Kingdom-wide evolutionary characterization of RNA editing factors in Archaeplastida

Ming Chen[1, 2, 3], Zhang Zhang[1, 2, 3, *]

[1]National Genomics Data Center, China National Center for Bioinformation, Beijing, China

[2]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

*Corresponding author: Zhang Zhang (zhangzhang@big.ac.cn)

Abstract

RNA editing, a post-transcriptional modification in plant mitochondria and plastids, is essential for plant terrestrialization. Editosome, mediating RNA editing, comprises nuclear-encoded factors, primarily including pentatricopeptide repeat (PPR), multiple organelle RNA editing factor (MORF), organelle RNA recognition motif-containing (ORRM), and organelle zinc finger (OZ) proteins. However, the origin, evolution, and diversification of these editing factors largely remain enigmatic. Here, we analyze 364 high-quality proteomes spanning Archaeplastida to identify 212,779 PPR, 2,759 MORF, 7,760 ORRM and 829 OZ proteins, and chart their evolutionary landscapes. We find that PPR genes in terrestrial plants undergo substantial expansions, mainly driven by dispersed duplication and retroposition. Further analysis shows that DYW subgroup PPR genes, encoding RNA-editing enzymes, are largely transferred from plants to rotifers, potentially enhancing ultraviolet irradiation resistance of these recipients. Meanwhile, we find that MORF proteins, accessory factors of PPR proteins, are restricted to seed plants. Notably, their hallmark MORF domains share sequence and structural similarity with peptidase S8 propeptide/proteinase inhibitor I9 domains that mediate pro-enzyme folding, implying MORFs regulate PPR folding. Additionally, we reclassify MORF, ORRM, and OZ proteins based on domain architecture, revealing hallmark domain fusions occur among them and with PPR proteins, except between MORF and OZ. Collectively, our findings shed light on the evolution of plant RNA editing factors.

Key words: PPR; MORF; ORRM; OZ; evolution

Introduction

RNA editing, a post-transcriptional modification primarily involving cytidine-to-uridine (C-to-U) and uridine-to-cytidine (U-to-C) alterations in plant organelle (mitochondria and plastid), is essential for adaptation to terrestrial environments[1-4], as well as various physiological processes[5], including development[6], growth[7], flowering[8], fruit ripening[9], and resistance to stresses[10,11]. Editing factors, a group of nuclear-encoded proteins, are assembled into an editing complex to mediate RNA editing[4]. Given the diversity of editing factors in plants, therefore, deciphering their origin, evolution, and diversification is critical for understanding the machinery underlying RNA editing[1,12-14].

Over the past decades, great efforts have been made to identify RNA editing factors, primarily including Pentatricopeptide Repeat (PPR), Multiple Organelle RNA Editing Factor (MORF) (also known as RNA Editing Factor Interacting Protein, RIP), Organelle RNA Recognition Motif-containing (ORRM), and Organelle Zinc (OZ) finger proteins[12], each defined by their hallmark domains/motifs that contribute distinctly to RNA editing[15]. Among them, PPR proteins are essential editing factors, specifically binding RNA bases through their P/PLS motif tracts and catalyzing C-to-U and U-to-C editing via their carboxyl-terminal (C-terminal) DYW and DW:KP domains, respectively[16,17]. Furthermore, PPR proteins are classified into two subfamilies, P and PLS, with the PLS subfamily further divided into PLS, E, E+, and DYW subgroups[15]. PPRs' accessory proteins, MORF, ORRM, and OZ, named after their hallmark MORF, RRM, Ran Binding Protein 2 (RanBP2) type Zinc finger (Znf) domains, respectively, interact with PPR proteins to coordinate RNA editing[4,8,18-25]. In addition to their hallmark domains, these proteins also incorporate diverse

66 additional domains, such as Glycine-Rich (GR) domain in MORF proteins[26-28]
67 and PPR, MORF, or CRM domains in ORRM protein[29].
68
69 Building on these advances, studies have increasingly focused on the origin,
70 evolution, and diversification of PPR, MORF, ORRM, and OZ genes[4,12,13,30,31].
71 The PPR gene family is one of the largest gene families in plants[17], ranging
72 from tens of members in early-diverging aquatic chlorophytes to hundreds or
73 thousands in terrestrial plants[2,4] and thereby reflecting massive expansions of
74 the PPR family[2,32]. Besides, PPR genes are also present beyond plants, with P
75 subfamily members widespread in eukaryotes and occasionally in
76 prokaryotes[33,34]. Notably, DYW subgroup PPR genes have also been detected
77 in multiple non-plant eukaryotes[35-38], including Amoebozoa, Discoba, Fungi,
78 Metazoa, and Malawimonadidae, presumably resulting from horizontal gene
79 transfer (HGT)[39]. In contrast, MORF genes have been found only in
80 angiosperms and gymnosperms[32], ORRM genes occur across Viridiplantae
81 lineages including chlorophytes, mosses, liverworts, and angiosperms[29], and
82 OZ genes are sporadically distributed across gymnosperms, mosses, and
83 lycophytes, but absent from chlorophytes[23]. Despite this, the origin, evolution,
84 and diversification of these RNA editing factors across Archaeplastida remain
85 poorly understood, particularly without systematic association with the
86 expansion driver of PPR genes, evolutionary trajectory of PPR genes between
87 plants and non-plants, and types of MORFs, ORRMs, and OZs.
88
89 Toward this end, we collect a comprehensive list of genomes across
90 Archaeplastida, systematically identify PPR, MORF, ORRM, and OZ genes,
91 and construct an evolutionary landscape of RNA editing factors in plants. Based
92 on the landscape, we perform phylogenetic and comparative genomic analyses
93 to explore mechanisms driving PPR expansions, scale and impact of HGT in

DYW subgroup PPR genes, evolution and function of MORF domains, and reclassification and hallmark domain/motif fusion among PPR, MORF, ORRM, and OZ proteins, which collectively provide new insights into the plant RNA editing machinery.

Results

Evolutionary landscape of PPR, MORF, ORRM, and OZ genes across Archaeplastida

To systematically investigate the evolution of PPR, MORF, ORRM, and OZ genes, we obtain a total of 364 high-quality genomes across Archaeplastida that span early-diverging aquatic algae to highly derived terrestrial plants (for details see Materials and Methods), including glaucophytes (*n*=1), rhodophytes (*n*=7), prasinodermophytes (*n*=1), chlorophytes (*n*=33), charophytes (*n*=7), hornworts (*n*=22), mosses (*n*=13), liverworts (*n*=4), lycophytes (*n*=6), ferns (*n*=5), gymnosperms (*n*=13), and angiosperms (*n*=262) (Supplementary Table 1). Based on these genomes, we first identify 138 single-copy orthologous gene families (Supplementary Table 2) and then build a plant phylogenetic tree (Figure 1A and Supplementary Figure 1) that is topologically congruent with previous studies[40,41]. Next, we systematically identify 212,779 PPR, 2,759 MORF, 7,760 ORRM, and 829 OZ genes (Supplementary Tables 3 and 4), and evolutionarily characterize these genes in the phylogenetic tree as built above. Noticeably, we find that PPR genes are scarcely present in glaucophytes, rhodophytes, prasinodermophytes, chlorophytes, and charophytes, ranging only from 8 to 182 (Figure 1B and Supplementary Table 4). In contrast, hundreds to thousands of PPR genes are detected in terrestrial plants, particularly some hornworts, mosses, and lycophytes, followed by gymnosperms, ferns, and angiosperms. Remarkably, two lycophytes, namely, *Selaginella moellendorffii* and *Isoetes sinensis*, have the highest number of

PPR genes (*n*=3,861 and *n*=3,256, respectively) among the 364 Archaeplastida species in our study (Supplementary Table 4). On the contrary, liverworts are short of PPR genes, less than 100 in count (Supplementary Table 4).

Compared to PPR genes, the other three types of genes present distinct patterns. Specifically, MORF genes are found only in seed plants, including angiosperms and gymnosperms (Figure 1C), in agreement with a previous study[32]. Additionally, seed plants generally harbor about 10 MORF genes, whereas some species, such as *Brassica napus*, *Pogostemon cablin*, and *Cissus rotundifolia*, have more MORF genes (*n*=40, 39, and 37, respectively; Supplementary Table 4). Conversely, ORRM genes are ubiquitous across Archaeplastida and exhibit a gradual increase from glaucophytes to angiosperms, indicating their ancient origin (Figure 1D). Albeit much lower in count than the rest three types of genes, OZ genes are enriched in angiosperms, with a minor presence in chlorophytes, hornworts, mosses, liverworts, lycophytes, ferns, and gymnosperms (Figure 1E). Interestingly, some angiosperms, such as *Pogostemon cablin*, simultaneously possess numerous MORF, ORRM, and OZ genes, as well as PPR genes (Figure 1 and Supplementary Table 4). Overall, PPR genes exhibit dramatic expansions in terrestrial plants, in contrast to MORF, ORRM, and OZ genes.

Dispersed duplication and retroposition driving PPR gene expansions in early terrestrial plants

To explore the underlying mechanisms of PPR gene expansions following plant terrestrialization, we select high-quality chromosome-level genomes, including four hornworts (*Anthoceros agrestis*, 2,308 PPRs; *Notothylas orbicularis*, 1,772 PPRs; *Paraphymatoceros pearsonii*, 1,583 PPRs; and *Phymatoceros phymatodes*, 1,627 PPRs), one moss (*Takakia lepidozioides*, 2,016 PPRs), and

149 two lycophytes (*I. sinensis*, 3,256 PPRs; and *Huperzia asiatica*, 1,583 PPRs)
150 (Supplementary Table 4). In addition, we choose a charophyte, *Zygnema*
151 *circumcarinatum* (67 PPRs), as a phylogenetic outgroup. It is known that gene
152 expansion can be achieved by gene duplication with four different modes[42-44],
153 viz., whole genome duplication (WGD), tandem duplication (TD), proximal
154 duplication (PD) and dispersed duplication (DD). To determine genomic loci of
155 PPR genes and investigate the contribution of the four duplication modes to
156 their expansions, we retain PPR genes annotated on chromosomes,
157 accordingly accounting for 93.40~100% of our identified PPR genes in each
158 genome (Supplementary Table 5; for details see Materials and Methods).
159 Furthermore, we find that 90.04~95.89% of these chromosomal PPR genes
160 form duplicate gene pairs, each comprising a PPR gene and its duplicate
161 (Supplementary Table 5). Before exploring causes of duplicated PPR genes in
162 these pairs, we infer WGDs of these selected plants through comprehensive
163 analyses of gene synteny and synonymous substitution rates (Ks) between
164 homologous genes (see Materials and Methods). We observe that *Z.*
165 *circumcarinatum* and four selected hornworts are absent from WGD, whereas
166 *T. lepidozioides* and *I. sinensis* are detected to experience one WGD, and *H.*
167 *asiatica* undergoes two rounds of WGD (Supplementary Figures 2-9). These
168 results are consistent with previous studies[3,45-49], except that *I. sinensis* has
169 been reported to experience two WGDs based solely only on Ks distributions[50].
170 We examine genomic loci of PPR genes in these duplicate pairs (for details see
171 Materials and Methods), and find that DD is primarily responsible for PPR gene
172 expansions in the selected species, followed by WGD in *I. sinensis* and *H.*
173 *asiatica* (Figure 2A and Supplementary Table 5).
174
175 It has been reported that DD is related to retroposition[42,44], a process that the
176 complementary DNA transcript of a gene is inserted back into the genome, thus

177 producing a retrogene copy[51]. Generally, retrogene features intron loss, a
178 polyadenylic acid (polyA) tail, and flanking short direct repeats (e.g.
179 transposable element, TE)[51,52]. Thus, we examine these features of DD PPR
180 genes as indicators of retroposition. First, we compare intron counts of PPR
181 genes in each dispersed duplicate pair, and find that 74.78~93.21% of DD PPR
182 gene pairs in hornworts and lycophytes are intronless, followed by the moss *T.*
183 *lepidozioides* with 51.98% (Supplementary Table 6 and Figure 2B).
184 Furthermore, we investigate polyA tails downstream of the 3' untranslated
185 regions (3'UTRs) of intronless DD PPR genes and observe that 75.22~88%
186 have polyA tails in lycophytes, 44.29~62.45% in hornworts, and 61.66% in the
187 moss *T. lepidozioides* (Figure 2C), implying that polyA tails are partially
188 degenerated in hornworts and *T. lepidozioides*, followed by lycophytes.
189 Meanwhile, we search for TEs within 20,000 base pairs upstream and
190 downstream of intronless DD PPR gene pairs, and show that 70.52~99.93% of
191 them are flanked by TEs in lycophytes, 26.56~68.31% in hornworts, and 80.11%
192 in *T. lepidozioides* (Figure 2D). In addition, the majority of these TEs belong to
193 long terminal repeats (LTRs), including LTR/Copia, LTR/Gypsy, and
194 LTR/unknown (Supplementary Table 7). Furthermore, we provide two examples
195 that intronless PPR genes in one DD pair are flanked by the same type of LTR
196 with high sequence similarity, viz., *HaA01G1373* and *HaA43G0090* in *H.*
197 *asiatica* flanked by LTR/Copia (96.43% upstream and 100% downstream)
198 (Figure 2E), and *Tle3c03864.1* and *Tle3c03878.1* in *T. lepidozioides* by
199 LTR/Gypsy (99.86% upstream, 100% downstream) (Figure 2F). Collectively,
200 these results indicate that PPR gene expansions following plant
201 terrestrialization mainly result from DD and retroposition.

202 Horizontal transfer of DYW subgroup PPR genes from plants to 203 non-plants

204 As PPR genes consist of the P and PLS subfamilies, we explore their

phylogenetic patterns by constructing an evolutionary landscape across Archaeplastida (Supplementary Figure 10). We find that the P subfamily PPR genes are ubiquitously distributed in Archaeplastida, with their numbers progressively increasing from early-diverging glaucophytes to derived angiosperms. Conversely, the PLS subfamily, including PLS, E, E+, DYW, and DYW:KP subgroups, is confined to terrestrial plants, with an absence in liverworts and some mosses. In addition, the PLS and E subgroup PPR genes are numerous in lycophytes, while the E+ subgroup is exclusively enriched in hornworts. The DYW subgroup is most abundant in the moss *T. lepidozioides*, with 750 homologs in count; its variant, the DYW:KP subgroup, is restricted to hornworts, lycophytes, and ferns (Supplementary Table 3 and Supplementary Figure 10), conforming well with a previous study[32].

Inspired by findings that the DYW subgroup PPR genes have also been detected in non-plants by previous studies[35,39], we next explore the evolutionary trajectory of the DYW and DYW:KP subgroup PPR genes in plants and non-plants. Given that repetitive and variable PLS motif triplets may hinder reliable phylogenetic inference and the DYW/DYW:KP domains serve as the core functional region of PPR proteins[53], we extract 6,646 DYW and 3,761 DYW:KP domains of 68 terrestrial plants as representatives of DYW/DYW:KP subgroup proteins (Supplementary Table 8). We further search for their homologs in 40 non-plant lineages spanning eukaryotes, prokaryotes, and viruses (Supplementary Table 9). As a result, we detect 1,666 DYW and 6 DYW:KP domain homologs in Metazoa, 31 DYW in Rhizaria, 22 DYW and 1 DYW:KP in Discoba, 15 DYW in Fungi, 9 DYW in Alveolata, 3 DYW in Haptista, and 1 DYW in Amoebozoa (Supplementary Table 10), exhibiting 21.21~80% protein sequence similarity to DYW/DYW:KP domains (Supplementary Table 11). To minimize the domain redundancy, we further select 1,151 DYW and 356

233 DYW:KP domains from 6 species for phylogenetic tree construction, including
234 *P. pearsonii* (hornwort), *T. lepidozioides* (moss), *I. sinensis* (lycophyte),
235 *Alsophila spinulosa* (fern), *Gnetum montanum* (gymnosperm), and *Amborella*
236 *trichopoda* (angiosperm) along with all DYW/DYW:KP domain homologs
237 (Supplementary Table 10). As a result, four non-plant DYW/DYW:KP homolog
238 clusters are interspersed by plant DYW/DYW:KP domains, implying that
239 multiple waves of HGT have occurred from plants to non-plants (Figure 3).
240 Furthermore, we randomly select one DYW domain homolog from each non-
241 plant lineage and compare sequences and structures of these homologs to
242 those of their plant-derived counterparts. On the one hand, most of these
243 homologs contain HxE(x)nCxxC, while PG-boxes (approximately 15 amino acid
244 residues (aa) beginning with proline and glycine[54]) and DYW tripeptides are
245 variable in both plant DYW domains and their non-plant homologs (Figure 4A
246 and 4B; Supplementary Figure 11). On the other hand, structures of these
247 homologs are significantly similar with those of their plant donors (Root Mean
248 Square Deviation, RMSD ≤ 2.5Å[55]), such as the DYW domain in the moss *T.*
249 *lepidozioides* and its homologs in the *Adineta vaga* from Metazoa (Figure 4B
250 and Supplementary Figure 11).
251
252 Furthermore, to explore potential biological functions of DYW domains in non-
253 plants, we analyze expressions of DYW subgroup PPR genes in Metazoa,
254 obtaining the most DYW domains from plants (Figure 3). Surprisingly, all of
255 DYW domains in Metazoa are derived from rotifers, including *A. vaga*, *Adineta*
256 *ricciae*, *Adineta steineri*, *Didymodactylos carnosus*, *Philodina roseola*, *Rotaria*
257 *magnacalcarata*, *Rotaria socialis*, *Rotaria sordida*, *Rotaria sp. Silwood1*,
258 *Rotaria sp. Silwood2*, which are resistant to ultraviolet (UV) irradiation and
259 colonize in a variety of habitats, such as mosses, lichens, soils, and deserts[56,57].
260 Specifically, we find that *A. vaga* has lots of PPR genes (*n*=191), nearly one

261 third of which (*n*=60) fall into the DYW subgroup (Supplementary Table 11).
262 Importantly, the majority of these DYW subgroup PPR genes (58/60) are
263 expressed, indicating their potential importance in biological processes (Figure
264 4C). Coincidentally, the moss *T. lepidozioides*, a potential donor of DYW
265 subgroup PPR genes in *A. vaga*, is also resistant to UV irradiation, and its DYW
266 subgroup PPR genes present elevated expressions after UV irradiation
267 treatment (Figure 4D and Supplementary Table 12), indicating the importance
268 of DYW subgroup PPR genes in response to UV irradiation. Taken together,
269 these results suggest that rotifers, such as *A. vaga*, may inhabit mosses like *T.*
270 *lepidozioides* and obtain the DYW subgroup PPR genes as an adaptation to
271 high UV irradiation (Figure 4E).
272

273 Evolutionary conservation and diversification of MORF 274 domains

275 MORF proteins, essential for PPR-mediated RNA editing and exclusive to seed
276 plants (gymnosperms and angiosperms) (Figure 1C), are defined by their
277 hallmark MORF domains (MORF1~9) and a C-terminal Glycine-Rich (GR)
278 domain[26-28]. To explore the evolutionary conservation and diversification of
279 MORF domains, we extract 334 instances from 33 representative seed plants
280 (Supplementary Table 13) and search for their homologs across 45 lineages
281 spanning eukaryotes, prokaryotes, and viruses (Supplementary Table 9). We
282 find that MORF domains are homologous to peptidase S8
283 propeptide/proteinase inhibitor I9 domains (also known as PA domain[26]) in
284 metazoa, fungi, bryophytes, lycophytes, and ferns (Figure 5A and
285 Supplementary Table 14). Furthermore, we select MORF domains and their PA
286 domain homologs with the highest sequence similarity within the above five
287 lineages and compare their structures. We find that although they have lower
288 sequence similarity (33.33~41.94%), their structures are significantly similar

289 (RMSD ≤ 2.5Å), with RMSD values varying from 0.705~1.039Å (Figure 5B, 5C
290 and Supplementary Figure 12). Notably, MORF3 and MORF8 domains share
291 the same α-helix and β-strand architecture as PA domains (Figure 5C and
292 Supplementary Figure 12). As PA domain is important in assisting fold
293 zymogens or mature peptidases[58,59], these results imply that MORF3 and
294 MORF8 domains potentially help fold nascent PPR proteins to regulate RNA
295 editing.
296
297 To investigate the evolutionary diversifications of MORF domains, we classify
298 them into MORF1 to MORF9 as identified in *A. thaliana* and survey PA domains
299 across our collected 364 high-quality Archaeplastida proteomes. Interestingly,
300 PA domains are ubiquitous in glaucophytes, charophytes, mosses, liverworts,
301 hornworts, lycophytes, ferns, gymnosperms, and angiosperms (Supplementary
302 Figure 13). We choose five PA domains from non-seed plants with the highest
303 similarity to MORF domains as outgroups for phylogenetic tree construction and
304 reveal the early origins of MORF2, MORF5, and MORF6 (Figure 5D).
305 Additionally, to investigate potential biological functions of MORF domains in
306 seed plants, we analyze the expression of genes containing only MORF
307 domains across multiple tissues of the gymnosperm *Torreya grandis* and the
308 angiosperm *A. thaliana*. Interestingly, genes only containing MORF8 and
309 MORF9 are highly expressed in reproductive tissues (Supplementary Figure
310 14), including inflorescence axis tissue, carpel, flower, male and female cone,
311 ovule, and stigmatic tissue. Intriguingly, their expression pattern resembles that
312 of *de novo* genes, which are also preferentially expressed in reproductive
313 tissues[60,61].
314

315 **Diversity and hallmark domain fusions of MORF, ORRM, and**
316 **OZ proteins**

317 Although over 3,000 genomes from more than 1,000 plant species have been
318 sequenced[62], MORF proteins are restricted to only nine members
319 MORF1~MORF9, defined by the hallmark MORF domain in *A. thaliana*. In
320 reality, however, MORF proteins can also contain additional domains, such as
321 GR and RRM[22,27], alongside MORF domains, emphasizing the need for a more
322 precise classification. Thus, based on domain compositions, we reclassify our
323 identified 2,759 MORF proteins into four groups: (1) Single-Pure MORF (SP-
324 MORF), containing a single MORF domain; (2) Single-Mixed MORF (SM-
325 MORF), containing a single MORF domain plus one or more non-MORF
326 domains; (3) Multiple-Pure MORF (MP-MORF), containing multiple MORF
327 domains; and (4) Multiple-Mixed MORF (MM-MORF), containing multiple
328 MORF domains plus one or more non-MORF domains (Figure 6A).
329 Consequently, those 2,759 MORF proteins are classified into 1,916 SP-MORF,
330 544 SM-MORF, 42 MP-MORF, and 257 MM-MORF (Supplementary Table 16).
331 Among them, 493 SM-MORF and 9 MM-MORF proteins feature a GR domain
332 co-occurring with MORF domains (MORF1~6, 8, 9), a combination previously
333 observed only in MORF1, MORF4, and MORF8[27]. To explore the evolutionary
334 distribution of our reclassified MORF proteins, we construct their phylogenetic
335 landscapes across 262 angiosperms and 13 gymnosperms (Figure 6B).
336 Specifically, SP-MORF proteins are widespread in gymnosperms and
337 angiosperms, whereas SM-MORF proteins are abundant in angiosperms but
338 rare in gymnosperms. Conversely, MP-MORF proteins occur sparsely in both
339 lineages and MM-MORF proteins are only observed in angiosperms.
340 Surprisingly, we find that MORF domains and P motifs are fused in two SM-
341 MORF proteins, viz., *AY_017782_RA* in the angiosperm *Aquilaria yunnanensis*
342 and KAH9320946.1 in the gymnosperm *Taxus chinensis* (Supplementary Table

343 16). Meanwhile, MORF domain and the hallmark RRM domain of ORRM, are
344 fused in SM- and MM-MORF proteins in angiosperms. Collectively, these
345 domains' fusions are observed to fall into four types: 1-MORF & 1-RRM (*n*=23),
346 1-MORF & 3-RRM (*n*=1), 2-MORF & 1-RRM (*n*=246), 3-MORF & 1-RRM (*n*=3)
347 (Supplementary Figure 15 and Supplementary Table 16), where the leading
348 number indicates the copy number of each corresponding domain. Notably, the
349 most abundant fused protein, 2-MORF & 1-RRM, has been identified as the
350 RNA editing factor ORRM1[63].
351
352 ORRM proteins, specifically, ORRM1 to ORRM6, have been found in several
353 angiosperms as mediators of RNA editing[21,64]. To investigate their diversity,
354 likewise, we reclassify our identified 7,760 ORRM proteins into SP-ORRM
355 (*n*=2,689), SM-ORRM (*n*=1,751), MP-ORRM (*n*=2,948), and MM-ORRM
356 (*n*=372) (Supplementary Table 17), and build their evolutionary landscapes
357 covering 364 species in Archaeplastida. In addition to previously identified
358 ORRM1 to ORRM6 with a single RRM domain, MP-ORRM proteins, such as
359 CP29A, CP31A, and CP31B, each containing two RRM domains, also affect
360 RNA editing[21]. Intriguingly, RRM domain and motifs/domains (P, E, and DYW)
361 of PPR proteins are fused in SM- and MM-ORRMs across Archaeplastida
362 (Supplementary Figure 17). Specifically, angiosperms have 171 SM-ORRMs,
363 featuring a fusion of RRM and P motifs, far more than any other fusions
364 observed in other lineages (Supplementary Table 17). In addition, RRM domain
365 and RanBP2 Znf domain, the hallmark of OZ proteins, are fused in RRM
366 domain-containing proteins that are not localized to mitochondria or plastids
367 (Supplementary Table 18).
368
369 Similarly, we reclassify OZ proteins and explore their evolutionary distribution
370 based on the same methods used for MORF and ORRM proteins. As a result,

371 we identify 829 OZ proteins, with MP-OZ (*n*=761) overwhelmingly predominant,
372 followed by MM-OZ (*n*=61), SP-OZ (*n*=6), and SM-OZ (*n*=1) (Supplementary
373 Table 19). Based on their evolutionary landscapes across Archaeplastida, we
374 find that these reclassified OZ proteins are enriched in angiosperms, including
375 725 MP-OZ, 60 MM-OZ, 6 SP-OZ, and 1 SM-OZ members (Supplementary
376 Figure 18). Besides, a few MP-OZ proteins are found in chlorophytes (*n*=3),
377 mosses (*n*=14), liverworts (*n*=3), lycophytes (*n*=6), ferns (*n*=6), and
378 gymnosperms (*n*=4), with a single MM-OZ present in hornworts. Interestingly,
379 RanBP2 Znf domains are also fused with motifs/domains (P and E) of PPR
380 proteins in four angiosperm proteins, despite these proteins not being localized
381 to mitochondria or plastids (Supplementary Table 20). Collectively, our results
382 indicate that hallmark motifs/domains of PPR, MORF, ORRM, and OZ proteins
383 can be fused with more diverse patterns than previously thought, with the
384 exception of MORF and OZ (Figure 6C).

385 Discussion

386 Here, we systematically identify PPR, MORF, RRM, and RanBP2 Znf domain-
387 containing proteins and construct their evolutionary landscapes across
388 Archaeplastida. Furthermore, we find that PPR genes greatly expand in
389 hornworts, mosses, lycophytes, mainly due to dispersed duplications and
390 retroposition. Considering the key role of DYW subgroup PPR proteins in
391 mediating RNA editing and their presence in plants and certain non-plant
392 lineages, we investigate their evolutionary trajectories and find extensive
393 horizontal transfer from plants to rotifers, exhibiting sequence and structure
394 similarities, and high expressions. Next, we find that MORF proteins are
395 restricted to seed plants with high expression in reproductive tissues, and their
396 hallmark MORF domains exhibit sequence and structure similarity to PA domain,
397 helping folding proteins. Furthermore, we reclassify MORF, ORRM, and OZ

398 proteins, and find that hallmark domain fusions occur among them and PPR
399 proteins, with the exception of MORF and OZ. Based on these results, we
400 propose a model for the evolution of plant RNA editosomes: RNA editosome,
401 comprised of PPR, MORF, ORRM, OZ, and other proteins, has evolved through
402 diversification of its components in terrestrial plants, (Figure 7), potentially
403 enhancing RNA editing efficiency[65].
404
405 High-quality genome is the key to accurately estimate the size and evolution of
406 gene families. We use BUSCO to estimate the completeness of gene sets and
407 obtain 364 high-quality genomes of Archaeplastida, paving the way for
408 analyses of the evolutionary dynamics of RNA editing factors. For instance, the
409 genome of *T. lepidozioides* indicates that moss harbors numerous PPR genes[3],
410 whereas previously only a few PPR genes have been found in mosses[4,32].
411 However, this example also indicates that more high-quality genomes in
412 Archaeplastida are urgently to be deciphered, especially for non-seed plants,
413 which is the key to depict the more precise evolutionary landscape of RNA
414 editing factors and elucidate their molecular mechanisms.
415
416 PPR proteins, the primary RNA editing factor and one of the largest gene
417 families in terrestrial plants, significantly expand in hornworts, mosses, and
418 lycophytes. We select four hornworts (*A. agrestis*, *N. orbicularis*, *P. pearsonii*,
419 and *P. phymatodes*), one moss (*T. lepidozioides*), and two lycophytes (*H.*
420 *asiatica* and *I. sinensis*) to explore the mechanisms of expansions of PPR
421 genes, and find they are mainly driven by DD in these six plants, followed by
422 WGD in *H. asiatica* and *I. sinensis*. Furthermore, DD PPR genes are likely
423 related to retroposition, characterized by intron loss, the presence of polyA tails,
424 and flanked LTRs. However, WGD in *T. lepidozioides* is scarcely responsible
425 for expansions of PPR genes (Figure 2A) and approximately half of DD PPR

426 pairs experience intron loss (Figure 2B), which potentially results from its
427 ancient WGD event and large-scale gene fractionation[66], and a partial
428 contribution from retroposition. In addition, polyA tails occur at relatively low
429 frequencies in intronless DD PPR genes of *P. pearsonii*, *A. agrestis*, and *N.*
430 *orbicularis* (44.29%-48.25%; Figure 2C), and LTRs flanking intronless DD PPR
431 genes of *N. orbicularis* are rarer (26.56%), suggesting that polyA tails and LTRs
432 are largely degenerated in some hornworts.
433
434 Given that DYW domains in PPR proteins are responsible for catalyzing RNA
435 editing[67], we explore their evolutionary dynamics, and find they are largely
436 transferred from plants to non-plants, including Metazoa (rotifers), Alveolata,
437 Amoebozoa, Discoba, Fungi, Haptista, and Rhizaria (Figure 3A), implying C-to-
438 U RNA editing occur in these non-plants and regulated by PPR proteins.
439 Accordingly, RNA editing events has been found in rotifers and Discoba[39,68,69].
440 Interestingly, we find that rotifers and their DYW domain donors, such as *T.*
441 *lepidozioides*, possess highly expressed DYW domains and high resistance to
442 UV-irradiation, implying that RNA editing mediated by DYW domains repairs
443 damaged DNA at the RNA level. Furthermore, DYW subgroup PPR genes have
444 been found to play a key role for plants in adapting to high/low temperatures
445 and drought[70-72], and rotifers also have these traits[56], suggesting that DYW
446 domain is also important for rotifers in adapting to these extreme environments.
447 In addition, DYW subgroup PPR genes in plants have been proposed to
448 originate via HGT[36] and one member has been found in the chlorophyte *N.*
449 *hyalina*[35]; however, our results do not support the hypothesis or finding,
450 highlighting the need for more high-quality genomes from ancient organisms.
451 We find MORF domains, the hallmarks of PPR accessory proteins, only exist
452 in seed plants and are homologous to PA domains based on sequence and high
453 structure similarity (Figure 5A, 5B, and 5C; Supplementary Figure 8), helping

454 fold proteins. On the contrary, MORF domain has been found to share the
455 highest structure similarities with N-terminal ferredoxin-like domains, rather
456 than PA domains[26]. The difference is likely attributable to domain alignment,
457 with MORF3 domain and more similar PA domains in Metazoa identified in our
458 study differing from MORF1 domain and PA domains in human described in the
459 previous study. Furthermore, these results also indicate that MORF domains
460 fulfill multiple functions in the regulation of RNA editing. Interestingly, we find
461 that MORF genes in seed plants are highly expressed in reproductive tissues
462 and share the same expression pattern with *de novo* genes[60,61,73], suggesting
463 their *de novo* generation.
464
465 As studies of non-PPR editing factors, including MORF, ORRM, and OZ
466 proteins, are limited in a few angiosperms, we reclassify these proteins based
467 on domain compositions and explore their evolution dynamics across
468 Archaeplastida (Figure 6B; Figure Supplementary 13 and 15). Importantly,
469 these reclassified proteins, in addition to the previously identified editing factors,
470 are potential candidates to participate in the regulation of RNA editing. Besides,
471 we also observe other domains also are fused with PPR, MORF, ORRM, and
472 OZ proteins (Supplementary Table 16, 17, and 19). As protein functions and
473 interactions are related to gene/domain fusions[74,75], proteins composed of
474 these fused domains are likely to provide valuable resources to discovering new
475 RNA editing factors.

476 # Materials and Methods

477 ## Collection and quality control of Archaeplastida genomes

478 To estimate the numbers of PPR, MORF, ORRM, and OZ genes and construct
479 their evolutionary landscapes, 364 Archaeplastida high-quality genomes were
480 retrieved from public databases (Supplementary Table 1). To ensure accurate

481 gene counts, only the longest protein isoform of each gene was retained. The
482 completeness of annotated gene sets was estimated by Benchmarking
483 Universal Single-Copy Orthologue (BUSCO) (version 5.8.2) with three lineage
484 datasets chlorophyta_odb10, viridiplantae_odb10, and eukaryota_odb10[76]. As
485 these datasets contain many sequences of chlorophytes and angiosperms,
486 genomes in these two lineages with BUSCO scores ("complete") greater than
487 90% were retained. For other lineages, chromosome-level genomes with
488 BUSCO scores ("complete") greater than 85%; otherwise, the top 20%
489 genomes with the highest BUSCO scores within one lineage were retained.

490 Phylogenetic tree construction of Archaeplastida

491 To construct the phylogenetic tree of 364 Archaeplastida, single-copy
492 orthologous gene families were inferred by OrthoFinder (version 3.0.1b1)[77], and
493 138 single-copy orthologous gene families covering at least 75% species were
494 retained. These single-copy orthologous gene families were aligned by MAFFT
495 (version 7.453)[78] with the parameter "--auto", and poorly aligned regions were
496 identified and trimmed by BMGE (version 1.12) with BLOSUM30[79]. Gene trees
497 were constructed by IQ-TREE (version 2.1.2)[80] with parameters "-B 1000 -bnni
498 -m MFP -T AUTO -ntmax 20" and merged. Taking *Cyanophora paradoxa* as an
499 outgroup, the species tree was inferred by ASTRAL (version 5.7.8)[81].

500 Identification of PPR, MORF, ORRM, and OZ proteins

501 The hallmark domains/motifs of PPR (P, P1, P2, S1, S2, SS, L1, L2, E1, E2,
502 E+, DYW, and DYW:KP), MORF (MORF), ORRM (RRM), and OZ (RanBP2 Znf)
503 proteins were identified using hmmsearch in HMMER (version 3.2.1)[82] with
504 corresponding domain/motif profiles. PPR hallmark domain/motif profiles were
505 obtained from a previous study[32]. PPR hallmark domains/motifs were searched
506 using hmmsearch with parameters "--domtblout -noali -E 0.1". All hmmsearch
507 scores for P, P1, P2, S1, S2, L1, L2, E1, E2, and E+ domains/motifs were

508 greater than 0, with SS motif greater than 10 and DYW/DYW:KP domain greater
509 than 30. As most PPR hallmark domain/motif-containing proteins are localized
510 to mitochondria and/or plastids[17,83], they are defined as PPR proteins. In
511 addition, PPR proteins were required to contain at least three of P, P1, P2, L1,
512 L2, or SS motifs, unless they contained a DYW or DYW:KP domain.
513
514 The MORF domain profile was constructed using hmmbuild in HMMER (version
515 3.2.1) with nine MORF domains from *A. thaliana*. MORF domain was searched
516 using hmmsearch with parameters "--domtblout -noali -E 1e-10". As MORF
517 domain is similar to PA domain, only hits with hmmsearch score greater than
518 36 were kept. The RRM profile was obtained from PF00076.26 in Pfam[84]. RRM
519 domain was searched using hmmsearch with parameters "--domtblout -noali -
520 E 1e-10", and the hmmsearch score was greater than 0. As some MORF and
521 RRM domain-containing proteins harbor a GR domain, it was defined as
522 containing at least 8 glycines within a 20-amino-acid sliding window. The
523 RanBP2 Znf profile was obtained from PF00641.22 in Pfam. RanBP2 Znf
524 domain was searched using hmmsearch with the same parameters as used for
525 RRM domain.
526
527 The domain composition of MORF, RRM, and RanBP2 Znf domain-containing
528 proteins were identified using InterPro (https://www.ebi.ac.uk/interpro)[85].
529 Furthermore, these proteins localized to mitochondrion and/or plastids were
530 defined as MORF, ORRM, and OZ proteins, respectively. The subcellular
531 localization was predicted using DeepLoc (version 2.1)[86] with "Accurate" mode,
532 TargetP (version 2)[87], Plant-mSub[88] (https://bioinfo.usu.edu/Plant-mSubP), and
533 LOCALIZER[89] (https://localizer.csiro.au). Only localization results supported by
534 at least two tools were retained.

535 ## WGD Inference

536 To infer WGD of a plant, intra-genome alignment was conducted using BLASTP
537 in the Basic Local Alignment Search Tool (BLAST, version 2.12.0+)[90] with an E-
538 value cutoff of 1e-5. Based on the alignment results, homologous dotplots were
539 generated using WGDI[91], and synteny blocks were inferred from the same
540 alignments using MCScanX[92]. Ks values of homologs in synteny blocks were
541 calculated using KaKs_Calulator (version 3.0)[93] and ParaAT (version 2.0)[94] with
542 parameters "-p proc -m mafft -f axt -g -k". Next, the median Ks value for each
543 synteny block was calculated. Collectively, Ks peak counts and synteny blocks
544 in the homologous dotplot were used to infer the number of WGD events.

545 ## Identification and classification of gene duplicate pairs

546 To identify gene duplicates, all protein sequences within a genome were aligned
547 by BLASTP (BLAST, version 2.12.0+) with an E-value cutoff of 1e-5.
548 Furthermore, the best alignment hit of a gene was considered a duplicate pair.
549 If the alignment covered more than 60% of the longer gene (query or subject)
550 and the sequence similarity should exceed 20%. Alternatively, if the alignment
551 covered less than 60% of the longer gene, the alignment had to span at least
552 105 amino acids and the sequence similarity exceeded 40%. Otherwise, the
553 query and subject genes were regarded as singletons. Based on genomic
554 locations, gene duplicate pairs were classified into four categories: tandem
555 duplicate pair (adjacent genes), proximal duplicate pair (within 10 genetic loci
556 on the same chromosome), dispersed duplicate pair (more than10 gene loci
557 apart on the same chromosome or on different chromosomes), and WGD
558 duplicate pair (located in synteny blocks).

559 ## TE annotation

560 TEs were identified by a combination of *ab initio* and homology-based methods.
561 For *ab initio* prediction, a consensus sequence library was built by

562 RepeatModeler (version 2.0.6)[95] with parameters "-LTRStruct". LTR_Finder
563 (version 1.07)[96] / LTR_FINDER_parallel (version 1.1)[97], LTRharvest (version
564 1.6.6)[98] with parameters "-minlenltr 100 -maxlenltr 7000", and LTR_retriever
565 (version 2.9.0)[99] were used to build an LTR library. Both libraries were used to
566 annotate TEs by RepeatMasker (version 4.1.8)[100]. For homology-based
567 detection, Dfam (version 3.9)[101] was used to identify TEs by RepeatMasker
568 (version 4.1.8). Collectively, TE annotations obtained from the two methods
569 above were combined.

570 **Retroposition inference**

571 Evidence for retroposition-mediated gene duplication included intron loss,
572 polyA enrichment downstream of the 3'UTR, and TEs flanking duplicate pairs.
573 First, intron loss should satisfy one of following criteria: (1) both genes in
574 dispersed duplicate pair had no intron; (2) one gene had no intron, while the
575 other had at least one intron; (3) considering intron retention, one gene had one
576 intron, while the other had at least one intron; (4) both genes had more than
577 one intron, differing by more than 10 introns. Second, polyA enrichment
578 downstream of the 3'UTR (or polyT enrichment upstream of the 5'UTR) was
579 defined as the presence of at least 8 adenines within each 10-base window.
580 Genomic annotations were used to detect polyA/T enrichment, considering
581 regions extending 1,000 base pairs (bp) downstream of the 3'UTR (or upstream
582 of the 5'UTR). If the 3'UTR (or 5'UTR) of a gene was not annotated, the shortest
583 3'UTR (or 5'UTR) in the genome was used. If no 3'UTR (or 5'UTR) was
584 annotated in the genome, the 100-bp region downstream of the stop codon was
585 taken as the 3'UTR (or the 100-bp region upstream of the start codon was taken
586 as the 5'UTR). Third, genes with TEs located within 20,000 bp upstream and
587 downstream were considered flanked by TEs.

588 **Identification of DYW/DYW:KP domain homologs and inference**
589 **of HGT**

590 To explore HGT of plant DYW/DYW:KP domains, 10,407 protein sequences
591 (hmmsearch score ≥ 30, length ≥ 130 aa) of 68 representative species were
592 obtained (Supplementary Table 8). The NCBI Non-redundant Protein Sequence
593 Database (NR) (acquisition date: 2025.01.19) was used to search these
594 DYW/DYW:KP domain homologs by BLASTP (BLAST, version 2.12.0+) with an
595 E-value cutoff of 1e-5. The searched protein sequences in NR came from 40
596 lineages (Supplementary Table 9). Homologs exhibiting greater than 20%
597 similarity with plant DYW/DYW:KP domains were retained. To avoid
598 contamination, homologs were required to occur in at least two species of the
599 corresponding lineage. Otherwise, a homolog found in only one species of the
600 lineage had to be found in at least one additional genome assembly.
601
602 To construct the phylogenetic tree of plant DYW/DYW:KP domains and their
603 homologs, DYW/DYW:KP domains of six representative species (*P*. *pearsonii*,
604 *T*. *lepidozioides*, *I*. *sinensis*, *A. spinulosa*, *G*. *montanum*, and *A*. *trichopoda*)
605 were further selected. DYW domains and their homologs were aligned by
606 MAFFT (version 7.525), and the poor aligned regions were trimmed by trimAl
607 (version 1.5.rev0)[102] with parameters "-gt 0.7 -cons 60 -st 0.1". The
608 phylogenetic tree was inferred by IQ-TREE (version 2.1.2) with parameters "-B
609 1000 -bnni -m MFP -T AUTO --safe" and visualized by iTOL (version 7)[103]. In
610 addition, multiple sequence alignment of DYW domains from plants and non-
611 plants was conduct by MEGA (version 11)[104] and visualized by Jalview (version
612 2.11.4.0)[105].

613 **Identification of MORF domain homologs**

614 To identify homologs of plant MORF domains, 334 protein sequences
615 (hmmsearch score ≥ 0, length ≥ 86 aa) from 33 representative species were

616 obtained (Supplementary Table 13). The NCBI NR was used to search these
617 MORF domain homologs by BLASTP (BLAST, version 2.12.0+) with an E-value
618 cutoff of 1e-5. The searched protein sequences in NR came from 45 lineages
619 (Supplementary Table 9). Homologs exhibiting greater than 20% similarity to
620 plant MORF domains were retained. The MORF domains and their homologs
621 were aligned by MAFFT (version 7.525), and the phylogenetic tree was
622 constructed by FastTree[106] with parameters "-pseudo -gamma", and visualized
623 by iTOL (version 7).

624 Phylogenetic analysis of MORF domains

625 To construct the phylogenetic tree of MORF domains, all MORF domains of
626 gymnosperms and angiosperms were selected. These domains were aligned
627 to MORF1~MORF9 in *A. thaliana* to determine their types. Five MORF domain
628 homologs from non-plants (Dicom.22G070000.1.p, HaB36G042,
629 Phcar.S3G189600.t1, Phsp.C2G192800.t1, XP_024533311.1), which exhibited
630 the highest similarity to MORF domains, were used as the outgroup. All MORF
631 domains, along with the outgroup sequences, were aligned by MAFFT (version
632 7.525). The phylogenetic tree was inferred with IQ-TREE (version 2.1.2) using
633 the parameters "-B 1000 -bnni -m MFP -T AUTO" and visualized by iTOL
634 (version 7)[103].

635 Analysis of PPR and MORF gene expression

636 To calculate gene expression, reference genome indexes were built by hisat2-
637 build in HISAT2 (version 2.0.5)[107] and rsem-prepare-reference in RSEM
638 (version 1.3.1)[108]. Raw RNA sequencing data were cleaned by fastp (version
639 0.20.1)[109] with parameters "-q 20 -u 40 -l 15 -g -x -r -W 4 -M 20 -w 12". Strand
640 specificity was inferred by infer_experiment.py in RSeQC (version 2.6.4)[110].
641 Reads mapping was conducted by STAR (version 2.7.1a)[111]. Gene expression
642 matrix was generated by rsem-calculate-expression in RSEM (version 1.3.1)

643 with parameters "--keep-intermediate-files --time --star --append-names --
644 output-genome-bam --sort-bam-by-coordinate".

645 Protein structure prediction and alignment

646 Protein structures were predicted by ColabFold
647 (https://colab.research.google.com/github/sokrypton/ColabFold)[112], and
648 different protein structures were aligned by "Matchmaker" of ChimeraX (version
649 1.9)[113].

650 Code availability

651 Computational scripts for evolutionary analyses of plant RNA editing factors are
652 available at GitHub (https://github.com/Bioinfo-Ming/PlantEditosomeEvolution).

653 ACKNOWLEDGEMENTS

654 These authors would like to thank Shuhui Song, Lina Ma, Zhao Li, Yang Zhang,
655 Tongtong Zhu, Xing Zheng, Yue Qi, Xinyu Zhou, Yuxin Qin, Zhixiang Yuan and
656 Zhenxian Han for discussions and advice. This work was supported by grants
657 from National Natural Science Foundation of China [T2425005 and 32030021]
658 and International Partnership Program of the Chinese Academy of Sciences
659 [153F11KYSB20160008].

660 CONTRIBUTOR INFORMATION

661 Z.Z. conceived, initiated and supervised this study. M.C. wrote the manuscript.
662 M.C. collected and analyzed the data.

663 FUNDING

664 Conflict of interest statement. None declared.

665 **REFERENCES**

666 1 Takenaka, M., Zehrmann, A., Verbitskiy, D., Haertel, B. & Brennicke, A.
667 RNA editing in plants and its evolution. *Annual Review of Genetics* **47**, 335-
668 352 (2013).
669 2 Fujii, S. & Small, I. The evolution of RNA editing and pentatricopeptide
670 repeat genes. *New Phytologist* **191**, 37-47 (2011).
671 3 Hu, R. *et al.* Adaptive evolution of the enigmatic Takakia now facing climate
672 change in Tibet. *Cell* **186**, 3558-3576. e3517 (2023).
673 4 Small, I. D., Schallenberg‐Rüdinger, M., Takenaka, M., Mireau, H. &
674 Ostersetzer‐Biran, O. Plant organellar RNA editing: what 30 years of
675 research has revealed. *The Plant Journal* **101**, 1040-1056 (2020).
676 5 Wang, Y. *et al.* Empty Pericarp21 encodes a novel PPR-DYW protein that
677 is required for mitochondrial RNA editing at multiple sites, complexes I and
678 V biogenesis, and seed development in maize. *PLoS Genetics* **15**,
679 e1008305 (2019).
680 6 Sun, Y. K., Gutmann, B., Yap, A., Kindgren, P. & Small, I. Editing of
681 chloroplast rps14 by PPR editing factor EMB2261 is essential for
682 Arabidopsis development. *Frontiers in Plant Science* **9**, 841 (2018).
683 7 Li, J. *et al.* The Pentatricopeptide Repeat Protein OsPPR674 Regulates
684 Rice Growth and Drought Sensitivity by Modulating RNA Editing of the
685 Mitochondrial Transcript ccmC. *International Journal of Molecular Sciences*
686 **26**, 2646 (2025).
687 8 Shi, X., Germain, A., Hanson, M. R. & Bentolila, S. RNA recognition motif-
688 containing protein ORRM4 broadly affects mitochondrial RNA editing and
689 impacts plant development and flowering. *Plant Physiology* **170**, 294-309
690 (2016).
691 9 Yang, Y. *et al.* The RNA editing factor SlORRM4 is required for normal fruit

692 ripening in tomato. *Plant Physiology* **175**, 1690-1702 (2017).
693 10 Luo, Z. *et al.* Pentatricopeptide repeat gene-mediated mitochondrial RNA
694 editing impacts on rice drought tolerance. *Frontiers in Plant Science* **13**,
695 926285 (2022).
696 11 Yuan, H. & Liu, D. Functional disruption of the pentatricopeptide protein
697 SLG1 affects mitochondrial RNA editing, plant development, and
698 responses to abiotic stresses in Arabidopsis. *The Plant Journal* **70**, 432-
699 444 (2012).
700 12 Sun, T., Bentolila, S. & Hanson, M. R. The unexpected diversity of plant
701 organelle RNA editosomes. *Trends in Plant Science* **21**, 962-973 (2016).
702 13 Wang, T. & Takenaka, M. The molecular basis and evolution of the
703 organellar RNA editosome by complementary DYW deaminases in seed
704 plants. *Plant Physiology* **197**, kiaf142 (2025).
705 14 Knoop, V. C-to-U and U-to-C: RNA editing in plant organelles and beyond.
706 *Journal of Experimental Botany* **74**, 2273-2294 (2023).
707 15 Yan, J., Zhang, Q. & Yin, P. RNA editing machinery in plant organelles.
708 *Science China Life Sciences* **61**, 162-169 (2018).
709 16 Ichinose, M. & Sugita, M. RNA editing and its molecular mechanism in plant
710 organelles. *Genes* **8**, 5 (2016).
711 17 Wang, Y. & Tan, B.-C. Pentatricopeptide repeat proteins in plants: Cellular
712 functions, action mechanisms, and potential applications. *Plant*
713 *Communications* **6** (2025).
714 18 Yan, J. *et al.* MORF9 increases the RNA-binding activity of PLS-type
715 pentatricopeptide repeat protein in plastid RNA editing. *Nature Plants* **3**, 1-
716 8 (2017).
717 19 Wang, Y. *et al.* Maize PPR-E proteins mediate RNA C-to-U editing in
718 mitochondria by recruiting the trans deaminase PCW1. *The Plant Cell* **35**,
719 529-551 (2023).

720 20 Yang, Y.-Z. *et al.* GRP23 plays a core role in E-type editosomes via
721 interacting with MORFs and atypical PPR-DYWs in Arabidopsis
722 mitochondria. *Proceedings of the National Academy of Sciences* **119**,
723 e2210978119 (2022).
724 21 Shi, X., Bentolila, S. & Hanson, M. R. Organelle RNA recognition motif-
725 containing (ORRM) proteins are plastid and mitochondrial editing factors in
726 Arabidopsis. *Plant Signaling & Behavior* **11**, 294-309 (2016).
727 22 Sun, T. *et al.* An RNA recognition motif-containing protein is required for
728 plastid RNA editing in Arabidopsis and maize. *Proceedings of the National*
729 *Academy of Sciences* **110**, E1169-E1178 (2013).
730 23 Sun, T. *et al.* A zinc finger motif-containing protein is essential for
731 chloroplast RNA editing. *PLoS Genetics* **11**, e1005028 (2015).
732 24 Gipson, A. B., Hanson, M. R. & Bentolila, S. The RanBP2 zinc finger
733 domains of chloroplast RNA editing factor OZ1 are required for protein–
734 protein interactions and conversion of C to U. *The Plant Journal* **109**, 215-
735 226 (2022).
736 25 Wang, Z. *et al.* OsOZ1, a zinc finger motif-containing protein, is essential
737 for chloroplast development through RNA editing and intron splicing in rice.
738 *Plant Growth Regulation* **105**, 449-461 (2025).
739 26 Haag, S. *et al.* Crystal structures of the Arabidopsis thaliana organellar RNA
740 editing factors MORF1 and MORF9. *Nucleic Acids Research* **45**, 4915-
741 4928 (2017).
742 27 Li, J., Yuan, J., Jing, Y. & Lin, R. MORF proteins: A small family regulating
743 organellar RNA editing and beyond. *Journal of Integrative Plant Biology*
744 (2025).
745 28 Takenaka, M. *et al.* Multiple organellar RNA editing factor (MORF) family
746 proteins are required for RNA editing in mitochondria and plastids of plants.
747 *Proceedings of the National Academy of Sciences* **109**, 5104-5109 (2012).

748 29 Sugita, M., Uchiyama, H. & Ichinose, M. RRM and PPR proteins network
749 as a global regulator for plastid RNA metabolism. *Endocytobiosis Cell Res.*
750 **27**, 41-44 (2016).
751 30 Baudry, K., Monachello, D., Castandet, B., Majeran, W. & Lurin, C.
752 Dissecting the molecular puzzle of the editosome core in Arabidopsis
753 organelles. *Plant Science* **344**, 112101 (2024).
754 31 Shikanai, T. RNA editing in plant organelles: machinery, physiological
755 function and evolution. *Cellular and Molecular Life Sciences CMLS* **63**, 698-
756 708 (2006).
757 32 Gutmann, B. *et al.* The expansion and diversification of pentatricopeptide
758 repeat RNA-editing factors in plants. *Molecular plant* **13**, 215-230 (2020).
759 33 Manna, S. An overview of pentatricopeptide repeat proteins and their
760 applications. *Biochimie* **113**, 93-99 (2015).
761 34 Manna, S. & Barth, C. Identification of a novel pentatricopeptide repeat
762 subfamily with a C-terminal domain of bacterial origin acquired via ancient
763 horizontal gene transfer. *BMC Research Notes* **6**, 525 (2013).
764 35 Schallenberg-Rüdinger, M., Lenz, H., Polsakiewicz, M., Gott, J. M. & Knoop,
765 V. A survey of PPR proteins identifies DYW domains like those of land plant
766 RNA editing factors in diverse eukaryotes. *RNA Biology* **10**, 1549-1556
767 (2013).
768 36 Knoop, V. & Rüdinger, M. DYW-type PPR proteins in a heterolobosean
769 protist: plant RNA editing factors involved in an ancient horizontal gene
770 transfer? *FEBS Letters* **584**, 4287-4291 (2010).
771 37 Schaap, P. *et al.* The Physarum polycephalum genome reveals extensive
772 use of prokaryotic two-component and metazoan-type tyrosine kinase
773 signaling. *Genome Biology and Evolution* **8**, 109-125 (2016).
774 38 Manna, S., Brewster, J. & Barth, C. Identification of pentatricopeptide
775 repeat proteins in the model organism Dictyostelium discoideum.

776 *International Journal of Genomics* **2013**, 586498 (2013).
777 39 Fu, C.-J., Sheikh, S., Miao, W., Andersson, S. G. & Baldauf, S. L. Missing
778 genes, multiple ORFs, and C-to-U type RNA editing in Acrasis kona
779 (Heterolobosea, Excavata) mitochondrial DNA. *Genome Biology and*
780 *Evolution* **6**, 2240-2257 (2014).
781 40 Puttick, M. N. *et al.* The interrelationships of land plants and the nature of
782 the ancestral embryophyte. *Current Biology* **28**, 733-745. e732 (2018).
783 41 Wu, W. *et al.* Genomic convergence in terrestrial root plants through
784 tandem duplication in response to soil microbial pressures. *Cell Reports* **43**
785 (2024).
786 42 Panchy, N., Lehti-Shiu, M. & Shiu, S.-H. Evolution of gene duplication in
787 plants. *Plant Physiology* **171**, 2294-2316 (2016).
788 43 Freeling, M. Bias in plant gene content following different sorts of
789 duplication: tandem, whole-genome, segmental, or by transposition.
790 *Annual Review of Plant Biology* **60**, 433-453 (2009).
791 44 Wang, X. *et al.* A recent burst of gene duplications in Triticeae. *Plant*
792 *Communications* **3** (2022).
793 45 Feng, X. *et al.* Genomes of multicellular algal sisters to land plants
794 illuminate signaling network evolution. *Nature Genetics* **56**, 1018-1031
795 (2024).
796 46 Zhang, J. *et al.* (2020).
797 47 Li, F.-W. *et al.* Anthoceros genomes illuminate the origin of land plants and
798 the unique biology of hornworts. *Nature Plants* **6**, 259-272 (2020).
799 48 Schafran, P. *et al.* Pan-phylum genomes of hornworts reveal conserved
800 autosomes but dynamic accessory and sex chromosomes. *Nature Plants*
801 **11**, 49-62 (2025).
802 49 Li, C. *et al.* Extraordinary preservation of gene collinearity over three
803 hundred million years revealed in homosporous lycophytes. *Proceedings*

804 *of the National Academy of Sciences* **121**, e2312607121 (2024).
805 50 Cui, J. *et al.* Chromosome-level reference genome of tetraploid Isoetes
806 sinensis provides insights into evolution and adaption of lycophytes.
807 *GigaScience* **12**, giad079 (2023).
808 51 Zhang, C., Gschwend, A. R., Ouyang, Y. & Long, M. Evolution of gene
809 structural complexity: an alternative-splicing-based model accounts for
810 intron-containing retrogenes. *Plant Physiology* **165**, 412-423 (2014).
811 52 Schrader, L. & Schmitz, J. The impact of transposable elements in adaptive
812 evolution. *Molecular Ecology* **28**, 1537-1549 (2019).
813 53 Wu, C. S. & Chaw, S. M. Evolution of mitochondrial RNA editing in extant
814 gymnosperms. *The Plant Journal* **111**, 1676-1687 (2022).
815 54 Hayes, M. L., Giang, K., Berhane, B. & Mulligan, R. M. Identification of two
816 pentatricopeptide repeat genes required for RNA editing and zinc binding
817 by C-terminal cytidine deaminase-like domains. *Journal of Biological*
818 *Chemistry* **288**, 36519-36529 (2013).
819 55 Bolhuis, P. Sampling kinetic protein folding pathways using all-atom models.
820 *Lecture Notes in Physics*, 703, 393-433 (2006).
821 56 Hespeels, B. *et al.* Back to the roots, desiccation and radiation resistances
822 are ancestral characters in bdelloid rotifers. *BMC Biology* **21**, 72 (2023).
823 57 Wang, W., Yang, Y., Örstan, A., He, Z. & Wang, Q. Species diversity of
824 bdelloid rotifer (Rotifera, Bdelloidea) in different areas in China, with a
825 description of two new species. *ZooKeys* **1229**, 1 (2025).
826 58 Jain, S. C., Shinde, U., Li, Y., Inouye, M. & Berman, H. M. The crystal
827 structure of an autoprocessed Ser221Cys-subtilisin E-propeptide complex
828 at 2.0 Å resolution. *Journal of Molecular Biology* **284**, 137-144 (1998).
829 59 Li, Y., Hu, Z., Jordan, F. & Inouye, M. Functional Analysis of the Propeptide
830 of Subtilisin E as an Intramolecular Chaperone for Protein Folding:
831 Refolding And Inhibitory Abilities of Propeptide Mutants (∗). *Journal of*

832 *Biological Chemistry* **270**, 25127-25132 (1995).
833 60 Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in
834 plants. *Trends in Plant Science* **19**, 698-708 (2014).
835 61 Li, L. *et al.* Identification of the novel protein QQS as a component of the
836 starch metabolic network in Arabidopsis leaves. *The Plant Journal* **58**, 485-
837 498 (2009).
838 62 Xie, L. *et al.* Technology-enabled great leap in deciphering plant genomes.
839 *Nature Plants* **10**, 551-566 (2024).
840 63 Searing, A. M., Satyanarayan, M. B., O′ Donnell, J. P. & Lu, Y. Two
841 organelle RNA recognition motif proteins affect distinct sets of RNA editing
842 sites in the Arabidopsis thaliana plastid. *Plant Direct* **4**, e00213 (2020).
843 64 Shi, X., Hanson, M. R. & Bentolila, S. Two RNA recognition motif-containing
844 proteins are plant mitochondrial editing factors. *Nucleic Acids Research* **43**,
845 3814-3825 (2015).
846 65 Lombana, J. M., Hanson, M. R. & Bentolila, S. Deciphering the role of
847 accessory proteins in Arabidopsis chloroplast editosomes via interaction
848 with a synthetic PPR-PLS factor in E. coli. *Nucleic Acids Research* **53**,
849 gkaf483 (2025).
850 66 Liang, Z. & Schnable, J. C. Functional divergence between subgenomes
851 and gene pairs after whole genome duplications. *Molecular Plant* **11**, 388-
852 397 (2018).
853 67 Takenaka, M. *et al.* DYW domain structures imply an unusual regulation
854 principle in plant organellar RNA editing catalysis. *Nature catalysis* **4**, 510-
855 522 (2021).
856 68 Rüdinger, M., Fritz-Laylin, L., Polsakiewicz, M. & Knoop, V. Plant-type
857 mitochondrial RNA editing in the protist Naegleria gruberi. *RNA* **17**, 2058-
858 2062 (2011).
859 69 Yang, J., Harding, T., Kamikawa, R., Simpson, A. G. & Roger, A. J.

860 Mitochondrial genome evolution and a novel RNA editing system in deep-
861 branching heteroloboseids. *Genome Biology and Evolution* **9**, 1161-1174
862 (2017).
863 70 Su, H.-G. *et al.* Genome-wide analysis of the DYW subgroup PPR gene
864 family and identification of GmPPR4 responses to drought stress.
865 *International Journal of Molecular Sciences* **20**, 5667 (2019).
866 71 Cui, X. *et al.* The RNA editing factor DUA 1 is crucial to chloroplast
867 development at low temperature in rice. *New Phytologist* **221**, 834-849
868 (2019).
869 72 Huang, C. *et al.* The PPR protein RARE1-mediated editing of chloroplast
870 accD transcripts is required for fatty acid biosynthesis and heat tolerance
871 in Arabidopsis. *Plant Communications* **4** (2023).
872 73 Zhang, L. *et al.* Rapid evolution of protein diversity by de novo origination
873 in Oryza. *Nature Ecology & Evolution* **3**, 679-690 (2019).
874 74 Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein
875 interaction maps for complete genomes based on gene fusion events.
876 *Nature* **402**, 86-90 (1999).
877 75 Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg,
878 D. A combined algorithm for genome-wide prediction of protein function.
879 *Nature* **402**, 83-86 (1999).
880 76 Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M.
881 BUSCO update: novel and streamlined workflows along with broader and
882 deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and
883 viral genomes. *Molecular Biology and Evolution* **38**, 4647-4654 (2021).
884 77 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for
885 comparative genomics. *Genome Biology* **20**, 1-14 (2019).
886 78 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
887 version 7: improvements in performance and usability. *Molecular Biology*

888 *and Evolution* **30**, 772-780 (2013).
889 79 Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with
890 Entropy): a new software for selection of phylogenetic informative regions
891 from multiple sequence alignments. *BMC Evolutionary Biology* **10**, 1-21
892 (2010).
893 80 Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for
894 phylogenetic inference in the genomic era. *Molecular Biology and Evolution*
895 **37**, 1530-1534 (2020).
896 81 Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree
897 estimation. *Bioinformatics* **30**, i541-i548 (2014).
898 82 Eddy, S. Sequence analysis using profile hidden Markov models.
899 http://hmmer. *wustl. edu* (2004).
900 83 Meng, L. *et al.* PPR proteins in plants: Roles, mechanisms, and prospects
901 for rice research. *Frontiers in Plant Science* **15**, 1416742 (2024).
902 84 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids*
903 *Research* **49**, D412-D419 (2021).
904 85 Blum, M. *et al.* InterPro: the protein sequence classification resource in
905 2025. *Nucleic Acids Research* **53**, D444-D456 (2025).
906 86 Ødum, M. T. *et al.* DeepLoc 2.1: multi-label membrane protein type
907 prediction using protein language models. *Nucleic Acids Research* **52**,
908 W215-W220 (2024).
909 87 Armenteros, J. J. A. *et al.* Detecting sequence signals in targeting peptides
910 using deep learning. *Life Science Alliance* **2** (2019).
911 88 Sahu, S. S., Loaiza, C. D. & Kaundal, R. Plant-mSubP: a computational
912 framework for the prediction of single-and multi-target protein subcellular
913 localization using integrated machine-learning approaches. *AoB Plants* **12**,
914 plz068 (2020).
915 89 Sperschneider, J. *et al.* LOCALIZER: subcellular localization prediction of

916 both plant and effector proteins in the plant cell. *Scientific Reports* **7**, 44598
917 (2017).
918 90 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC*
919 *Bioinformatics* **10**, 1-9 (2009).
920 91 Sun, P. *et al.* WGDI: a user-friendly toolkit for evolutionary analyses of
921 whole-genome duplications and ancestral karyotypes. *Molecular Plant* **15**,
922 1841-1851 (2022).
923 92 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis
924 of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49-e49
925 (2012).
926 93 Zhang, Z. KaKs_Calculator 3.0: calculating selective pressure on coding
927 and non-coding sequences. *Genomics, Proteomics and Bioinformatics* **20**,
928 536-540 (2022).
929 94 Zhang, Z. *et al.* ParaAT: a parallel tool for constructing multiple protein-
930 coding DNA alignments. *Biochemical and Biophysical Research*
931 *Communications* **419**, 779-781 (2012).
932 95 Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of
933 transposable element families. *Proceedings of the National Academy of*
934 *Sciences* **117**, 9451-9457 (2020).
935 96 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-
936 length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-W268
937 (2007).
938 97 Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER
939 enabling rapid identification of long terminal repeat retrotransposons.
940 *Mobile DNA* **10**, 48 (2019).
941 98 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible
942 software for de novo detection of LTR retrotransposons. *BMC*
943 *Bioinformatics* **9**, 1-14 (2008).

944 99 Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program
945 for identification of long terminal repeat retrotransposons. *Plant Physiology*
946 **176**, 1410-1422 (2018).
947 100 Chen, N. Using RepeatMasker to identify repetitive elements in genomic
948 sequences. *Current Protocols in Bioinformatics* **5**, 4.10. 11-14.10. 14 (2004).
949 101 Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam
950 community resource of transposable element families, sequence models,
951 and genome annotations. *Mobile DNA* **12**, 2 (2021).
952 102 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for
953 automated alignment trimming in large-scale phylogenetic analyses.
954 *Bioinformatics* **25**, 1972-1973 (2009).
955 103 Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to
956 the phylogenetic tree display and annotation tool. *Nucleic Acids Research*
957 **52**, W78-W82 (2024).
958 104 Tamura, K., Stecher, G. & Kumar, S. MEGA11: molecular evolutionary
959 genetics analysis version 11. *Molecular Biology and Evolution* **38**, 3022-
960 3027 (2021).
961 105 Procter, J. B. *et al.* Alignment of biological sequences with Jalview. *Methods*
962 *in Molecular Biology* **2231**, 203-224 (2021).
963 106 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately
964 maximum-likelihood trees for large alignments. *PloS One* **5**, e9490 (2010).
965 107 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based
966 genome alignment and genotyping with HISAT2 and HISAT-genotype.
967 *Nature Biotechnology* **37**, 907-915 (2019).
968 108 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-
969 Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 1-
970 16 (2011).
971 109 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ

preprocessor. *Bioinformatics* **34**, i884-i890 (2018).

110 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185 (2012).

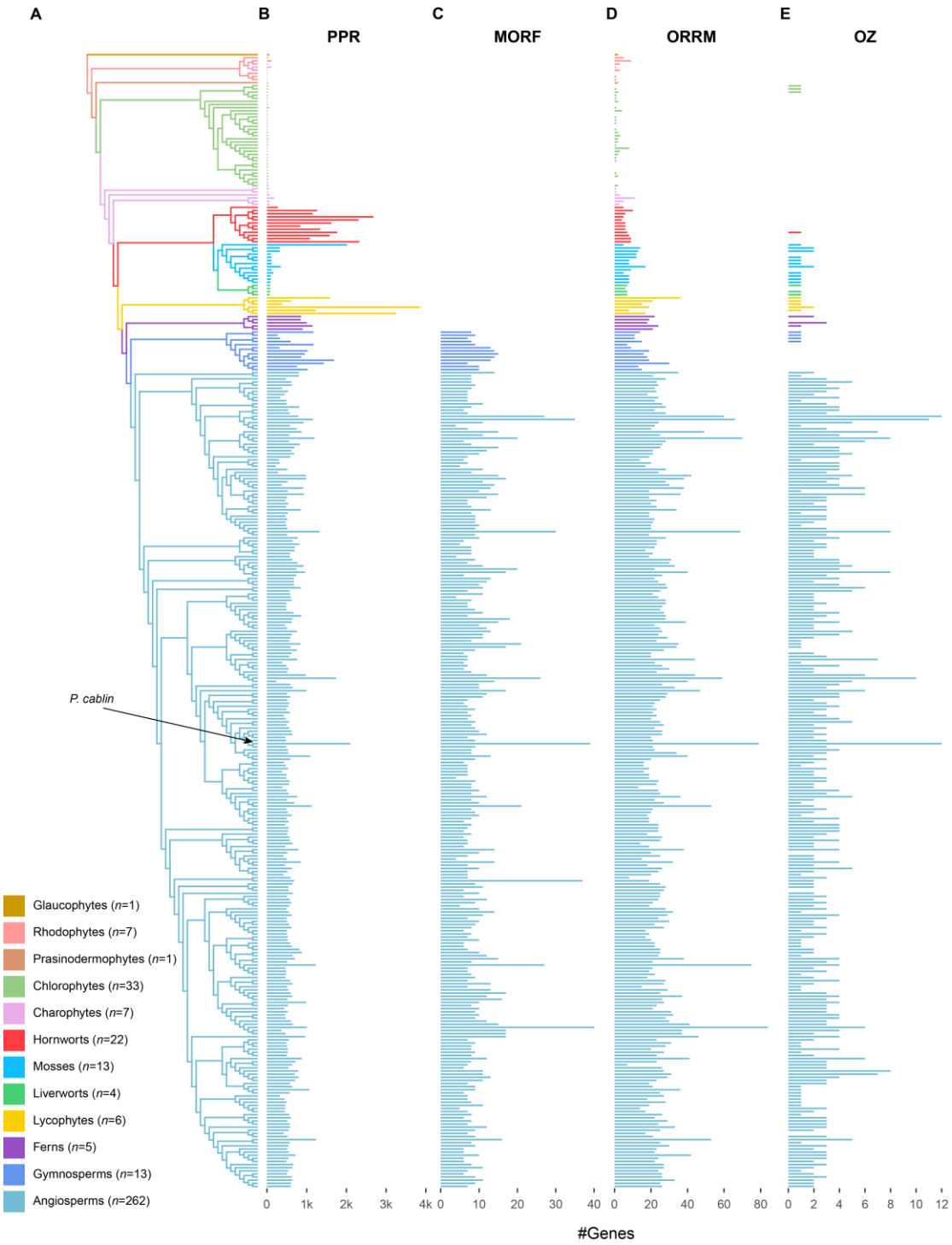
111 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

112 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679-682 (2022).

113 Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Science* **32**, e4792 (2023).

981 # Figures

982 ## Figure 1

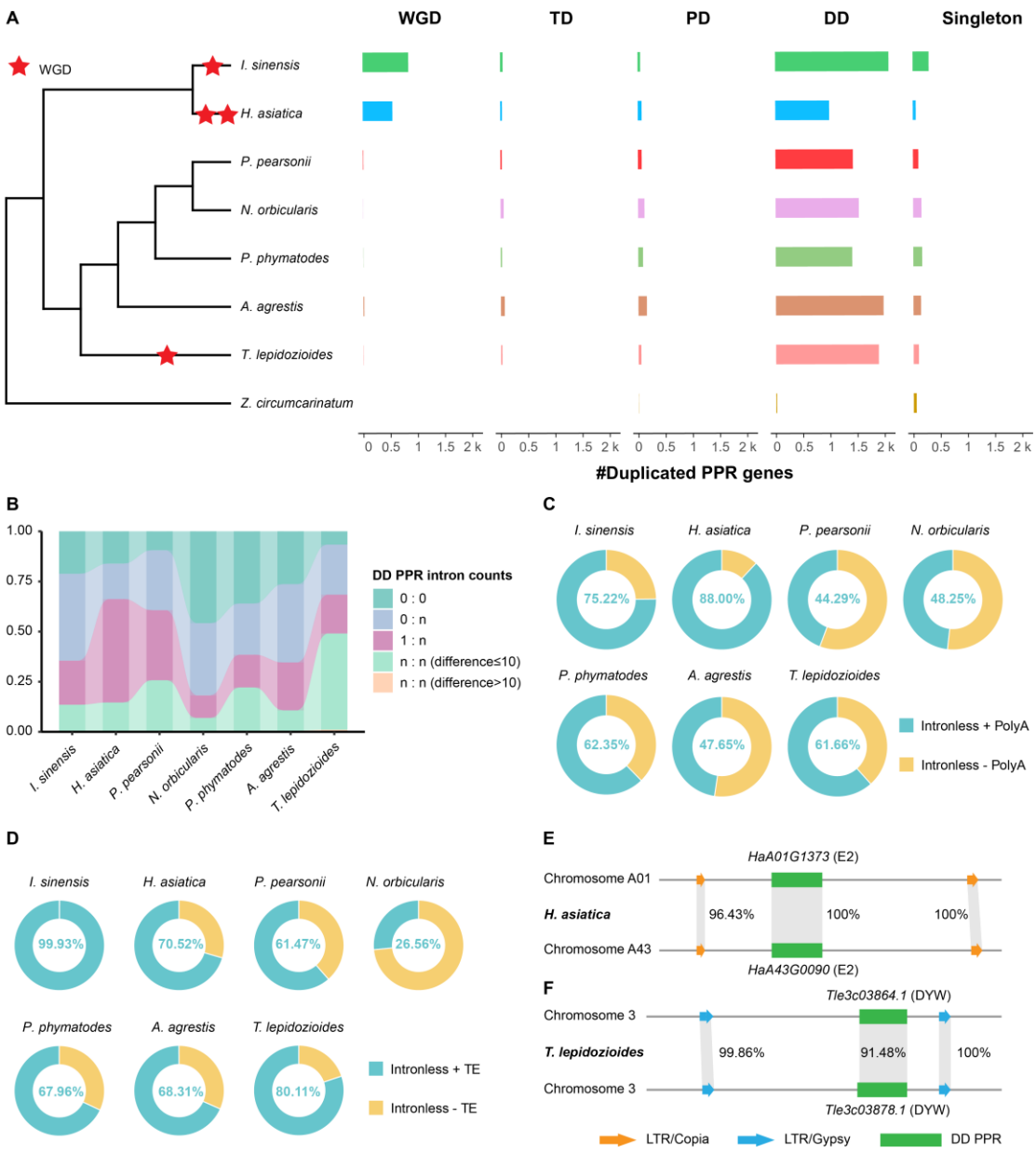


983

984 **Figure 1 The evolutionary landscape of PPR, MORF, ORRM, and OZ genes**
985 **across Archaeplastida. (A)** Species tree of 364 Archaeplastida, including
986 glaucophytes (*n*=1), rhodophytes (*n*=7), prasinodermophytes (*n*=1),

987 chlorophytes (*n*=33), charophytes (*n*=7), hornworts (*n*=22), mosses (*n*=13),
988 liverworts (*n*=4), lycophytes (*n*=6), ferns (*n*=5), gymnosperms (*n*=13), and
989 angiosperms (*n*=262). Distribution of **(B)** PPR, **(C)** MORF, **(D)** ORRM, and **(E)**
990 OZ gene counts across Archaeplastida lineages. *P*. *cablin* simultaneously
991 possess numerous MORF, ORRM, and OZ genes, as well as PPR genes.
992

993 **Figure 2**

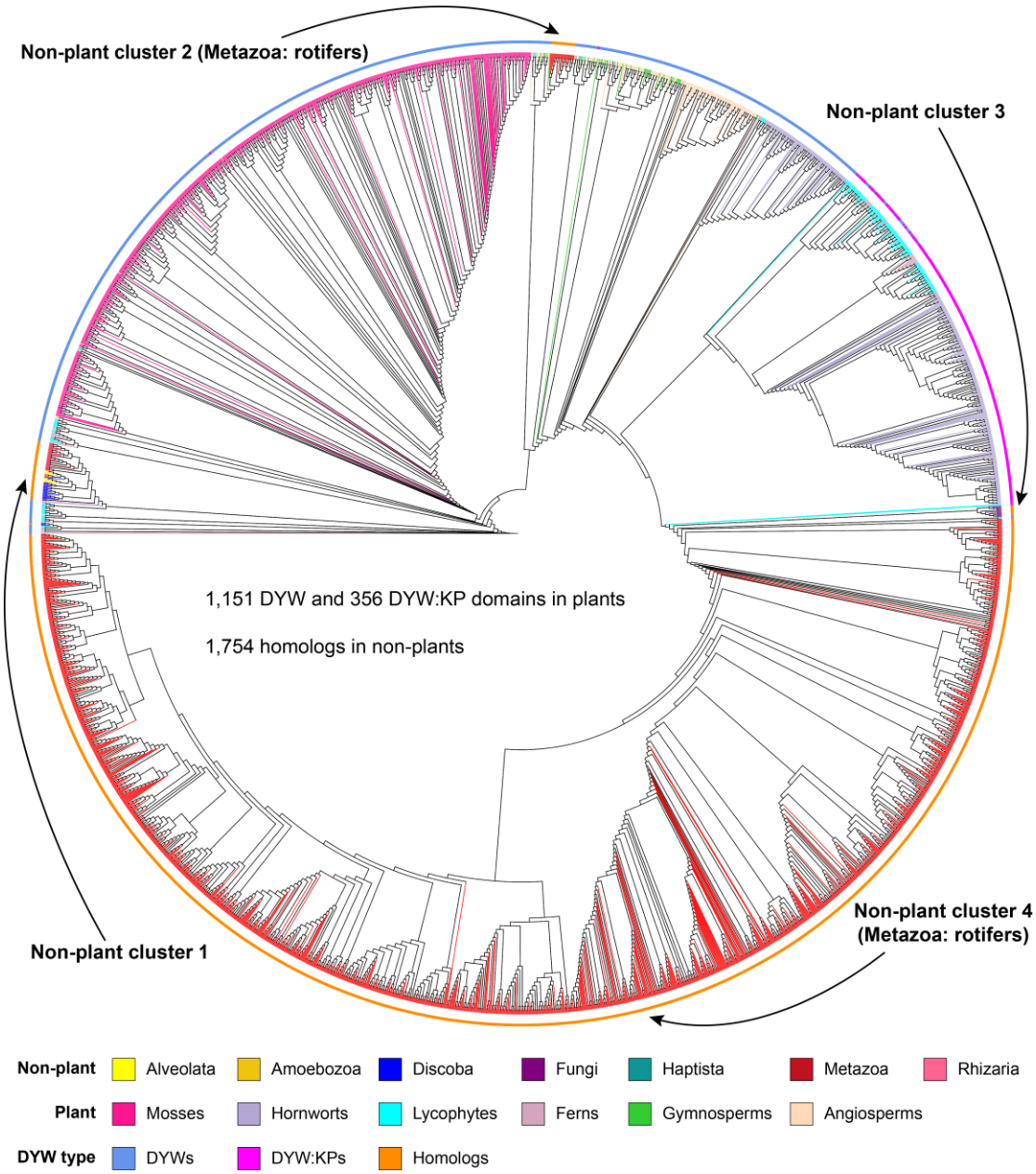


994

995 **Figure 2. Mechanisms of PPR gene expansions. (A)** Numbers of PPR genes
996 (original and duplicated) arising from WGD, TD, PD, and DD in seven
997 representative species (*A. agrestis*, *N. orbicularis*, *P. pearsonii*, *P. phymatodes*,
998 *T. lepidozioides*, *I. sinensis*, and *H. asiatica*) and one outgroup (*Z.*
999 *circumcarinatum*). One red five-pointed star denotes one whole genome
1000 duplication. WGD: Whole Genome Duplication, TD: Tandem Duplication, PD:
1001 Proximal Duplication, DD: Dispersed Duplication. **(B)** Percentage of intronless
1002 DD PPR gene pairs. Numbers to the left and right of the colon indicate intron

1003 DD PPR genes in each pair. Categories “0:0”, “0:n”, “1:n”, and “n:n (difference >
1004 10)” denote intron loss in DD PPR gene pairs. Percentages of intronless DD
1005 PPR gene pairs with **(C)** polyA tails and **(D)** flanked by LTRs. Examples of **(E)**
1006 E2 and **(F)** DYW subgroup DD PPR gene pairs flanked by highly similar
1007 LTR/Copia and LTR/Gypsy, respectively.
1008

1009 **Figure 3**



1010

1011 **Figure 3. Phylogenetic tree of DYW and DYW:KP domains in plants and**

1012 **their non-plant homologs.**

1013

Figure 4

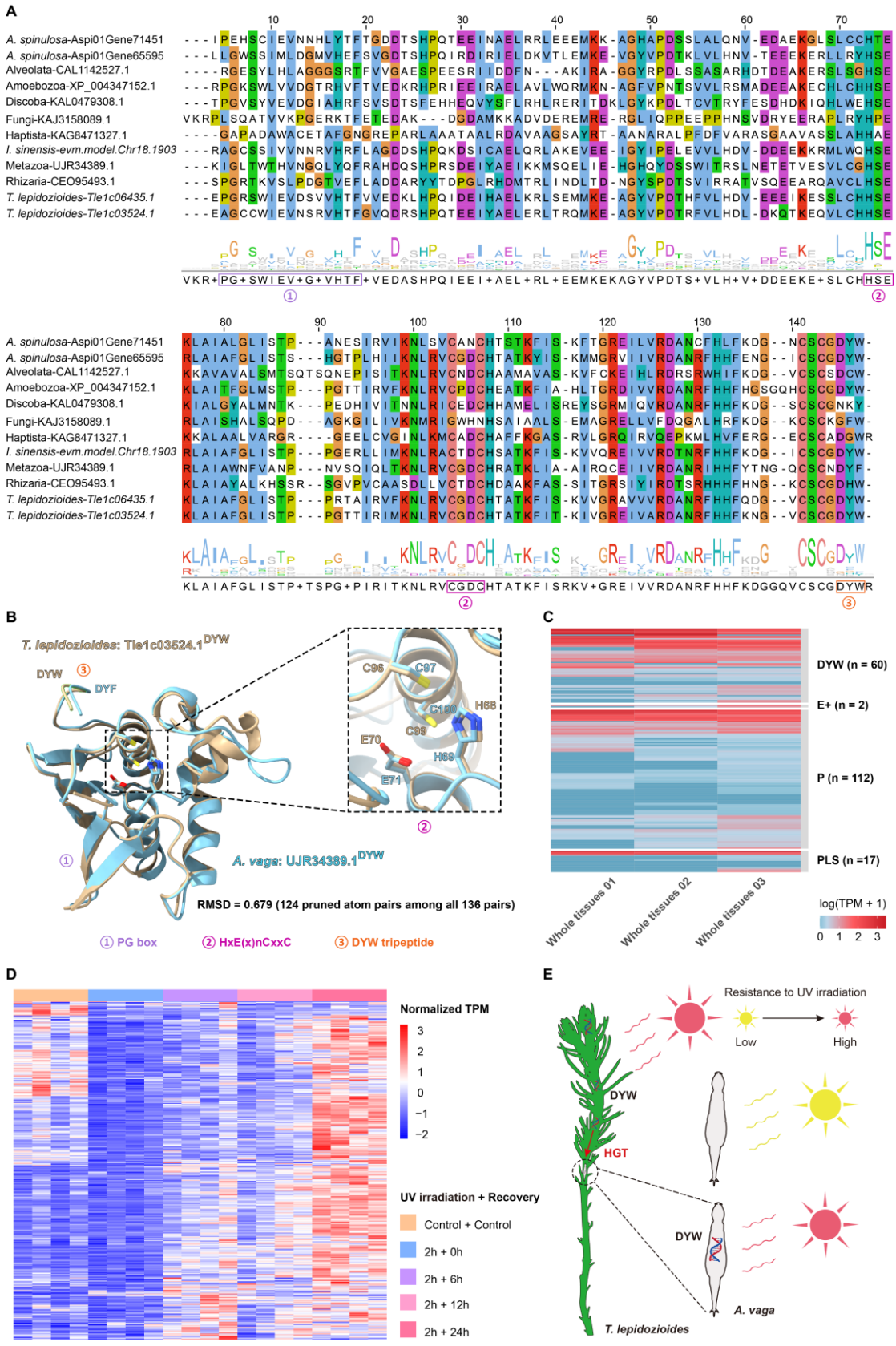
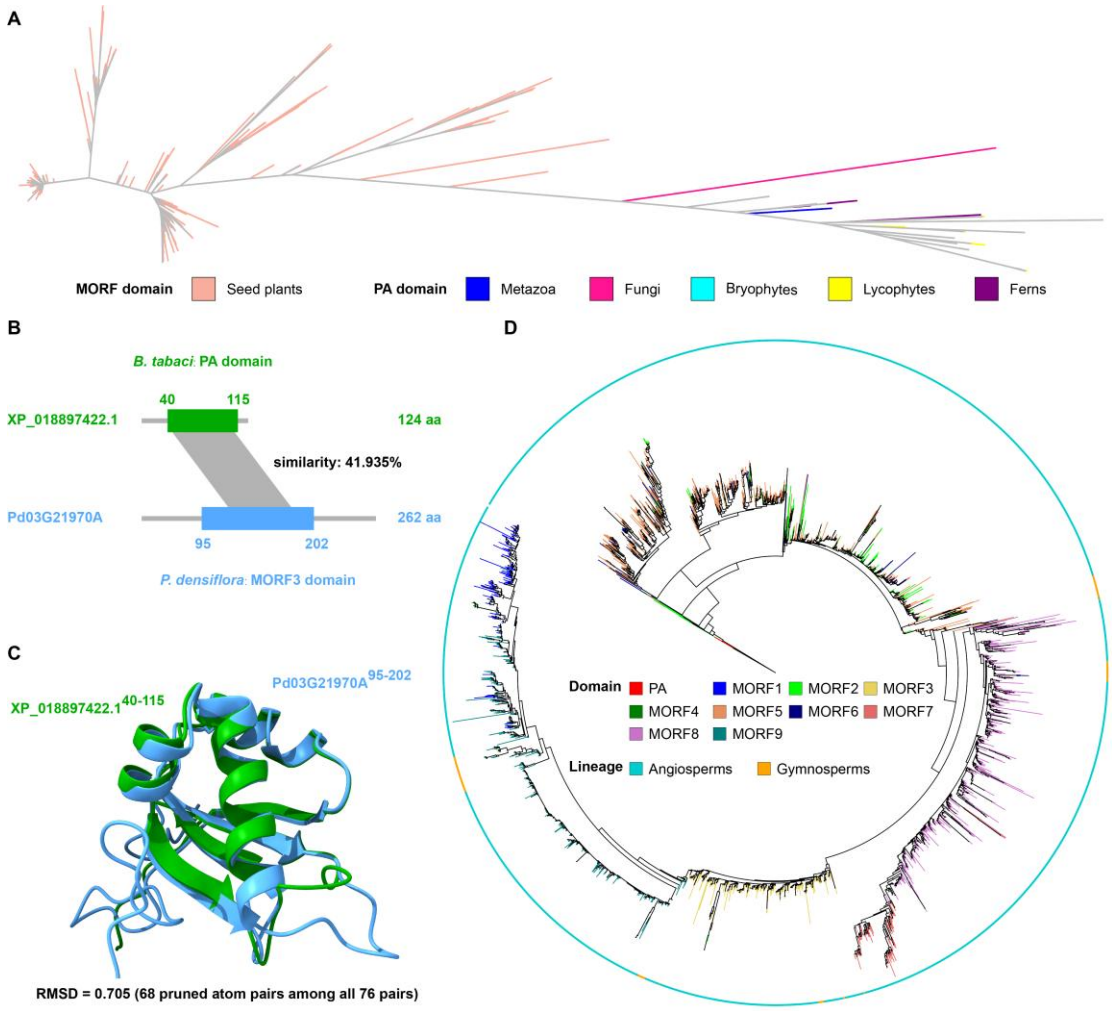


Figure 4. Comparison, expression, and HGT of DYW domains and their

1017 **non-plant homologs. (A)** Sequence alignments between DYW domains and
1018 their non-plant homologs. Amino acid residues in rectangles denote PG box,
1019 HxE(x)nCxxC, and DYW tripeptide. **(B)** Protein structure alignment between
1020 DYW domain (protein ID: Tle1c03524.1) in *T. lepidozioides* and its homolog
1021 (protein ID: UJR34389.1) in the rotifer *A. vage*. **(C)** Expression of PPR genes
1022 in whole tissues of *A. vage*. TPM: Transcripts Per Million. **(D)** Expression of
1023 DYW subgroup PPR genes in *T. lepidozioides* under control condition, and after
1024 2-hour UV irradiation treatment with 0, 6, 12, and 24-hour recovery. **(E)**
1025 Proposed model of DYW subgroup PPR gene transfer from mosses (e.g. *T.*
1026 *lepidozioides*) to rotifers (e.g. *A. vage*).
1027

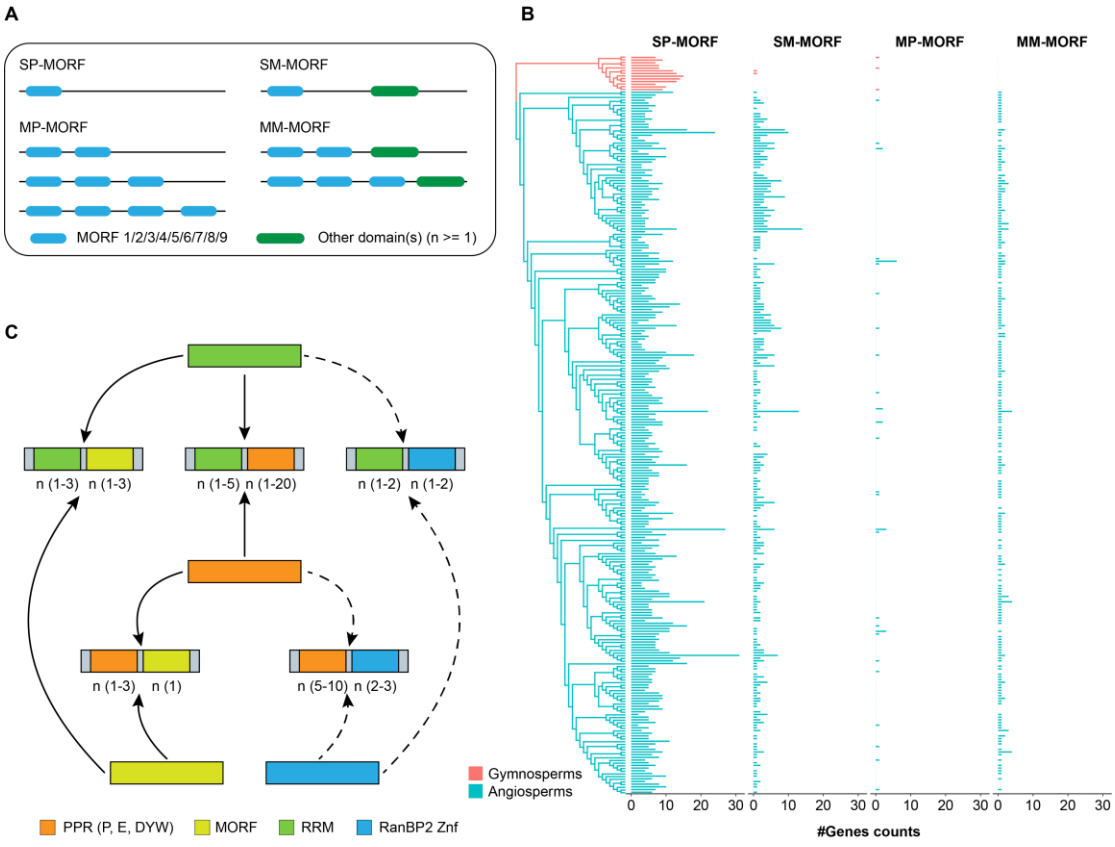
1028 **Figure 5**



1029

1030 **Figure 5. Homology between MORF and PA domains and MORF domain**
1031 **phylogeny. (A)** Phylogenetic tree of MORF and PA domains. **(B)** Sequence
1032 and **(C)** structure alignments between MORF3 domain (protein ID:
1033 Pd03G21970A) of *Pinus densiflora* and its homolog PA domain (protein ID:
1034 XP_018897422.1) of *B. tabaci*. **(D)** Phylogenetic tree of MORF1~MORF9
1035 domains in angiosperms and gymnosperms. Five PA domains (protein IDs:
1036 Dicom.22G070000.1.p, HaB36G042, Phcar.S3G189600.t1,
1037 Phsp.C2G192800.t1, and XP_024533311.1) showing highest similarity to
1038 MORF domains are used as outgroups.
1039

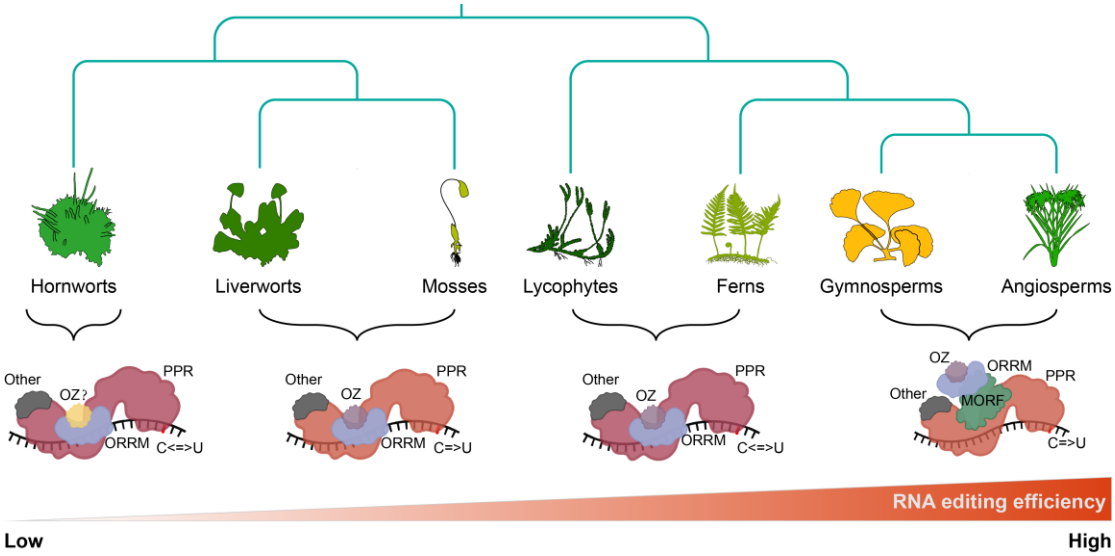
1040 **Figure 6**



1041

1042 **Figure 6. Reclassification and diversification of PPR accessory proteins.**
1043 **(A)** Reclassified MORF proteins based on domain compositions. Blue and
1044 green bars represent MORF1~MORF9 and other domains, respectively. SP-
1045 MORF: Single-Pure MORF, SM-MORF: Single-Mixed MORF; MP-MORF:
1046 Multiple-Pure MORF; MM-MORF: Multiple-Mixed MORF. **(B)** Phylogenetic tree
1047 of reclassified MORF proteins in angiosperms and gymnosperms. **(C)** Fusions
1048 among PPR, MORF, RRM, and RanBP2 Znf domains/motifs. Numbers in
1049 parentheses indicate domain counts. The proteins marked with solid and
1050 dashed arrows are localized to mitochondria/plastids and other subcellular
1051 compartments, respectively.
1052

1053 **Figure 7**



1054

1055 **Figure 7. Proposed model for the evolutionary diversification of plant RNA**

1056 **editosome.** Editosome composition varies from hornworts to angiosperms,

1057 driving increased RNA editing efficiency.