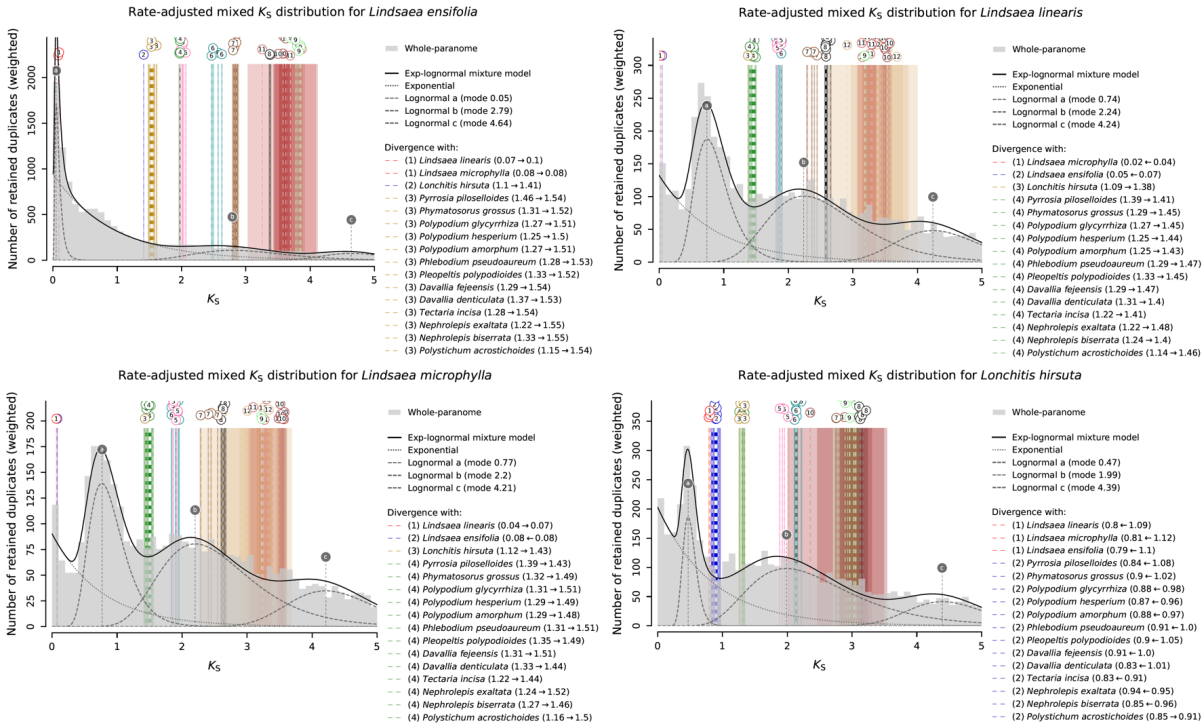


Fig. SM7 (continued)

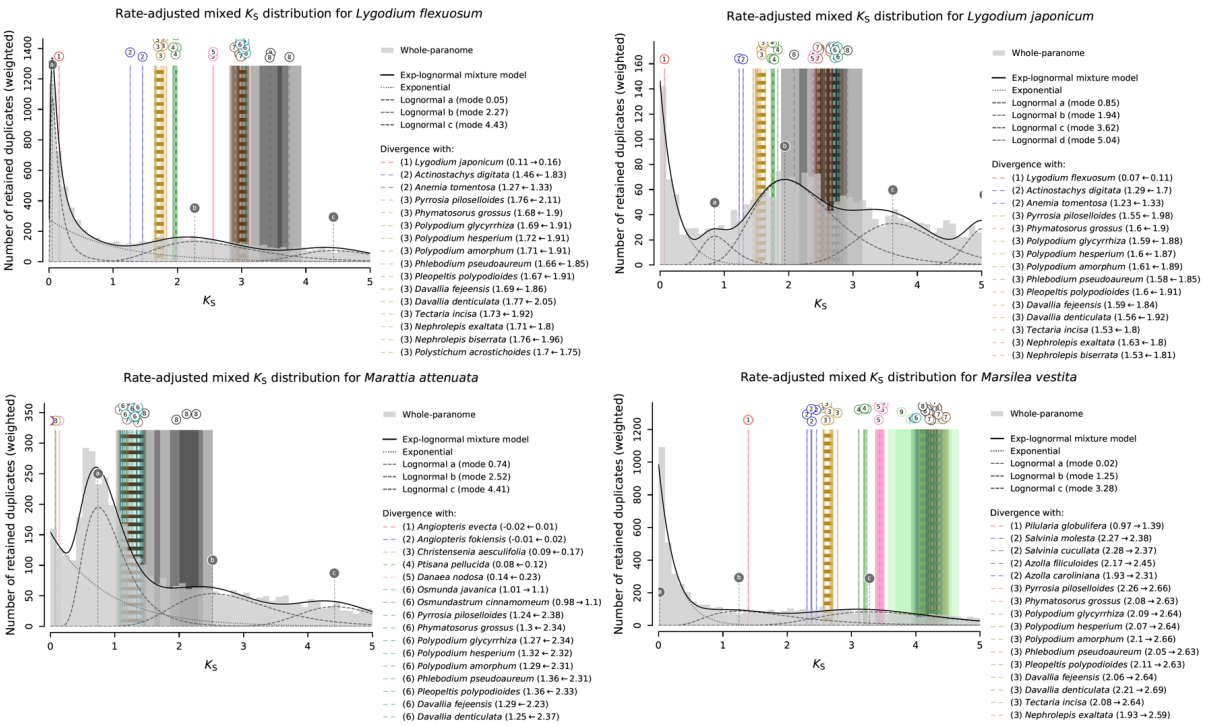


Fig. SM7 (continued)

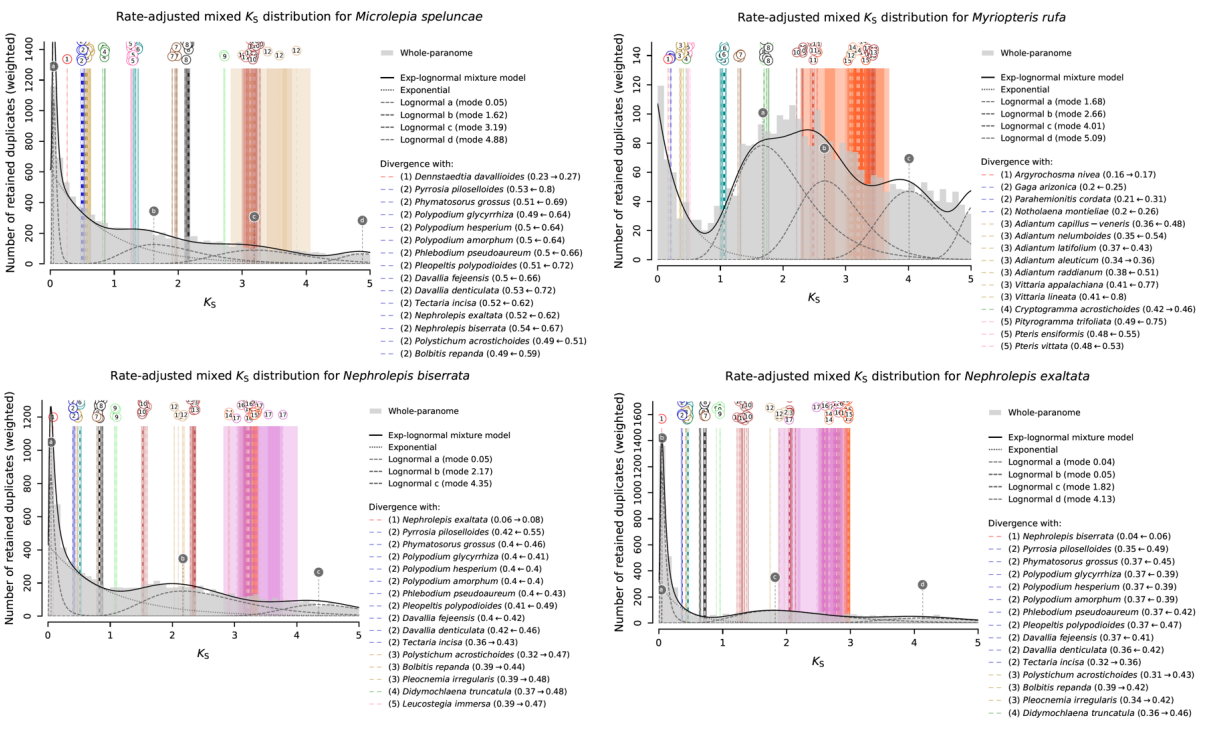


Fig. SM7 (continued)

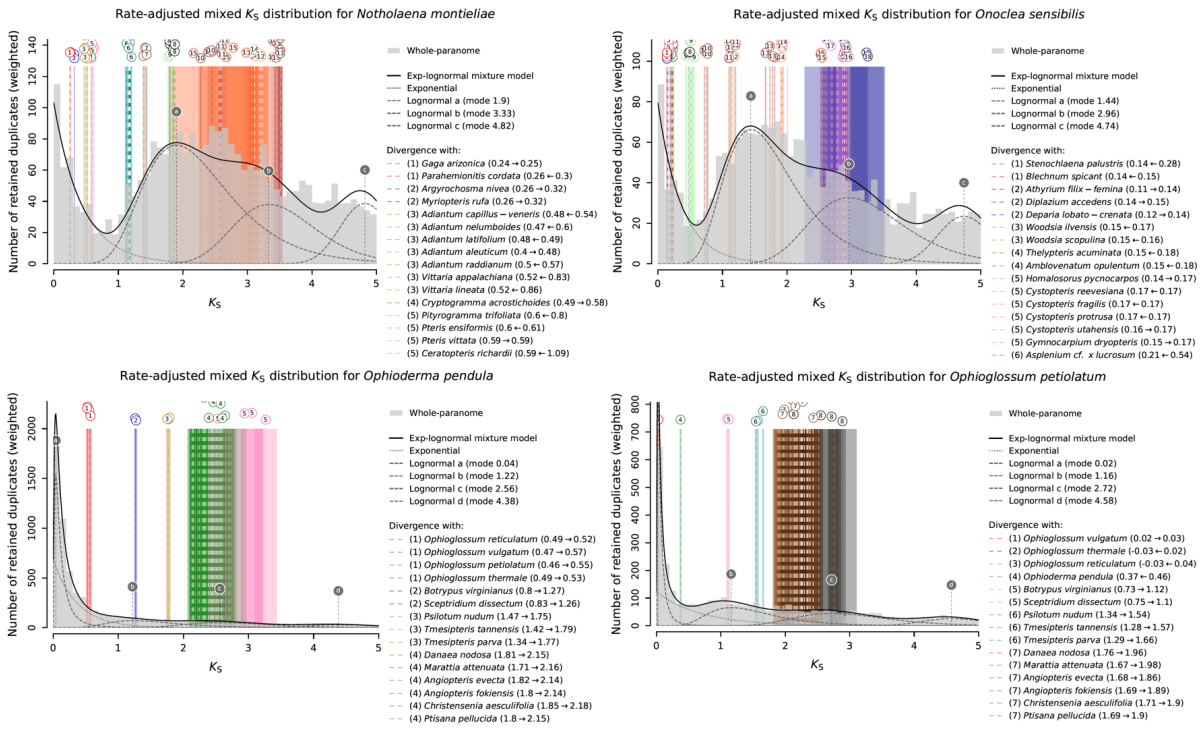


Fig. SM7 (continued)

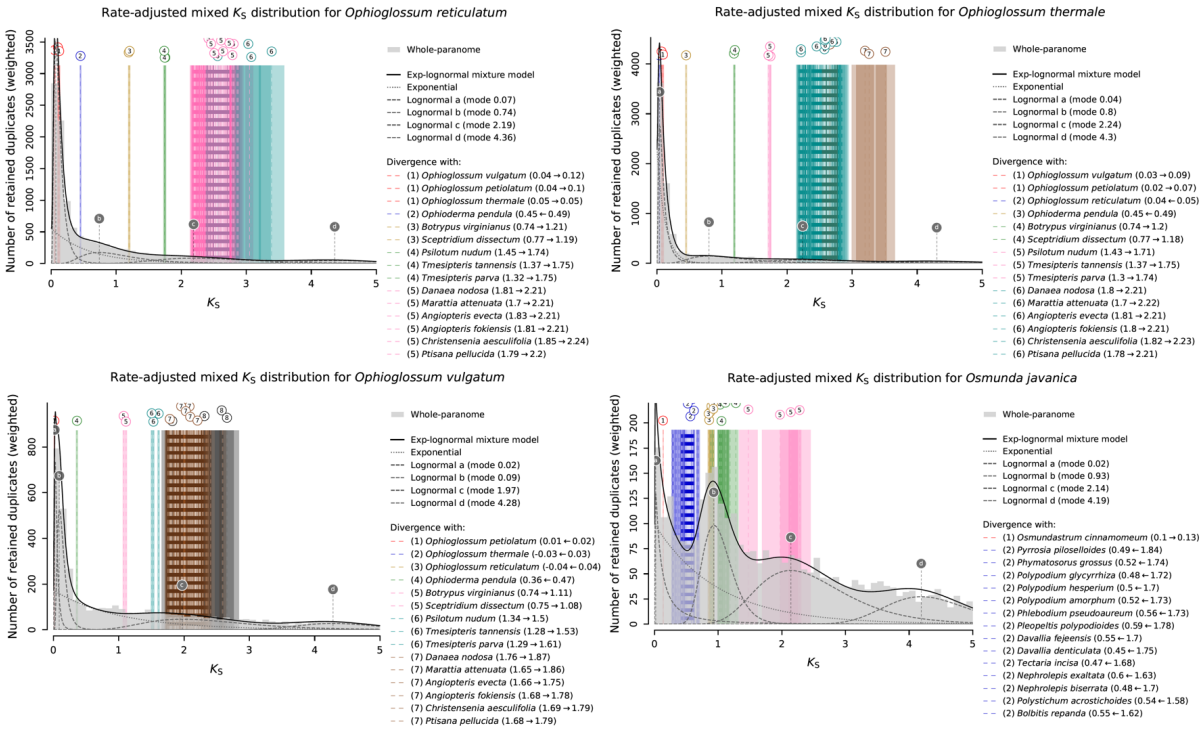


Fig. SM7 (continued)

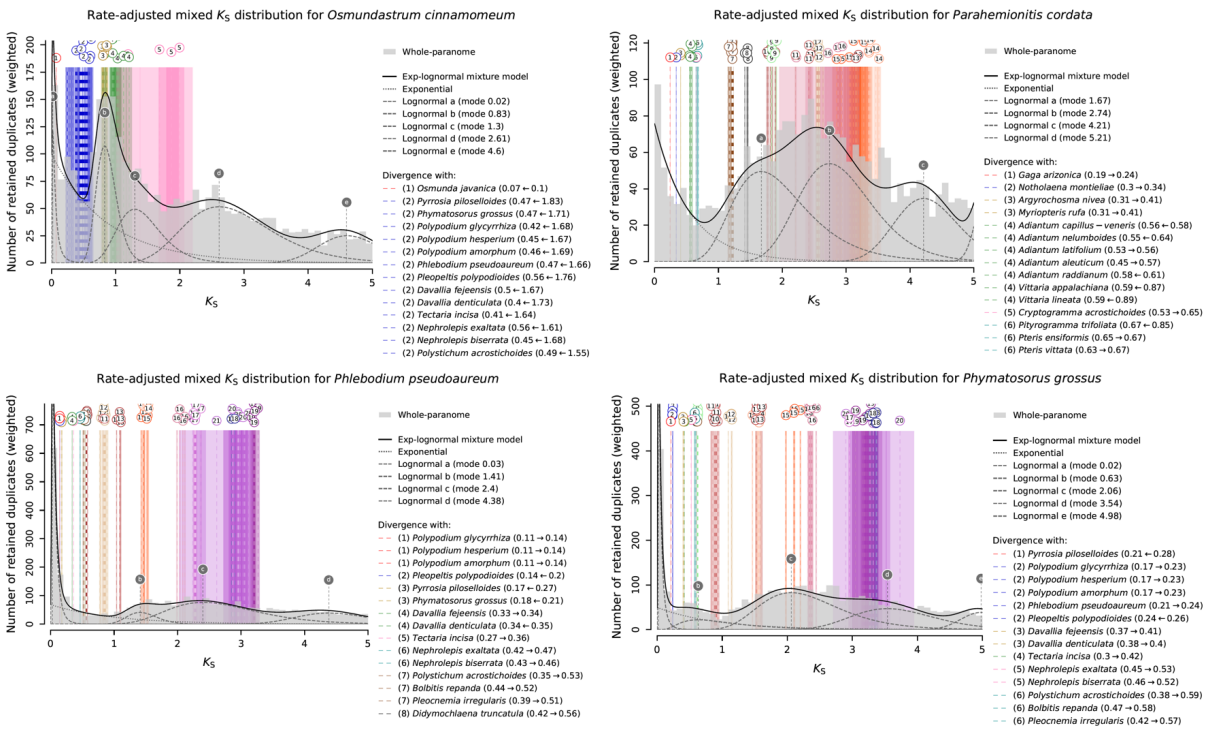


Fig. SM7 (continued)

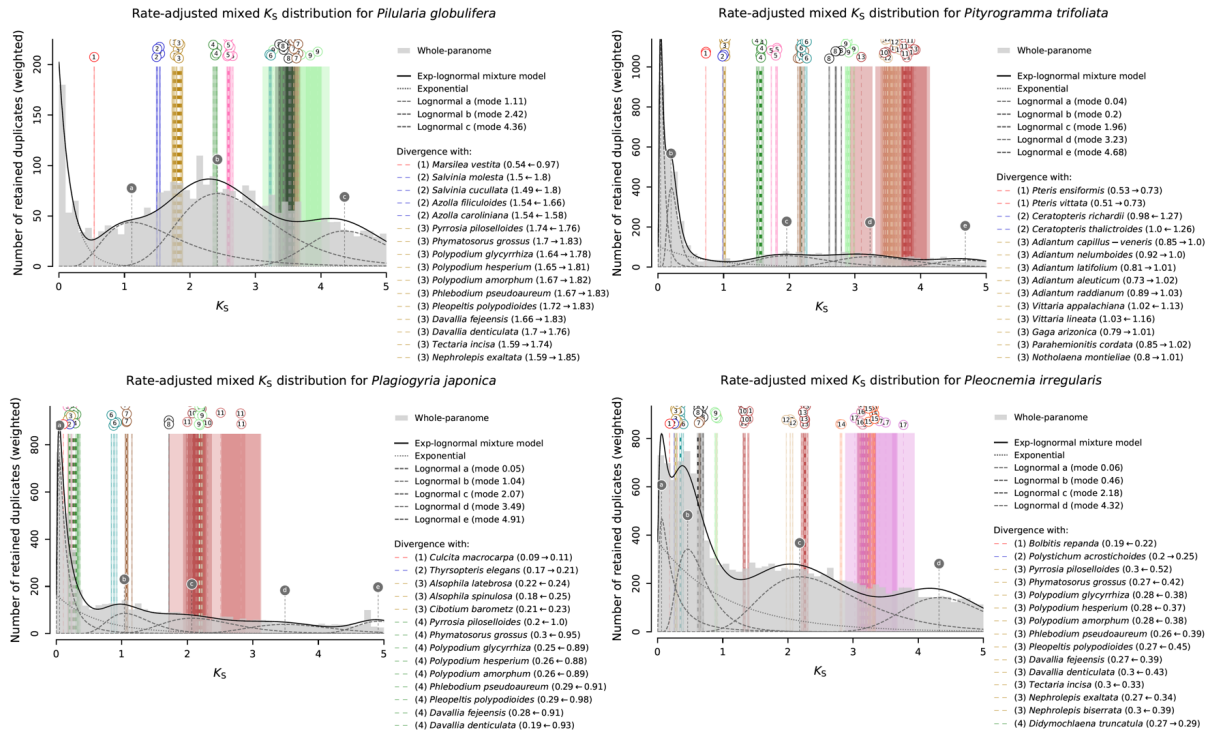


Fig. SM7 (continued)

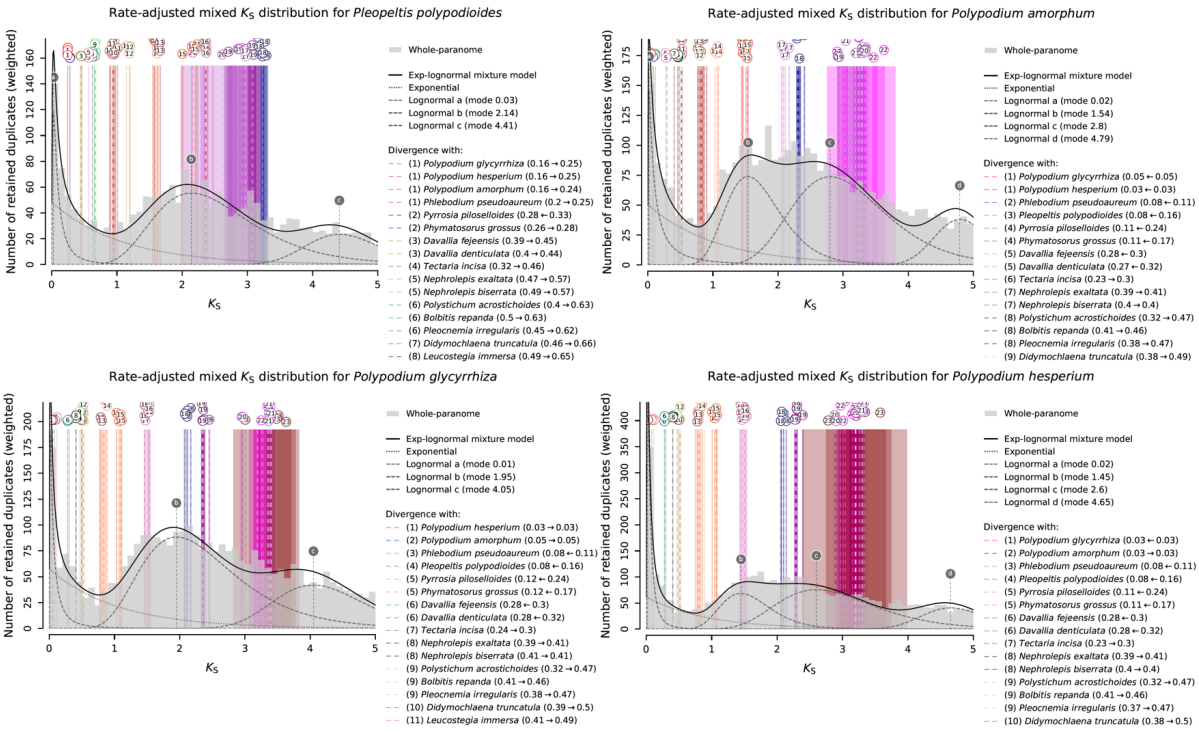


Fig. SM7 (continued)

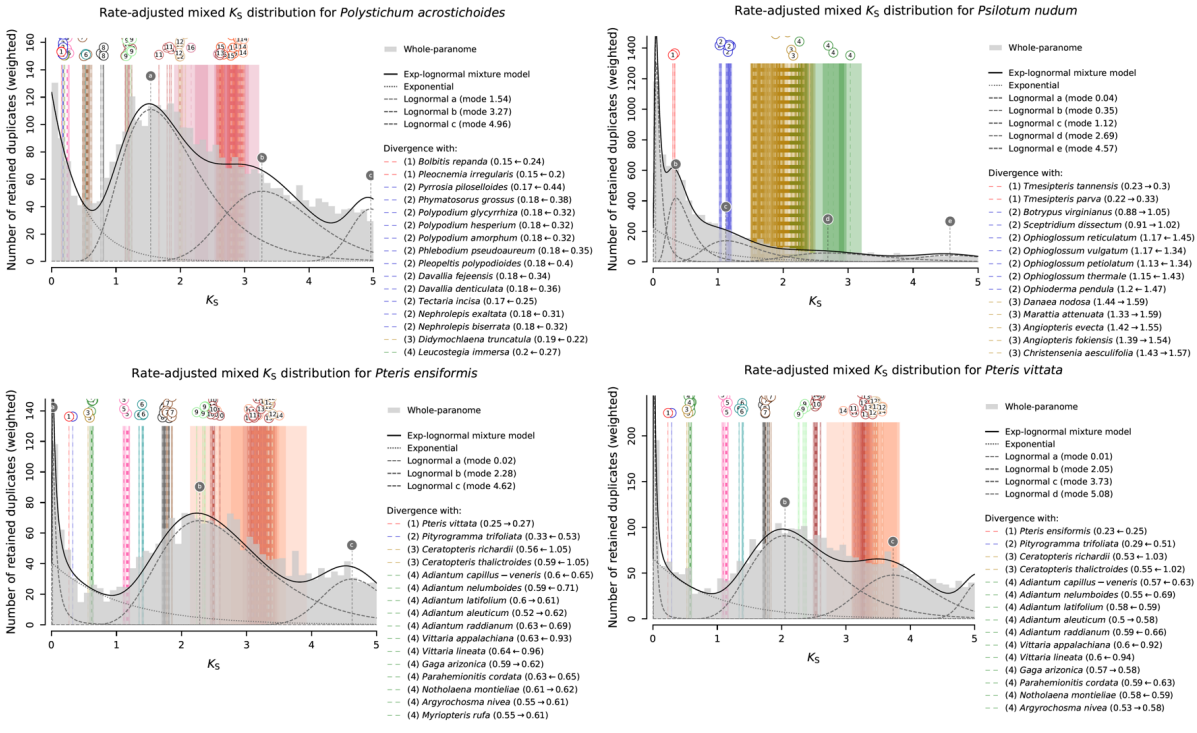


Fig. SM7 (continued)

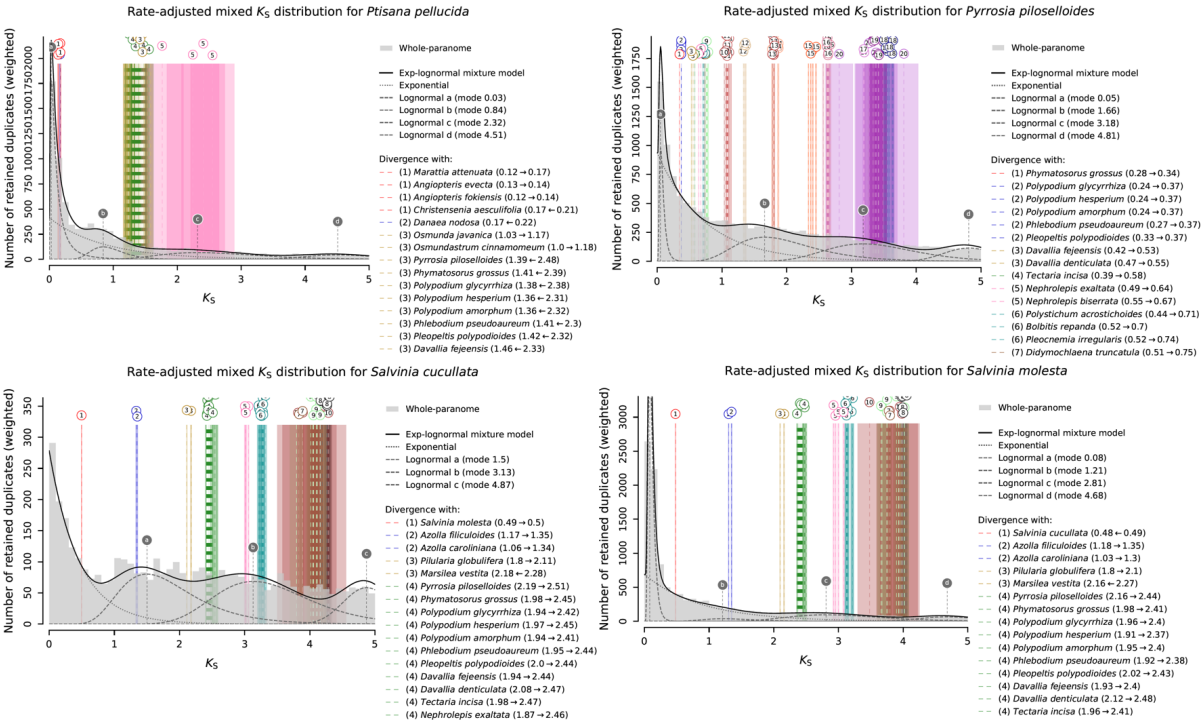


Fig. SM7 (continued)

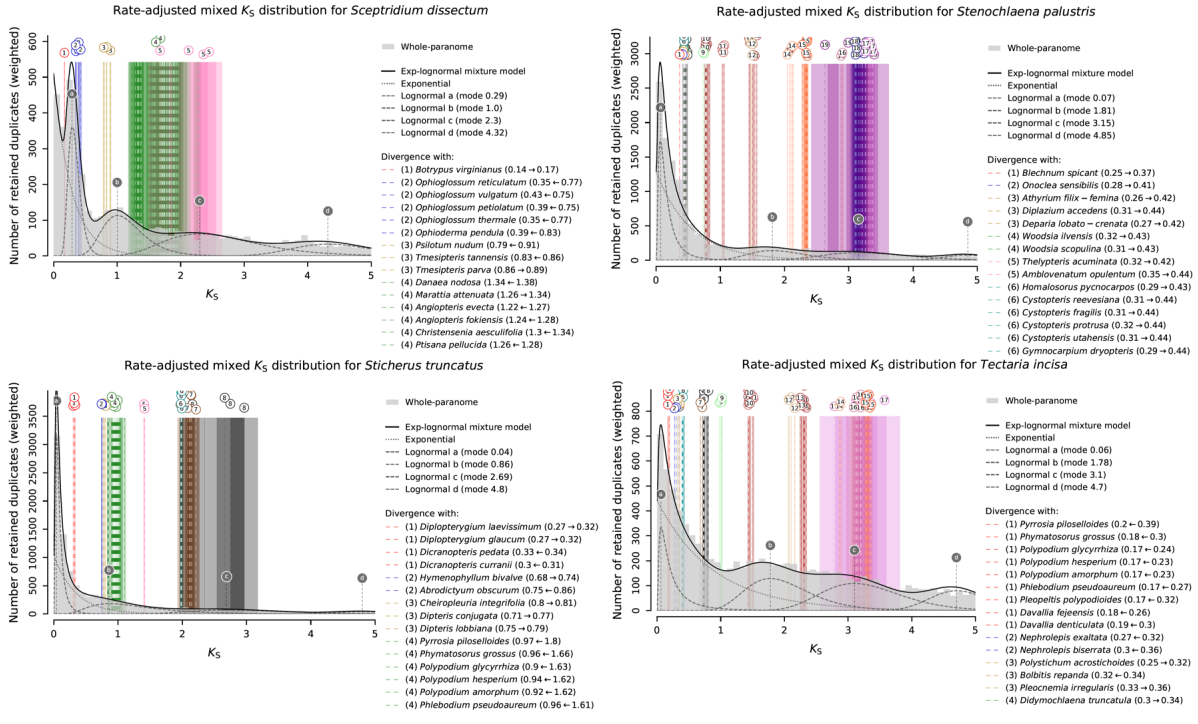


Fig. SM7 (continued)

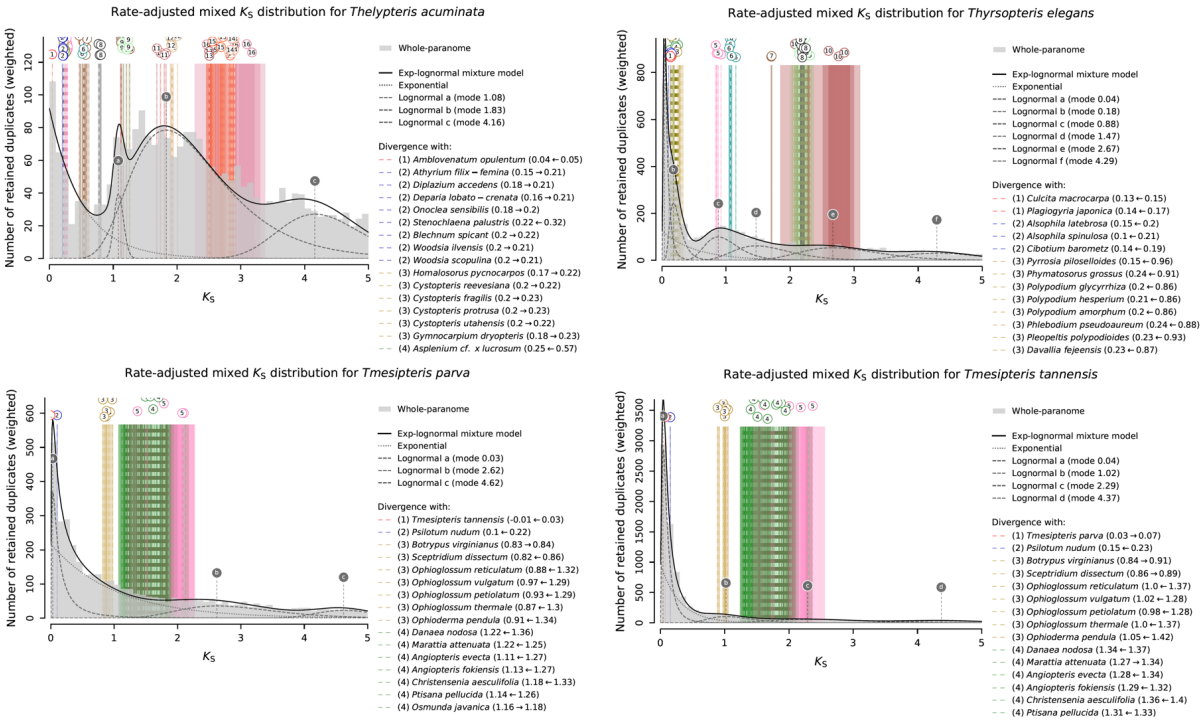


Fig. SM7 (continued)

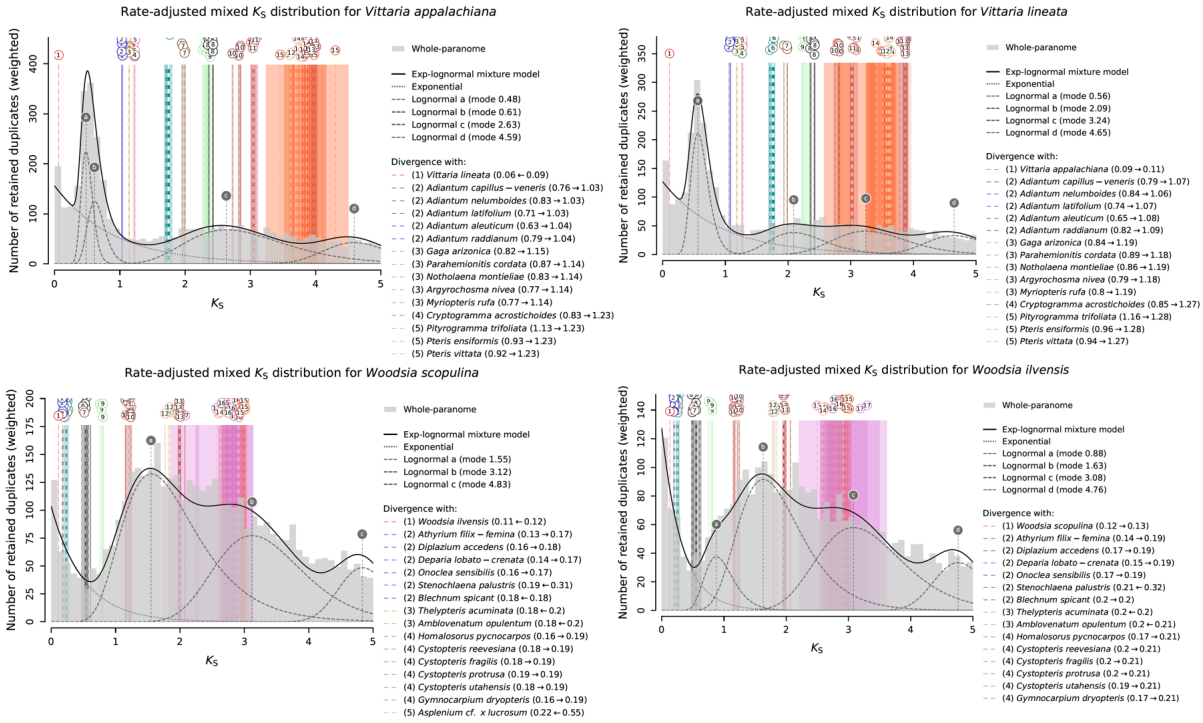


Fig. SM7 (continued)

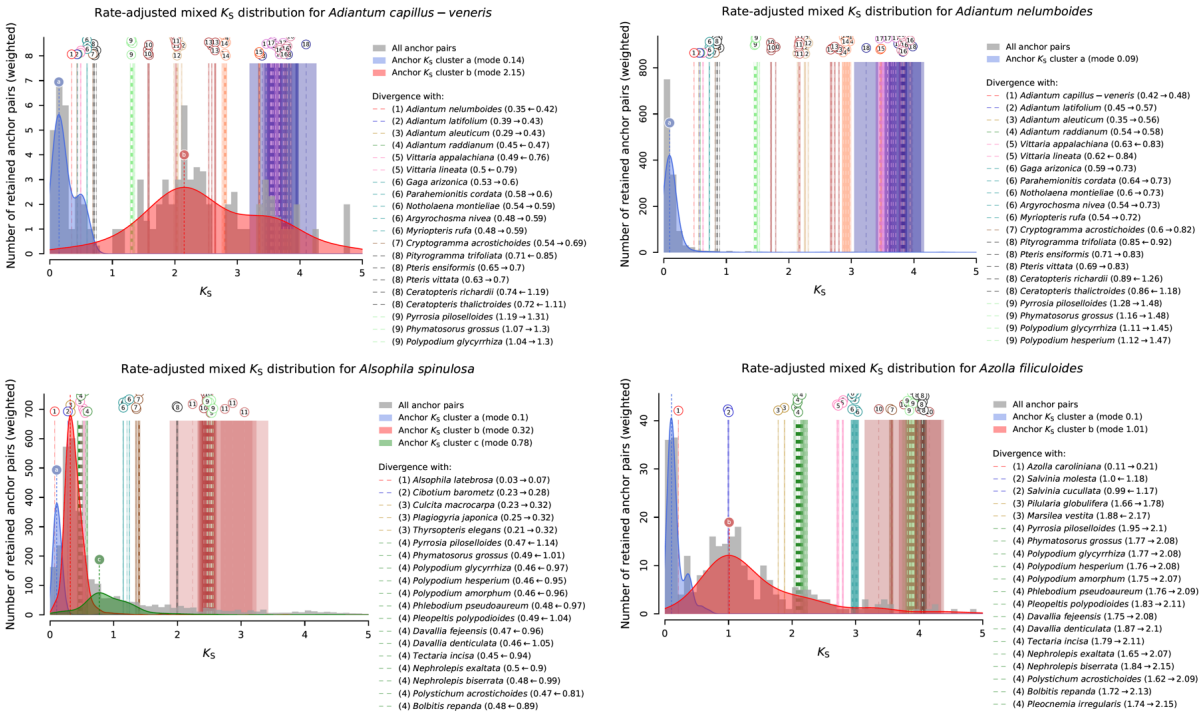


Fig. SM8 | The rate-adjusted mixed K_S-age distributions of anchor pairs for 7 ferns with genome assembly. The K_S distributions of anchor clusters overlaid by colored KDE curves and labelled by letters, with vertical lines denoting associated peaks derived from the lognormal mixture modeling of median K_S values for the collinear segment pairs, were shown as a grey histogram. Rate-corrected modes of orthologous K_S-age distributions between focal species and sister species, representing speciation events, were drawn as numbered vertical long-dashed lines denoting the mean of estimated KDE mode and colored boxes denoting the associated STD. Lines representing the same speciation event in the phylogeny share color and numbering. Horizontal arrows in figure legends indicated the K_S shifts resulted from the substitution rate correction. Speciation events were truncated for presenting to fit the space while complete representation of speciation events is available in supplementary documents. No clusters were left after filtering due to poor K_S content for *Marsilea vestita* hence only the raw K_S distribution of anchor pairs and whole paranome was plotted.

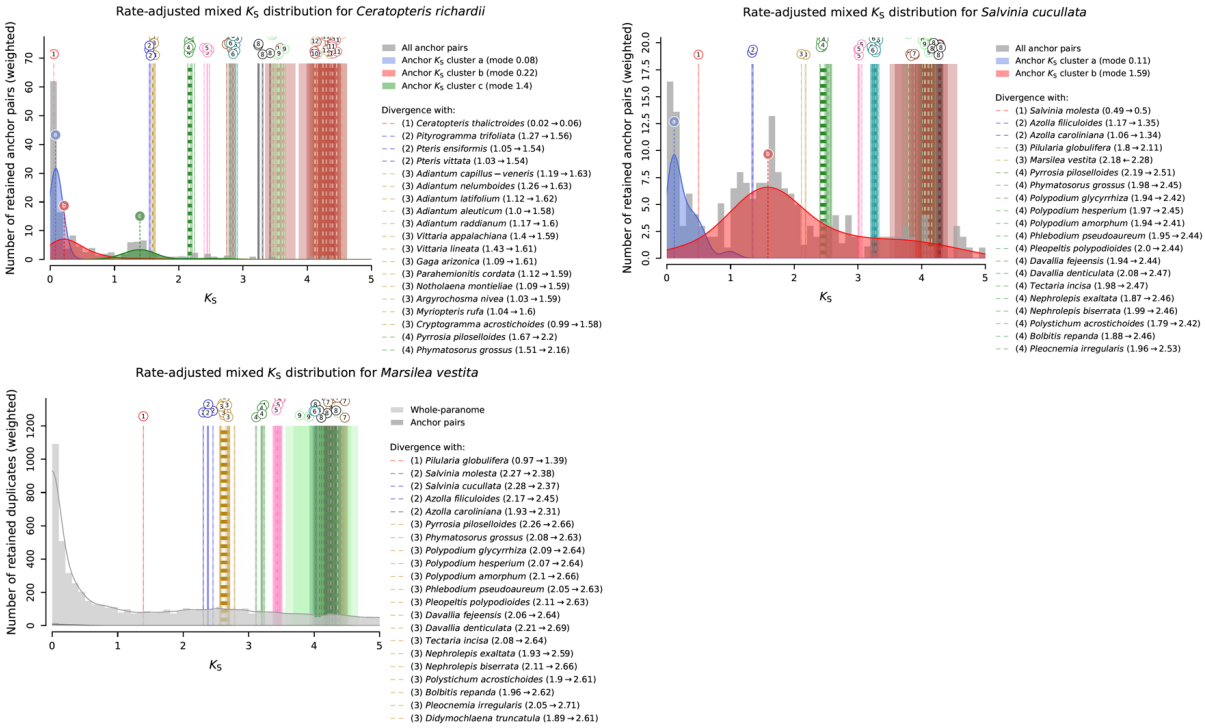


Fig. SM8 (continued).

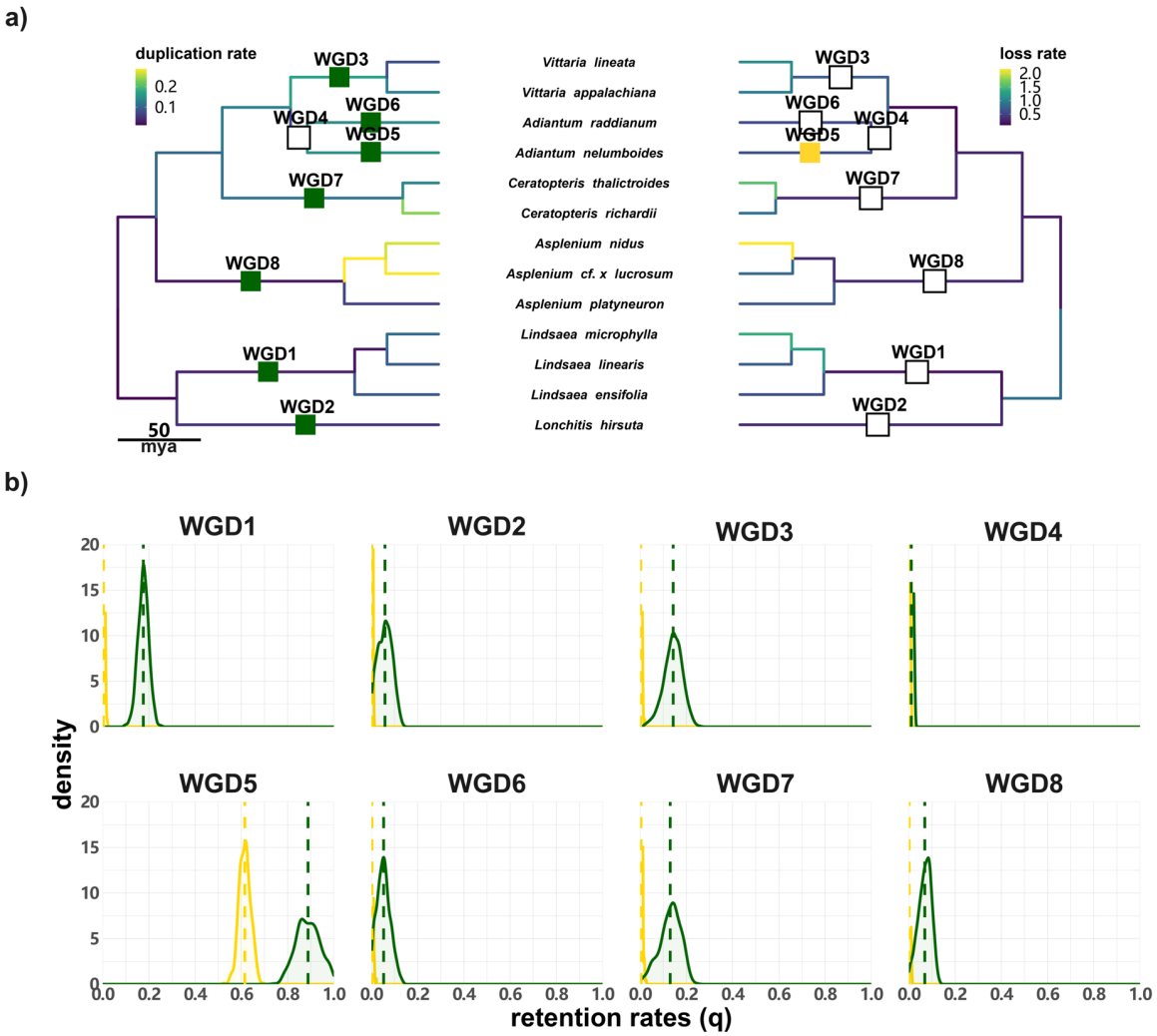


Fig. SM9 | Phylogenomic analysis testing eight hypothetical whole-genome duplications (WGDs) using the DL + WGD model in WHALE. (a) The species tree with eight putative WGD events implied by K_S results. The branch lengths represented the estimated divergence time as shown in Fig 1. The squares in green on the left species tree and the squares in yellow on the right species tree indicate supported WGDs with retention rate (q) higher than 0.05 under the relaxed or critical branch-specific model, respectively. While the hollow squares indicate WGDs with retention rate lower than 0.05. The color upon each branch represents the estimated duplication and loss rates on left and right species tree, respectively. (b) The posterior distributions of WGD retention rates for the eight putative WGDs under the relaxed (green) and critical (yellow) branch-specific model. The dotted lines showed the posterior mean of each posterior distribution.

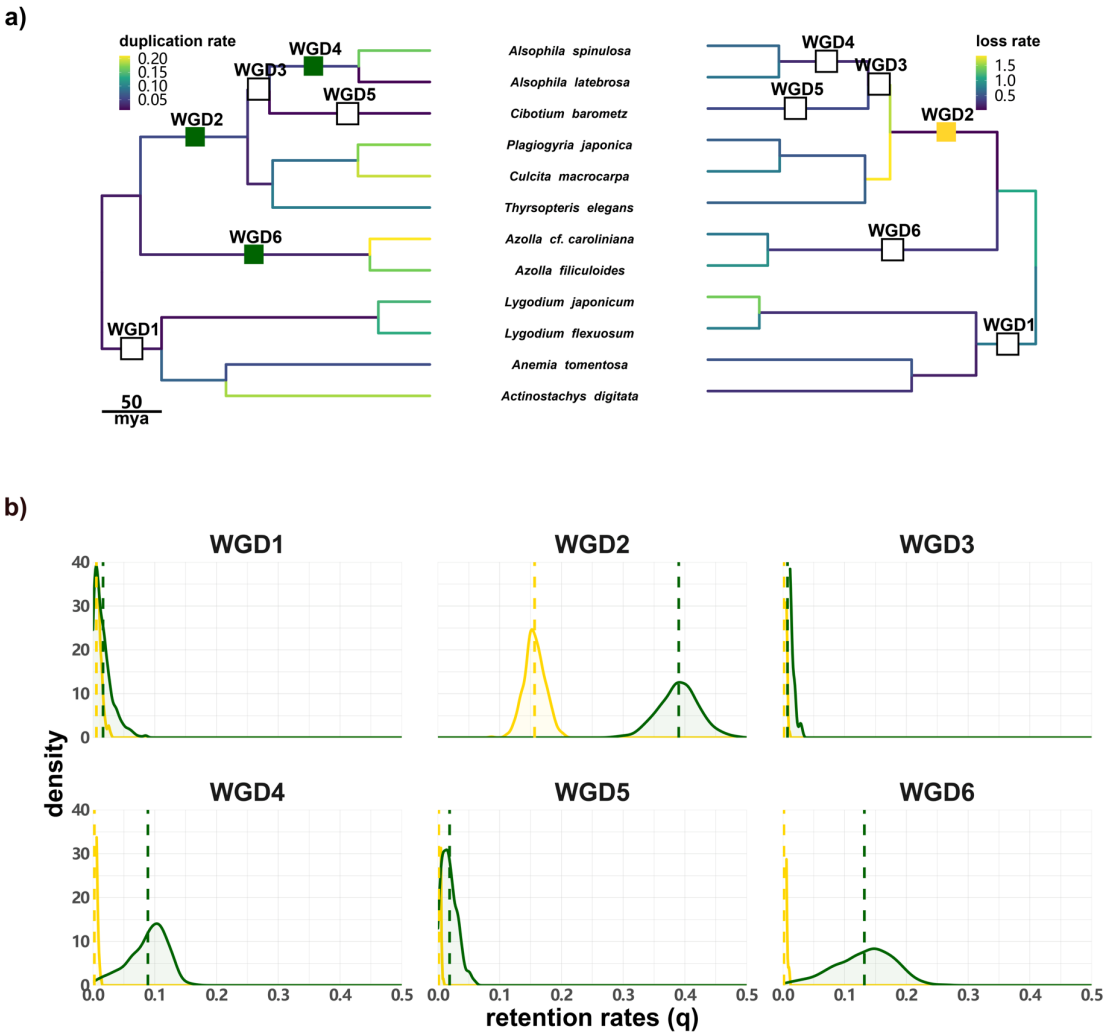


Fig. SM10 | Phylogenomic analysis testing six hypothetical whole-genome duplications (WGDs) using the DL + WGD model in WHALE. (a) The species tree with six putative WGD events implied by K_S results. The branch lengths represented the estimated divergence time as shown in Fig 1. The squares in green on the left species tree and the squares in yellow on the right species tree indicate supported WGDs with retention rate (q) higher than 0.05 under the relaxed or critical branch-specific model, respectively. While the hollow squares indicate WGDs with retention rate lower than 0.05. The color upon each branch represents the estimated duplication and loss rates on left and right species tree, respectively. (b) The posterior distributions of WGD retention rates for the six putative WGDs under the relaxed (green) and critical (yellow) branch-specific model. The dotted lines showed the posterior mean of each posterior distribution.

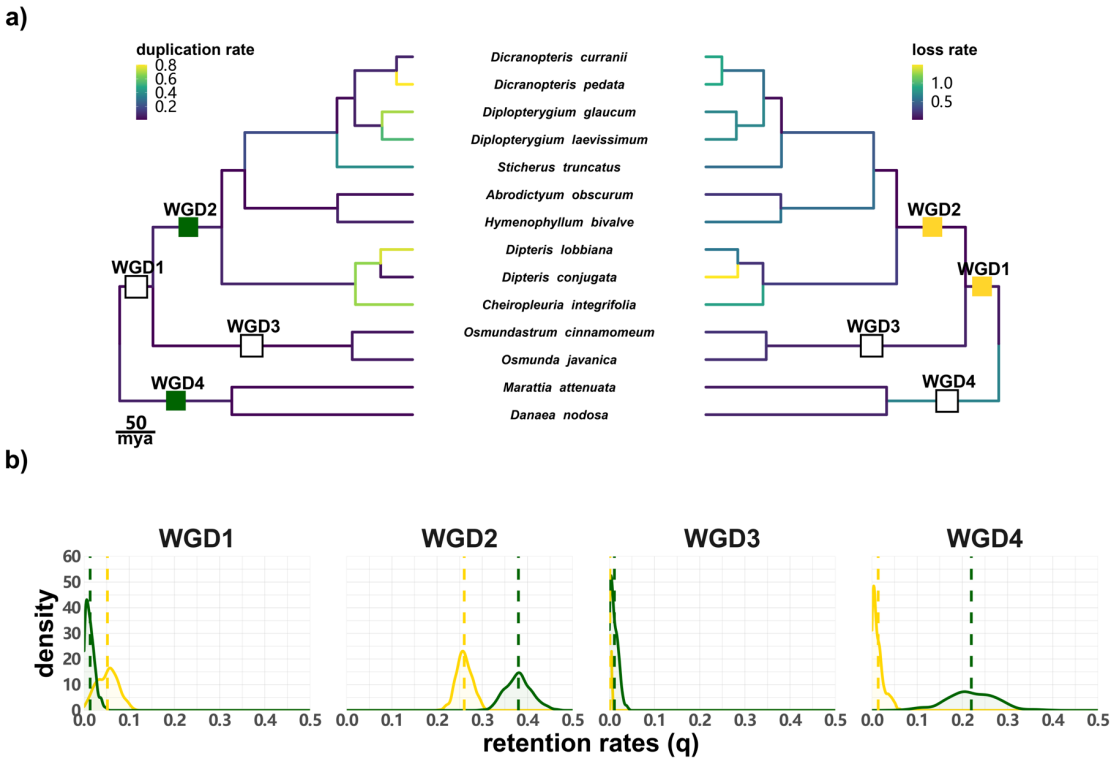


Fig. SM11 | Phylogenomic analysis testing four hypothetical whole-genome duplications (WGDs) using the DL + WGD model in WHALE. (a) The species tree with four putative WGD events implied by K_S results. The branch lengths represented the estimated divergence time as shown in Fig 1. The squares in green on the left species tree and the squares in yellow on the right species tree indicate supported WGDs with retention rate (q) higher than 0.05 under the relaxed or critical branch-specific model, respectively. While the hollow squares indicate WGDs with retention rate lower than 0.05. The color upon each branch represents the estimated duplication and loss rates on left and right species tree, respectively. (b) The posterior distributions of WGD retention rates for the four putative WGDs under the relaxed (green) and critical (yellow) branch-specific model. The dotted lines showed the posterior mean of each posterior distribution.

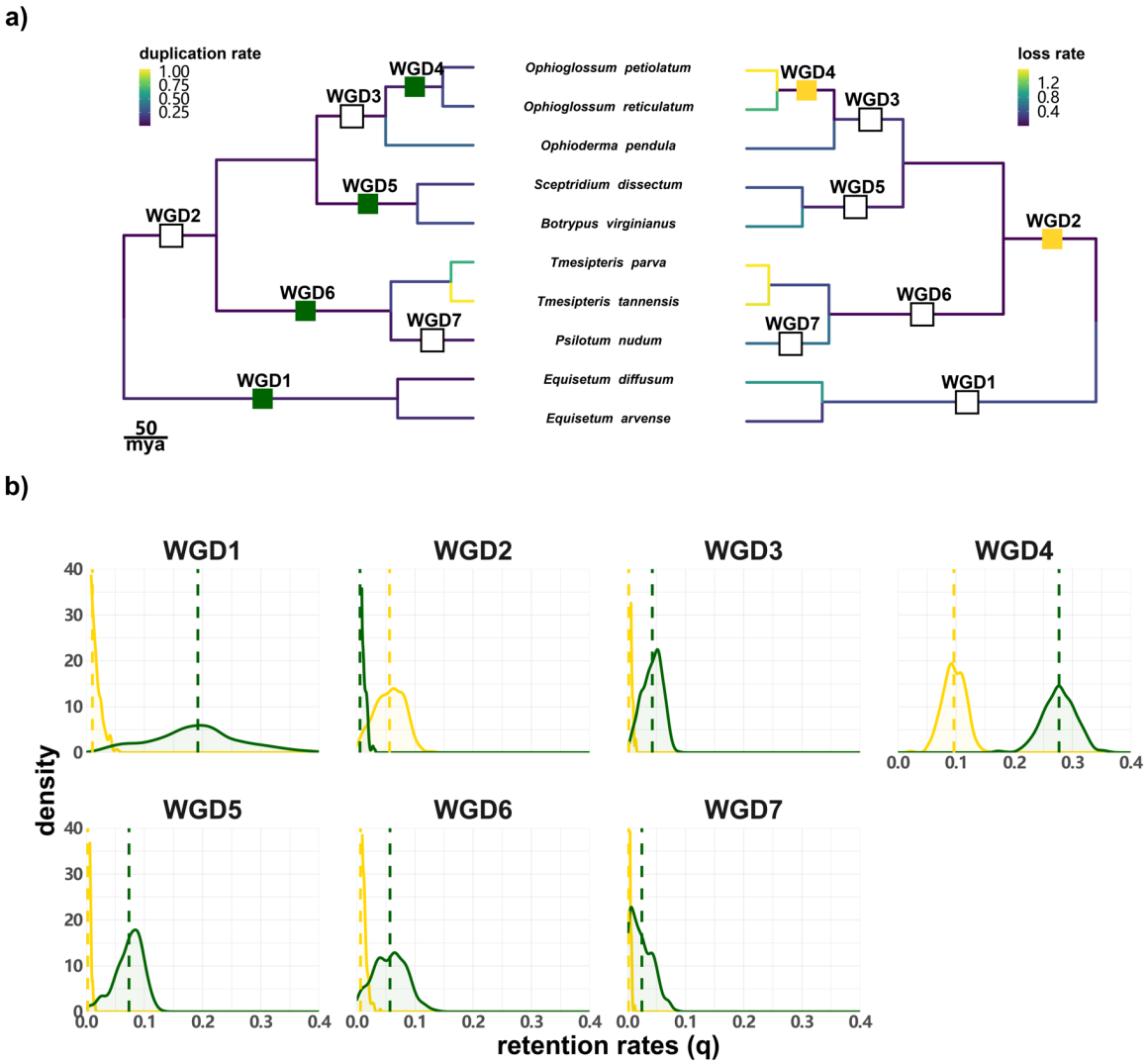


Fig. SM12 | Phylogenomic analysis testing seven hypothetical whole-genome duplications (WGDs) using the DL + WGD model in WHALE. (a) The species tree with seven putative WGD events implied by K_S results. The branch lengths represented the estimated divergence time as shown in Fig 1. The squares in green on the left species tree and the squares in yellow on the right species tree indicate supported WGDs with retention rate (q) higher than 0.05 under the relaxed or critical branch-specific model, respectively. While the hollow squares indicate WGDs with retention rate lower than 0.05. The color upon each branch represents the estimated duplication and loss rates on left and right species tree, respectively. (b) The posterior distributions of WGD retention rates for the seven putative WGDs under the relaxed (green) and critical (yellow) branch-specific model. The dotted lines showed the posterior mean of each posterior distribution.

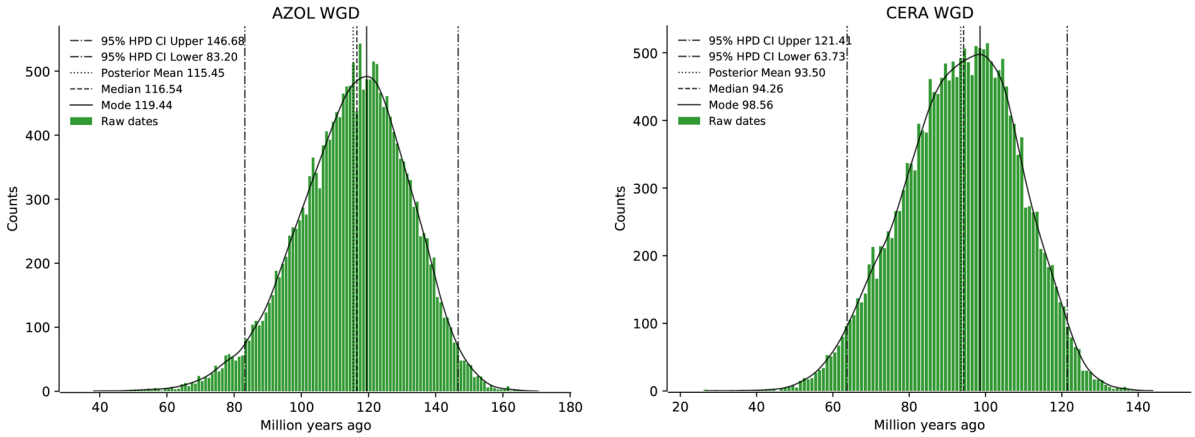


Fig. SM13 | Absolute dating of AZOL and CERA WGD. The posterior distribution of dates for two WGDs with collinear information. The 95% HPD, posterior mean, median and mode were 83.20 - 146.68, 115.45, 116.54 and 119.44 mya for AZOL WGD and 63.73 - 121.41, 93.50, 94.26 and 98.56 mya for CERA WGD, respectively.

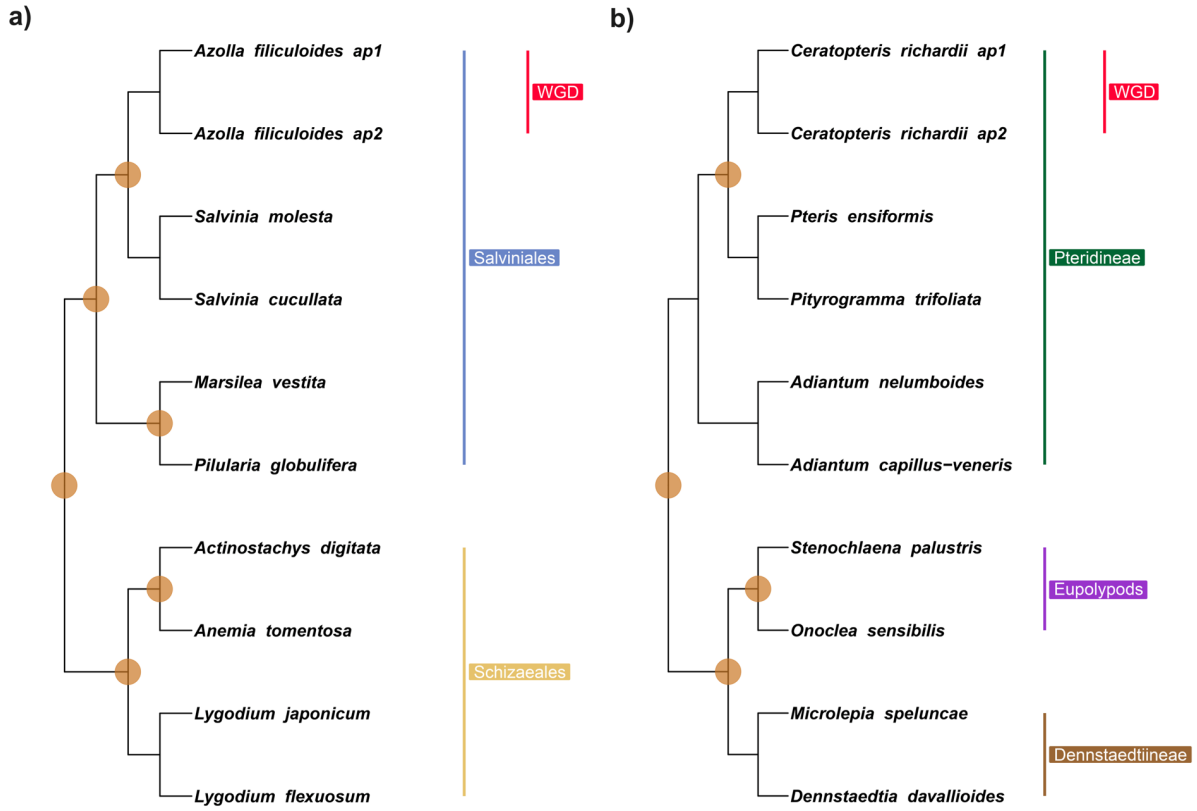


Fig. SM14 | Cladogram of constructed orthogroup in absolute dating of WGDs. The left and right panels showed the cladograms of constructed orthogroups in absolute dating for AZOL and CERA WGD, respectively. The fossil calibrations were showed as ochre circles at corresponding phylogenetic node. The clade strips denoting affiliations were displayed as vertical bars with distinct colors. The nodes joining anchor pairs which were assumed to be retained from WGD were highlighted in red clade strip.

Salicylic acid- and jasmonic acid-mediated signaling

An initial presence/absence analysis suggested that some types of salicylic acid (SA) and jasmonic acid (JA) signaling and downstream components are lacking (at least in part) across all or most here sequenced ferns. We have thus investigated this further for by focusing on the presence of candidates for NON-EXPRESSOR OF PATHOGENESIS-RELATED1 (NPR1), the signaling module SAG101-PAD4-EDS1, the JA receptor CORONATINE INSENSITIVE 1 (COI1) and downstream repressors JASMONATE-ZIM DOMAIN (JAZ).

NPR proteins belong to the BTB/POZ domain family and have been reported throughout land plants with exception of the hornwort representatives of the genus *Anthoceros* (Li et al. 2020). Yet, the functional separation of NPR1, 3 and 4 has only happened later due to a duplication in flowering plants (Li et al. 2020). Model Bryophytes except hornwort models include only one or two NPR genes that are in part able to complement *Arabidopsis thaliana* NPR1 (Peng et al. 2017, Jeon et al. 2024); but within *Marchantia polymorpha* their endogenous function regarding immunity resembles that of negative regulators NPR3/4 (Jeon et al. 2024)

Material and Methods

NPR candidate search

We used CD search (Marchler-Bauer et al. 2017) to predict candidates encoding NPR and other BTB/POZ family members. We used the predicted proteome sequences from all 22 fern species with transcriptomes and the 4 fern species with genomes as a database and the NPR1, 3 and 4 protein sequences from Arabidopsis thaliana as a query using BLASTp (*e* value cutoff of 10^{-7}). The output was funneled into CD search and screened for canonical NPR (NPR1-likeC superfamily domain, a BTB/POZ NPR plant or BTB/POZ superfamily domain and an ANKYRIN domain) and other BTB/POZ family members (BTB/POZ superfamily domain and an ANKYRIN domain) as well as partial candidates (i.e. with a reduced domain set, where partial NPR candidates were required to contain a NPR1-likeC superfamily domain and partial BTB/POZ candidates only a BTB/POZ superfamily domain). Additionally, we counted protein sequences with non-canonical domain setups (i.e. canonical domains with another non-canonical domain combination) separately. Only genes and not isoforms were counted.

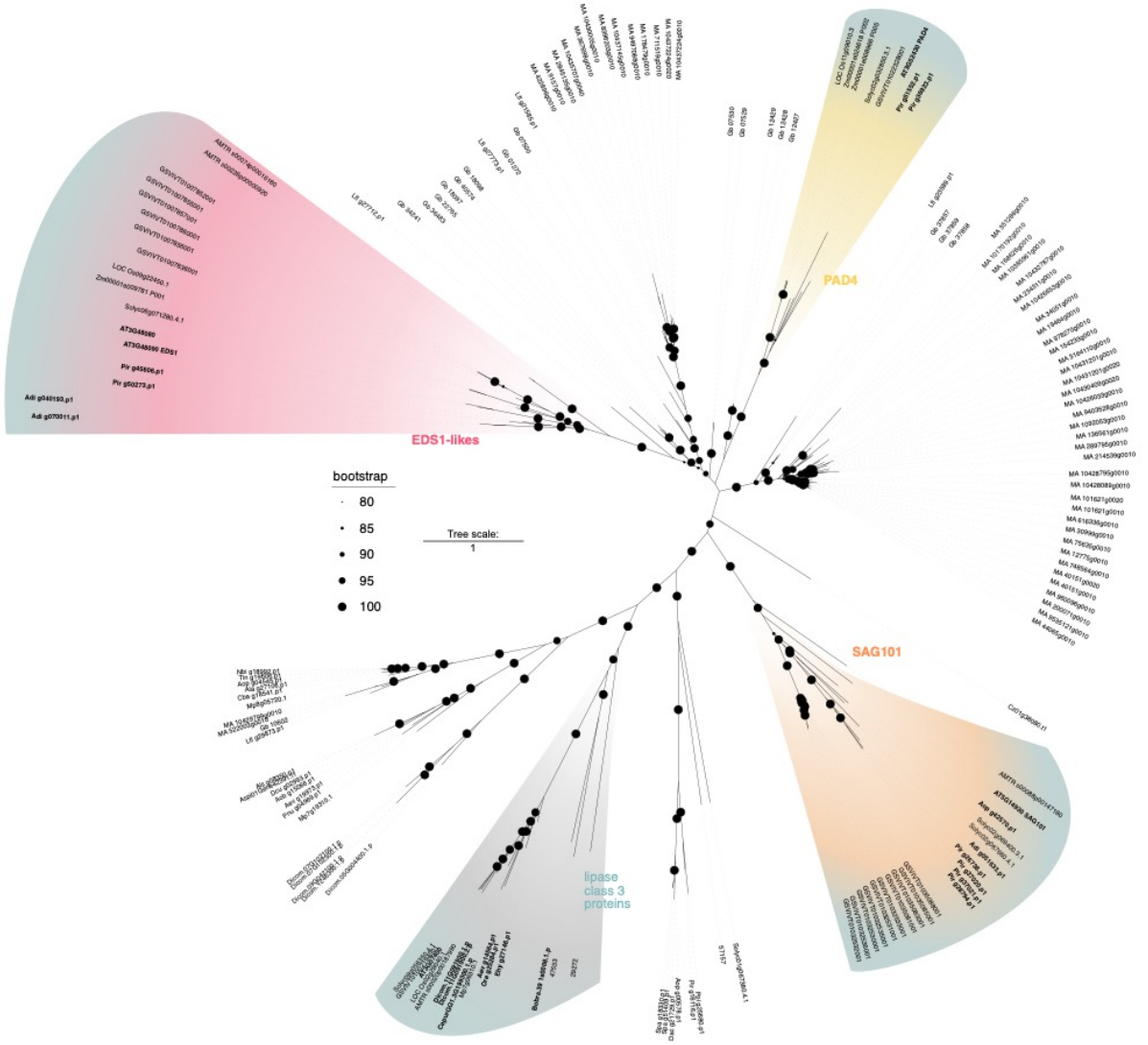
Phylogenetic analysis

To understand the evolutionary history of genes salient to plant immunity, sequences of well-characterized Arabidopsis protein families were used as a query in a BLASTp search against the protein database that contained the predicted fern proteins and proteins inferred from genomes of representative species across the Archaeplastida. All significant hits (*e* value cutoff of 10^{-5}) were aligned with MAFFT v7.490 (L-INS-I, Katoh et al. 2013). We computed maximum likelihood trees using IQ-Tree v. 1.5.5 (Nguyen et al. 2015) with 1000 ultrafast bootstrap pseudo-replicates (Minh et al. 2013). The best models for protein evolution were determined by ModelFinder (Kalyaanamoorthy et al. 2017) and the best models according to Bayesian Information Criterion were used.

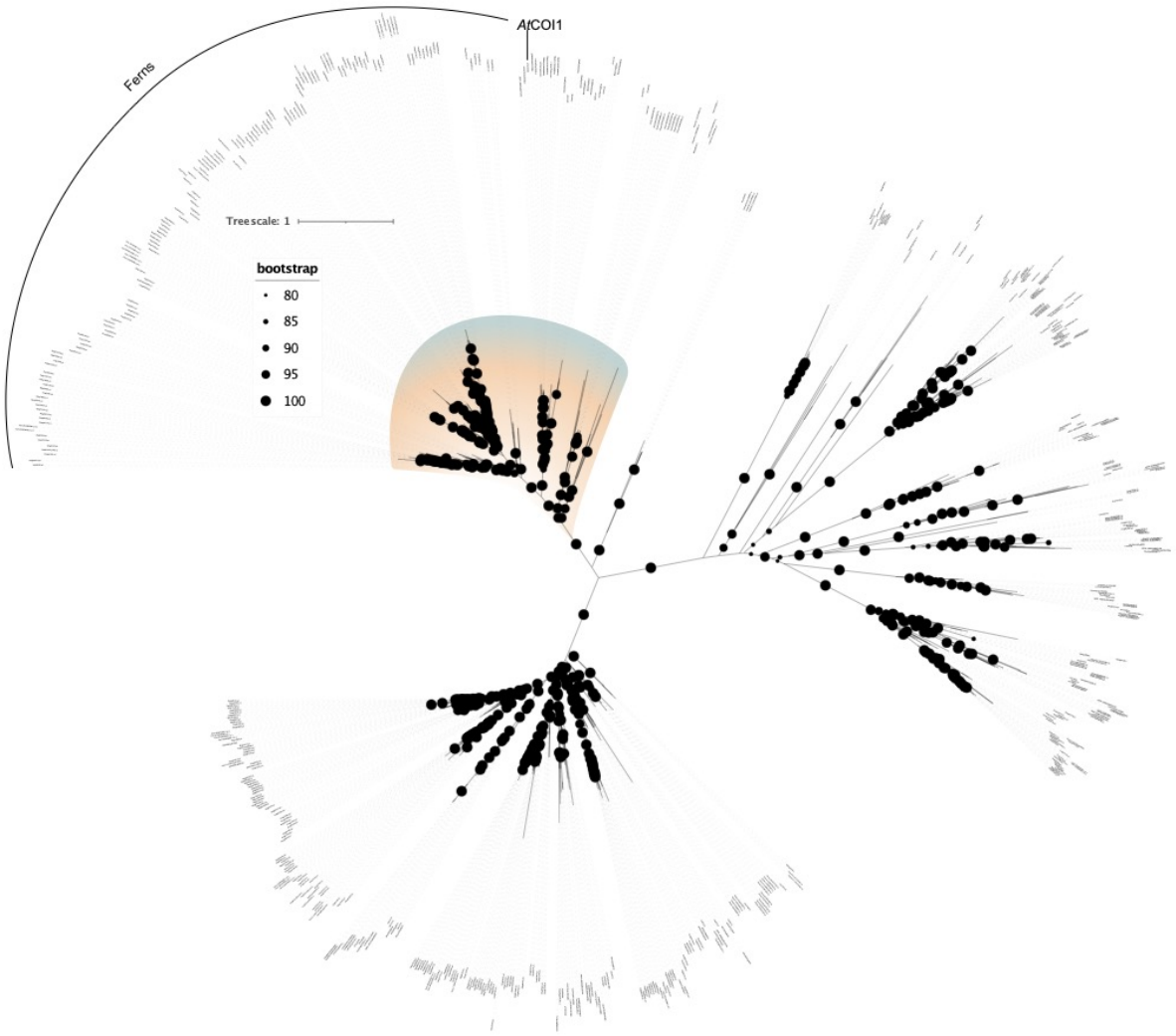
Figures

	complete		partial		with DUF3420		novel domain combination		
Species	NPR	BTB/POZ	NPR	BTB/POZ	NPR	BTB/POZ	NPR	BTB/POZ	Species full name
Adi	1	1	0	0	0	0	0	0	*Actinostachys digitata*
Aev	1	1	0	0	0	0	0	0	*Angiopteris evecta*
Azfi	2	1	0	0	0	0	0	0	*Azolla filiculoides*
Ala	2	1	0	0	0	0	0	0	*Adiantum latifolium*
Als	2	3	0	0	0	0	0	0	*Alsophila latebrosa*
Aob	2	1	0	0	0	0	0	0	*Abrodictyum obscurum*
Aop	4	2	0	0	0	0	0	0	*Amblovenatum opulentum*
Aspi	0	4	1	1	0	0	1	1	*Alsophila spinulosa*
Cric	2	2	0	0	0	0	0	0	*Ceratopteris richardii*
Dcu	5	1	1	1	0	0	1	0	*Dicranopteris curranii*
Dde	3	0	0	0	0	0	0	0	*Davallia denticulata*
Ehy	1	0	0	0	0	0	0	0	*Equisetum hyemale*
Len	3	0	0	0	0	0	0	2	*Lindsaea ensifolia*
Lfl	4	1	0	0	0	0	0	2	*Lygodium flexuosum*
Msp	2	0	0	0	0	0	0	1	*Microlepia speluncae*
Nbi	3	2	0	0	0	0	0	0	*Nephrolepis biserrata*
Ore	1	1	1	0	0	0	0	0	*Ophioglossum reticulatum*
Pir	3	0	0	0	3	0	0	0	*Pleocnemia irregularis*
Ppi	3	1	0	0	0	0	0	0	*Pyrrosia piloselloides*
Sam	3	0	0	0	0	0	0	0	*Salvinia molesta*
Sacu	1	2	1	0	0	0	0	0	*Salvinia cucullata*
Spa	3	0	0	0	0	0	0	0	*Stenochlaena palustris*
Tin	4	2	0	0	0	0	0	0	*Tectaria incisa*

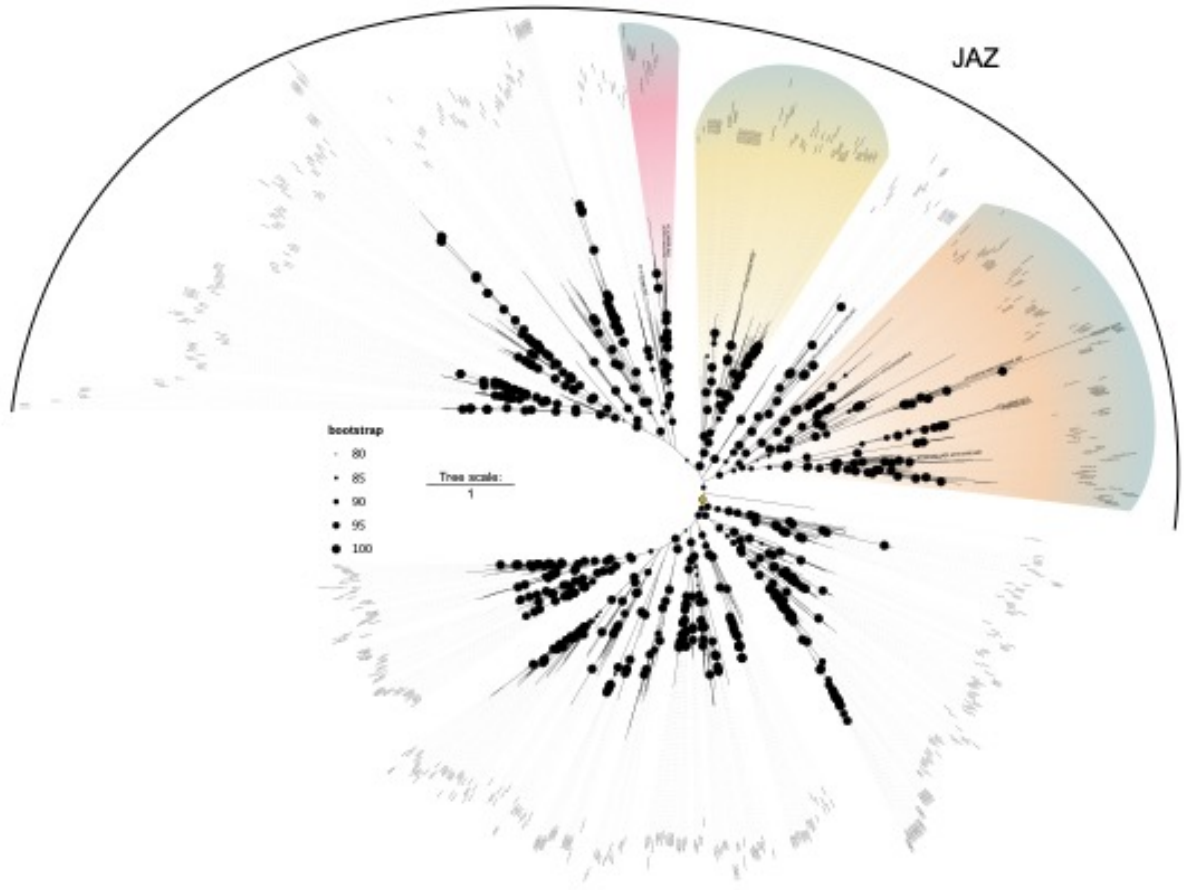
Supplemental Figure SM7. NPR and other BTB/POZ candidates in ferns. A domain search across all fern datasets (22 transcriptomic, 4 genomic) was conducted using CD search. The canonical NPR domain set includes an NPR1-likeC superfamily domain, a BTB/POZ NPR plant or BTB/POZ superfamily domain and an ANKYRIN domain. Other BTB/POZ candidates include a BTB/POZ superfamily domain and an ANKYRIN domain. Genes including these domains were characterized as complete NPR or complete BTB/POZ candidates. Partial NPR candidates have an NPR1-likeC superfamily domain and one of the other two domains. Partial BTB/POZ candidates have only a BTB/POZ superfamily domain. In several ferns we identified NPR or BTB/POZ candidates with a DUF3420 domain or NPR or BTB/POZ candidates with another domain. Unlike the combination with the DUF3420 domain these other domain combinations were only occurring in few cases (indicated as novel domain combinations). The numbers indicate the number of identified gene candidates (isoforms not included) and the color code correspondence to the detected abundance with light blue = 0 (absence) to pink = 5 gene candidates.



Supplemental Figure SM8. Phylogenetic analyses of the SAG101/PAD4/EDS1 module. SAG101, PAD4 and EDS1 candidates form three distinct and supported clades (Ultrafast bootstrap support 100). Additionally, another fully supported clade for a lipase, lipase class 3 proteins, was recovered. All clades include a few but not many fern species. Noteworthy is *Pleocnemia irregularis*, which encodes sequences clustering with all three clades, suggesting that it possess orthologs to all components of the module. However, EDS1 originates from a duplication not shared by all angiosperms. Thus, homologs found in ferns or other angiosperms can only be identified as homologs to EDS1 and EDS1-like.



Supplemental Figure SM9. Phylogenetic analyses of COI1. A distinct fully supported COI1 clade was recovered (Ultrafast bootstrap support 100) that included both the functional COI1 of *Arabidopsis thaliana* and *Marchantia polymorpha* as well as several fern sequences. We identified both recent and more ancestral duplications of COI1 homologs in the ferns, speaking to a yet under investigated diversification of the jasmonate receptor.



Supplemental Figure SM10. Phylogenetic analyses of JAZ. The JAZ repressor negatively controls jasmonate signaling and repression is released upon perception of jasmonates by COI1. *A. thaliana* encodes 13 different JAZ and additional TIFY domain containing proteins in its genome, while *M. polymorpha* has one JAZ candidate, whose function is conserved in the model liverwort and model angiosperm. We recovered one large JAZ protein containing clade with an ultrafast bootstrap support of 95 that includes all 13 JAZ members from *A. thaliana.* Interspersed with the sequences from *A. thaliana,* several sequences from different fern species cluster in this clade. This suggests that also ferns possess JAZ homologs.

References

Carrillo-Carrasco VP, et al. (2023) The birth of a giant: evolutionary insights into the origin of auxin responses in plants. The EMBO Journal, 42: e113018.

Chini A, et al. (2023) Evolution of the jasmonate ligands and their biosynthetic pathways. New Phytologist, 238: 2236-2246.

de Vries S, et al. (2018) Jasmonic and salicylic acid response in the fern Azolla filiculoides and its cyanobiont. Plant Cell and Environment, 41: 2530-2548.

Jeon HW, et al. (2024) Contrasting and conserved roles of NPR pathways in diverged land plant lineages. New Phytologist, doi: 10.1111/nph.19981.

Li FW, et al. (2020) Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. Nature Plants, 6: 259–272.

Kalyaanamoorthy S, et al. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods, 14: 587-589.
Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30: 772-780.
Marchler-Bauer A, et al. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Research, 45: D200–D203.
Minh BQ, et al. (2013) Ultrafast approximation for phylogenetic bootstrap. Molecular Biology and Evolution, 30: 1188-1195.
Monte I, et al. (2018) Ligand-receptor co-evolution shaped the jasmonate pathway in land plants. Nature Chemical Biology, 14: 480-488.
Monte I, et al. (2019) A single JAZ repressor controls the jasmonate pathway in Marchantia polymorpha. Molecular Plant, 12: 185-198.
Nguyen LT, et al. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution, 32: 268-274.
Peng Y, et al. (2017) Perception of Salicylic Acid in Physcomitrella patens. Frontiers in Plant Science, 8: 2145.
Thines B, et al. (2007) JAZ repressor proteins are targets of the SCF^{COI1} complex during jasmonate signalling. Nature, 448: 661–665.

Chemical synthesis of 3-*O*-methyl-D-galactopyarnose (5a), 2-*O*-methyl-D-galactopyarnose (5b) and 2-*O*-methyl-D-glucopyranose (10) (Figure S14)

The commercially available methyl β-D-galactopyranoside (**1**) was converted into methyl 4,6-*O*-benzylididene-β-D-galactopyranoside (**2**) according to published procedures.[1,2] Compound **2** was reacted with 4-methoxybenzyl chloride to produce methyl 4,6-*O*-benzylidene-2-*O*-(4-methoxybenzyl)-β-D-galactopyranoside (**3a**) and methyl 4,6-*O*-benzylidene-3-*O*-(4-methoxybenzyl)-β-D-galactopyranoside (**3b**) and the remaining free 3-OH and/or 2-OH of compounds **3a** and **3b** was conventionally methylated to obtain the corresponding 3-*O*-methyl-derivative **4a** and **4b**. Removal of the remaining substituents by acid hydrolysis of **4a** and **4b** produced 3-*O*-methyl-D-galactopranose (**5a**) and 2-*O*-methyl-D-galactopranose (**5b**). The same protocol was used for the synthesis of compound **10**. Thus, the commercially available methyl α-D-glucopyranoside (**6**) was converted into methyl 4,6-*O*-benzylidene-α-D-glucopyranoside (**7**) according to published procedures.[2,7,8] Compound **7** was reacted with 4-methoxybenzyl chloride to produce methyl 4,6-*O*-benzylidene-3-*O*-(4-methoxybenzyl)-α-D-glucopyranoside (**8**)[9] and the remaining free 3-OH of compound **8** was conventionally methylated to obtain the 2-*O*-methyl-derivative **9**. Removal of the remaining substituents by acid hydrolysis of **9** produced 2-*O*-methyl-D-glucopyranose (**10**). The structures of compounds **2**,[1] **3a,b**,[3] **5a,b,**[5-6,13-15] **7**,[7,8] **8**,[9,10] and **10**[11,12,15] were ascertained by NMR spectroscopy and were in agreement with reported data.

References

1)-Synthesis of the tetrasaccharide glycoside moiety of Solaradixine and rapid NMR-based structure verification using the program CASPER. Thibault Angles d'Ortoli, Goran Widmalm *Tetrahedron* 72 (**2016)** 912-927.

2)-Efficient Iodine-Catalyzed Preparation of Benzylidene Acetals of Carbohydrate Derivatives. Rajib Panchadhayee and Anup Kumar Misra *Journal of Carbohydrate Chemistry* 27(**2008**) 148–155.

3)- Structure-Based Design of a Monosaccharide Ligand Targeting Galectin-8 Mohammad H. Bohari,[a] Xing Yu, Chandan Kishor, Brijesh Patel, Rob Marc Go, Hadieh A. Eslampanah Seyedi, Yaron Vinik, I. Darren Grice, Yehiel Zick, and Helen Blanchard. *ChemMedChem* 13, (**2018**)1664 –1672.

4)-Sugar composition of the pectic polysaccharides of charophytes, the closest algal relatives of land-plants: presence of 3-O-methyl-D-galactose residues Christina O'Rourke, Timothy Gregson, Lorna Murray, Ian H. Sadler and Stephen C. Fry. *Annals of Botany* 116 (**2015**) 225–236.

5)-3-O-Methyl-d-galactose residues in lycophyte primary cell walls Zoe¨ A. Popper , Ian H. Sadler, Stephen C. Fry. *Phytochemistr*y 57 (**2001**) 711–719.

6)-An improved procedure for the synthesis of 3-O-methyl-D-galactose E. G. Gros and I.O. Mastronardi. *Carbohyd. Res.*, 10 **(1969**) 325-327.

7)-Benzylidene Acetal Protecting Group as Carboxylic Acid Surrogate: Synthesis of Functionalized Uronic Acids and Sugar Amino Acids Amit Banerjee, Soundararasu Senthilkumar, and Sundarababu Baskaran, *Chem. Eur. J.* **22**, (19016), 902 – 906.
8)-Acceleration and Deceleration Factors on the Hydrolysis Reaction of 4,6-O-Benzylidene Acetal Group. Yuta Maki, Kota Nomura, Ryo Okamoto, Masayuki Izumi, Yasuhisa Mizutani, and Yasuhiro Kajihara, *J. Org. Chem.* 85 (2020),15849−15856.
9)-Asymmetric Total Synthesis of Phosphatidylinositol 3-Phosphate and 4-Phosphate Derivatives Jian Chen, Li Feng, and Glenn D. Prestwich, *J. Org. Chem.* 63 (**1998**) 6511-6522.
10)-Selective removal of the (2-naphthyl)methyl protecting group in the presence of *p*-methoxybenzyl group by catalytic hydrogenation. László Lázár, Lóránt Jánossy, Magdolna Csávás, Mihály Herczeg, Anikó Borbása, and Sándor Antus, ARKIVOC (v), (**2012**) 312-325.
11)-^{13}C Nuclear Magnetic Resonance Spectra of Glucobioses, Glucotrioses, and Glucans. Taichi Usui, Naotaka Yamaoka, Kazuo Matsuda, and Katura Tuzimura, ***J. Chem. Soc., Perkin Trans. 1***, **1973**, 2425-2432.
12)-A study of ^{13}CH coupling constants in hexopyranoses. Klaus Bock and Christian Pedersen, ***J. Chem. Soc., Perkin Trans. 2***, **1974**, 293-297.
13)-P.M.R. Spectroscopy of monomethyl ethers of d-galactopyranose and its derivatives. E. B. Ratheone, A. M. Stephen, *Carbohyd. Res.* **20**, (1971) 357-367.
14)-^{1}H-N.m.r. and 13C-n.m.r. spectroscopy of methyl ethers of D-galactopyranose. Daphne C. Vogt, Alistair M. Stephen, and Graham E. Jackson, *Carbohyd. Res.* **206**, (1990) 333-339.
15)-P.m.r. spectral assignments for b-methyl groups in mono-methylated D-hexoses Eduardo G. Gros, Irma O. Masironardi, Andadolfo R. Frasca, *Carbohyd. Res.* **16**, (1971) 232-234.