Pentatricopeptide repeat (PPR) proteins form a huge family in plants (450 members in Arabidopsis and 477 in rice) defined by tandem repetitions of a degenerate 35 amino acid motif, discovered in silico (Small and Peeters 2000) during a search of the then incomplete Arabidopsis thaliana genome sequence for genes predicted to be targeted to mitochondria or plastids. Although no PPR structures are known, the motif is predicted to fold into a helix-turn-helix structure (Small and Peeters 2000) similar to those found in “solenoid” proteins such as widespread tetratricopeptide repeat family. However, the sequence characteristics of the motif clearly distinguish PPR proteins from other solenoid proteins (Small and Peeters 2000; Karpenahalli et al. 2007). Solenoid proteins generally form protein-binding surfaces, but current evidence suggests that PPR proteins bind RNA rather than, or as well as, proteins (reviewed in Nakamura et al. 2004; Delannoy et al. 2007). The complete nuclear genome of A. thaliana contains 450 distinct genes encoding PPR proteins, separated into 2 subfamilies and 4 subclasses based on their C-terminal domain structure (Lurin et al. 2004). Evidence from expressed sequence tag (EST) data suggests that many other land plants also contain hundreds of PPR genes (Hattori et al. 2004; Lurin et al. 2004; Salone et al. 2007).

Recent years have seen many experimental investigations of PPR function, motivated by the finding that mutants of several PPR genes in plants display embryo lethal or otherwise spectacular phenotypes (for recent reviews, see Andrés et al. 2007; Saha et al. 2007). PPR proteins have been shown to play crucial roles in virtually all stages of organellar gene expression. For example, PPR proteins are associated with both the transcription (Ikeda and Gray 1999; Pfälz et al. 2006) and translation machinery (Pusnik et al. 2007) and involved in many stages of mRNA processing including splicing (Schmitz-Linneweber et al. 2006; Falcon de Langevialle et al. 2007), endonucleolytic cleavage (Hashimoto et al. 2003), and RNA editing (Kotera et al. 2005; Okuda et al. 2007). PPR proteins from various higher plants also act to suppress the expression of mitochondrial genes associated with cytoplasmic male sterility (e.g., Desloire et al. 2003; Gillman et al. 2007). The growing body of experimental results on PPR protein functions is consistent with the fact that the majority of Arabidopsis PPR proteins are predicted to be targeted to mitochondria or chloroplasts (Lurin et al. 2004; Small et al. 2004). The common thread to the various roles implicated for PPR proteins is an RNA-binding activity, demonstrated in several cases (Lahmy et al. 2000; Mancebo et al. 2001; Nakamura et al. 2003; Schmitz-Linneweber et al. 2005, 2006; Okuda et al. 2006).

Computational scans of complete genome sequences reveal that nonplant organisms contain very few PPR-encoding genes (Lurin et al. 2004; Andrés et al. 2007). For example, PPR genes are virtually absent from prokaryotes (Pusnik et al. 2007), and moreover the yeast, Drosophila and human genomes are predicted to contain only 5, 2, and 6 PPR genes, respectively. One of the yeast PPR genes, PET309, was the first PPR gene to be functionally described. It plays an essential role in translation of the mitochondrial COX1 gene (Manthey and McEwen 1995; Tavares-Carreon et al. 2008). Mutations in a human PPR protein, LRPPRC, give rise to Leigh syndrome French Canadian variant (Mootha et al. 2003). LRPPRC has been demonstrated to be a mitochondrial mRNA stabilization factor (Xu et al. 2004). The nonplant organism with the largest number of predicted PPR genes is the parasitic protozoan Trypanosoma brucei, with 28. Several of these have been shown to be essential for mitochondrial function (Pusnik et al. 2007). Thus, evidence suggests that the few PPR proteins in nonplants also play roles in organellar (i.e., mitochondrial) gene expression.

The vast difference between the numbers of PPR genes in higher plants and nonplant organisms indicates
that a massive expansion of the PPR gene family occurred during the evolution of plants. The expansion was commented on in earlier studies based on Arabidopsis PPR genes (Lurin et al. 2004; Rivals et al. 2006); however, its origins and significance remain largely a mystery. In the present study, we take advantage of newly completed genome sequencing results to perform a systematic genome-wide comparison of the PPR genes in 3 organisms, widely separate along the plant lineage: the dicot Arabidopsis, the monocot Oryza sativa (rice), and the moss Physcomitrella patens (Rensing et al. 2008), hereafter often referred to as “moss” for simplicity. Our results enable us to draw firm conclusions on the causes and timing of the expansion of PPR genes in higher plants. The dramatic differences in numbers of different subclasses of PPR proteins between species also give insight into the evolution and mechanism of RNA editing in plant organelles.

Materials and Methods
Identification of PPR-Encoding Genes from Genomic Data

The genome sequence data and gene annotations used in this work were for Arabidopsis thaliana: Release 6 of the Arabidopsis annotation from The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org), O. sativa: the Osa1 Release 3 (Yuan et al. 2005) from The Institute for Genomic Research (TIGR; http://www.tigr.org), and P. patens: sequence data from the v.1.1 release produced by the US Department of Energy Joint Genome Institute (JGI; http://www.jgi.doe.gov).

The hmmsearch program from the HMMER package (Eddy 1998) was used to detect PPR motifs in protein and translated genomic sequences. The hidden Markov models used were identical to those used for detection of Arabidopsis PPR motifs (Lurin et al. 2004). As multiple models based on PPR variants were used, multiple overlapping hits were usually obtained. In these cases, the highest scoring chain of nonoverlapping hits was retained and the alternative overlapping hits discarded.

As an initial screen, genomic nucleotide sequence data in the pseudochromosomes of Arabidopsis and rice and scaffolds of Physcomitrella were translated in all 6 frames. The hmmsearch program was applied to the translated sequence data to identify clusters of all PPR motifs (P, L, S, L2, E, E+, and DYW) separated by fewer than 200 base pairs. These clusters correspond to putative PPR genes and were visualized in modified versions of the FlagDB++ (Samson et al. 2004) and GBrowse (Stein et al. 2002), genome browsers. In the case of Physcomitrella, these clusters of motifs correspond to already annotated genes from the JGI genome release and these were used in this study, except where a JGI gene model did not include either a start or stop codon (or both). In these cases, the open reading frames (ORFs) were extended to include these. It is expected that future sequencing results and refinements will significantly improve many of the Physcomitrella models. Following our earlier study on Arabidopsis (Lurin et al. 2004), several clusters of rice PPR motifs found in the translated genome-wide search fell outside existing gene models. Modified gene models were constructed using the PPR motif data and Genemark.hmm trained on rice sequences (Yuan et al. 2005) to verify possible alternative exon–intron structures. Several of the Arabidopsis models we proposed earlier (Lurin et al. 2004) have also been revised after reviewing the new data from rice. Approximately 20% of the Arabidopsis and rice PPR models differ from the current gene models released by TAIR and TIGR, respectively. The most common errors found in the earlier gene models were insertions of extra introns, leading to the noninclusion of sequences encoding PPR motifs, and fusions to downstream exons of a neighboring gene.

The nomenclature of our final rice PPR gene models follows the schema OsPPR##g##### where the first 2-digit number indicates the chromosome and the second 5-digit number corresponds wherever possible to the equivalent number from the Osa1 gene model. Where this is not possible for models that have been split in 2 or lie entirely in a region predicted to be intergenic in the Osa1 annotation, an appropriate number lying between those of the adjacent Osa1 gene models has been chosen. The nomenclature of the Arabidopsis models follows the same logic except for the single digit chromosome number (AtPPR##g####). For Physcomitrella, PPR gene models are named PpPPR##, numbered sequentially. All gene models can be browsed using the “PPR Genome Browser” based on the Gbrowse software (Stein et al. 2002) at http://www.plantenergy.uwa.edu.au/applications/osatppr/index.html. This site also contains GFF files describing all gene models.

Sequence Comparisons
Clustal W (Thompson et al. 1994) was used for protein sequence alignments and for calculating distance trees using the Neighbor-Joining (NJ) method. All the figures in the paper were obtained using default “slow, accurate” parameters; variations in gap opening and extension parameters were tested but made only minor differences to the trees and did not affect any of the conclusions reached in this work. Distance trees were visualized with A Tree Viewer (ATV) (Zmasek and Eddy 2001) and drawn using a modified version of ATV to produce scalable vector graphics output.

Results and Discussion
The PPR Content of the Arabidopsis, Rice, and Moss Genomes

Complete sets of genes encoding PPR proteins in the genomes of Arabidopsis, rice, and moss were identified using techniques described in the Materials and Methods. Final results identified 450 PPR genes in Arabidopsis, 477 in rice, and 103 in moss. The raw numbers of PPR genes in these species are informative: moss diverged early in the evolution of land plants and the much smaller number of PPR genes encoded by this genome compared with those of Arabidopsis and rice suggests at face value that the bulk of the expansion of the PPR family occurred following the divergence of moss and the lineage leading to vascular plants. The number of PPR genes in Arabidopsis and rice are strikingly similar, particularly, so given that there are approximately twice as many predicted protein-coding genes in rice than Arabidopsis. We return to a discussion of this similarity later.
PPR genes can be divided into 4 subclasses based on their C-terminal domain structure and the presence of longer (L) or shorter (S) variant PPR motifs within the tandem arrays of the classic P PPR. Figure 1 displays the numbers of PPR genes divided by species and subclass. A large majority of moss PPR proteins belong to the P subclass, apart from a small set of DYW subclass proteins. No E subclass PPR genes are found in moss, compared with over 100 in Arabidopsis and rice. Note the similar number of genes in Arabidopsis and rice in the various PPR subclasses.

A peculiarity of PPR genes in Arabidopsis is that the large majority do not contain any introns (Lurin et al. 2004; Rivals et al. 2006). Most of the rice PPR genes are also intron less. Figure 2 displays the proportions of PPR genes in each species with no introns, 1 intron, 2–5 introns, and 6 or more introns. Approximately 80% of Arabidopsis and rice PPR genes are intron less. Once again, we find a striking similarity in the proportions of Arabidopsis and rice PPR genes in the intron number categories. In stark contrast, moss PPR genes are divided into these categories of intron content in roughly equal proportions.

Phylogenetic Comparison of PPR Proteins

The predicted protein sequences of all PPR genes in this study were aligned by ClustalW to produce a NJ tree, displayed in figure 3. Tree branches are colored by species, numbers of introns, PPR subfamily, and predicted targeting to organelles. The trees are available as high-resolution vectorial figures and in standard New Hampshire format as supplementary material (Supplementary Material online). It is evident from figure 3A that there are few species-specific clusters of PPR genes. In those regions of the tree dominated by Arabidopsis and rice genes, there generally appears an even mix of genes from both species. On the other hand, there is a clear separation between genes in the PPR subclasses (fig. 3C), which group in separate regions in the tree.

Targeting to plant organelles on the basis of N-terminal protein sequences for all PPR sequences was predicted by the program Predotar (Small et al. 2004) and as expected, a high proportion of the proteins were predicted to be targeted to mitochondria or chloroplasts. These are colored in figure 3D. We find that sequences do not broadly group according to targeting predictions in the phylogenetic tree, although small clusters of predicted mitochondrial or plastid proteins can be identified.

Evidence That the PPR Gene Family Expanded via Retrotransposition

One of the mechanisms of new gene formation in eukaryotes is retrotransposition, wherein a mature messenger RNA, associated with a retrotransposon, is reverse transcribed and integrated into the genome. For a recent review on the subject, see Babushok et al. (2007). Retrotranscribed copies of genes that originally contained introns are thus intron less. Noting the largely intron-less nature of PPR genes in Arabidopsis, Lurin et al. (2004) suggested that retrotransposition may be responsible for the expansion of the PPR gene family in higher plants. The results of the present study provide compelling evidence that this is indeed the
case. We find that the majority of PPR coding sequences in rice are also intron less, whereas the PPR genes of moss typically contain many introns (fig. 2). If retrotransposition was responsible for the expansion of the PPR gene family, the few PPR genes in Arabidopsis and rice with many introns would represent “ancient” PPR genes that predated, and provided the template for, the expanded number of intron-less genes. Support for this hypothesis is immediately revealed by inspection of figure 3A and B, where intron-rich Arabidopsis and rice PPR genes cluster among the intron-rich PPR genes of moss, which diverged early in the plant lineage. An example of a triplet of orthologous intron-rich Arabidopsis, rice, and moss PPR genes is displayed in figure 4A. In figure 4B and C, we also display examples of intron-less Arabidopsis and rice PPR genes along with their intron-containing orthologs in moss. These figures also contain paralogous pairs of moss and rice PPR genes arising from genome duplication events, to be discussed in a later section.

Further evidence of retrotransposon-mediated expansion of the PPR family is provided by the distribution of intron positions in intron-poor PPR genes. It has been noted that in eukaryotes, that in eukaryotes that are intron poor, introns are preferred by these authors and others (Lin and Zhang 2005; Roy and Gilbert 2005) that this bias is due to a mechanism of intron loss mediated by reverse transcription: cDNAs reverse transcribed from the 3′ polyadenosine end of mRNA molecules are generally truncated before the 5′ end. Subsequent homologous recombination of these molecules with genomic DNA would preferentially remove introns from the 3′ end of genes, resulting in the observed 5′ biased location of introns.

Relative intron positions, defined as the length of the ORF upstream of the start of the intron divided by the full length of the ORF, were calculated for all introns in PPR genes of Arabidopsis and rice. The distribution of all intron positions is uniform (data not shown); however, these data are dominated by the introns of the few intron-rich PPRs mentioned above and which, under the hypothesis of retrotransposition-mediated PPR family expansion, can be considered as “ancestral” genes predating the expansion. Following Sakurai et al. (2002), figure 5 shows the distribution of intron positions for those PPR genes with a single intron. There is a clear overrepresentation of introns in the 5′ end of these genes, as is observed for all introns in intron-poor eukaryotes, and for the introns of those genes with a single intron in several other intron-rich eukaryotes (Sakurai et al. 2002). In contrast, the distribution of intron position for all introns in Arabidopsis is uniform (Mourier and Jeffares 2003). Similar distributions to that in figure 5 are found for Arabidopsis, rice, and moss separately, suggesting that intron-poor moss PPR genes have also been generated by reverse transcription. The fact that there are over 100 moss PPR genes and that several are intron less and orthologous to Arabidopsis and rice PPR genes indicates that at least some of the retrotransposition-mediated expansion of the PPR gene family occurred prior to the divergence of moss and vascular plants.

The PPR Gene Family Expanded prior to the Monocot/Dicot Divergence

As noted earlier, there is a striking similarity in the number of PPR genes in Arabidopsis and rice. These similarities extend to the breakdown of the number of these genes by subclass (fig. 1), intron content (fig. 2), and are also reflected in the topology of the phylogenetic tree for PPR genes in figure 3A. An extraordinarily large proportion of outmost branches in the phylogenetic tree are pairs of probably orthologous Arabidopsis and rice PPR genes (e.g., AtPPR_5g27270 and OsPPR_06g02120 in fig. 4A). Bootstrap support for these pairs is very strong (generally 100%) and the branch lengths for all these pairs are similar, consistent with the idea that these pairs diverged from each other at roughly the same time, presumably the date of the last common ancestor of rice and Arabidopsis. Recently, several genome-wide phylogenetic analyses of other protein families from rice and Arabidopsis have been conducted. From these studies, pairs of proteins from the outmost branches of phylogenetic trees presented, with maximum bootstrap support, were identified. The proportion of orthologous Arabidopsis/rice protein pairs among all pairs in these studies are compared with that found for PPR proteins in figure 6. It is clear that the PPR protein family stands out in its exceptionally high degree of interspecies conservation of individual proteins. These data on the conserved number and nature of PPR genes in Arabidopsis and rice and their phylogenetic relationships provide
strong evidence that the complement of PPRs in these organisms existed prior to the divergence of monocots and dicots, with few examples of gain or loss of PPR genes since that event.

It should be noted that 2 regions of the phylogenetic tree in figure 3A do not conform to the general trend observed above and consist of groups of rice-specific and Arabidopsis-specific paralogs. These proteins in these regions are homologous to the restorer-of-fertility (Rf) genes mentioned in the Introduction and found in several plant species (Chase 2007). These genes cluster in chromosomes 1 and 10 of Arabidopsis and rice, respectively, and recently (Geddy and Brown 2007) have demonstrated that radish Rf genes have been subject to diversifying selection. The unusual evolutionary relationships of Rf genes and their functional implications will be treated elsewhere.

Effects of Ancient Genome Duplication Events

The history of plant genomes is one of duplications (Sterck et al. 2007). Over 70% of the Arabidopsis genome contains regions that are remnants of a genome duplication that occurred between 20 and 60 MYA, after the split of monocots and dicots (Blanc et al. 2003; Bowers et al. 2003). There is also evidence of earlier genome duplication events in the Arabidopsis lineage (Maere et al. 2005). Similarly, over 65% of the rice genome is composed of remnants of whole-genome or segmental duplications,
including a recent (ca. 8 MYA) segmental duplication between regions of chromosomes 11 and 12 (Yu et al. 2005), and recently, a genome duplication approximately 30–60 MYA has been detected for *Physcomitrella* using EST sequence data (Rensing et al. 2007). The approximate times of these duplication events are marked in figure 7.

The high proportion of orthologous *Arabidopsis* and rice pairs in our phylogenetic analysis indicates that the retention of new PPR genes following genome duplication is very rare in both species. Indeed, using the data of (Blanc et al. 2003) and (Yu et al. 2005) for *Arabidopsis* and rice, respectively, we find that 92% of the orthologous *Arabidopsis*/rice pairs in the phylogenetic analysis comprise at least one member located in a duplicated segment of its genome, indicating widespread loss of PPR genes following segmental or whole-genome duplications postdating the monocot/dicot divergence. The phylogenetic trees based on a comparison of *Arabidopsis* and rice genes in most other protein families (see, e.g., the references accompanying fig. 6) typically contain pairs of paralogous genes created by these genome duplications. Among the 26 paralogous pairs of rice PPRs with 100% bootstrap support in our phylogenetic analysis, we find 9 pairs on chromosomes 11 and 12 that can be attributed to the recent segmental duplication event between these regions (Yu et al. 2005). One of these pairs is displayed in figure 4C. In contrast to the trends found for the angiosperms, a total of 42 of the 103 *Physcomitrella* PPR proteins in the phylogenetic tree of figure 5 are present as paralogous pairs with 100% bootstrap support and branch lengths shorter than those of equivalent *Arabidopsis* and rice orthologous pairs. One of these pairs is displayed in figure 4B. These pairs presumably arose during the *Physcomitrella* genome duplication event detected by Rensing et al. (2007) of approximately 50 MYA. Thus, a much higher proportion of duplicate moss PPR genes are retained than is found for *Arabidopsis* or rice.

It has been demonstrated that the extent to which both pairs of genes formed by genome duplication are retained varies depending on their function (Blanc and Wolfe 2004). Rensing et al. (2007) noted that patterns of retention of duplicate moss genes based on functional categorization are markedly different than for seed plants. The extent to which the differing levels of duplicate PPR gene retention found in moss and the flowering plants is a reflection of a functional divergence of PPR genes in either set remains to be seen.

**Implications for PPR Function**

Several of the findings in this study are consistent with recent experimental evidence on the functions of PPR genes. The extraordinary conservation of the number of orthologous PPR genes from *Arabidopsis* and rice over hundreds of millions of years argues for their essentiality—something that is confirmed by gene knockout studies (Lurin et al. 2004; Cushing et al. 2005). This conservation also suggests that orthologous PPR proteins in these organisms carry out the same function in *Arabidopsis* and rice and, presumably, other flowering plants.

Numerous PPR proteins have been shown to bind avidly to RNA (reviewed in Delannoy et al. 2007) and at least some do so with a high degree of specificity in vitro and in vivo (Nakamura et al. 2003; Schmitz-Linneweber et al. 2005, 2006; Okuda et al. 2006). Domain-swapping experiments implicate the PPR motifs in RNA recognition (Okuda et al. 2007), and indeed, apart from a very few exceptions (Schmitz-Linneweber et al. 2006), PPR proteins...
do not contain other known RNA-binding motifs. The various extra domains associated with PPR proteins (generally C-terminal) are thought to determine the function of them once bound (splicing, cleavage, editing, RNA stability, enhancing, or blocking translation) (Lurin et al. 2004). These extra domains can be used to subdivide the PPR family as shown in figure 1 for the major classes. There is a notable difference in the relative proportions of PPR subfamilies among the PPR genes of the angiosperms and moss. For example, *Physcomitrella* does not contain any E subclass PPRs (fig. 1), whereas these are abundant in *Arabidopsis* and rice. To date, the molecular function of only 2 E-type PPR proteins is known, and both are implicated in RNA editing. CRR4 (Kotera et al. 2005) and CRR21 (Okuda et al. 2007) are required for editing of independent sites on *ndhD* transcripts in *Arabidopsis* chloroplasts and are thought to play a role as specificity factors, with the PPR repeats binding the target transcript while the E domain is needed to recruit the (as yet unknown) editing enzyme (Okuda et al. 2006, 2007). RNA editing is prevalent in most angiosperm organelles, with over 400 sites catalogued in *Arabidopsis* (Giege and Brennicke 1999) and rice (Notsu et al. 2002) mitochondria and a further 30 or more in plastids (Tsudzuki et al. 2001; Chateigner-Boutin and Small 2007). Editing in *Physcomitrella* is a much rarer event but does occur (Miyata et al. 2002; Miyata and Sugita 2004); approximately 9 sites are predicted between the 2 organelles based on their genome sequences (Sugitana et al. 2003; Terasawa et al. 2007). As *Physcomitrella* contains no E-type PPR proteins, something else must take this role in moss. The most likely candidates are the 10 related DYW domain proteins and indeed we have recently proposed, based on sequence similarity and phylogenetic distribution, that the DYW domain could carry the catalytic activity need for cytosine deamination (Salone et al. 2003). The correlation between the diversity of E and DYW PPR proteins and the number of editing sites in these 3 species is further support for this hypothesis.

**Conclusions**

Our study provides comprehensive genome-wide data on the content, nature, and evolutionary relationship of the PPR genes in 3 organisms: *Arabidopsis*, rice, and the moss *P. patens*. Here we have focused on the remarkable expansion of the family in plants and form 3 main conclusions: 1) that the expansion of the family is likely to have been mediated by retrotransposition; 2) that the expansion of the family occurred prior to the monocot/dicot divergence, 3) that since this time little gene loss or gain has occurred, implying considerable conservation of function; and 4) that the expansion of the E and DYW classes in rice and *Arabidopsis* is correlated with a similar striking increase in the number of RNA editing sites in these 2 species.

**Supplementary Material**

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe. oxfordjournals.org/) and at http://www.plantenergy.uwa.edu.au/applications/osatppr/index.html. GFF format files of all gene models used in this analysis. FASTA format files of all predicted protein sequences. NHX format file of the NJ tree. Adobe Illustrator files of the trees depicted in figure 3.

**Acknowledgments**

This work was supported by grants from the French Ministry of Education and Research, the French Australian S and T Programme (FR060030), the Australian Research Council (CE0561495), and the Western Australian State Government Centres of Excellence scheme.

**Literature Cited**


Geddy R, Brown GG. 2007. Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. BMC Genomics. 8:130.
Comparing the PPR Genes of Moss and Flowering Plants


pentatricopeptide repeat protein facilitates the trans-splicing of the maize chloroplast rps12 pre-mRNA. Plant Cell. 18:2650–2663.

Geoffrey McFadden, Associate Editor

Accepted February 26, 2008