Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution

Sota Fujii[a], Charles S. Bond[b], and Ian D. Small[a,1]

[a]Australian Research Council Centre of Excellence in Plant Energy Biology and [b]School of Biomedical, Biomolecular, and Chemical Sciences, University of Western Australia, Crawley, WA 6009, Australia

Edited by Michael Lynch, Indiana University, Bloomington, IN, and approved December 12, 2010 (received for review May 31, 2010)

Eukaryotic cells have harbored mitochondria for at least 1.5 billion years in an apparently mutually beneficial symbiosis. Studies on the agronomically important crop trait cytoplasmic male sterility (CMS) have suggested the semblance of a host–parasite relationship between the nuclear and mitochondrial genomes, but molecular evidence for this is lacking. Key players in CMS systems are the fertility restorer (*Rf*) genes required for the development of a functional male gametophyte in plants carrying a mitochondrial CMS gene. In the majority of cases, *Rf* genes encode pentatricopeptide repeat (PPR) proteins. We show that most angiosperms for which extensive genomic sequence data exist contain multiple PPR genes related to *Rf* genes. These *Rf*-like genes show a number of characteristic features compared with other PPR genes, including chromosomal clustering and unique patterns of evolution, notably high rates of nonsynonymous to synonymous substitutions, suggesting diversifying selection. The highest probabilities of diversifying selection were seen for amino acid residues 1, 3, and 6 within the PPR motif. PPR proteins are involved in RNA processing, and mapping the selection data to a predicted consensus structure of an array of PPR motifs suggests that these residues are likely to form base-specific contacts to the RNA ligand. We suggest that the selection patterns on *Rf*-like genes reveal a molecular "arms-race" between the nuclear and mitochondrial genomes that has persisted throughout most of the evolutionary history of angiosperms.

gene-for-gene relationships | nuclear–mitochondrial interaction | RNA-binding protein | adaptive evolution

The longest-lasting and most intimate symbiotic relationship known is that between eukaryotic cells and their mitochondria. It is widely accepted that eukaryotes obtained mitochondria via an ancient symbiosis with a bacterium (1) and generally assumed that the nuclear and mitochondrial genomes now share a common evolutionary path. However, it has been suggested that under some circumstances the nuclear and cytoplasmic genomes might be in conflict (2, 3). There is a fine line between parasitic and symbiotic relationships, but gene selection patterns within apparent host–symbiont pairs can indicate the exact nature of the relationship. Host–parasite interactions often comprise gene-for-gene relationships, wherein the host's resistance against the parasite's ability to cause disease is controlled by coevolving genes in each partner (4). In plants, the rapid evolution of leucine-rich repeat (LRR) *R* genes provides the best examples of such coevolutionary "arms races" (5, 6). Interactions in host-parasite systems are important forces driving genome diversity within species, and may even contribute to the formation of new species.

One of the most studied nuclear-mitochondrial interactions is cytoplasmic male sterility (CMS), wherein normally hermaphrodite plants fail to generate viable pollen and thus become effectively female. CMS has been widely observed in nature in over 150 diverse plant species (7–10). Quite apart from its importance in the genetics of natural populations of many plant species, CMS is an indispensable resource for commercial hybrid seed production in many crops (7–10). The CMS trait is carried by the mitochondrial genome and many different CMS genes have been identified in a range of disparate plant species (7, 8, 10–13). The products of CMS genes accumulate in the inner mitochondrial membrane, perturbing mitochondrial function, and possibly triggering premature programmed cell death during male gametogenesis (14). Theoretically, male-sterile plants can reallocate considerable resources from pollen development to seed production, thus potentially increasing female fertility and favoring dissemination of the maternally inherited mitochondrial genome (15, 16). In populations with CMS, nuclear genotypes carrying "restorer of fertility" (*Rf*) genes that can suppress the expression of the mitochondrial CMS genes are favored (8, 17–19). Each *Rf* gene prevents the expression of a single specific mitochondrial CMS gene (8, 17–19). Populations with polymorphic *Rf* alleles can display gynodioecy, where the population consists of a mixture of hermaphrodites and females. Gynodioecy is fairly common in flowering plants (observed in about 7% of the species), although modeling has shown that under most plausible conditions, the female state provides little or no advantage for nuclear genes (20), leading to the proposition that the mitochondrial and nuclear genomes are in conflict (3, 21). A key prediction of CMS theory is rapid fixation of both CMS genes and the corresponding *Rf* allele, which means that many populations will be carriers of both CMS and matched restorers despite being hermaphroditic (15). This theory is supported by the fact that most molecularly characterized CMS-*Rf* systems are in hermaphroditic taxa and are only revealed in wide crosses (8). Thus, the importance of *Rf* evolution is much broader than its role in gynodioecy. Despite the general acceptance of these theoretical models, there has been little or no hard evidence provided to test them. We reasoned that if the arms-race hypothesis was correct the selection pressures operating on the *Rf* genes involved should be visible in their patterns of sequence divergence.

In most cases, *Rf* genes encode pentatricopeptide repeat (PPR) proteins (17, 22–29), members of a large family of eukaryotic RNA-binding proteins required for many posttranscriptional processes in organelles (30). We will designate PPR proteins encoded by *Rf* genes as *Rf-PPRs* hereafter. *Rf-PPR*s are targeted to mitochondria, where they prevent the accumulation of the CMS-specific gene product (17–19, 28). *Rf-PPR* genes are present in clusters of similar *Rf-PPR-like* (*RFL*) genes in almost all cases (17–19, 31). To test our hypothesis of *Rf-PPR* adaptive evolution for CMS gene silencing, we have carried out a systematic analysis

Author contributions: S.F. and I.D.S. designed research; S.F. and C.S.B. performed research; S.F., C.S.B., and I.D.S. analyzed data; and S.F. and I.D.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: iansmall@cyllene.uwa.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1007667108/-/DCSupplemental.

of substitution rates in PPR genes across the entire family from 11 diverse angiosperm species.

Results

***RFL* Genes Have Diverged from a Single Origin.** *RFL* genes were identified from *Arabidopsis lyrata* (rock cress), *Arabidopsis thaliana* (thale cress), *Brachypodium distachyon* (purple false brome), *Glycine max* (soybean), *Mimulus guttatus* (common monkeyflower), *Oryza sativa* ssp. japonica (japonica rice), *Oryza sativa* ssp. indica (indica rice), *Populus trichocarpa* (poplar), *Sorghum bicolor* (sorghum), *Vitis vinifera* (grape), and *Zea mays* (maize) by the criteria described in *Materials and Methods*. No *RFL* genes were found in the sequenced genomes of *Carica papaya* (papaya), *Physcomitrella patens* (a moss), or *Selaginella moellendorffii* (gemmiferous spikemoss). Overall, 33 genes from *A. lyrata*, 26 from *A. thaliana*, 9 from *B. distachyon*, 41 from *G. max*, 13 from *M. guttatus,* 13 from *O. sativa* ssp. japonica, 14 from *O. sativa* ssp. indica, 20 from *P. trichocarpa*, 21 from *S. bicolor*, 10 from *V. vinifera*, and 5 from *Z. mays* were determined as *RFL* genes (Table S1). Including the four known *Rf-PPR* genes (*OsRf1a*, *OsRf1b*, *PhRfPPR592*, and *RsPPRB_Rf*) and three known *RFL* genes (*PhRfPPR591*, *RsPPRA*, and *RsPPRC*), 212 *RFL* genes were identified from 13 species.

All P-class PPR proteins including 212 *RFL* proteins were aligned and a tree based on sequence distance was generated (Fig. 1*A*). Most of the non-*RFL* PPR genes form orthologous clusters, suggesting that they were present before these species diverged. In contrast, *RFL* genes form species-specific paralogous clusters, indicating that these genes have extensively evolved since these species diverged (Fig. 1*B*). Despite the rapid evolution of *RFL* genes, cereal and dicot *RFL* genes form a single clade in this tree. To investigate this finding more accurately, we realigned 873 predicted mitochondrial P-class PPR protein sequences with 212 *RFL* sequences and constructed a phylogenetic tree using maximum-likelihood inference. This analysis provides strong support for the hypothesis that all *RFL* genes have an ancient common origin. Considering that no *RFL* genes were identified in *P. patens* (32) nor in *S. moellendorffii* (present study), we conclude that *RFL* genes appeared in the flowering plant lineage history after the divergence of lycophytes but before the divergence of monocots and dicots.

Rapid Evolution of *Arabidopsis RFL* Genes. A clear example of the rapid evolution of *RFL* genes comes from the comparison of *A. lyrata* and *A. thaliana*, two related species that diverged only 4 to 5 million years ago (33). *Arabidopsis RFL* genes are subdivided into three subgroups, which were designated A_1, A_2, and A_3 (Fig. 2). A_3 shows massive species-specific divergence and no clear orthologous relationships between any of the *RFL* genes, only large paralogous clusters. In contrast, in subgroup A_2, seven pairs of orthologous *RFL* genes are conserved between the two genomes (Fig. 2).

There was a clear correlation between each of the subgroups and the chromosomal location of the genes comprising them. Three of the four *thaliana* A_1 genes (*AtRFL2, AtRFL3*, and *AtRFL4*) are clustered on the long arm of chromosome 1 within a 143-kb region (nucleotides 4,183,066–4,326,197, also mentioned in ref. 34), whereas members of A_2 are dispersed widely across the genome (see AGI numbers in Table S1). All 11 members of A_3 are within a 332-kb region of the short arm of chromosome 1 (nucleotides 23,176,930–23,509,053, also mentioned in ref. 34). Most of the *A. lyrata RFL* genes in A_1 are located within a 221-kb region of scaffold_1 (nucleotides 4,996,673–5,295,534), and all of the *A. lyrata* members in A_3 are located within a 4.4 Mb region of scaffold_2 (nucleotides 628,493–5,039,588), mirroring the situation in *A. thaliana*. The ancestor of *Arabidopsis* experienced three major genome duplication events (35), and the regions surrounding *RFL* clusters A_1 and A_3 are considered as duplicate copies arising from one of these events (36). The conservation of these duplicated genes is unusual; conservation of duplicated non-*RFL* PPR genes is rare (32).

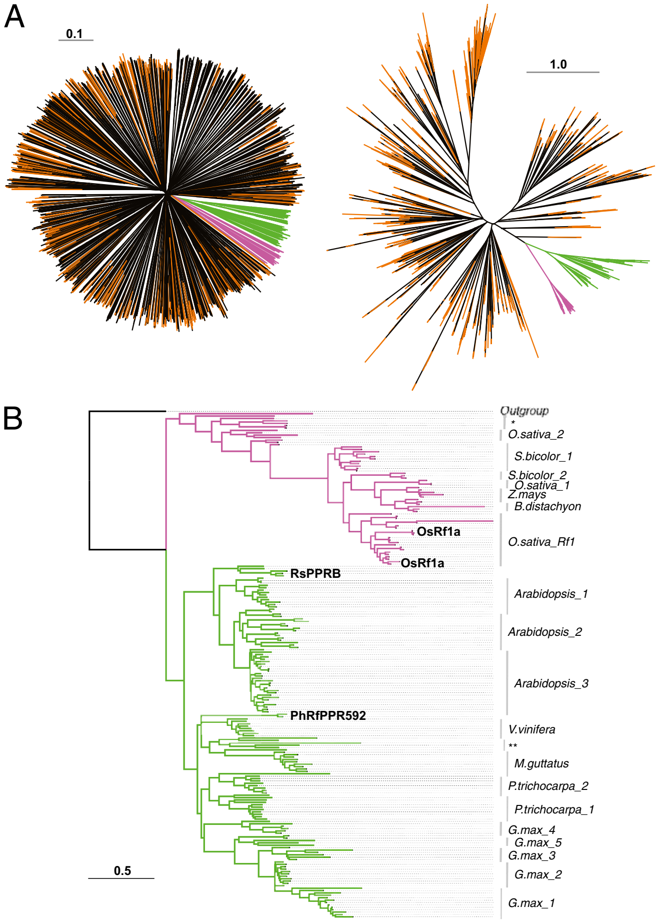

Fig. 1. Phylogenetic relationships of RFL proteins. Branch length represents the estimated rate of amino acid substitution. (*A*) Sequence distance neighbor-joining tree of 2409 P-class PPR proteins from *A. lyrata*, *A. thaliana*, *B. distachyon*, *G. max*, *M. guttatus*, *O. sativa* ssp. japonica, *P. trichocarpa*, *S. bicolor*, *V. vinifera*, and *Z. mays* (*Left*) compared with a phylogeny based on maximum-likelihood inference of 1,085 P-class PPR proteins predicted to be targeted to mitochondria (including 212 RFL proteins) (*Right*). Orange indicates predicted mitochondrial PPR proteins; magenta indicates monocot RFL proteins; green indicates dicot RFL proteins. Bootstrap support for the monophylogeny of the *RFL* clade was 86/100. (*B*) An expansion of the maximum-likelihood tree to show details of the phylogeny of the 212 RFL proteins. Monospecific clades are indicated by the species name. *A mixed cluster of *Rf* genes from *B. distachyon*, *O. sativa*, and *S. bicolor*. **A mixed cluster of *Rf* genes from *V. vinifera* and *M. guttatus*. Known Rf proteins are labeled in bold.

Diversifying Selection on *RFL* Genes. Nonsynonymous versus synonymous nucleotide substitution ratio (dN/dS), frequently used as a marker for diversifying selection, was used to evaluate the selection acting on *RFL* genes. Protein-encoding sequences with a high dN/dS ratio (>1) are considered to exhibit diversifying selection (positive selection for variability at some sites), whereas a low dN/dS ratio (<1) indicates purifying (negative) selection. To calculate the probabilities of diversifying selection, dN/dS ratios within each paralogous *RFL* sequence set were compared with predictions from two types of codon substitution models: those that only assume purifying or neutral evolution (the null hypothesis) or those that also allow for diversifying selection (37, 38)

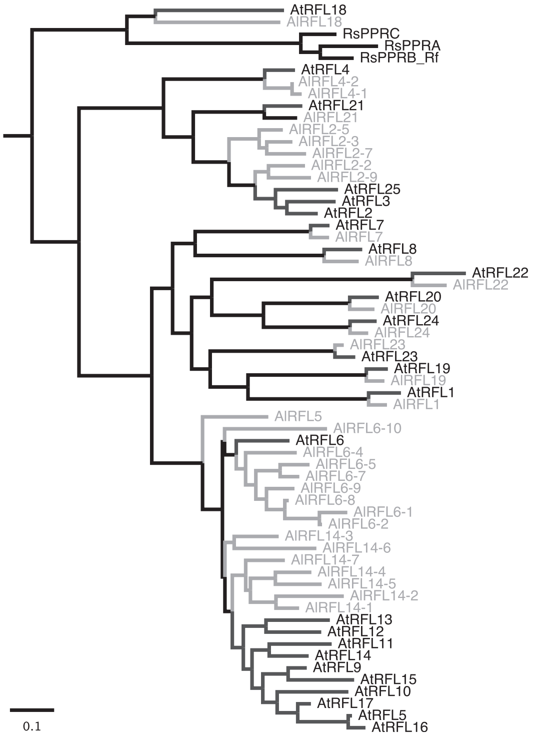

Fig. 2. Maximum-likelihood method based phylogeny of 59 *Arabidopsis* RFLs and three *R. sativus RFLs*. Petunia *Rf-PPR* and rice *OsRf1a* were chosen as outgroups. Gray edges are derived from *A. lyrata*, and black edges are derived from *A. thaliana*, respectively. Branch length represents the estimated amino acid substitution rate.

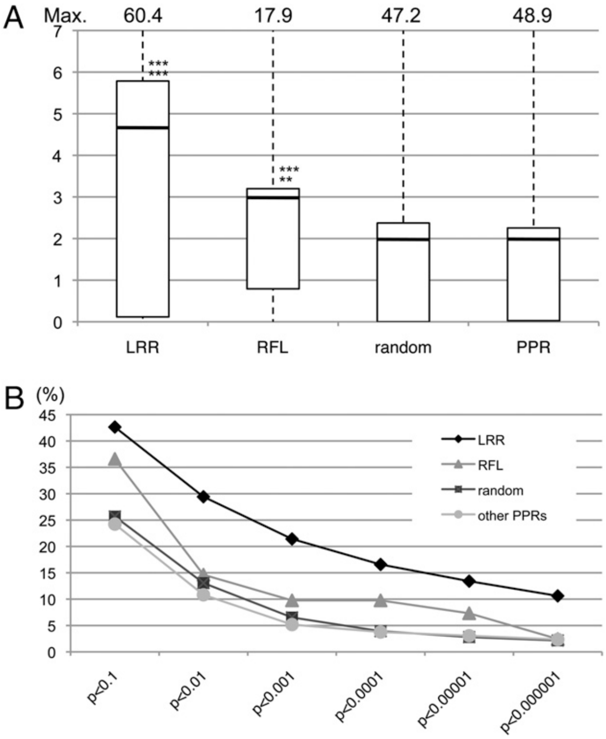

Fig. 3. Diversifying selection in *RFL* genes. (*A*) Mean difference plots of likelihood scores (codon substitution models M8–M7, where M8 includes the possibility of diversifying selection but M7 does not). Random is a dataset derived from 3,539 randomly chosen *Arabidopsis* genes (with PPR or LRR genes removed); PPR, non-*RFL* pentatricopeptide repeat genes. Horizontal bars indicate means, and boxes include 50% of the distributions. Upper asterisks indicate the comparison with the random gene set, lower asterisks indicate the comparison with non-*RFL* PPR genes (PPR): $**P < 0.005$; $***P < 0.001$. (*B*) Cumulative proportions of genes that fit M8 better than M7 at different *P* values.

(see *Materials and Methods* for details). Likelihood ratio tests were used to measure which of the models best fitted the observed values (37, 38).

In all species with *RFL* genes, strong indications of diversifying selection (at worst $P = 0.00174$ in *B. distachyon*) were observed for genes in at least one subclade of *RFL* genes (Tables S2 and S3). In *Arabidopsis*, the highest probabilities of diversifying selection were seen in subgroup A_3, which is also the most recently diverged between *A. thaliana* and *A. lyrata* (Fig. 2). Overall, 0.3 to 32.4% of sites were predicted to be under positive selection in paralogous *RFL* clades (Table S3).

To estimate whether this degree of apparent diversifying selection was exceptional for plant genes, we compared the probability of diversifying selection within *RFL* genes to that observed within LRR genes [including *R* genes implicated in pathogen defense and known to exhibit extremely high rates of diversifying selection (5)], within non-*RFL* PPR genes, and within a large set of genes selected at random (Fig. 3). LRR genes showed the highest frequency of diversifying selection (Fig. 3*A*) as expected, but a significantly greater proportion of *RFL* genes showed diversifying selection than that observed in non-*RFL* PPR genes (Fig. 3*A*) ($P < 0.005$). Thus, about 10% of *RFL* genes show high probabilities of diversifying selection compared with most other genes in the genome, and in particular, compared with other PPR genes.

Residues Under Diversifying Selection Selected in *RFL* Genes Are Probably in Contact with the RNA Ligand. *RFL* genes are comprised primarily of tandem arrays of 15 to 20 PPR motifs (each composed of 35 amino acids). We were curious to see if there is a preference in the position of positively selected sites within a PPR motif. We mapped the mean Bayes-Empirical-Bayes (BEB) probability of positive selection at each amino acid residue (Fig. 4) and found strikingly high scores at residues 1, 3, and 6 of the motif. We carried out the same calculations using 3,520 PPR motifs from non-*RFL* PPR genes (Fig. 4). BEB-positive selection probabilities at nine residues were significantly ($P <$ 1.0e-6) higher in *RFL* genes compared with that of non-*RFL* PPR genes (Fig. 4). In particular, the probability of diversifying selection at residues 1, 3, and 6 was 5 to 15 times higher in *RFL* genes. As an example, two of the four amino acid substitutions in a recessive nonfunctional allele of *Rfk* in radish CMS are located

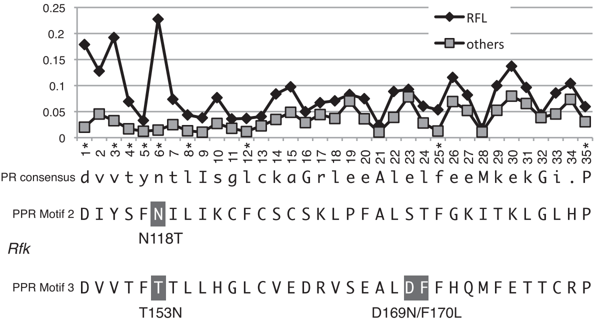

Fig. 4. BEB probabilities of positive selection mapped onto PPR motifs. The line chart displays mean positive-selection probabilities at each amino acid position within the PPR motif. Asterisks indicate residues where the probabilities in *RFL* genes are significantly higher than that of non-*RFL* (others) PPR genes (Wilcoxon-Mann-Whitney test, $P <$ 1.0e-6).

at residue 6 (26), providing evidence that this residue is indeed important for restorer function.

No experimentally determined structures are available for PPR motifs, but we have built a consensus structural model based on predicted structural contacts from amino acid covariance data. Amino acid residues 1, 3, and 6 are predicted to be located on the internal face of the first helix of the motif (Fig. 5). By analogy with proteins thought to form similar structures [e.g., PUF domain proteins (39)], this helix is likely to form the ligand-binding face of the PPR motif array. PPR proteins are sequence-specific RNA-binding proteins (40, 41), although how they discriminate between sequences is still only partially understood (42). The identification of residues 1, 3, and 6 as the most likely specificity-determining contact points was used to constrain modeling of the likely interactions with the RNA ligand (Fig. 5). The properties of the model establish the plausibility of an approximate 1:1 correspondence of bases to PPR motifs. The model reveals no contradictions between the predicted PPR structure and the hypothesis that residues 1, 3, and 6 help determine RNA-binding specificity.

Discussion

We have shown that *RFL* genes can be readily distinguished from the bulk of other PPR genes in plant genomes. Although previous studies pointed out the shared similarities of *Rf-PPR* genes (25, 28, 31), a rigorous demonstration of the monophyly of angiosperm *RFL* genes was lacking. That the majority of known *Rf* genes come from the same small clade of PPR genes among the much larger family of similar genes might seem surprising, given the apparent functional similarities between Rf proteins and other PPR proteins. Binding of PPR proteins to mRNAs influences RNA processing and translation (30, 40, 41). The mode of action of Rf proteins appears to be identical: Rf-PPR592 of *Petunia hybrida* interacts with the transcript of the CMS gene *pcf* (43), *Raphanus sativus* PPRB interacts with the transcripts from the CMS gene *orf138* (19) and *O. sativa* Rf1a physically interacts with the 5′ UTR sequence of the CMS gene *orf79* (18); each of these interactions affect processing and translation of these CMS transcripts. It is important to note that the various apparent activities of PPR proteins (including Rf proteins) may well all result from simple passive binding to their RNA target, thus altering RNA secondary structure or access by other proteins (e.g., see model proposed for transcript end formation in ref. 41).

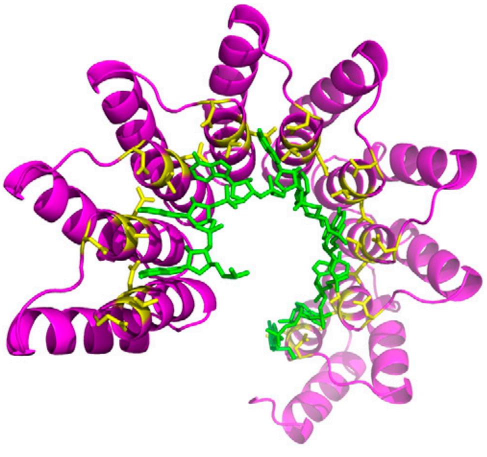

Fig. 5. Amino acid residues 1, 3, and 6 mapped onto a predicted array of consensus PPR motifs. Nine modeled PPR motifs are displayed (magenta), with the sidechains of residues 1, 3, and 6 highlighted in yellow. The modeled path of a poly(A) RNA ligand is indicated in green showing physically plausible proximity between the RNA bases and residues 1, 3, and 6 of the PPR motifs.

In this context it is clear that the molecular action of a PPR protein (including Rf and RFL proteins) is determined by which RNA sequence it binds; thus new *Rf* alleles are expected to arise by mutations that alter target recognition. We presume that such mutations are more likely to arise in *RFL* genes than other PPR genes because of the clustering of *RFL* alleles that promotes unequal crossovers and gene conversion. The linear, modular structure of PPR proteins (Fig. 5) implies that motif duplication, deletion, or exchange by recombination can all give rise to functional variants with altered target recognition. The expectation that *Rf* alleles evolve primarily by protein-coding changes as opposed to changes in expression levels or patterns must hold for the analysis of diversifying selection undertaken here to be informative. The fact that strong signals for diversifying selection were observed indicates that the starting expectations are probably correct.

The ability to distinguish *RFL* genes from the hundreds of other PPR genes in the genome is of potential utility in identifying candidate *Rf* genes. Furthermore, the analysis of selection patterns within *RFL* genes should indicate which genes have been under recent diversifying selection and therefore are most likely to be active restorer genes. The analyses used here are applicable to any plant genome sequence and should accelerate the molecular cloning of *Rf-PPR* genes for use in hybrid breeding programs. Given the shared and distinctive evolutionary behavior of all *RFL* genes and the known functional similarities of *Rf* genes in cereals and dicots, this finding strongly suggests that *Rf* genes and CMS have coexisted since before the monocot-dicot split. In other words, CMS-driven reproduction systems such as gynodioecy have probably been a core reproductive strategy in flowering plants for at least 120 million years.

The species used in this study include examples of contrasting reproductive strategies among the angiosperms: dioecious (*P. trichocarpa*, *C. papaya*), predominantly autogamous hermaphoditic (*A. thaliana*, *B. distachyon*, *O. sativa*, *G. max*), or predominantly allogamous hermaphoditic (*A. lyrata*, *S. bicolor*, *Z. mays*, *V. vinifera*, *M. guttatus*). One would expect the type of reproductive strategy to influence selection on CMS and *Rf* alleles, as only in allogamous hermaphrodites are there theoretical advantages to either (15, 16). In this context it is interesting that we were unable to identify *RFL* genes from *C. papaya*, a primarily dioecious species with defined sex chromosomes (44). However, the reported *C. papaya* genome sequence only contains about 75% of the genome (90% of the euchromatic regions) (45) and we cannot rule out that one or more *RFL* clusters lie in unsequenced regions.

It is not immediately obvious what the function of the *RFL* genes might be in *A. thaliana* or other autogamous species (we are unaware of any reported observation of CMS in natural populations of *A. thaliana*). The only one of these genes for whom a function has been ascribed is *RNA PROCESSING FACTOR 2* (*RPF2*) (46), identical to *AtRFL6* in this study. *RPF2* promotes the 5′ end formation of mRNA from two mitochondrial genes, *nad9* and *cox3* (46). Thus, *RFL* genes in *A. thaliana* may have been recruited to participate in functions other than suppressing CMS. It should also be noted that despite the fact that *O. sativa* is a strongly autogamous like *A. thaliana*, several *Rf*/CMS systems are known in *O. sativa* from backcrossing of distantly related cultivars. Most autogamous species have recent allogamous ancestors (47) and thus, are potential carriers of cryptic *Rf*/CMS systems.

The presence of *RFL* genes in *P. trichocarpa* is harder to explain. The role of these genes is unlikely to pertain to CMS as it can have no advantage in an exclusively dioecious species (48). Data on *RFL* genes from other members of the Salicaceae would be interesting to see if the molecular evolution of *RFL* genes in the family differs from that in other plant families. One possibility is that *RFL* genes have a general role in suppressing gene expression from the many small, nonconserved ORFs present in the large and continually rearranging plant mitochondrial genomes (30).

The RNA targets of non-*RFL* proteins are known to be highly conserved (49, 50), as mutations in the target, which disrupt protein binding, will lead to problems with organelle biogenesis and function, leading to decreased reproductive fitness for both the organelle genome and the nuclear genome in the same cell. Natural selection is thus likely to lead to purifying selection on both the RNA target and the codons that determine RNA specificity within the corresponding PPR gene, exactly as we observed for non-*RFL* genes. In contrast, if the hypothesis that CMS has a selective advantage for the transmission of mitochondrial genome is correct (2, 21), the RNA targets of *Rf* genes are expected to be under selection for variability, in an analogous manner to pathogen avirulence genes. This process in turn will drive positive selection on the codons determining RNA-binding specificity within the *Rf* genes. We propose that this result is exactly what we have shown; the positive-selection signal within *RFL* genes is not only the "smoking gun" showing that the nuclear genome is truly in an arms-race against its own mitochondria, it is also telling us which residues in *RFL* proteins determine target specificity.

Previous reports that PPR genes show signs of positive selection either did not consider *RFL* genes (51) or did not clearly distinguish *RFL* genes from the rest (the vast majority) of the PPR family (34). These previous studies were also limited to a few genes in one or two species, and made no attempt to draw functional inferences from the patterns of nucleotide changes they observed. We feel that the differential rates of divergence between PPR genes and the highly position-specific signal in the divergence rates are vital clues to the function of these proteins. It is quickly apparent that these results will be useful for guiding future experimentation into target recognition by PPR proteins. Given the crucial roles that PPR proteins play in all eukaryotes with organelle genomes, an understanding of the mechanism by which they recognize their RNA targets will have a wide impact.

Materials and Methods

Sequences and Alignment. Identification of PPR sequences for *A. thaliana* and *O. sativa* ssp. japonica was described previously (32). Amino acid sequences and nucleotide sequences for the coding regions for *A. lyrata* (v1.0) and *P. tricocharpa* (v1.1) (52) were retrieved from gene models available at the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/). Gene models available from the Genoscope Web site (http://www.genoscope.cns.fr/spip/Vitis-vinifera-whole-genome.html) were used for *V. vinifera* (v1.0) (53). Phytozome v4.0 (http://www.phytozome.net/index.php) was the source for *B. distachyon* (54), *C. papaya* (45), *G. max* (55), *M. guttatus*, *S. moellendorffii* (v1.0), *S. bicolor* (v1.4) (56), *Z. mays* (57), and lastly, assembled shotgun contigs available at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) were used to identify *RFL* genes from *O. sativa* ssp. indica.

To identify as many likely *RFL* genes within these assembled genomes as possible, BLAST searches (http://www.ncbi.nlm.nih.gov/Tools/) were used to pick up sequences overlooked in the public annotation. The sequence of *Rf1a* from *O. sativa* (28) was used for BLAST searches against monocot genomes, *PPRB* (24) from *R. sativa* was used for BLAST searches against the *A. lyrata* genome and *Rf-PPR592* (17) from *P. hybrida* was used for BLAST searches against other dicot genomes. Hits with an estimated E-value under 1e^{-100} were considered as significant. Putative genes identified in this way were modeled using GlimmerHMM (58) and six-frame translation, as described previously (32). This manual annotation procedure was primarily required for *A. lyrata* and *O. sativa* ssp. indica.

Detection of PPR motifs within gene models was performed as described previously (32). To obtain high-quality alignments, genes that were predicted to encode 6 or PPR motifs were selected, and regions outside the PPR motifs were discarded, thus ensuring that the analysis was only influenced by sequence divergence within the PPR region. LRR proteins were identified by hmmsearch (http://www.csb.yale.edu/userguides/seq/hmmer/) using the LRR_1 hmm matrix (PF00560.25) from the Pfam database (59). MUSCLE v3.6 (60) was used for DNA and protein sequence alignments throughout the study, followed by manual refinement of the alignments to make sure that gaps corresponded to entire PPR motifs. Mis-annotated gene models with incorrect start codons, gene fusions, or truncations were manually corrected or removed. Programs included in the EMBOSS v6.0.0 package (http://emboss.sourceforge.net/index.html) and in-house PERL scripts were used to handle the sequence information.

Phylogenetic Analysis. ClustalW v1.8.3 (61) was used to generate distance matrix-based neighbor-joining trees. RAxML v7.0.4 (62) was used to conduct maximum-likelihood inference based phylogeny construction. Bootstrap analysis was performed to search for the best maximum-likelihood scoring tree. For Fig. 1*B*, At5g61990 was chosen as the outgroup as it was the closest *A. thaliana* non-*RFL* gene in Fig. 1*A*. Phylogenetic trees were depicted using FigTree v1.2.2 (http://tree.bio.ed.ac.uk/).

Establishing Gene Sets for PAML Analysis. Paralogous *RFL* gene sets were constructed based on their phylogeny (Fig. 1*B*; subdivision listed in Table S1). To establish a comprehensive set of non-*RFL* PPR genes, the 424 non-RFL PPR genes from *A. thaliana* were matched against *A. lyrata*, *P. trichocarpa*, and *V. vinifera* using BLAST, and the top hits retained as putative orthologs. To make a direct comparison against non-*RFL* PPRs (Fig. 3), the same BLAST-based ortholog selection strategy was chosen to identify quartets for 26 *A. thaliana RFL*s. The same method was also used to identify orthologs for 537 LRR genes and 3,539 randomly chosen genes that do not include LRR and PPR motifs.

Detection of Positive Selection. For calculation of dN/dS ratios, codons were aligned using PAL2NAL v12.1 (63). The dN/dS ratio was calculated with the four codon substitution models M1, M2, M7, and M8 in codeml from PAML v4.2 (37, 38). M1 is a neutral model in which two discrete categories of codon substitutions are assumed, dN/dS = 0 (purifying selection) or dN/dS = 1 (neutral evolution). M2 is a positive-selection model with an extra category with dN/dS > 1. M7 is a neutral model but unlike M1, a continuous β distribution of dN/dS values is assumed. M8 matches M7, except that it allows dN/dS > 1. Therefore M1 and M7 can be used as the null hypothesis against M2 and M8, respectively. If M1 or M7 can be rejected in favor of M2 or M8, the set of genes is likely to possess codons under positive selection. See Yang et al. (38) for further details of the codon substitution models. The fitness of codon substitution models was evaluated with likelihood ratio statistics as described in Yang et al. (38). The Wilcoxon-Mann-Whitney test was applied to compare the mean differences of likelihood values between M7 and M8. Positively selected sites were identified under M8 with BEB statistics implemented in codeml (38). Mean BEB probabilities for each codon mapped to a PPR motif consensus were calculated from 151 PPR motifs taken from *RFL* genes and *Rf-PPR* genes from 11 species, and 3,520 PPR motifs taken from 424 non-*RFL* PPR genes.

Development of a Predicted Structural Model. The structure of a consensus protein (10 tandem motifs) was derived from sequence covariation data obtained from a sequence alignment of 8,068 PPR motifs. Pairs of covarying residues were restrained to be close in space in a distance-geometry minimization carried out with X-PLOR-NIH (64) using a solely α-helical starting model. Coordinates for an RNA decamer with arbitrary sequence (rA10) were generated, and a distance-geometry minimization was performed where the N1 atom of each adjacent base was restrained to adopt a position close (4-6 Å) to the sidechain atoms of residues 1, 3, and 6 of adjacent PPR motifs.

ACKNOWLEDGMENTS. We thank David Schibeci of the Australian Bioinformatics Facility at Murdoch University for access to high-performance computing facilities and Profs. Christopher Leaver and Kinya Toriyama for helpful comments on the manuscript.

1. Sagan L (1967) On the origin of mitosing cells. *J Theor Biol* 14:255–274.
2. McCauley DE, Olson MS (2008) Do recent findings in plant mitochondrial molecular and population genetics have implications for the study of gynodioecy and cytonuclear conflict? *Evolution* 62:1013–1025.
3. Touzet P, Budar F (2004) Unveiling the molecular arms race between two conflicting genomes in cytoplasmic male sterility? *Trends Plant Sci* 9:568–570.
4. Flor HH (1971) Current status of the gene-for gene concept. *Annu Rev Phytopathol* 9: 275–296.

5. Bergelson J, Kreitman M, Stahl EA, Tian D (2001) Evolutionary dynamics of plant R-genes. *Science* 292:2281–2285.
6. Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res* 8:1113–1130.
7. Chase CD (2007) Cytoplasmic male sterility: A window to the world of plant mitochondrial-nuclear interactions. *Trends Genet* 23:81–90.
8. Hanson MR, Bentolila S (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. *Plant Cell* 16 (Suppl):S154–S169.
9. Pelletier G, Budar F (2007) The molecular biology of cytoplasmically inherited male sterility and prospects for its engineering. *Curr Opin Biotechnol* 18:121–125.
10. Schnable PS, Wise RP (1998) The molecular basis of cytoplasmic male sterility and fertility restoration. *Trends Plant Sci* 3:175–180.
11. Dewey RE, Siedow JN, Timothy DH, Levings CS, 3rd (1988) A 13-kilodalton maize mitochondrial protein in E. coli confers sensitivity to *Bipolaris maydis* toxin. *Science* 239:293–295.
12. Iwabuchi M, Kyozuka J, Shimamoto K (1993) Processing followed by complete editing of an altered mitochondrial atp6 RNA restores fertility of cytoplasmic male sterile rice. *EMBO J* 12:1437–1446.
13. Nivison HT, Hanson MR (1989) Identification of a mitochondrial protein associated with cytoplasmic male sterility in petunia. *Plant Cell* 1:1121–1130.
14. Balk J, Leaver CJ (2001) The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release. *Plant Cell* 13:1803–1818.
15. Charlesworth D, Ganders FR (1979) The population genetics of gynodioecy with cytoplasmic-genic male-sterility. *Heredity* 43:213–218.
16. Frank SA (1989) The evolutionary dynamics of cytoplasmic male sterility. *Am Nat* 133: 345–376.
17. Bentolila S, Alfonso AA, Hanson MR (2002) A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proc Natl Acad Sci USA* 99: 10887–10892.
18. Kazama T, Nakamura T, Watanabe M, Sugita M, Toriyama K (2008) Suppression mechanism of mitochondrial ORF79 accumulation by Rf1 protein in BT-type cytoplasmic male sterile rice. *Plant J* 55:619–628.
19. Uyttewaal M, et al. (2008) Characterization of *Raphanus sativus* pentatricopeptide repeat proteins encoded by the fertility restorer locus for Ogura cytoplasmic male sterility. *Plant Cell* 20:3331–3345.
20. Lewis D (1941) Male sterility in natural populations of hermaphrodite plants. *New Phytol* 40:56–63.
21. Delph LF, Touzet P, Bailey MF (2007) Merging theory and mechanism in studies of gynodioecy. *Trends Ecol Evol* 22:17–24.
22. Akagi H, et al. (2004) Positional cloning of the rice *Rf-1* gene, a restorer of BT-type cytoplasmic male sterility that encodes a mitochondria-targeting PPR protein. *Theor Appl Genet* 108:1449–1457.
23. Brown GG, et al. (2003) The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. *Plant J* 35: 262–272.
24. Desloire S, et al. (2003) Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family. *EMBO Rep* 4:588–594.
25. Kazama T, Toriyama K (2003) A pentatricopeptide repeat-containing gene that promotes the processing of aberrant atp6 RNA of cytoplasmic male-sterile rice. *FEBS Lett* 544:99–102.
26. Koizuka N, et al. (2003) Genetic characterization of a pentatricopeptide repeat protein gene, orf687, that restores fertility in the cytoplasmic male-sterile Kosena radish. *Plant J* 34:407–415.
27. Komori T, et al. (2004) Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa* L.). *Plant J* 37:315–325.
28. Wang Z, et al. (2006) Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. *Plant Cell* 18:676–687.
29. Klein RR, et al. (2005) Fertility restorer locus *Rf1* [corrected] of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12. *Theor Appl Genet* 111:994–1012.
30. Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: A socket set for organelle gene expression. *Trends Plant Sci* 13:663–670.
31. Barr CM, Fishman L (2010) The nuclear component of a cytonuclear hybrid incompatibility in Mimulus maps to a cluster of pentatricopeptide repeat genes. *Genetics* 184:455–465.
32. O'Toole N, et al. (2008) On the expansion of the pentatricopeptide repeat gene family in plants. *Mol Biol Evol* 25:1120–1128.
33. Kuittinen H, Aguadé M (2000) Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana*. *Genetics* 155:863–872.
34. 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.
35. De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20:591–597.
36. Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438.
37. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
38. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
39. Wang X, McLachlan J, Zamore PD, Hall TM (2002) Modular recognition of RNA by a human pumilio-homology domain. *Cell* 110:501–512.
40. Delannoy E, Stanley WA, Bond CS, Small ID (2007) Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochem Soc Trans* 35:1643–1647.
41. Pfalz J, Bayraktar OA, Prikryl J, Barkan A (2009) Site-specific binding of a PPR protein defines and stabilizes 5′ and 3′ mRNA termini in chloroplasts. *EMBO J* 28:2042–2052.
42. Hammani K, et al. (2009) A study of new *Arabidopsis* chloroplast RNA editing mutants reveals general features of editing factors and their target sites. *Plant Cell* 21: 3686–3699.
43. Gillman JD, Bentolila S, Hanson MR (2007) The petunia restorer of fertility protein is part of a large mitochondrial complex that interacts with transcripts of the CMS-associated locus. *Plant J* 49:217–227.
44. Vyskot B, Hobza R (2004) Gender in plants: Sex chromosomes are emerging from the fog. *Trends Genet* 20:432–438.
45. Ming R, et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991–996.
46. Jonietz C, Forner J, Hölzle A, Thuss S, Binder S (2010) RNA PROCESSING FACTOR2 is required for 5′ end processing of nad9 and cox3 mRNAs in mitochondria of *Arabidopsis thaliana*. *Plant Cell* 22:443–453.
47. Eckert CG, et al. (2010) Plant mating systems in a changing world. *Trends Ecol Evol* 25: 35–43.
48. Wade MJ, Goodnight CJ (2006) Cyto-nuclear epistasis: Two-locus random genetic drift in hermaphroditic and dioecious species. *Evolution* 60:643–659.
49. Johnson X, et al. (2010) MRL1, a conserved Pentatricopeptide repeat protein, is required for stabilization of rbcL mRNA in *Chlamydomonas* and *Arabidopsis*. *Plant Cell* 22:234–248.
50. Okuda K, Habata Y, Kobayashi Y, Shikanai T (2008) Amino acid sequence variations in Nicotiana CRR4 orthologs determine the species-specific efficiency of RNA editing in plastids. *Nucleic Acids Res* 36:6155–6164.
51. Foxe JP, Wright SI (2009) Signature of diversifying selection on members of the pentatricopeptide repeat protein family in *Arabidopsis lyrata*. *Genetics* 183:663–672, 1SI—8SI.
52. Tuskan GA, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–1604.
53. Jaillon O, et al. (2007) French-Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
54. Vogel JP, et al.; International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768.
55. Schmutz J, et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183.
56. Paterson AH, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–556.
57. Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326:1112–1115.
58. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20:2878–2879.
59. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36 (Database issue):D281–D288.
60. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
61. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
62. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
63. Suyama M, Torrents D, Bork P (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34 (Web Server issue):W609–W612.
64. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* 160:65–73.