



Article

Comparative Genomic Analysis of Asian Cultivated Rice and Its Wild Progenitor (*Oryza rufipogon*) Has Revealed Evolutionary Innovation of the Pentatricopeptide Repeat Gene Family through Gene Duplication

Li-Ying Feng ^{1,†} , Pei-Fan Lin ^{1,†}, Rong-Jing Xu ², Hai-Qi Kang ² and Li-Zhi Gao ^{1,2,*}

¹ Institution of Genomics and Bioinformatics, South China Agricultural University, Guangzhou 510642, China; lyfengad@163.com (L.-Y.F.); lamevan0754@gmail.com (P.-F.L.)

² Tropical Biodiversity and Genomics Research Center, Hainan University, Haikou 570228, China; Xuuuurj@163.com (R.-J.X.); kanghq@126.com (H.-Q.K.)

* Correspondence: 995322@hainanu.edu.cn

† These authors contributed equally to this work.

Abstract: The pentatricopeptide repeat (PPR) gene family is one of the largest gene families in land plants. However, current knowledge about the evolution of the PPR gene family remains largely limited. In this study, we performed a comparative genomic analysis of the PPR gene family in *O. sativa* and its wild progenitor, *O. rufipogon*, and outlined a comprehensive landscape of gene duplications. Our findings suggest that the majority of PPR genes originated from dispersed duplications. Although segmental duplications have only expanded approximately 11.30% and 13.57% of the PPR gene families in the *O. sativa* and *O. rufipogon* genomes, we interestingly obtained evidence that segmental duplication promotes the structural diversity of PPR genes through incomplete gene duplications. In the *O. sativa* and *O. rufipogon* genomes, 10 (~33.33%) and 22 pairs of gene duplications (~45.83%) had non-PPR paralogous genes through incomplete gene duplication. Segmental duplications leading to incomplete gene duplications might result in the acquisition of domains, thus promoting functional innovation and structural diversification of PPR genes. This study offers a unique perspective on the evolution of PPR gene structures and underscores the potential role of segmental duplications in PPR gene structural diversity.

Keywords: rice; *Oryza*; pentatricopeptide repeat; innovation; evolution; segmental duplication



Citation: Feng, L.-Y.; Lin, P.-F.; Xu, R.-J.; Kang, H.-Q.; Gao, L.-Z. Comparative Genomic Analysis of Asian Cultivated Rice and Its Wild Progenitor (*Oryza rufipogon*) Has Revealed Evolutionary Innovation of the Pentatricopeptide Repeat Gene Family through Gene Duplication. *Int. J. Mol. Sci.* **2023**, *24*, 16313. <https://doi.org/10.3390/ijms242216313>

Academic Editor: Yong-Gu Cho

Received: 20 September 2023

Revised: 10 November 2023

Accepted: 12 November 2023

Published: 14 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Since the discovery of the pentatricopeptide repeat (PPR) protein motif 23 years ago [1], this protein family has played an immense role in plant organelle biology [2]. As one of the largest gene families reported in land plants [3], it is exceedingly varied in land plants with hundreds or even thousands of members [4], each of which seems to have definite targets in the transcriptomes of chloroplasts or mitochondria. PPR genes are characterized by the presence of pentatricopeptide repeat motifs, which are approximately 35 amino acid sequences that often form a helical structure [5–8]. These repeat motifs are responsible for protein–protein interactions and binding to target RNA molecules [9–12]. The PPR gene family is involved in various biological processes, including post-transcriptional regulation of gene expression [2,13,14], RNA editing [15,16], RNA splicing [17,18], and RNA stability [19]. In plants, PPR proteins play critical roles in organelle gene expression, particularly in chloroplasts and mitochondria [16,17,20].

PPR proteins are often classified into two major subfamilies, P and PLS, based on the characteristic properties of their PPR motifs [21,22]. The P subfamily is composed of multiple P motifs, the involvement of P-class PPR proteins in RNA stability, processing, and mRNA translation initiation [18,23,24], whereas the PLS subfamily is unique to land

plants and typically encompasses P motifs along with two derived variants, namely S motifs (short) and L motifs (long). Moreover, the majority of PLS subfamily members are characterized by the presence of additional C-terminal domains, including E, E+, and DYW domains [25]. Meanwhile, the primary function assigned to PLS-class PPR proteins is the C-to-U editing of organellar transcripts [26].

Heterosis, widely applied in crops, aims to enhance yield potential and improve quality by harnessing the phenomenon of hybrid vigor [27,28]. The male sterility of plants is induced by the Cytoplasmic Male Sterility (CMS) genes in the mitochondria, whereas fertility restoration is achieved through the interaction of *RF* genes in the nucleus. Thus, the CMS genes in the mitochondria result in male sterility, while the *RF* genes in the nucleus act reciprocally to restore plant fertility [29–31]. The majority of known restorer genes belong to the P-class of the PPR protein family [32,33].

The first comparative genomic analyses of *Arabidopsis thaliana* and *Oryza sativa* and *Physcomitrella patens* revealed the expansion of the PPR gene family, becoming a huge family in the two flowering plants (450 members in *A. thaliana* and 477 in rice) in sharp contrast to 103 PPR genes in the moss [7,34,35]. The availability of an increasing number of plant genomes permits one to recognize the PPR gene family to be one of the largest families in angiosperms. The lowest number of PPR genes is reported in *Medicago truncatula* (365), while the highest number is found in *Glycine max* (629). The partially assembled *Selaginella moellendorffii* (spikemoss, lycophyte) genome is suggested to contain >1000 PPR genes [36]. It is clear that a wide range of angiosperms have a large number of PPR genes whilst the moss *Physcomitrella patens* [37] contains 103 [34,35], and the green alga *Chlamydomonas reinhardtii* has only 12 PPR genes [38]. These results suggest that the initial expansion of PPR genes occurred after the separation of the land plants from green algae such as *Chlamydomonas* but before the divergence of seed plants from bryophytes [36]. Recent comprehensive comparative transcriptomic analysis of OneKP datasets including over 1000 transcriptomes from diverse plants and algae has established a clear picture of the evolution of this massive gene family [4]. They interestingly found that, across the plant kingdom, rather than a single expansion, most land plant lineages with high numbers of editing factors have continued to generate novel sequence diversity.

The PPR gene family provides an unprecedented opportunity to understand the evolutionary nature of driving a 50-fold expansion of such a single-gene family. Up to 80% of the PPR genes in *A. thaliana* and rice were intron-less [7,34,39], while >75% of *P. patens* PPR genes contained introns. Evolutionarily older PPR genes in rice and *A. thaliana* had more introns [34] and some of the putative orthologues to the *P. patens* intron-containing genes lack introns in angiosperms [34]. Such results offer compelling evidence that one or more waves of retrotransposition were responsible for the expansion of the PPR gene family in flowering plants [34]. Thus, the duplication of genomic sequences stands as one of the fundamental mechanisms for generating novel genes [40–42]. Gene duplication, which mainly originates from unequal crossing over, retrotransposition, and segmental duplication, has been reported to play a critical role in gene family expansion [36,43,44]. Gene families can also increase in size by whole-genome duplication events. Angiosperm genomes have undergone multiple rounds of whole-genome duplication events [45,46], which have a great impact on the number of large gene families. It was suggested that PPR gene content seemed quite conserved between *A. thaliana* and rice [34] and that the PPR genes duplicated by independent whole-genome duplication events were rarely retained [36]. However, current knowledge about the evolution of the PPR gene family remains largely limited, particularly in the context of the role of gene duplications.

It deserves mentioning that segmental duplications (SDs) are large (≥ 1 kb), nearly identical copies of genomic DNA sequences that map to two or more genomic locations [47–49]. These duplications can encompass various genomic elements, including gene sequences and common repetitive elements, and can be organized either in tandem or dispersed throughout the genome [50]. In plants, the impacts of segmental duplications include genome size expansion, nuclear DNA structure modeling, and adaptive evolution [51–53].

Many studies have explored how SDs contribute to the emergence of new genes especially in the evolution of primate genomes [54–56]. Not only the duplication of an entire gene but also the exon shuffling and fusion transcripts that result from juxtaposing unrelated SDs appear to be the mechanism by which SDs give rise to novel genes [50,57]. In addition, due to the high sequence identity, duplication serves as the foundation for subsequent rearrangements facilitated by non-allelic homologous recombination. Thus, it is expected to aid in gaining novel insights into the structural complexity of *PPR* genes, which will provide more evidence for us to further understand the evolution of this gene family.

In this study, we performed genome-wide identification and characterization of the *PPR* gene family in *O. sativa* and its wild progenitor, *O. rufipogon*. In addition, we presented a comparative genomic analysis of these two rice genomes and outlined a comprehensive landscape of gene duplications towards understanding the molecular evolutionary mechanisms of *PPR* genes. We finally obtained a full map of segmental duplications that greatly enhance structural variation and the innovation of *PPR* genes through incomplete gene duplications. The findings will also form a solid foundation for future functional studies on *PPR* genes such as *RFL* genes of agronomical importance for rice genetic improvement programs.

2. Results

2.1. Identification and Classification of *PPR* Genes in the *O. sativa* and *O. rufipogon* Genomes

In this study, a total of 460 and 479 *PPR* genes were identified in *O. sativa* and *O. rufipogon*, respectively (Figure 1a). Based on the protein structure, 460 *PPR* proteins consisted of 223 P-class and 237 PLS-class subfamilies in *O. sativa*. There were 148 members in the E subgroup and 73 members in the DYW subgroup within the PLS subfamily. In *O. rufipogon*, the 479 *PPR* proteins comprise 225 P-class and 254 PLS-class members. In the PLS subfamily, there were 160 members in the E subgroup and 80 members in the DYW subgroup. Information regarding the physical lengths, classification, and protein lengths of these genes can be found in Tables S1 and S2.

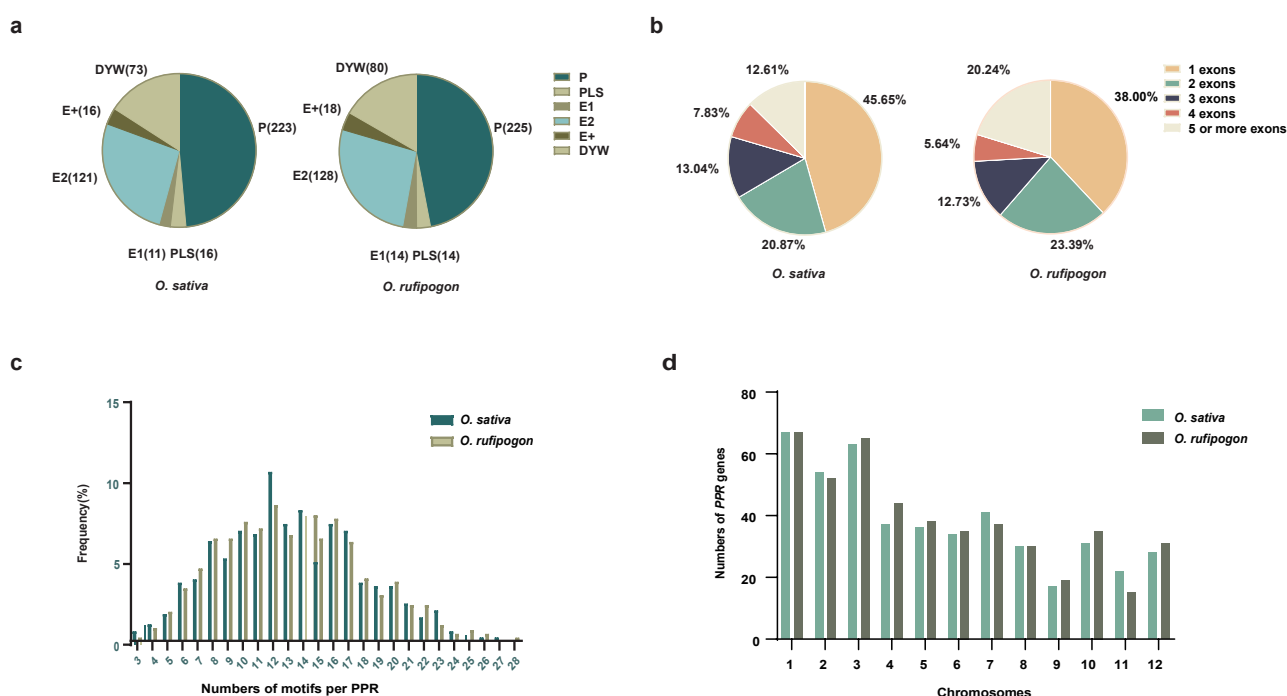


Figure 1. Classification of the *PPR* genes in *O. sativa* and *O. rufipogon*. (a) *PPR* protein numbers in *O. sativa* and *O. rufipogon*; (b) the distribution of exon numbers in two rice species; (c) the motif quantity of *PPR* genes in two rice species; (d) the distribution and quantity of *PPR* genes across 12 chromosomes in two rice species.

Previous findings suggested that *PPR* genes are predominantly composed of a single exon, accounting for the largest proportion [2,8]. The gene structure analysis indicated that in *O. sativa*, 45.65% of the genes had a single exon, while 20.87%, 13.04%, 7.83%, and 12.61% of the genes had two, three, four, and five or more exons, respectively (Figure 1b; Table S1). Our results showed that in *O. rufipogon*, 38.00% of the genes had a single exon, and 23.39%, 12.73%, 5.64%, and 20.25% of the genes had two, three, four, and five or more exons, respectively (Figure 1b; Table S2). Overall, the PLS subfamily gene is more abundant than the P subfamily gene in terms of individual exons of the two species. In *O. sativa*, there were 128 PLS subfamily genes and 82 P subfamily genes, while *O. rufipogon* harbored 106 PLS subfamily genes and 76 P subfamily genes (Figure S1a). The characteristic feature of PPR proteins was the tandem array of PPR motifs. For both rice species, the number of PPR motifs per protein ranged from 3 to 28, with the highest number of proteins having 12 motifs (Figure 1c; Tables S1 and S2).

It is noteworthy that the number of *PPR* genes with a single exon in *O. rufipogon* displayed a significant decrease compared to cultivated rice, while the number of *PPR* genes with five or more exons exhibited a significant increase. There were no significant changes observed in the quantities of PPR proteins with other numbers of exons. The results indicated that *PPR* genes in these two species were extensively and unevenly distributed across the 12 chromosomes (Figure 1d; Table S1). Both rice species exhibited the highest number of *PPR* genes on Chromosome 1; cultivated rice exhibited the fewest number on Chromosome 9, while *O. rufipogon* had the fewest number on Chromosome 11 (Figure 1d; Table S2). Our results showed that the P-subclass and PLS-subclass of the *PPR* genes were widely but unevenly positioned across 12 chromosomes (Figure 2). However, *O. sativa* and *O. rufipogon* exhibited similar distribution of *PPR* genes. Compared to eleven other chromosomes, they were densely distributed on Chromosome 10 for both rice species.

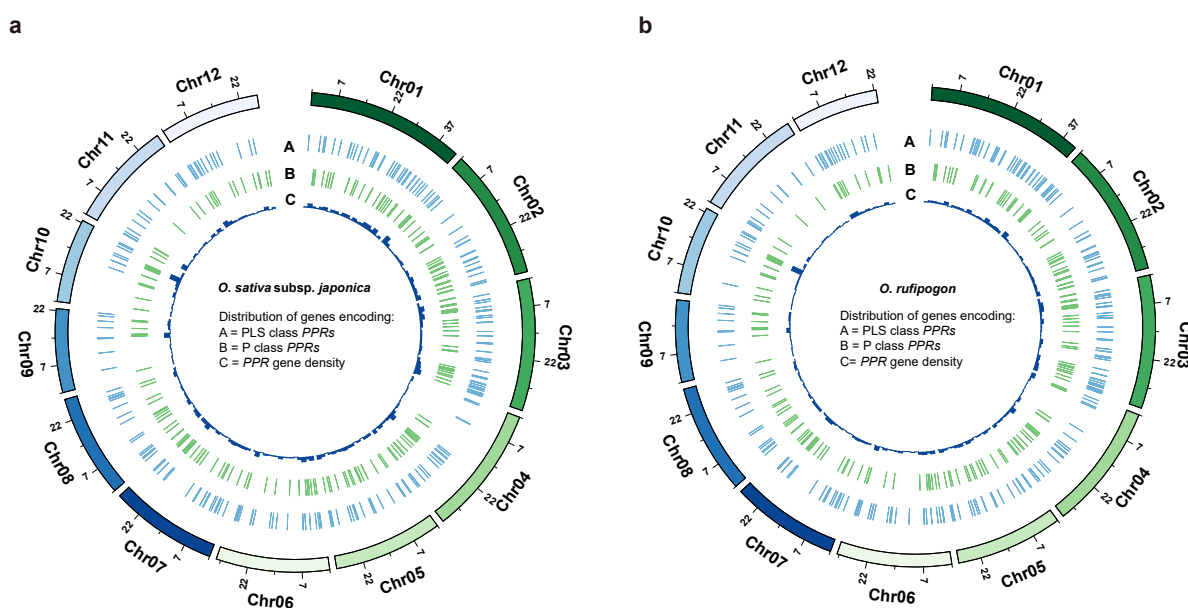


Figure 2. Genome-wide distribution of *PPR* genes. The density of *PPR* genes is presented along each chromosome in 2 Mb windows. (a) *O. sativa*; (b) *O. rufipogon*.

To investigate the regulatory elements of *PPR* genes, we predicted cis-regulatory elements for the promoter sequences of *PPR* genes in *O. sativa* and *O. rufipogon* (Figures S2 and S3). Our results showed that, for both species, the cis-regulatory elements on the promoters of *PPR* genes mainly included hormone response elements, growth and development elements, and stress response elements. Among the cis-acting elements associated with hormone response, both species had the highest number. These elements primarily included abscisic

acid response elements (ABRE), the CAAT-box transcription factor, and the core promoter element (TATA-box).

2.2. Phylogenetic Analysis of PPR Genes in the *O. sativa* and *O. rufipogon* Genomes

Based on the full amino acid sequences of PPR genes in the two species, a phylogenetic tree was constructed using the Maximum Likelihood (ML) method. The phylogenetic tree was divided into two distinct groups, known as the P and PLS subfamilies (Figure 3a,b). A small subset of P-class genes clustered within the PLS class subfamily, including genes such as *LOC_Os09g26190*, *RUF029470*, *LOC_Os07g46730*, and *RUF025381*. PLS-class genes also clustered within the P-class subfamily, including *RUF012986*, *LOC_Os03g62400*, *RUF002433*, and *LOC_Os01g42280*. Such an observation is similar to the previous phylogenetic analysis of the PPR gene family in poplar [5].

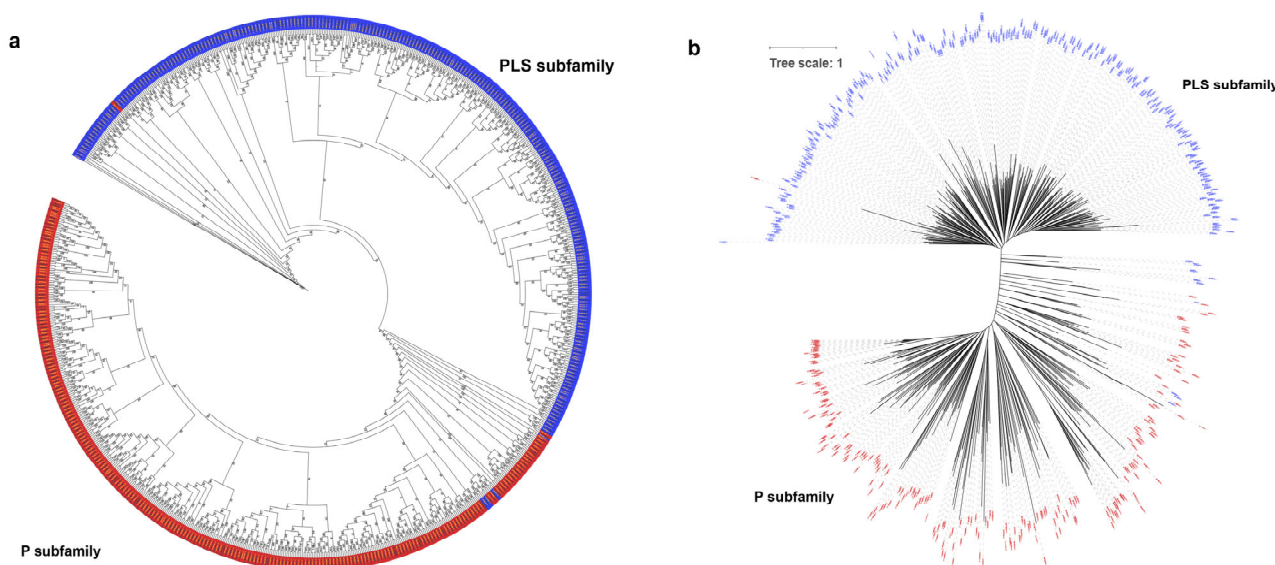


Figure 3. Phylogenetic trees of the PPR gene family in *O. sativa* and *O. rufipogon*. The maximum likelihood trees were constructed using IQ-TREE V2.2.3 software with 1000 bootstrap replicates. Blue represents the PLS subfamily, while red symbolizes the P subfamily. (a) Phylogenetic tree of PPR genes of two species with bootstrap values. The black font represents *O. sativa* and the gold font denotes *O. rufipogon*; (b) phylogenetic tree of PPR genes of two rice species with branch length. Before constructing the phylogenetic tree, the LG+F+R10 model was determined as the best model.

2.3. Comparative Genomic Analysis of PPR Gene Families between *O. sativa* and *O. rufipogon*

To investigate the origins and evolution of PPR gene families after the domestication of *O. sativa* from its wild progenitor, we employed a stringent ortholog identification strategy, including specifically reciprocal best-hit and synteny comparisons, to identify shared and specific PPR genes between the *O. sativa* and *O. rufipogon* genomes. Our results showed that a total of 390 gene pairs met both criteria. We also identified 10 gene pairs that were located on scaffolds not anchored to any chromosome, and 13 gene pairs were outside syntenic regions (Figure S4). Hence, approximately 84.78% to 89.78% of PPR genes were shared between the *O. sativa* and *O. rufipogon* genomes (Figure 4a,b).

The analysis of species-specific PPR genes showed that *O. sativa* possessed approximately 70 PPR genes, while *O. rufipogon* had 89 (Figure 4b). Statistical analysis was performed on the exons of shared and unique PPR genes in the two species. We observed that the largest proportion among the 390 shared genes was the 180 genes with only one exon. Among the 70 unique PPR genes in *O. sativa*, the highest proportion of 30 genes contained one exon. Interestingly, out of 89 unique PPR genes, 34 had 5 exons with the highest proportion in *O. rufipogon*, surpassing the number of 25 genes with only 1 exon (Figure S1b).

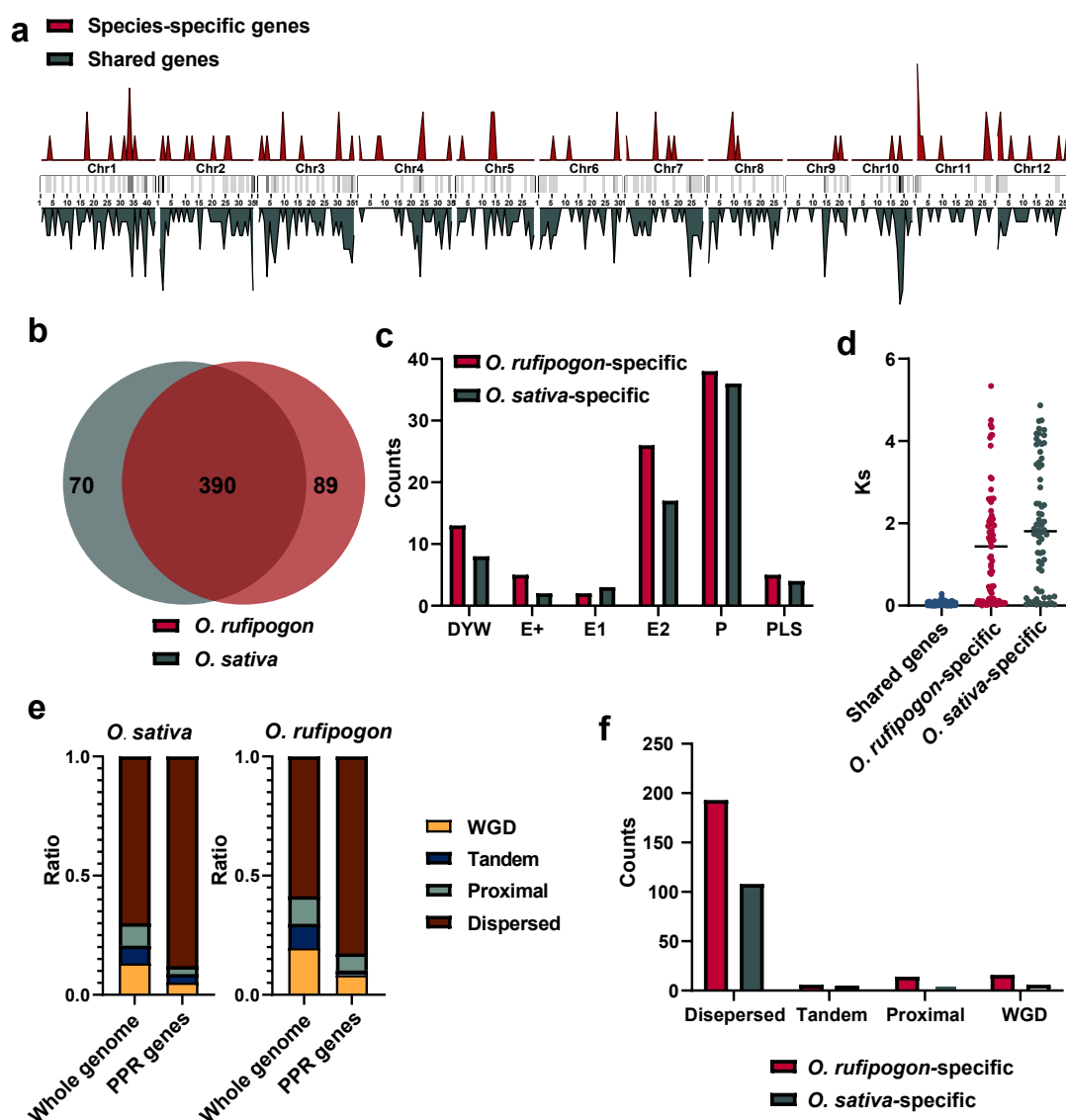


Figure 4. Shared and species-specific genes between *O. sativa* and *O. rufipogon*. (a) Chromosomal distribution of shared and species-specific PPR genes in the *O. sativa* and *O. rufipogon* genomes. The intensity of chromosome band color indicates the density of PPR genes. Red represents species-specific genes, while dark green denotes shared genes; (b) number of shared and species-specific PPR genes between *O. sativa* and *O. rufipogon*; (c) PPR subfamilies of specific genes in *O. sativa* and *O. rufipogon*; (d) Ks values of shared gene pairs and the smallest Ks values of paralogous genes for specific genes within their own genomes; (e) composition of gene duplication types for the whole genome and PPR genes in *O. sativa* and *O. rufipogon*; (f) composition of gene duplication types of species-specific genes in *O. sativa* and *O. rufipogon*.

To investigate the nature of species-specific PPR genes, we classified them into different subfamilies. Our results showed that the P subfamily harbored the most prominent different genes with 36 and 38 genes in *O. sativa* and *O. rufipogon*, respectively, followed by 17 and 26 genes in the E2 subgroup (Figure 4c). Species-specific genes may arise due to the quality of genome assembly and annotation, or potentially from the loss of one copy of old duplicated genes or newly duplicated genes. After excluding the 10 gene pairs on unanchored sequences and 13 gene pairs outside syntenic regions, we calculated the Ks values of these species-specific genes with all paralogous genes in their genomes and selected the minimum Ks values. By comparing them with the Ks values obtained from each pair of orthologous genes, their origins were determined to have resulted from the

loss of one copy of old duplicated genes or the emergence of newly duplicated genes. Our data showed that in *O. sativa*, 17 species-specific genes had *Ks* values lower than the average *Ks* value between orthologous genes. In *O. rufipogon*, 29 species-specific genes had *Ks* values below the average. Thus, it may be estimated that ~36.17% and 42.94% of *PPR* genes originated from new gene duplications in the *O. sativa* and *O. rufipogon* genomes, respectively (Figure 4d). Of them, the wild progenitor, *O. rufipogon*, had more newly duplicated genes than the cultivated rice.

We further examined which types of gene duplications have generated *PPR* genes in *O. sativa* and *O. rufipogon*. Our results showed that, in the *O. sativa* genome, approximately 5.10% of *PPR* genes were derived from whole-genome duplication, only approximately 3.60% of them were from tandem duplications, and around 3.40% of them resulted from proximal duplications, while the majority (~87.99%) of the *PPR* genes were from dispersed duplications. *PPR* genes were remarkably enriched in the dispersed duplication category (~87.99%), which was ~1.26 times higher than the genome-wide average (~69.99%). Similarly, *PPR* genes were outstandingly augmented in the category of dispersed duplication (~82.64%), which was ~1.40 times greater than the genome-wide average (~58.63%) in the *O. rufipogon* genome (Figure 4e; Table S3). Consequently, our findings strongly suggested that the majority of *PPR* genes originated from dispersed duplications for both rice species (Figure 4f).

We further conducted Gene Ontology (GO) enrichment analysis of *PPR* genes uniquely identified in *O. sativa* and *O. rufipogon* (Figure 5). Interestingly, *PPR* genes in *O. sativa* exhibited an enrichment of a unique biological process related to the termination of embryo development and seed dormancy, indicating that the presence of specific *PPR* genes may be functionally related to embryo development and reduced seed dormancy during the domestication from wild rice to cultivated rice. The genes involved in this process included *LOC_Os11g03850*, *LOC_Os12g42120*, *LOC_Os03g19650*, *LOC_Os11g43800*, *LOC_Os06g09880*, *LOC_Os02g43080*, and *LOC_Os05g30240* (Figure 5a). It is worth noting that *PPR* genes in *O. rufipogon* displayed enrichment of the mitochondria, as a cellular component. These genes contained *RUF036508*, *RUF037041*, *RUF037004*, *RUF006046*, *RUF002492*, *RUF004777*, and *RUF030818* (Figure 5b). GO enrichment analysis of a total of 390 shared *PPR* genes (Figure 5c) revealed that, compared to unique *PPR* genes, there were two additional components in the cellular component category (intracellular anatomical structure and obsolete cell).

2.4. Expression Patterns of *PPR* Genes in *O. sativa* and *O. rufipogon*

We examined the differential expression patterns of these *PPR* genes in the two rice species. To ensure consistency in transcriptome data, we obtained 24 samples across four tissues, including the root, leaf, seedling, and panicle of both rice species during the same growth phase (Figure S5). By mapping the quality-controlled RNA-seq data to their respective genomes, we garnered a comprehensive overview of *PPR* gene expression. Using TPM as a metric, we stratified gene expression into five levels (very high, $\text{TPM} \geq 50$; high, $10 \leq \text{TPM} < 50$; moderate, $2 \leq \text{TPM} < 10$; low, $0.1 \leq \text{TPM} < 2$; and very low, $0 < \text{TPM} < 0.1$). Our findings indicate that the majority of *PPR* genes in the seedling and root tissues exhibited an expression level exceeding the moderate range (*O. sativa*: ~65.87% in seedling, ~66.96% in root; *O. rufipogon*: ~58.87% in seedling, ~69.51% in root), while only ~10.65% and ~26.93% of *PPR* genes exhibited an expression level exceeding the moderate range in the leaf tissue of *O. sativa* and *O. rufipogon* (Figure S6). A pronounced disparity in *PPR* gene expression surfaced in the panicle tissue, particularly in *O. rufipogon*, which exhibited elevated expression levels (Figure S6).

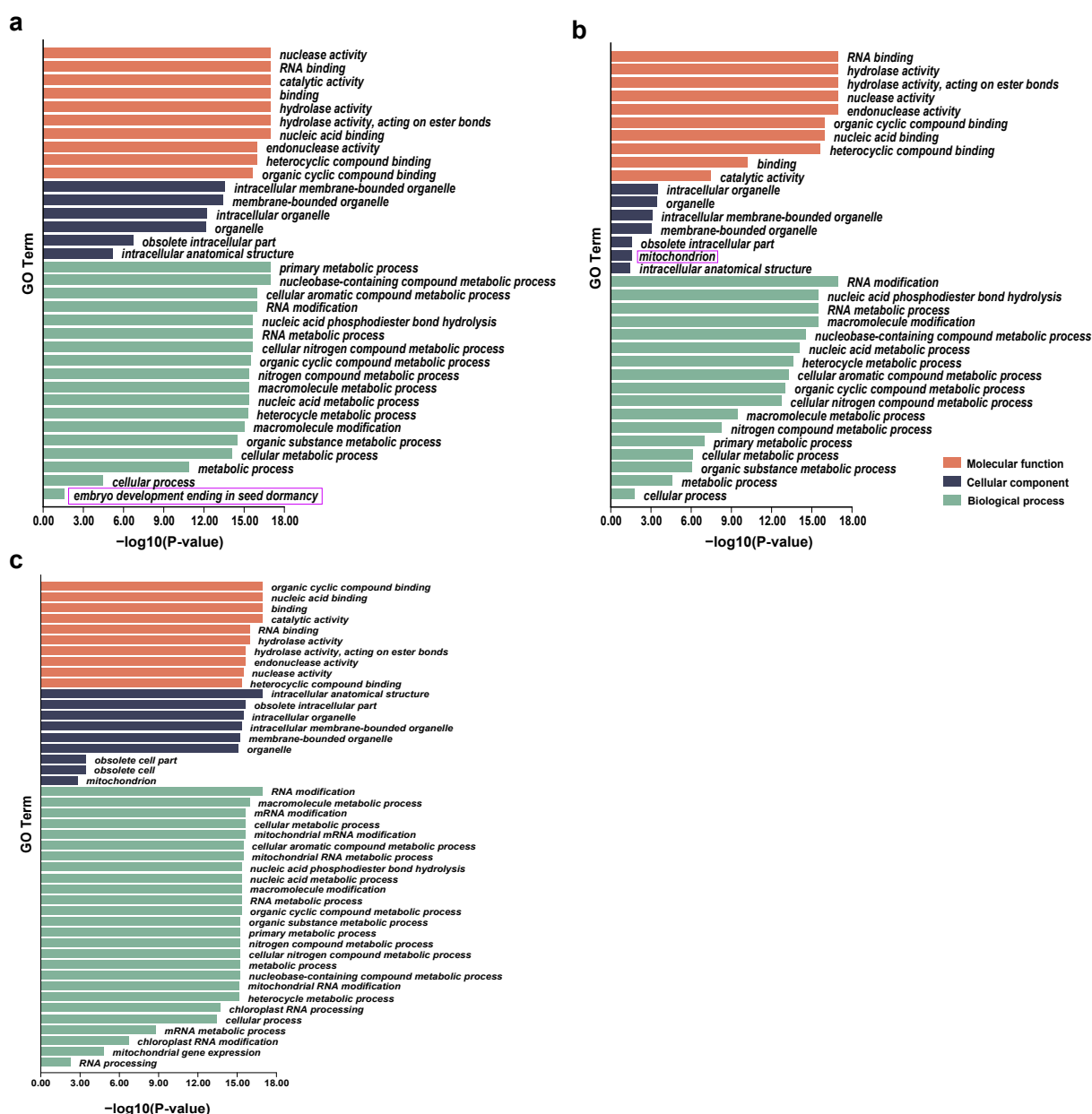


Figure 5. GO enrichment analyses for PPR genes. (a) *O. sativa*; (b) *O. rufipogon*; (c) shared PPR genes. In the purple box are the pathways specifically enriched by lineage-specific PPR genes in two species.

To measure the differential expression of PPR genes between the two rice species, we further mapped the RNA-Seq data to the MSU7 genome and identified differentially expressed genes (DEGs). We observed the different expression patterns of PPR genes in the panicle tissue in the heatmap of the expression matrix (Figure 6a). Based on the criteria of Fold Change > 2 and P-adj < 0.05, we identified 28 DEGs in the root tissue, 56 DEGs in the leaf tissue, 24 DEGs in the seedling, and 122 DEGs in the panicle (Figure 6b). Notably, within the panicle DEGs, the fertility restoration gene *Rf1* (*LOC_Os10g35436*) with the highest fold change, and another fertility restoration gene, *LOC_Os10g352140*, were also upregulated in *O. rufipogon*. We further categorized the DEGs into genes upregulated in *O. sativa* and upregulated in *O. rufipogon*. We found that, in the root and seedling tissues, *O. sativa* had more upregulated genes, with 17 and 13 genes showing increased expression in the root and seedling, respectively (Figure 6c). We subsequently carried out functional annotation and GO enrichment analysis on genes upregulated in both *O. sativa* and *O. rufipogon* across various tissues. We discovered that genes upregulated in the seedling tissue of

O. sativa were enriched in the biological process term “embryo development ending in seed dormancy”. However, all these genes were shared genes between *O. sativa* and *O. rufipogon*. In the leaf tissue of *O. rufipogon*, upregulated *PPR* genes were enriched in the “defense response to nematode” biological process term. In the seedling tissue, upregulated *PPR* genes were enriched in the “seed germination” biological process term (Figure 6d).

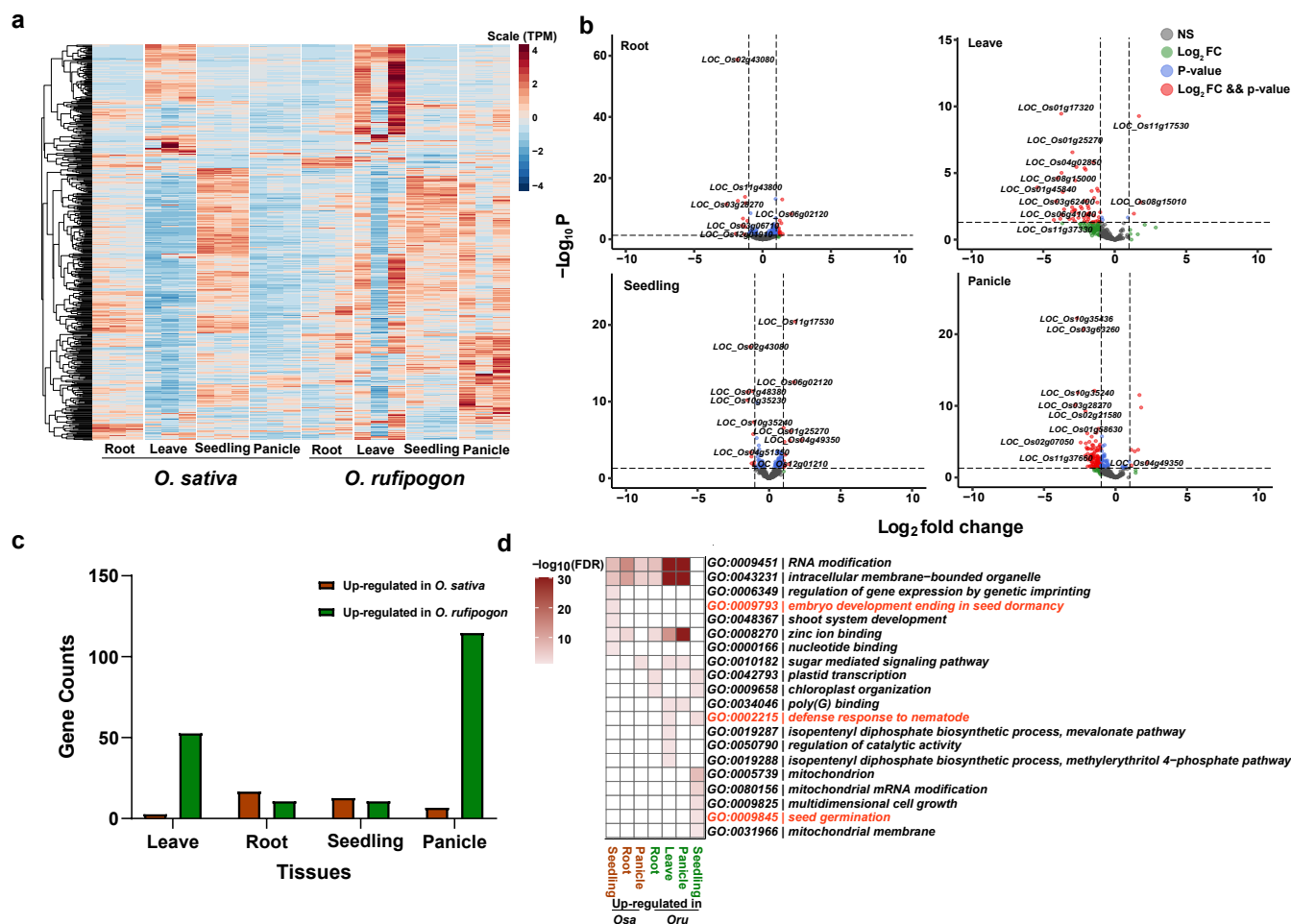


Figure 6. Expression patterns and DEGs of *PPR* genes in *O. sativa* and *O. rufipogon*. (a) Heatmap of *PPR* gene expression in four tissues of *O. sativa* and *O. rufipogon*. The four tissues include leaves, roots, seedlings, and panicles, each with three biological replicates. Heatmap was plotted based on normalized TPM values. (b) Volcano plots showing differentially expressed genes across tissues, based on a $\text{Log}_2\text{FoldChange} > 1$ and a $p\text{-value} < 0.05$. (c) The number of differentially expressed genes. Brown bars represent up-regulated genes in *O. sativa* while green bars signify *O. rufipogon* up-regulated genes (d) GO enrichment analyses of DEGs in various tissues.

2.5. Segmental Duplication Promotes the Structural Diversity of *PPR* Genes through Incomplete Gene Duplications

We further identified segmental duplications in the *O. sativa* and *O. rufipogon* genomes using WGAC (Whole Genome Assembly Comparison) and obtained the genome-wide landscape and impact of segmental duplications on *PPR* genes (Figure 7a). Our results suggested that, in the *O. sativa* and *O. rufipogon* genomes, 52 and 65 *PPR* genes, respectively, resulted from segmental duplications. This accounted for the expansion of approximately 11.30% and 13.57% of the *PPR* gene families (Figure 7c; Table S4). Among the lineage-specific *PPR* genes, 20 and 32 genes in the two genomes, respectively, were affected by segmental duplications. We observed that in the *O. sativa* and *O. rufipogon* genomes, identities of segmental duplications mainly ranged from ~94% to 96% (Figure 7b). Additionally, we observed that two pairs and four pairs of segmental duplications with an identity

exceeding 99%, representing the youngest segmental duplications, appeared in *O. sativa* and *O. rufipogon*, respectively.

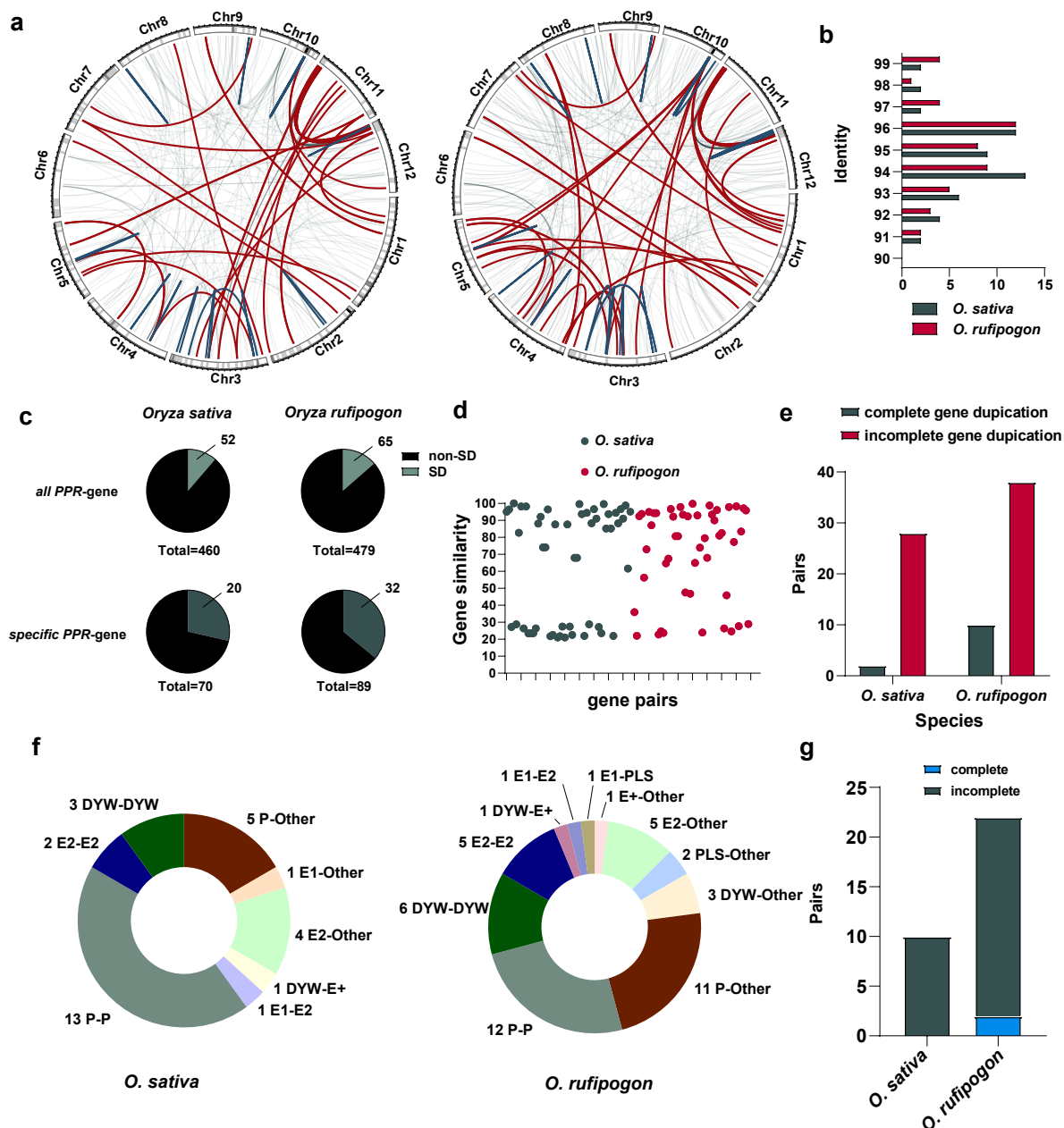


Figure 7. Segmental duplications facilitate incomplete gene duplications and promote the birth of new PPR genes. (a) PPR genes arising from segmental duplications in the *O. sativa* and *O. rufipogon* genomes. Left panel: *O. sativa*; right panel: *O. rufipogon*. Red lines represent inter-chromosomal duplications, dark green lines indicate intra-chromosomal duplications, and light gray lines denote other types of segmental duplications. 1 Mb sliding window with the shading indicates the density of PPR genes on each chromosome; (b) distribution of identities of segmental duplications. Dark green represents *O. sativa*, while red indicates *O. rufipogon*; (c) proportion of PPR genes derived from segmental duplications among all PPR genes and species-specific PPR genes; (d) protein similarity between duplicated genes arising from segmental duplications. Each dot represents the similarity between a pair of duplicated genes; (e) pairs of complete and incomplete gene duplications through segmental duplications; (f) numbers and proportions of duplicated genes across subfamilies with "Other" denoting non-PPR paralogous genes; (g) number of duplications between PPR genes and non-PPR genes in *O. sativa* and *O. rufipogon*.

The results showed that the similarity between protein products of homologous genes arising from segmental duplications was relatively lower. Specifically, only 19 and 20 gene pairs with sequence similarities were above 90%, accounting for approximately ~36.54% and ~41.6% of the duplicated gene pairs, respectively. The average similarity among all duplicated gene pairs was ~64.39% for *O. sativa* and ~71.34% for *O. rufipogon* (Figure 7d; Tables S5 and S6). Interestingly, our results suggested that, among the 30 and 48 pairs of segmental duplications in *O. sativa* and *O. rufipogon*, 28 and 38 pairs of segmental duplications, respectively, represented incomplete gene duplications. Only 2 and 10 pairs of segmental duplications were complete gene duplications (Figure 7e; Tables S5 and S6).

We further investigated paralogues of subfamilies of the identified segmental duplications. Our results indicated that 18 and 23 pairs of segmental duplications in *O. sativa* and *O. rufipogon*, respectively, had paralogous copies belonging to the same subfamily. In *O. sativa*, there were 13 pairs of P-P, 3 pairs of DYW-DYW, and 2 pairs of E2-E2, with P-P type gene duplications accounting for ~43.33% of all segmental duplications (Figure 7f). In *O. rufipogon*, there were 12 pairs of P-P, 6 pairs of DYW-DYW, and 5 pairs of E2-E2, with P-P type gene duplications accounting for ~25% of all segmental duplications (Figure 7f). There were two and three gene pairs in *O. sativa* and *O. rufipogon*, respectively, that had paralogous copies belonging to other subfamilies, for example, one pair of E1-E2 duplication and one pair of DYW-E2 duplication in *O. sativa* (Figure 7f). Interestingly, apart from the above-mentioned cases in which both paralogues belonged to the PPR gene family itself, a number of gene duplications had a non-PPR paralogous gene. In the *O. sativa* and *O. rufipogon* genomes, 10 (~33.33%) and 22 pairs of gene duplications (~45.83%) had non-PPR paralogous genes (Figure 7f,g). Specifically, in *O. rufipogon*, there were 11 pairs of P-Other type duplications, 5 pairs of E2-Other type duplications, and 3 pairs of DYW-Other type duplications. Of these duplicated gene pairs with non-PPR paralogous genes, *O. sativa* had no complete gene duplications, while *O. rufipogon* only had two complete gene duplications and the others were mainly incomplete gene duplications (~90.9%).

We investigated expression patterns of duplicated genes produced by segmental duplications and further classified them into three categories. When one copy exhibits significantly higher expression levels in two tissues compared to its sister gene and expression in other tissues is not lower than its sister copy, we categorize it as asymmetrically expressed duplicates (AEDs). A gene pair is defined as potentially sub- or neo-functionalized if each copy of the duplicates is expressed at a significantly higher level than the other in at least one tested tissue. The remaining pairs are classified as having no difference. Of 20 pairs in *O. sativa* and 26 pairs in *O. rufipogon*, where both copies belong to the PPR gene family, we identified 14 and 16 pairs as AEDs, and 6 and 10 pairs, respectively, as having no significant expression differences across tissues (Figures S7a, S8a and S9). In this study, we failed to find evidence of any sub- or neo-functionalized gene pairs. Within the AEDs, most genes with relatively lower expression than their sister gene maintained at least the “low” level of expression in three or more tissues. The only exception was *LOC_Os11g03850* with a “very-low” level in three tissues, possibly representing a pseudogene or silenced gene, which is a truncated copy of *LOC_Os03g40020* (Figures S7b and S8b).

We further predicted domain structures of the protein products of these duplicated genes with non-PPR paralogous genes using SMART. Among these incomplete genes produced by segmental duplications, our results showed that, in addition to PPR motifs, *O. sativa* had a total of 7 genes with other domains, while *O. rufipogon* had 26 genes (Figures 8a and 9a). Previously reported examples of PPR genes with other domains included PPR-SMRs, which had eight members in the *Arabidopsis* genome that contained both PPR motifs and SMR domains [58]. In addition, domains such as proteinaceous RNase P (PRORP) and LAGLIDADG motifs were also present in different subclasses of PPR genes [3]. We examined the structural origins of these PPR genes with other domains, which were mainly generated by incomplete duplications through segmental duplications. Our results showed that these homologous sequences were not PPR motifs but rather other domains. For instance, *LOC_Os03g59264* had a Calreticulin domain, *LOC_Os07g10400*

contained a S_TKc domain, and *LOC_Os07g20500* featured a Mem_trans domain. We then performed comparative analyses with other *Oryza* genomes to determine the direction of gene duplications. Our results intriguingly suggested that members of the *PPR* gene family often act as the “accepter” of other genes’ domains (Figure 8b,c). We further investigated the expression of these genes, and our results showed that five of them had moderate levels of expression in at least one tissue while *LOC_Os12g01910* and *LOC_Os12g04110* had low levels of expression in three and four tissues (Figure 8d). Of a greater number of duplicated genes with other domains in *O. rufipogon* than *O. sativa*, we interestingly found that some members even had multiple different domains, such as *RUF025888*, which had the Gal-bind_lectin, Galactosyl_T, and p450 domains (Figure 9a). In *O. rufipogon*, 23 of them had moderate-level expression in at least one tissue while *RUF010974*, *RUF03544*, and *RUF035442* had low levels of expression in all four tissues (Figure 9b).

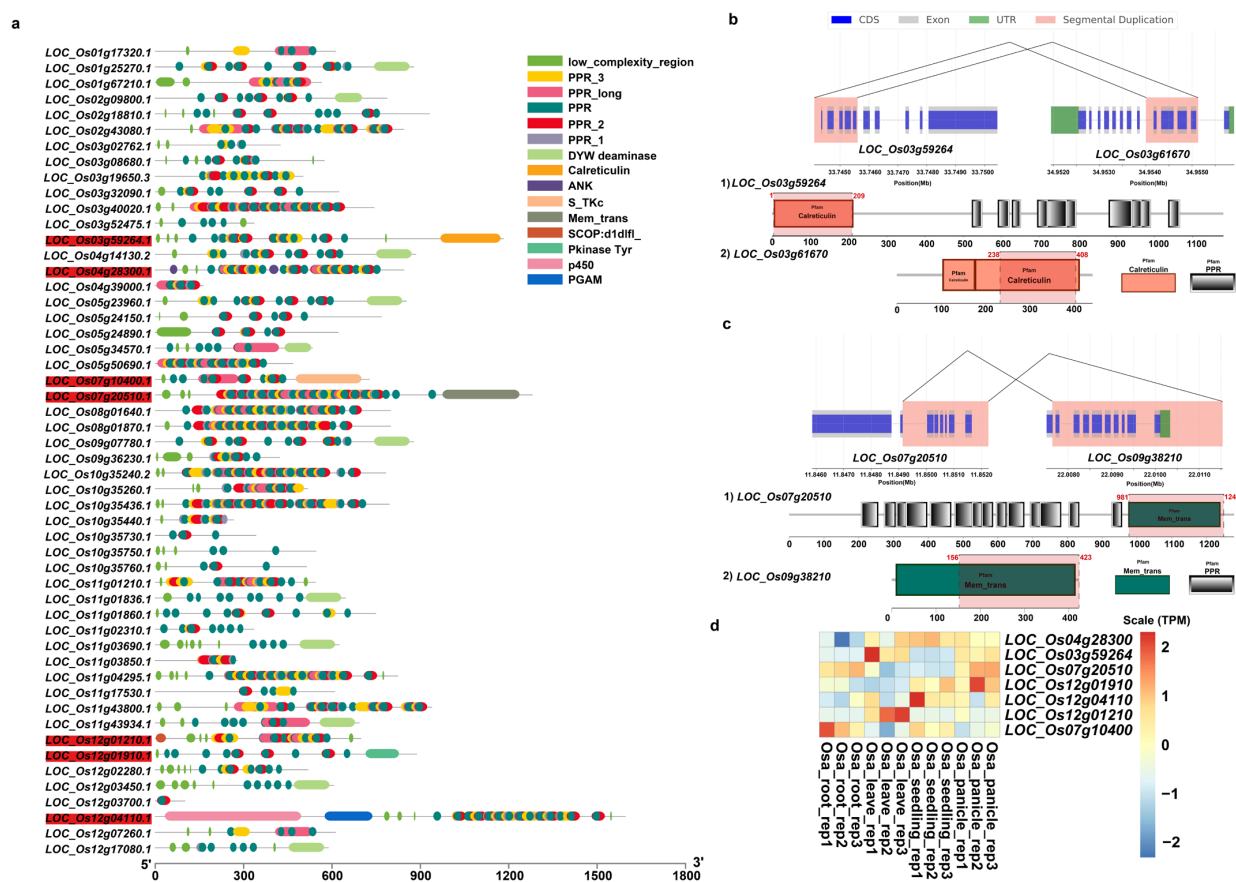


Figure 8. Domain composition of *PPR* genes produced by SDs in *O. sativa*. (a) Domain composition of 52 *PPR* genes generated by segmental duplications. Genes marked in red indicate that the protein product contained domains other than *PPR* domain; (b) incomplete gene duplications occurred in *PPR* genes that acquired domains from other proteins such as (b) *LOC_Os03g59264* and (c) *LOC_Os07g20510*. Pink represents homologous regions of segmental duplications. The structures of paralogous genes are depicted; coding sequences are represented in blue, exon sequences are represented in gray, and UTR regions are represented in green. The annotation of domain composition of the homologous genes is shown below; the gene structure with the homologous regions is marked in red and the location of homologous region is also labeled; (d) expression profiling of 7 *PPR* genes with other domains.

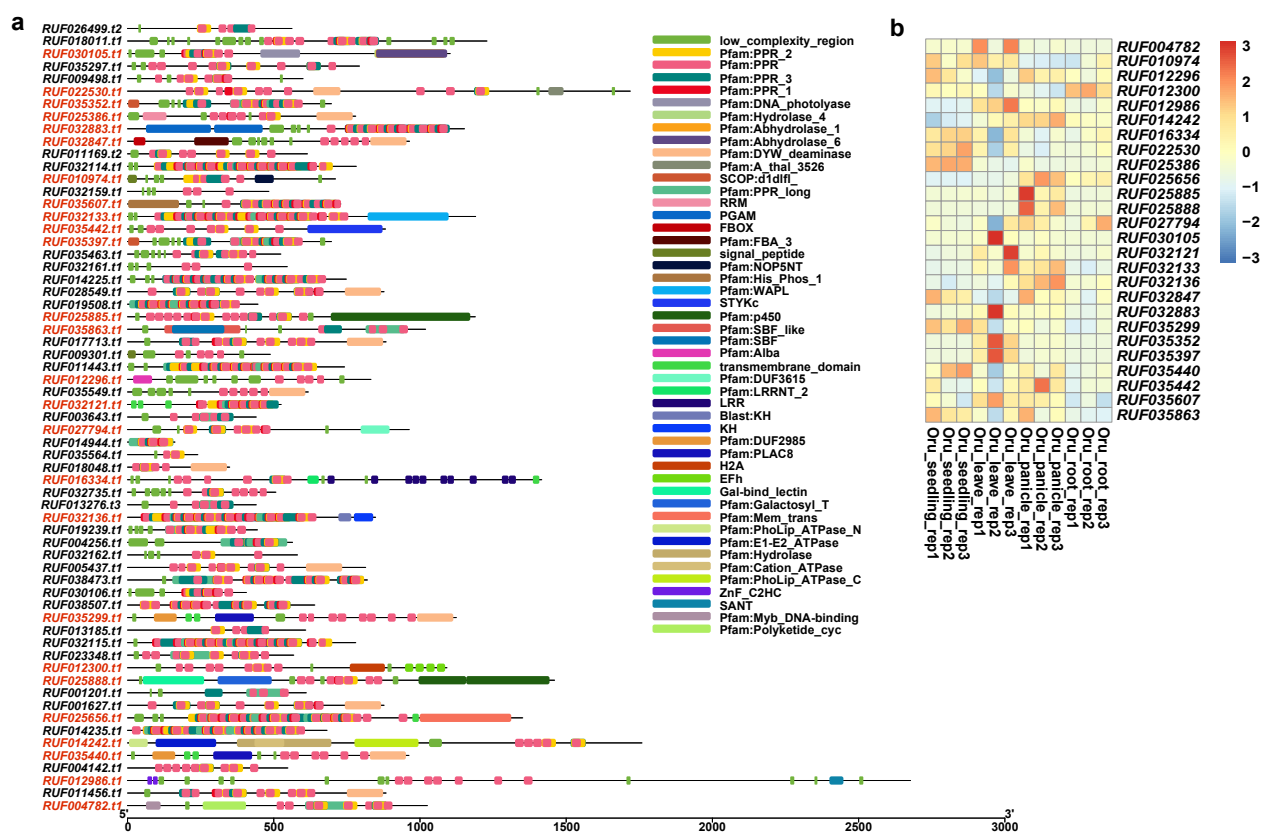


Figure 9. Domain composition of PPR genes produced by SDs in *O. rufipogon*. (a) Domain composition of 65 PPR genes duplicated by segmental duplications. Genes marked in red indicate that the protein product of the gene contains domains other than PPR domain; (b) expression of 26 PPR genes with other domains.

3. Discussion

PPR proteins are a protein family that widely exists in plants and some protists, and they regulate gene expression and RNA metabolism by binding to specific RNA sequences [1,25,59]. Recent decades have witnessed that the PPR gene families have been identified in multiple plant species, such as *O. sativa* [34,60], *A. thaliana* [7], *Populus trichocarpa* [5], *Citrullus lanatus* [61], and *Camellia sinensis* [62]. In this study, we performed whole-genome analysis and identified 460 and 479 PPR genes in *O. sativa* and *O. rufipogon*, respectively, which were then classified into six subgroups. Our findings are consistent with previous characterization of this gene family in *O. sativa* [34,60].

Based on the analysis of conserved domain alignment, PPR genes can be classified into two subfamilies, P and PLS, which were further confirmed by phylogenetic analysis. However, a few members within each subfamily were admixed. This phenomenon was also observed in previous studies [5,6]. The amino acid arrangement of PPR protein domains in these two rice species is highly consistent with other plant species [5], indicating that PPR proteins are highly conserved across flowering plants.

Previous studies have shown that in most species of PPR genes in angiosperms, over half of the individuals have only one exon [60,63]. The proportions of genes with only one exon were the highest with percentages of 45.65% and 38.00% in *O. sativa* and *O. rufipogon*, respectively, which agree with the observation that up to 80% of the PPR genes in rice were intron-less [34]. The results support that retrotransposition has resulted in the expansion of the PPR gene family in rice and its wild progenitor.

We investigated the origins and evolution of PPR gene families since the domestication of *O. sativa* from its wild progenitor. There were ~36.17% and ~42.94% of species-specific PPR genes from new gene duplications in the *O. sativa* and *O. rufipogon* genomes, respec-

tively. The results thus provide evidence that the evolution of the gene family is ongoing as a result of newly occurring gene duplications, although the *PPR* gene family is relatively conserved across land plants.

In spite of the recent split of rice and *O. rufipogon*, GO enrichment analysis of *PPR* genes uniquely identified in *O. sativa* and up-regulated genes in seedling tissue of *O. sativa* exhibited an enrichment of a unique biological process related to the embryo development ending in seed dormancy. We found a unique SMR domain in the included gene, *LOC_Os06g09880.1*. Genes with this domain were previously reported to be directly involved in RNA editing of chloroplast gene *rps8* transcripts [64]. *PPR-SMR1* is a *PPR* protein containing the SMR domain, and *Zm-mCSF1* is a protein containing the CRM domain, both of which target mitochondria. The loss-of-function mutations of these two genes severely hindered embryogenesis and endosperm development in maize [65]. Although we still cannot determine whether the upregulation of these *PPR* genes is associated with artificial selection during the domestication process, this finding may highlight the potential role of *PPR* genes in regulating seed embryo development and seed dormancy.

It is worthwhile noting that these *PPR* genes are agronomically very important functional genes, such as *RFL* (Restoration Fertility Like) genes for rice breeding programs (Table S7). Furthermore, our results showed that the genes upregulated in the leaf tissues of *O. rufipogon* were enriched entries related to biological processes for nematode defense. These loci may have potential breeding values for insect-resistant traits.

Although segmental duplications have only expanded approximately 11% of the *PPR* gene family members, we interestingly obtained evidence that segmental duplication promotes the structural diversity of *PPR* genes through incomplete gene duplications. Our results indicate that the similarity between protein products of homologous genes arising from segmental duplications was relatively lower. This could be explained by incomplete gene duplications, and only partial sequences are homologous between duplicated gene pairs. We interestingly observed that, among duplicated gene pairs with non-*PPR* paralogous genes, *O. sativa* had two complete gene duplications, while *O. rufipogon* only had ten complete gene duplications and the others were mainly incomplete gene duplications. Our results showed that most of the duplicated genes that both belong to the *PPR* gene resulting from segmental duplications exhibited asymmetrical expression. Except for two potentially silenced genes, the rest of the AEDs maintained low expression levels. These duplicated genes might be limited by dosage effects, as most of the recently duplicated genes tend to be functionally redundant, making them susceptible to loss-of-function mutations that could lead to the degradation of one copy into a pseudogene [66]. However, among the duplicated genes in which *PPR* genes contained other domains, which were mainly generated by incomplete duplication through segmental duplications, 5 out of 7 in *O. sativa* and 23 out of 26 in *O. rufipogon* exhibited expression levels above moderate in at least one tissue. These foreign domains might promote functional diversification, allowing them to escape the constraints of dosage effects.

As previously reported, *PPR-SMRs* are present in species representing major angiosperms, and recent gene duplication may account for the additional copies of some *PPR-SMRs* in *A. thaliana* [58]. In this study, we illustrate examples of *PPR* genes contained in other domains in the two closely related rice species, with incomplete gene duplications explaining the origins of these domains. While we are uncertain if these *PPR* genes with newly acquired domains follow the same mechanism in other plants, based on our observation here, we speculate that, during plant evolution, segmental duplications leading to incomplete gene duplications might result in the acquisition of domains, thus promoting the functional innovation and diversification of *PPR* genes. Accordingly, this study offers a unique perspective on the evolution of *PPR* gene structures and underscores the potential role of segmental duplications in *PPR* gene structural diversity.

4. Materials and Methods

4.1. Identification of PPR Genes

The rice genome of MSU7 was downloaded from <http://rice.uga.edu/>, accessed on 10 September 2023. The *O. rufipogon* genome data file can be found in the National Genomics Data Center under the accession number PRJCA002346 [67]. To comprehensively detect the PPR gene family in these two genomes, we downloaded the seed file (PF01535) of the PPR gene family based on the Hidden Markov Model (HMM) from the Pfam V31.0 database (<http://pfam.xfam.org/>, accessed on 9 September 2023). To identify PPR genes using the HMMER 3.0 program [68], we set the E-value < 10 as a cutoff to filter out fewer significant matches to increase the confidence in identifying true PPR genes. We used the SMART program (<http://smart.embl-heidelberg.de/>, accessed on 11 September 2023) to analyze the protein domains of the candidate PPR genes obtained from the two genomes. We finally used the HMMER 3.0 [68] matrix defined by conserved domains of PPR gene subfamilies (P, PLS, E1, E2, E+, and DYW) in *A. thaliana* to search, analyze, and classify these protein sequence domains. Proteins that do not have any PPR motifs or have only one PPR motif are typically discarded or excluded from further analyses.

4.2. Chromosomal Location and Gene Structure and Cis-Acting Element Analysis

We employed Weblogo (<http://weblogo.berkeley.edu/logo.cgi>, accessed on 11 September 2023) to create a motif recognition to assess the conservation of PPR motifs. The position of PPR genes in the genome was visualized in Circos V0.69 [69]. To analyze the cis-acting elements in the major promoter regions of the PPR genes, a sequence of 2000 base pairs upstream of the transcription start site was selected. An online tool named PlantCARE (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>, accessed on 12 September 2023) was used to obtain all the cis-acting elements in each gene's promoter region, and important responsive elements were selected through screening. The plots were created using the heatmap Package in R V4.2.3 [70].

4.3. Phylogenetic Analysis and Gene Ontology (GO) Enrichment Analysis

The PPR genes from the two rice species were aligned using the MAFFT V7.310 method [71]. Alignments were trimmed using TrimAL V1.4 [72] with a minimum conservation threshold and a gap threshold of 20%. The phylogenetic tree was built using the maximum likelihood method in IQ-TREE V2.2.3 [73], with the best evolutionary model "LG+R10". We constructed the phylogenetic tree using Maximum likelihood with 1000-replicate bootstrap analysis. iTOL is used to beautify and annotate evolutionary trees and add subgroup information [74]. We performed functional annotation of the PPR gene family in *O. sativa* and *O. rufipogon* using the InterProScan 5.54-87.0 program [75]. The results were visualized using Tbttools V2.008 [76].

4.4. Analysis of PPR Gene Expression and Classification of Expression Patterns of Sister Duplicates

Field-grown rice (*O. sativa* 'Nipponbare'; *O. rufipogon* 'R1') tissues including 7-d-old seedlings, mature plant flag leaves, roots, and panicles were harvested and frozen in liquid nitrogen for subsequent RNA sequencing and proteomic experiments. The RNeasy Mini Kit (Qiagen) bench protocol was used for plant total RNA extraction. To investigate patterns of PPR gene expression, we sequenced transcriptome data from four different tissues of *O. sativa* and *O. rufipogon*. The raw sequencing data were submitted to the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2017) in the BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number PRJCA020852. RNA-sequencing data were mapped using HISAT2 V2.2.1 [77] on the reference genome after building the index for *O. sativa* and *O. rufipogon*. The read counts matrix for 24 samples (4 tissues, each tissue with three biological replicates) was generated using featureCounts V2.0.1 [78]. Differentially expressed genes (DEGs) were evaluated by DESeq2 V1.42.0 [79]. The threshold for DEGs was Fold Change > 2 and p-adj < 0.05. For

the duplicated genes generated by segmental duplications, we selected gene pairs in which both copies belong to the *PPR* gene family for downstream analysis. Differential gene expression analyses among the duplicated gene pairs for each tissue were performed using DESeq2 V1.42.0 [79] with the same threshold. We divided the duplicated gene pairs into three classes according to their expression in four tissues: (1) sub- or neo-functionalized pairs, in which each of the two duplicates was significantly more highly expressed than the other in at least one tissue; (2) AEDs, when one copy exhibits significantly higher expression levels in two tissues compared to its sister gene and expression in other tissues is not lower than its sister copy; and (3) the remaining duplicates were classified as no-difference pairs [66]. A heatmap and volcano plot were plotted using pheatmap V1.0.12 and EnhancedVolcano V1.20.0 [70].

4.5. Identification of Shared and Species-Specific *PPR* Genes in *O. sativa* and *O. rufipogon* and *Ks* Analysis

To identify orthologous *PPR* genes in *O. sativa* and *O. rufipogon*, we first performed a reciprocal blast using Blastp 2.14 [80] on the genes retrieved from the two rice genomes. Subsequently, we selected the reciprocal best hits as candidates for orthologous genes. Next, we assessed whether the neighboring genes also had reciprocal best hits with the neighboring orthologous genes in the other genome. The criteria for the local synteny filter was based on the orthologous relationship of genes surrounding the *PPR* genes. Specifically, if there is at least one gene on each side within five upstream and downstream genes surrounding a *PPR* gene that has an orthologous relationship with counterpart genes in the other species, then the pair of genes passes the local synteny filter. If both conditions were met, these genes were categorized as shared genes; otherwise, they were considered species-specific genes. We further validated the results by using Orthofinder V2.5.2 [81] and MCScan V1.1.12 [82], which validated 387 gene pairs (Figures S10 and S11). For the three gene pairs that could not be identified as synteny genes by MCScan V1.1.12 [82], we performed sequence-level synteny analysis using Nucmer 4.0.0rc1 [83], which was certified (Figure S12). We conducted *Ks* analysis for orthologous gene pairs (orthogroups) and gene duplication groups (paralogous groups). We defined gene duplication groups as follows: firstly, we identified all duplicated gene pairs in both *O. sativa* and *O. rufipogon* using dupgen_finder. We then extracted putative duplicated pairs of lineage-specific *PPR* genes. Subsequently, we used paraAT2.0 [84] to generate aligned pairwise sequences in axt format through Muscle [85] alignment. We then calculated *Ks* values for each pair of genes, both orthologous and all duplicated gene pairs, using the YN model in KaKs_Calculator V2.0 [86]. For the following data processing, we took the minimum *Ks* value from all duplicated gene pairs as an approximation of the most recent gene duplication event. We then compared this value to the average *Ks* value of all orthologous *PPR* genes.

4.6. Identification of Gene Duplications and Segmental Duplications

We characterized different types of gene duplications using DupGen_finder [87] (https://github.com/qiao-xin/DupGen_finder, accessed on 14 September 2023) and counted the number and proportion of each type of *PPR* gene duplication in the whole genome. We identified segmental duplications in *O. sativa* and *O. rufipogon* with the Whole-Genome Assembly Comparison (WGAC) pipeline [49]. Firstly, the genome assemblies were split into trackable 400-kb segments after the TEs were removed leaving the unique genome for further searching all sequence similarity using BLAST 2.14 [80] and self-blast using webb_self program in SUN workstation. The BLAST results were parsed for alignments if they had >88% identity and >200 bp aligned length. After reinserting common repeats into resulting pairwise alignments and trimming the ends, the final set of segmental duplications was identified if global alignment bases were longer than 1 kb matched and had >90% identity.

4.7. Identification of Domain and Visualization of Segmental Duplication

We used SMART [88] (<http://smart.embl-heidelberg.de/smart/>, accessed on 11 September 2023) to identify the domains of the PPR gene and employed Tbtools V2.008 [76] to visualize the domains. KaryoploteR [89] and Circos V0.69 [69] were used to depict segmental duplication in the genomes. We employed our own script to plot the structures of duplicated genes.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ijms242216313/s1>.

Author Contributions: Conceptualization, L.-Y.F. and P.-F.L.; data curation, L.-Y.F., P.-F.L., R.-J.X. and L.-Z.G.; funding acquisition, L.-Z.G.; investigation, L.-Y.F., P.-F.L., R.-J.X. and H.-Q.K.; writing original draft, L.-Y.F., P.-F.L. and L.-Z.G.; writing—review and editing, L.-Y.F. and L.-Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: The project was supported by a start-up grant from Hainan University (Grant No. RZ210000 6631) to L.-Z.G.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article or Supplementary Material.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Small, I.D.; Peeters, N. The PPR motif—A TPR-related motif prevalent in plant organellar proteins. *Trends Biochem. Sci.* **2000**, *25*, 45–47. [CrossRef] [PubMed]
- Barkan, A.; Small, I. Pentatricopeptide Repeat Proteins in Plants. *Annu. Rev. Plant Biol.* **2014**, *65*, 415–442. [CrossRef] [PubMed]
- Manna, S. An overview of pentatricopeptide repeat proteins and their applications. *Biochimie* **2015**, *113*, 93–99. [CrossRef] [PubMed]
- Gutmann, B.; Royan, S.; Schallenberg-Rüdinger, M.; Lenz, H.; Castleden, I.R.; McDowell, R.; Vacher, M.A.; Tonti-Filippini, J.; Bond, C.S.; Knoop, V.; et al. The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants. *Mol. Plant* **2020**, *13*, 215–230. [CrossRef]
- Xing, H.; Fu, X.; Yang, C.; Tang, X.; Guo, L.; Li, C.; Xu, C.; Luo, K. Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. *Sci. Rep.* **2018**, *8*, 2817. [CrossRef]
- Ma, X.; Wang, D.; Xue, G.; Zheng, X.; Lu, Y.; Shi, J.; Hao, Z.; Chen, J. Characterization of the *Liriodendron chinense* Pentatricopeptide Repeat (PPR) Gene Family and Its Role in Osmotic Stress Response. *Genes* **2023**, *14*, 1125. [CrossRef]
- Lurin, C.; Andréas, C.; Aubourg, S.; Bellaoui, M.; Bitton, F.; Bruyère, C.; Caboche, M.; Debast, C.; Gualberto, J.; Hoffmann, B.; et al. Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. *Plant Cell* **2004**, *16*, 2089–2103. [CrossRef]
- Cheng, S.; Gutmann, B.; Zhong, X.; Ye, Y.; Fisher, M.F.; Bai, F.; Castleden, I.; Song, Y.; Song, B.; Huang, J.; et al. Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *Plant J.* **2016**, *85*, 532–547. [CrossRef]
- Coquille, S.; Filipovska, A.; Chia, T.; Rajappa, L.; Lingford, J.P.; Razif, M.F.M.; Thore, S.; Rackham, O. An artificial PPR scaffold for programmable RNA recognition. *Nat. Commun.* **2014**, *5*, 5729. [CrossRef]
- Barkan, A.; Rojas, M.; Fujii, S.; Yap, A.; Chong, Y.S.; Bond, C.S.; Small, I. A Combinatorial Amino Acid Code for RNA Recognition by Pentatricopeptide Repeat Proteins. *PLoS Genet.* **2012**, *8*, e1002910. [CrossRef]
- Yagi, Y.; Hayashi, S.; Kobayashi, K.; Hirayama, T.; Nakamura, T. Elucidation of the RNA Recognition Code for Pentatricopeptide Repeat Proteins Involved in Organelle RNA Editing in Plants. *PLoS ONE* **2013**, *8*, e57286. [CrossRef]
- Takenaka, M.; Zehrmann, A.; Brennicke, A.; Graichen, K. Improved Computational Target Site Prediction for Pentatricopeptide Repeat RNA Editing Factors. *PLoS ONE* **2013**, *8*, e65343. [CrossRef] [PubMed]
- Ichinose, M.; Tasaki, E.; Sugita, C.; Sugita, M. A PPR-DYW protein is required for splicing of a group II intron of *cox1* pre-mRNA in *Physcomitrella patens*. *Plant J.* **2012**, *70*, 271–278. [CrossRef] [PubMed]
- Zheng, S.; Dong, J.; Lu, J.; Li, J.; Jiang, D.; Yu, H.; Ye, S.; Bu, W.; Liu, Z.; Zhou, H.; et al. A cytosolic pentatricopeptide repeat protein is essential for tapetal plastid development by regulating *OsGLK1* transcript levels in rice. *New Phytol.* **2022**, *234*, 1678–1695. [CrossRef] [PubMed]
- Xiao, H.; Zhang, Q.; Qin, X.; Xu, Y.; Ni, C.; Huang, J.; Zhu, L.; Zhong, F.; Liu, W.; Yao, G.; et al. Rice *PPS1* encodes a DYW motif-containing pentatricopeptide repeat protein required for five consecutive RNA-editing sites of *nad3* in mitochondria. *New Phytol.* **2018**, *220*, 878–892. [CrossRef]

16. Wang, Y.; Li, H.; Huang, Z.; Ma, B.; Yang, Y.; Xiu, Z.; Wang, L.; Tan, B. Maize PPR-E proteins mediate RNA C-to-U editing in mitochondria by recruiting the trans deaminase PCW1. *Plant Cell* **2023**, *35*, 529–551. [\[CrossRef\]](#)
17. An, H.; Ke, X.; Li, L.; Liu, Y.; Yuan, S.; Wang, Q.; Hou, X.; Zhao, J. ALBINO EMBRYO AND SEEDLING is required for RNA splicing and chloroplast homeostasis in Arabidopsis. *Plant Physiol.* **2023**, *193*, 483–501. [\[CrossRef\]](#)
18. Meierhoff, K.; Felder, S.; Nakamura, T.; Bechtold, N.; Schuster, G. HCF152, an *Arabidopsis* RNA Binding Pentatricopeptide Repeat Protein Involved in the Processing of Chloroplast *psbB-psbT-psbH-petB-petD* RNAs. *Plant Cell* **2003**, *15*, 1480–1495. [\[CrossRef\]](#)
19. Nishimura, Y.; Shikanai, T. Maturation and longevity of mRNA in chloroplasts: The functions of PPR proteins in RNA editing and the molecular mechanisms of RNA stability control in chloroplasts. *Tanpakushitsu Kakusan Koso* **2009**, *54*, 2098–2101.
20. Macedo-Orsorio, K.S.; Martínez-Antonio, A.; Badillo-Corona, J.A. Pas de Trois: An Overview of Penta-, Tetra-, and Octo-Tricopeptide Repeat Proteins From *Chlamydomonas reinhardtii* and Their Role in Chloroplast Gene Expression. *Front. Plant Sci.* **2021**, *12*, 775366. [\[CrossRef\]](#)
21. Hammani, K.; Colas Des Francs-Small, C.; Takenaka, M.; Tanz, S.K.; Okuda, K.; Shikanai, T.; Brennicke, A.; Small, I. The Pentatricopeptide Repeat Protein OTP87 Is Essential for RNA Editing of *nad7* and *atp1* Transcripts in *Arabidopsis* Mitochondria. *J. Biol. Chem.* **2011**, *286*, 21361–21371. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Schmitz-Linneweber, C.; Williams-Carrier, R.; Barkan, A. RNA Immunoprecipitation and Microarray Analysis Show a Chloroplast Pentatricopeptide Repeat Protein to Be Associated with the 5' Region of mRNAs Whose Translation It Activates. *Plant Cell* **2005**, *17*, 2791–2804. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Pfalz, J.; Bayraktar, O.A.; Prikryl, J.; Barkan, A. Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *EMBO J.* **2009**, *28*, 2042–2052. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Prikryl, J.; Rojas, M.; Schuster, G.; Barkan, A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 415–420. [\[CrossRef\]](#)
25. Barkan, A. Expression of Plastid Genes: Organelle-Specific Elaborations on a Prokaryotic Scaffold. *Plant Physiol.* **2011**, *155*, 1520–1532. [\[CrossRef\]](#)
26. Melonek, J.; Zhou, R.; Bayer, P.E.; Edwards, D.; Stein, N.; Small, I. High intraspecific diversity of Restorer-of-fertility-like genes in barley. *Plant J.* **2019**, *97*, 281–295. [\[CrossRef\]](#)
27. Ma, Q.; Hedden, P.; Zhang, Q. Heterosis in rice seedlings: Its relationship to gibberellin content and expression of gibberellin metabolism and signaling genes. *Plant Physiol.* **2011**, *156*, 1905–1920. [\[CrossRef\]](#)
28. Labroo, M.R.; Studer, A.J.; Rutkoski, J.E. Heterosis and Hybrid Crop Breeding: A Multidisciplinary Review. *Front. Genet.* **2021**, *12*, 643761. [\[CrossRef\]](#)
29. Takatsuka, A.; Kazama, T.; Toriyama, K. Cytoplasmic Male Sterility-Associated Mitochondrial Gene *orf312* Derived from Rice (*Oryza sativa* L.) Cultivar Tadukan. *Rice* **2021**, *14*, 46. [\[CrossRef\]](#)
30. Chase, C.D. Cytoplasmic male sterility: A window to the world of plant mitochondria nuclear interactions. *Trends Genet.* **2007**, *23*, 81–90. [\[CrossRef\]](#)
31. Brown, G.G.; Formanová, N.; Jin, H.; Wargachuk, R.; Dendy, C.; Patil, P.; Laforest, M.; Zhang, J.; Cheung, W.Y.; Landry, B.S. The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. *Plant J.* **2003**, *35*, 262–272. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Chen, L.; Liu, Y. Male Sterility and Fertility Restoration in Crops. *Annu. Rev. Plant Biol.* **2014**, *65*, 579–606. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Hu, J.; Wang, K.; Huang, W.; Liu, G.; Gao, Y.; Wang, J.; Huang, Q.; Ji, Y.; Qin, X.; Wan, L.; et al. The Rice Pentatricopeptide Repeat Protein *RF5* Restores Fertility in Hong-Lian Cytoplasmic Male-Sterile Lines via a Complex with the Glycine-Rich Protein GRP162. *Plant Cell* **2012**, *24*, 109–122. [\[CrossRef\]](#) [\[PubMed\]](#)
34. O'Toole, N.; Hattori, M.; Andres, C.; Iida, K.; Lurin, C.; Schmitz-Linneweber, C.; Sugita, M.; Small, I. On the Expansion of the Pentatricopeptide Repeat Gene Family in Plants. *Mol. Biol. Evol.* **2008**, *25*, 1120–1128. [\[CrossRef\]](#)
35. Schmitz-Linneweber, C.; Small, I. Pentatricopeptide repeat proteins: A socket set for organelle gene expression. *Trends Plant Sci.* **2008**, *13*, 663–670. [\[CrossRef\]](#)
36. Fujii, S.; Small, I. The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytol.* **2011**, *191*, 37–47. [\[CrossRef\]](#)
37. Rensing, S.A.; Lang, D.; Zimmer, A.D.; Terry, A.; Salamov, A.; Shapiro, H.; Nishiyama, T.; Perroud, P.; Lindquist, E.A.; Kamisugi, Y.; et al. The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. *Science* **2008**, *319*, 64–69. [\[CrossRef\]](#)
38. Merchant, S.S.; Prochnik, S.E.; Vallon, O.; Harris, E.H.; Karpowicz, S.J.; Witman, G.B.; Terry, A.; Salamov, A.; Fritz-Laylin, L.K.; Maréchal-Drouard, L.; et al. The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. *Science* **2007**, *318*, 245–250. [\[CrossRef\]](#)
39. Rivals, E.; Bruyère, C.; Toffano-Nioche, C.; Lecharny, A. Formation of the *Arabidopsis* pentatricopeptide repeat family. *Plant Physiol.* **2006**, *141*, 825–839. [\[CrossRef\]](#)
40. Taylor, J.S.; Raes, J. Duplication and divergence: The evolution of new genes and old ideas. *Annu. Rev. Genet.* **2004**, *38*, 615–643. [\[CrossRef\]](#)
41. Muller, H.J. Bar duplication. *Science* **1936**, *83*, 528–530. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Ohno, S.; Wolf, U.; Atkin, N.B. Evolution from fish to mammals by gene duplication. *Hereditas* **1968**, *59*, 169–187. [\[CrossRef\]](#)
43. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **2003**, *18*, 292–298. [\[CrossRef\]](#)
44. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.* **2004**, *20*, 116–122. [\[CrossRef\]](#) [\[PubMed\]](#)

45. Bowers, J.E.; Chapman, B.A.; Rong, J.; Paterson, A.H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **2003**, *422*, 433–438. [\[CrossRef\]](#)
46. De Bodt, S.; Maere, S.; Van de Peer, Y. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **2005**, *20*, 591–597. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Vollger, M.R.; Guitart, X.; Dishuck, P.C.; Mercuri, L.; Harvey, W.T.; Gershman, A.; Diekhans, M.; Sulovari, A.; Munson, K.M.; Lewis, A.P.; et al. Segmental duplications and their variation in a complete human genome. *Science* **2022**, *376*, eabj6965. [\[CrossRef\]](#)
48. Bailey, J.A.; Gu, Z.; Clark, R.A.; Reinert, K.; Samonte, R.V.; Schwartz, S.; Adams, M.D.; Myers, E.W.; Li, P.W.; Eichler, E.E. Recent segmental duplications in the human genome. *Science* **2002**, *297*, 1003–1007. [\[CrossRef\]](#)
49. Bailey, J.A.; Yavor, A.M.; Massa, H.F.; Trask, B.J.; Eichler, E.E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **2001**, *11*, 1005–1017. [\[CrossRef\]](#)
50. Bailey, J.A.; Eichler, E.E. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **2006**, *7*, 552–564. [\[CrossRef\]](#)
51. Li, K.; Jiang, W.; Hui, Y.; Kong, M.; Feng, L.; Gao, L.; Li, P.; Lu, S. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Mol. Plant.* **2021**, *14*, 1745–1756. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Zhang, X.; Wang, G.; Zhang, S.; Chen, S.; Wang, Y.; Wen, P.; Ma, X.; Shi, Y.; Qi, R.; Yang, Y.; et al. Genomes of the Banyan Tree and Pollinator Wasp Provide Insights into Fig-Wasp Coevolution. *Cell* **2020**, *183*, 875–889. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Giannuzzi, G.; D’Addabbo, P.; Gasparro, M.; Martinelli, M.; Carelli, F.N.; Antonacci, D.; Ventura, M. Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genom.* **2011**, *12*, 436. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Antonacci, F.; Dennis, M.Y.; Huddleston, J.; Sudmant, P.H.; Steinberg, K.M.; Rosenfeld, J.A.; Miroballo, M.; Graves, T.A.; Vives, L.; Malig, M.; et al. Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat. Genet.* **2014**, *46*, 1293–1302. [\[CrossRef\]](#)
55. Dennis, M.Y.; Eichler, E.E. Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* **2016**, *41*, 44–52. [\[CrossRef\]](#)
56. Dennis, M.Y.; Nuttle, X.; Sudmant, P.H.; Antonacci, F.; Graves, T.A.; Nefedov, M.; Rosenfeld, J.A.; Sajjadian, S.; Malig, M.; Kotkiewicz, H.; et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **2012**, *149*, 912–922. [\[CrossRef\]](#)
57. Ciccarelli, F.D.; von Mering, C.; Suyama, M.; Harrington, E.D.; Izaurralde, E.; Bork, P. Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* **2005**, *15*, 343–351. [\[CrossRef\]](#)
58. Liu, S.; Melonek, J.; Boykin, L.M.; Small, I.; Howell, K.A. PPR-SMRs: Ancient proteins with enigmatic functions. *RNA Biol.* **2013**, *10*, 1501–1510. [\[CrossRef\]](#)
59. Chateigner-Boutin, A.; Small, I. Plant RNA editing. *RNA Biol.* **2010**, *7*, 213–219. [\[CrossRef\]](#)
60. Chen, G.; Zou, Y.; Hu, J.; Ding, Y. Genome-wide analysis of the rice PPR gene family and their expression profiles under different stress treatments. *BMC Genom.* **2018**, *19*, 720. [\[CrossRef\]](#)
61. Subburaj, S.; Tu, L.; Lee, K.; Park, G.; Lee, H.; Chun, J.; Lim, Y.; Park, M.; McGregor, C.; Lee, G. A Genome-Wide Analysis of the Pentatricopeptide Repeat (PPR) Gene Family and PPR-Derived Markers for Flesh Color in Watermelon (*Citrullus lanatus*). *Genes* **2020**, *11*, 1125. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Zhang, M.; Zhao, Y.; Meng, Y.; Xiao, Y.; Zhao, J.; Xiao, B.; An, C.; Gao, Y. PPR proteins in the tea plant (*Camellia sinensis*) and their potential roles in the leaf color changes. *Sci. Hortic.* **2022**, *293*, 110745. [\[CrossRef\]](#)
63. Wei, K.; Han, P. Pentatricopeptide repeat proteins in maize. *Mol. Breed.* **2016**, *36*, 170. [\[CrossRef\]](#)
64. Zhang, J.; Guo, Y.; Fang, Q.; Zhu, Y.; Zhang, Y.; Liu, X.; Lin, Y.; Barkan, A.; Zhou, F. The PPR-SMR Protein ATP4 Is Required for Editing the Chloroplast *rps8* mRNA in Rice and Maize. *Plant Physiol.* **2020**, *184*, 2011–2021. [\[CrossRef\]](#)
65. Chen, Z.; Wang, H.; Shen, J.; Sun, F.; Wang, M.; Xu, C.; Tan, B. PPR-SMR1 is required for the splicing of multiple mitochondrial introns, interacts with Zm-mCSF1, and is essential for seed development in maize. *J. Exp. Bot.* **2019**, *70*, 5245–5258. [\[CrossRef\]](#)
66. Lan, X.; Pritchard, J.K. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **2016**, *352*, 1009–1013. [\[CrossRef\]](#)
67. Li, W.; Li, K.; Huang, Y.; Shi, C.; Hu, W.; Zhang, Y.; Zhang, Q.; Xia, E.; Hutang, G.; Zhu, X.; et al. SMRT sequencing of the *Oryza rufipogon* genome reveals the genomic basis of rice adaptation. *Commun. Biol.* **2020**, *3*, 167. [\[CrossRef\]](#)
68. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic. Acids Res.* **2013**, *41*, e121. [\[CrossRef\]](#)
69. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [\[CrossRef\]](#)
70. Gu, Z.; Hübschmann, D. Make Interactive Complex Heatmaps in R. *Bioinformatics* **2022**, *38*, 1460–1462. [\[CrossRef\]](#)
71. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [\[CrossRef\]](#)
72. Capella-Gutiérrez, S.; Silla-Martínez, J.M.; Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25*, 1972–1973. [\[CrossRef\]](#) [\[PubMed\]](#)
73. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [\[CrossRef\]](#) [\[PubMed\]](#)

74. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256–W259. [[CrossRef](#)] [[PubMed](#)]
75. Jones, P.; Binns, D.; Chang, H.; Fraser, M.; Li, W.; Mcanulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [[CrossRef](#)] [[PubMed](#)]
76. Chen, C.; Chen, H.; Zhang, Y.; Thomas, H.R.; Frank, M.H.; He, Y.; Xia, R. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant.* **2020**, *13*, 1194–1202. [[CrossRef](#)]
77. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [[CrossRef](#)]
78. Liao, Y.; Smyth, G.K.; Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923–930. [[CrossRef](#)] [[PubMed](#)]
79. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [[CrossRef](#)]
80. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)]
81. Emms, D.M.; Kelly, S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **2015**, *16*, 157. [[CrossRef](#)] [[PubMed](#)]
82. Tang, H.; Bowers, J.E.; Wang, X.; Ming, R.; Alam, M.; Paterson, A.H. Synteny and Collinearity in Plant Genomes. *Science* **2008**, *320*, 486–488. [[CrossRef](#)] [[PubMed](#)]
83. Marçais, G.; Delcher, A.L.; Phillippy, A.M.; Coston, R.; Salzberg, S.L.; Zimin, A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **2018**, *14*, e1005944. [[CrossRef](#)] [[PubMed](#)]
84. Zhang, Z.; Xiao, J.; Wu, J.; Zhang, H.; Liu, G.; Wang, X.; Dai, L. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **2012**, *419*, 779–781. [[CrossRef](#)]
85. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [[CrossRef](#)]
86. Wang, D.; Zhang, Y.; Zhang, Z.; Zhu, J.; Yu, J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genom. Proteom. Bioinform.* **2010**, *8*, 77–80. [[CrossRef](#)]
87. Qiao, X.; Li, Q.; Yin, H.; Qi, K.; Li, L.; Wang, R.; Zhang, S.; Paterson, A.H. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **2019**, *20*, 38. [[CrossRef](#)]
88. Letunic, I.; Khedkar, S.; Bork, P. SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Res.* **2021**, *49*, D458–D460. [[CrossRef](#)]
89. Gel, B.; Serra, E. karyoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **2017**, *33*, 3088–3090. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.