Interspecific Diversity of a Restorer-of-Fertility-Like PPR Gene Cluster within the Maize NAM Founders

Jen Jaqueth

Introduction

Cytoplasmic male sterility, CMS, is a naturally occurring phenomenon found in many plant species caused by an incompatibility between the mitochondrial and nuclear genomes leading to accumulation of CMS products in the inner mitochondrial membrane, resulting in cell death. *Restorer-of-fertility-like* (*Rfl*) genes coded in the nuclear genome are responsible for reducing the accumulation of the CMS products. Frequently, these *Rfl* genes are pentatricopeptide repeat (PPR) encoded proteins, a class of RNA binding proteins involved in post-transcriptional processing in the mitochondria and chloroplast. PPR proteins contain tandem repeats of degenerate 35 amino acids motifs and are divided into two classes, the PLS-class and the P-class. The PLS-class of PPRs may have more variability in motif length, and this class of PPR is often involved in C-to-U RNA editing. The P-class of PPRs has only motifs 35 amino acids in length, may be involved in intron splicing, RNA folding or passive RNA binding (Gutmann, Royan et al. 2019). RFL proteins belong to the P-class of PPRs (Gaborieau, Brown et al. 2016).

Rfl PPR genes have characteristics which distinguish them from other PPR genes. Within the PPR motif, amino acids in positions 5 and 35 interact the RNA strand and are responsible for RNA-binding specificity (Figure1) (Manna 2015). The PPRCode of each RFL is the 5^{th} and 35^{th} amino acid of each repeat, which taken together targets a specific RNA sequence. RFL proteins have higher rates of diversifying selection, specifically at the 5^{th} and 35^{th} positions in the motif, which allow them to evolve more quickly to recognize different RNA target sequences (Dahan and Mireau 2013). *Rfl* sequences also are frequently found in gene clusters, which promotes rapid evolution of paralogous sequences. RFL PPR proteins match the PPR consensus better than other PPR proteins. Finally, RFL proteins have a high number of repeats, between 15 and 20 repeats (Fujii, Bond et al. 2011).

The *Zea mays,* maize, genome contains a significant RFL gene cluster on the long arm of chromosome 2. This cluster contains the *Rf3* gene, the major dominant restorer for CMS-S, found in the

NAM inbred Ky21 (Kamps and Chase 1997, Xu, Liu et al. 2009). It also contains *Rf8* for CMS-T (Meyer, Pei et al. 2011). In certain inbreds, this chromosome 2 cluster also contains an *Rf* gene for CMS-C cytoplasm. For this study, the chromosome 2 RFL cluster was analyzed within the set of 26 NAM lines, a diverse set of maize inbreds with fully sequenced genomes (McMullen, Kresovich et al. 2009). The phylogenetic relationship between these RFL sequences was studied, and the RFL sequences were evaluated for the presence of diversifying selection. Finally, the PPRCode and RFL gene content was compared across the NAM lines.

Materials & Methods

Prediction of PPR encoding genes in chr2 PPR cluster

Genomic sequences of the 26 NAM founder lines were acquired from MaizeDGB (https://www.maizegdb.org). Since physical coordinate locations will be different for each genome, two SNP markers C01689-3 and C00708-1, which flank the PPR cluster on chr2, were used to identify the location of the PPR cluster within each genome sequence (Supplementary Sequences). The marker sequences were located in each NAM genome using BLASTn, and then the sequences between the markers for each NAM line was extracted. FGENESH was run on each NAM sequence to predict the genes within this chromosome 2 interval. Predicted protein sequences for the longest ORF were created from the FGENESH gene predictions. Each gene was giving an identification number based on the order in which it was located in the sequence. *hmmsearch* from the HMMER 3.1 package was used to screen the proteins for the presence of PPR motifs based on the PPR domains (PF12854, PF1304, PF13812, PF01535 and PF17177) downloaded from the Pfam database from EMBL-EBI.

Two gene sequences, B73_236 and B73_113, which were annotated as *Rf1* orthologs and have the characteristics of RFL type PPRs, were used to identify closely related RFL homologs within the gene cluster. B73_236 and B73_113 were 2455 bp and 2446 bp in length, respectively. The DNA sequence of both genes were compared against the sequences of PPR genes using BLASTn, and PPRs with >94% identity for at least 1000 bp of alignment were considered homologous.

Phylogenetic relationships of RFL sequences

The 131 RFL homologous sequences were aligned MAFFT v7.031b. DNAdist from the PHYLIP package was used to compute a distance matrix between each pair of sequences under the Jukes-Cantor model of nucleotide substitution (nruns=2 nchains=4). A phylogenetic tree was generated using MrBayes 3.2.7 under the GTR model with gamma-distributed rate variation across sites with 300000 generations and tree sampling every 100 generations. The tree was visualized with FigTree v1.4.4.

Analysis of diversitying selection across RFL sequences

The CODEML package from the PAML program was used to estimate the effect of positive selection within the RLF sequences. First a subset of 98 sequences was selected which all had 19 repeats, few indels, and a conserved start site. These sequences were aligned with MAFFT, and then PAL2NAL was used to create protein to codon alignments for import into PAML. Then a maximum-likehood tree was created with MEGA X, and the Newick tree format was imported into PAML. The following settings were used for CODEML: runmode= 0, CodonFreq = 2:F3x4, model = 0, Nsites = 0 1 2 7 8. Bayes Empirical Bayes (BEB) probabilities were calculated for each position in the protein and then combined based on their position within the 35aa repeat. dN/dS was calculated using the M0 model, $\omega > 1$ indicating positive selection. The synonymous and non-synonymous substitutions were counted and combined for each position in the 35 amino acid motif. The aPPRove program (Harrison, Ruiz et al. 2016) was used to analyze the RFL sequences and extract the 5^{th} and 35^{th} amino acid within each PPR motif to produce the PPRCode of each RFL sequence.

Results

Identification of RFL sequences within the chromosome 2 PPR cluster

The RFL gene cluster was genetically mapped to a region on the long arm of chromosome 2 from 232.8 to 235.8 MB on the Zm-B73-REFERENCE-NAM-4.0 physical map. The SNP markers C01689-3

and C00708-1 flank this genetic mapping interval (Supplementary Sequences). By locating these markers against each NAM genome using BLASTn, the size of the interval in each NAM line was calculated. Ki11 had the smallest interval size, at 2.1 MB, and OH43 had the largest interval size at 3.8 MB (Table 1). The average interval size within the NAM founder set was 2.7 MB. Regions of high sequence repetitiveness, which can occur around gene clusters, may have experienced unequal crossing over during meiosis leading to structural variation and sequence rearrangements, as sometimes can be detected by sequence size differences across genomes. The sequence size variation between the NAM lines may suggest unequal crossing over has occurred, and the NAM lines may have different gene content.

The PPRs within each NAM sequence were discovered by screening the chr2 interval sequences with *hmmsearch* to find genes containing PPR motif. The NAM lines contain between 6 and 18 PPR encoding genes within this interval, with Mo18w containing the least and NC358 containing the most (Table 1). The NAM inbreds averaged 11 PPR encoding genes within this chr2 interval, and in total 290 PPRs were predicted across all NAM lines. After discovering all the PPR encoding genes, the PPRs were then filtered to identify the genes belonging to the RFL class of PPRs.

RFL PPRs have high similarity to RFL sequences in other species. Within this chr2 PPR cluster is a clade of genes annotated as *Rf1* homologs, suggesting they closely resemble a previously discovered RFL gene from *Oryza sativa* (Komori, Ohta et al. 2004). These putative RFL genes each contain 19 PPR repeats and have repeats of 35 amino acids in length. DNA sequences of two of these RFL genes, B73_236 and B73_113, were blasted against the 290 PPRs to identify homologs within the PPR cluster. Of the PPR sequences, 131 were determined to be RFL homologs sharing >94% identity over at least 1000 bp of aligned sequences. The NAM lines content varied from 3 to 9 RFL genes within the PPR cluster, with an average of 5 RFLs (Table 1).

Phylogenetic analysis of the RFL homologs

Through unequal crossing over, this RFL gene cluster may have expanded, contracted or rearranged leading to differing paralogous gene content between the NAM lines. To determine the RLF gene content of each NAM line, the 131 paralogous sequences were first grouped by sequence similarity and PPRCode to find identical and closely related gene sequences. DNAdist from the PHYLIP package

was used to make a sequence distance matrix of the 131 RFL sequences within the NAM lines (Figure 2). The maximum distance between the 131 paralogous sequences was 0.059. Genes were considered to be the same if the distance between two sequences was less than 0.004 and the PPRCode was identical. Some genes had the same PPRCode; however the genes could be distinguished by sequence differences or by physical location within the chromosome interval. Most of these RFLs had 19 repeats (107); however 3 had 20 repeats, 14 had 18 repeats, 3 had 17 repeats, 1 had 15 repeats, and 3 had 14 repeats. For the sequences with 18 repeats, it was the 3rd repeat that was deleted within the sequence. A phylogenetic tree was generated using MrBayes 3.2.7 to visualize the relatedness of the RFL sequences. Some sequences were present in most of the NAM lines and were highly conserved with no sequences differences. Other sequences were present in one or few NAM lines.

Diversifying selection affects RNA-binding specific sites

RFL proteins have higher rates of diversifying selection which allows them to evolve faster to recognize new RNA targets in the mitochondria and chloroplast. Comparisons of the synonymous and non-synonymous substitutions across the protein showed higher rates of non-synonymous substitutions (Figure 1a). The average ω (dN/dS) for these sequences was 1.60614 which indicates positive selection (calculated using Model 0 in CODEML). Since positions 5 and 35 interact directly with the RNA and are responsible for RNA-binding specificity, these amino acid positions may show higher rates of positive selection. The Bayes Empirical Bayes (BEB) analysis in CODEML was used to predict amino acid sites with positive selection (P>95%). Amino acid positions 5 and 35 had the highest number of sites with positive selection and sites 1 and 30 also showed higher positive selection compared to other sites within the PPR motif (Figure 4b). The PPRCode for each RFL was extracted using the aPPRove program. Of the 141 RFLs, there were 31 different combinations of amino acids at the 5th and 35th sites within the repeats indicating high diversifying selection at the RNA binding sites (Figure 5). The most common PPRCode was shared by 33 RFL sequences. The high diversity in the PPRCode could arise from nucleotide substitutions or from intergenic recombination between paralogs.

Discussion

PPR gene clusters contain multiple copies of highly similar, repetitive sequences which can result in unequal crossing over during meiosis. These meiotic mistakes can lead to large sequence rearrangements and increases or decreases in gene content. Through the analysis of the chromosome 2 PPR cluster, it was shown that the NAM founder lines have up to a 74% difference in physical size of this interval. The gene content within this PPR cluster shows that large segments of sequence were duplicated and recombined resulting in differing PPR paralog content within the NAM lines. The NAM lines have between 3 and 8 RFL genes within this cluster, sometimes with two copies of the same gene. Of the 26 NAM lines, only B73 and B97 had identical gene content throughout the interval, suggesting these are more closely related inbreds. Since all other NAM lines have different combinations of paralogous RFL genes, this NAM set screen likely did not capture all of the diversity present within maize germplasm.

Nuclear PPR genes interact with mitochondrial and chloroplast RNA to edit or modify RNA post-transcriptionally. An unfavorable mutation with the mitochondrial sequence may be mitigated by a PPR gene reducing the effect of the mutation. PPR genes show high rates of positive selection in order to evolve quickly to control mutations in the mitochondrial. For this reason, RFL genes, which mitigate unfavorable mitochondrial sequences, are characterized by a high dN/dS ratio ($\omega > 1$). RFL sequences within this chromosome 2 cluster had a $\omega=1.60614$ indicating these gene are under diversifying selection. Since amino acid positions 5 and 35 within the PPR motif are responsible for specifying the RNA sequence for RNA binding, these positions are often under the strongest selection. Within this PPR dataset, positions 5 and 35 showed the highest number of positively selected sites, but also positions 1 and 30 showed more positive selection compared to other positions in the motif.

Since these sequences were predicted using FGENESH, it is not known how many of the sequences are pseudo-genes instead of transcribed genes. Transcripts of these public 26 NAM are available, but due to the repetitiveness and high copy number, the short RNA-seq reads are being mis-mapped to the wrong sequences causing incorrectly annotated transcripts. With the next release of NAM transcript data, it may be possible to determine which RFLs are expressed. Surprisingly, these genes do not show a mitochondrial targeting sequence (MTS) using SignalP. The Rf3 gene from the inbred Ky21

has been mapped to this RFL cluster, so it is likely that one of those four RFLs does target the mitochondria even though there is not a predicted MTS.

References

Dahan, J. and H. Mireau (2013). "The Rf and Rf-like PPR in higher plants, a fast-evolving subclass of PPR genes." RNA Biol **10**(9): 1469-1476.

Fujii, S., C. S. Bond and I. D. Small (2011). "Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution." Proc Natl Acad Sci U S A **108**(4): 1723-1728.

Gaborieau, L., G. G. Brown and H. Mireau (2016). "The Propensity of Pentatricopeptide Repeat Genes to Evolve into Restorers of Cytoplasmic Male Sterility." Front Plant Sci **7**: 1816.

Gutmann, B., S. Royan, M. Schallenberg-Rudinger, H. Lenz, I. R. Castleden, R. McDowell, M. A. Vacher, J. Tonti-Filippini, C. S. Bond, V. Knoop and I. D. Small (2019). "The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants." Mol Plant.

Harrison, T., J. Ruiz, D. B. Sloan, A. Ben-Hur and C. Boucher (2016). "aPPRove: An HMM-Based Method for Accurate Prediction of RNA-Pentatricopeptide Repeat Protein Binding Events." Plos One **11**(8): e0160645.

Kamps, T. L. and C. D. Chase (1997). "RFLP mapping of the maize gametophytic restorer-of-fertility locus (rf3) and aberrant pollen transmission of the nonrestoring rf3 allele." Theor Appl Genet **95**(4): 525-531.

Komori, T., S. Ohta, N. Murai, Y. Takakura, Y. Kuraya, S. Suzuki, Y. Hiei, H. Imaseki and N. Nitta (2004). "Map-based cloning of a fertility restorer gene, Rf-1, in rice (Oryza sativa L.)." The Plant Journal **37**(3): 315-325.

Manna, S. (2015). "An overview of pentatricopeptide repeat proteins and their applications." Biochimie **113**: 93-99.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li, Q. Sun, S. Flint-Garcia, J. Thornsberry, C. Acharya, C. Bottoms, P. Brown, C. Browne, M. Eller, K. Guill, C. Harjes, D. Kroon, N. Lepak, S. E. Mitchell, B. Peterson, G. Pressoir, S. Romero, M. Oropeza Rosas, S. Salvo, H. Yates, M. Hanson, E. Jones, S. Smith, J. C. Glaubitz, M. Goodman, D. Ware, J. B. Holland and E. S. Buckler (2009). "Genetic properties of the maize nested association mapping population." Science **325**(5941): 737-740.

Meyer, J., D. Pei and R. P. Wise (2011). "Mediated T- Transcript Accumulation Coincides with a Pentatricopeptide Repeat Cluster on Maize Chromosome 2L." The Plant Genome Journal **4**(3): 283.

Xu, X.-B., Z.-X. Liu, D.-F. Zhang, Y. Liu, W.-B. Song, J.-S. Li and J.-R. Dai (2009). "Isolation and Analysis of Rice Rf1-Orthologus PPR Genes Co-segregating with Rf3 in Maize." Plant Molecular Biology Reporter **27**(4): 511-517.

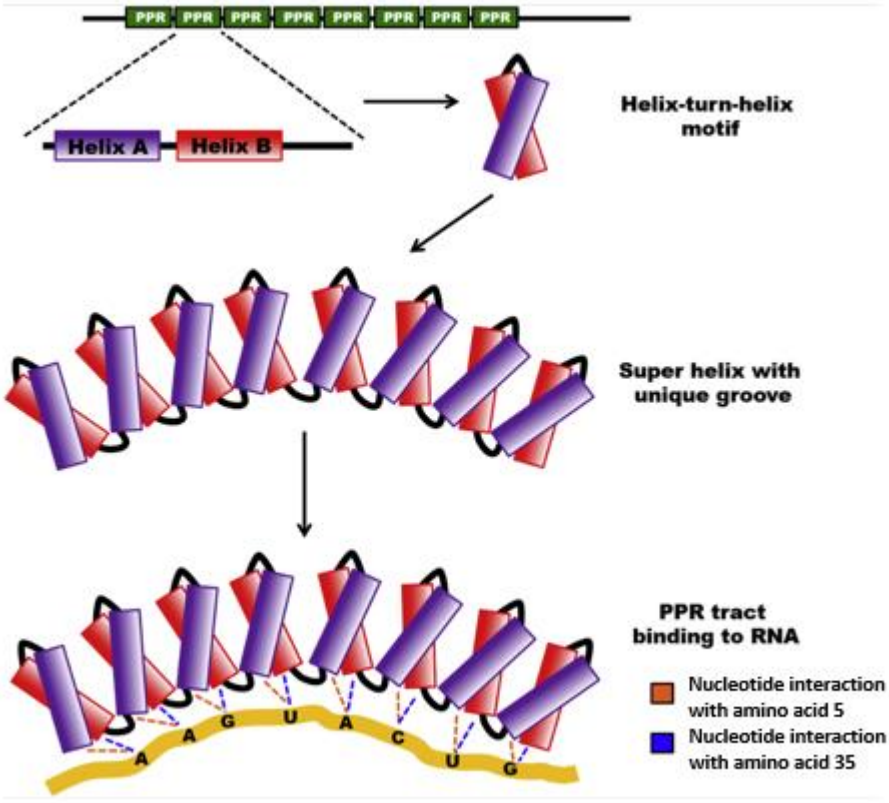

Figure 1. Diagram taken from Manna, Biochimie, 2015 (Manna 2015). PPR protein structure and the mechanism of transcript recognition. Each PPR motif forms two a-helices, which interact to form a helix turn helix motif. The series of helix turn helix motifs throughout the protein are stacked together to form a superhelix with an RNA binding groove. Modular recognition of transcripts is mediated by nucleotide interactions with the amino acids at positions 5 and 35 of each PPR motif.

Table 1. Sequence size and PPR gene content comparisons between the 26 NAM lines at the chromosome 2 PPR cluster. The total number of PPRs includes all gene sequences containing a PPR motif, and the RFL homologs are the subset of PPRs containing RFL gene characteristics.

NAM Inbred	Size of chr2 Inverval (bp)	Total # PPRs	# RFL PPR Homologs
B73	3,117,329	15	7
B97	3,107,053	16	7
CML103	2,490,446	7	3
CML228	2,597,411	9	5
CML247	2,373,169	7	3
CML277	2,288,785	7	3
CML322	2,205,063	9	4
CML333	2,583,269	11	5
CML52	2,747,804	12	5
CML69	2,621,040	8	4
HP301	2,561,320	9	4
Il14H	2,546,169	10	4
Ki11	2,113,111	10	4
Ki3	2,352,658	9	4
KY21	2,974,413	9	4
M162W	2,616,232	10	5
M37W	2,314,185	8	3
Mo18w	2,519,817	6	3
MS71	2,982,104	17	7
NC350	2,209,862	7	3
NC358	3,010,853	18	8
OH43	3,838,049	15	7
OH7B	3,093,792	16	7
P39	3,024,603	17	8
TX303	2,894,359	11	6
Tzi8	2,869,114	17	8
Grand Total		290	131

Figure 2. Distance matrix of 131 RFL sequences from DNAdist under the Jukes-Cantor model of nucleotide substitution. The darkest green has a distance of 0, indicating no sequence differences between the pairs. The darkest red pairs have a distance value of 0.059 or less. The sequences separated into paralogous groups by sequence similarity and PPRCode. Sequences were considered to be paralogous if the distance was < 0.004 and had the same PPRCode.

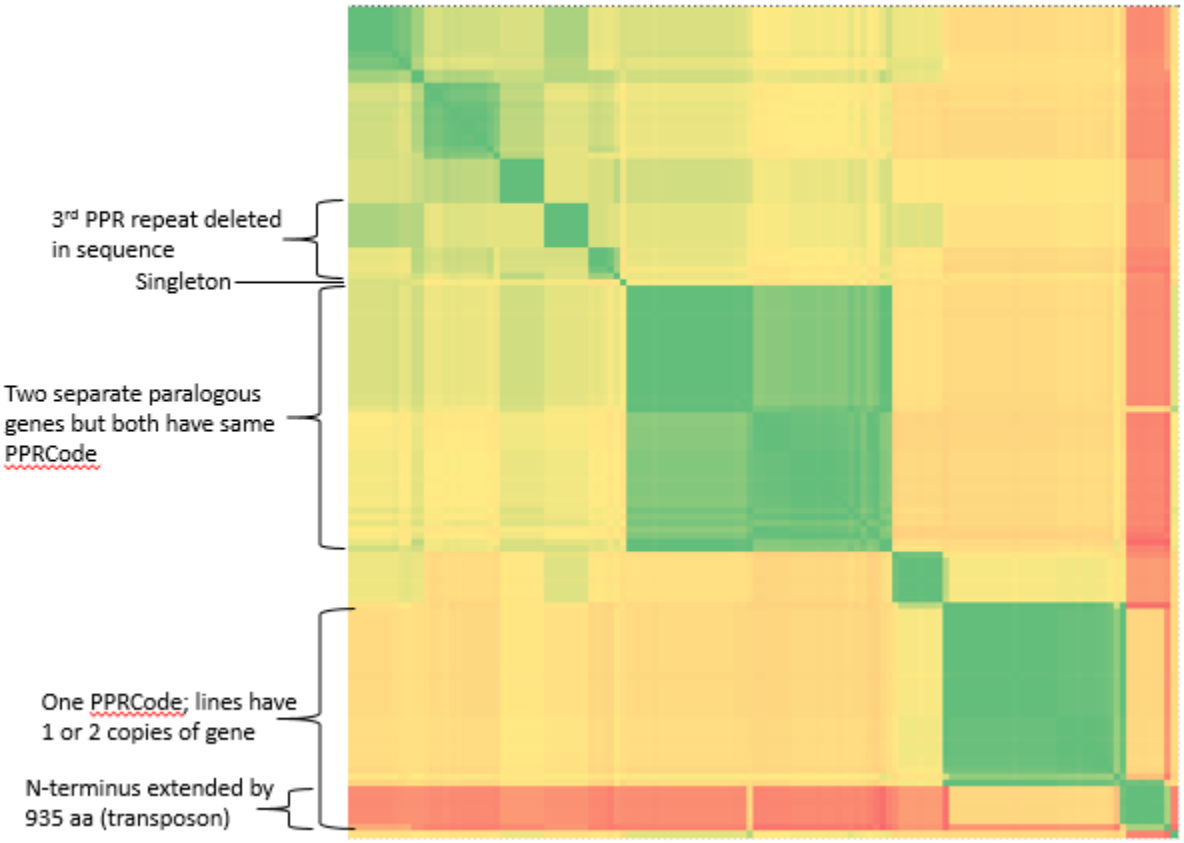

Figure 3. Phylogenetic relationships of RFL sequences within chr2 PPR cluster for NAM founder set. The tree was generated with MrBayes 3.2.7 and visualized with FigTree v1.4.4.

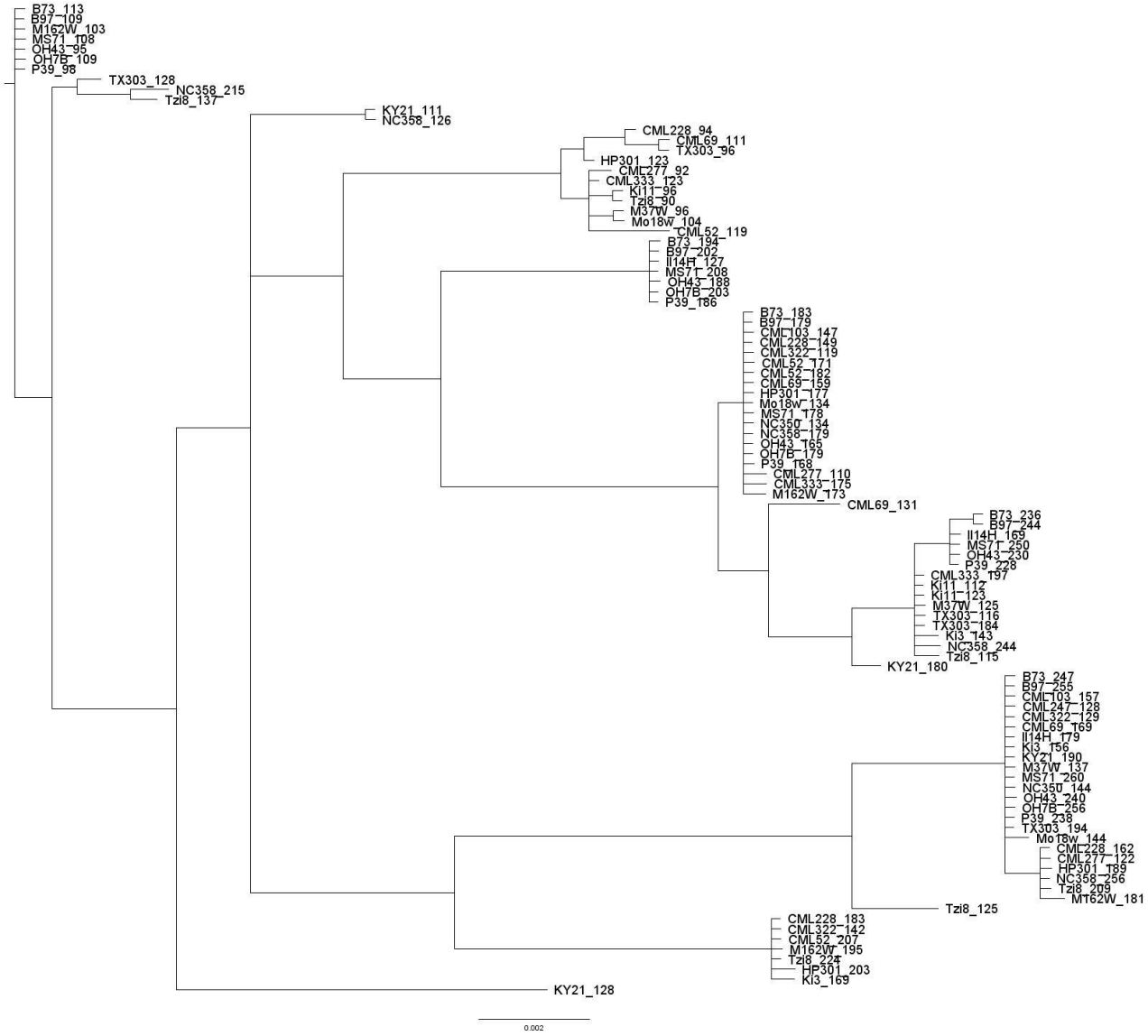

Figure 4. Locations of positive selection across the 35 amino acid motifs for the 98 RFLs with 19 repeats. (a) The sum of synonymous and non-synonymous substitutions found with the alignment of RNL sequences. (b) Positively selected sites (P>95%) within in the 35 amino acid motif as predicted by Bayes Empirical Bayes (BEB) probabilities in CODEML.

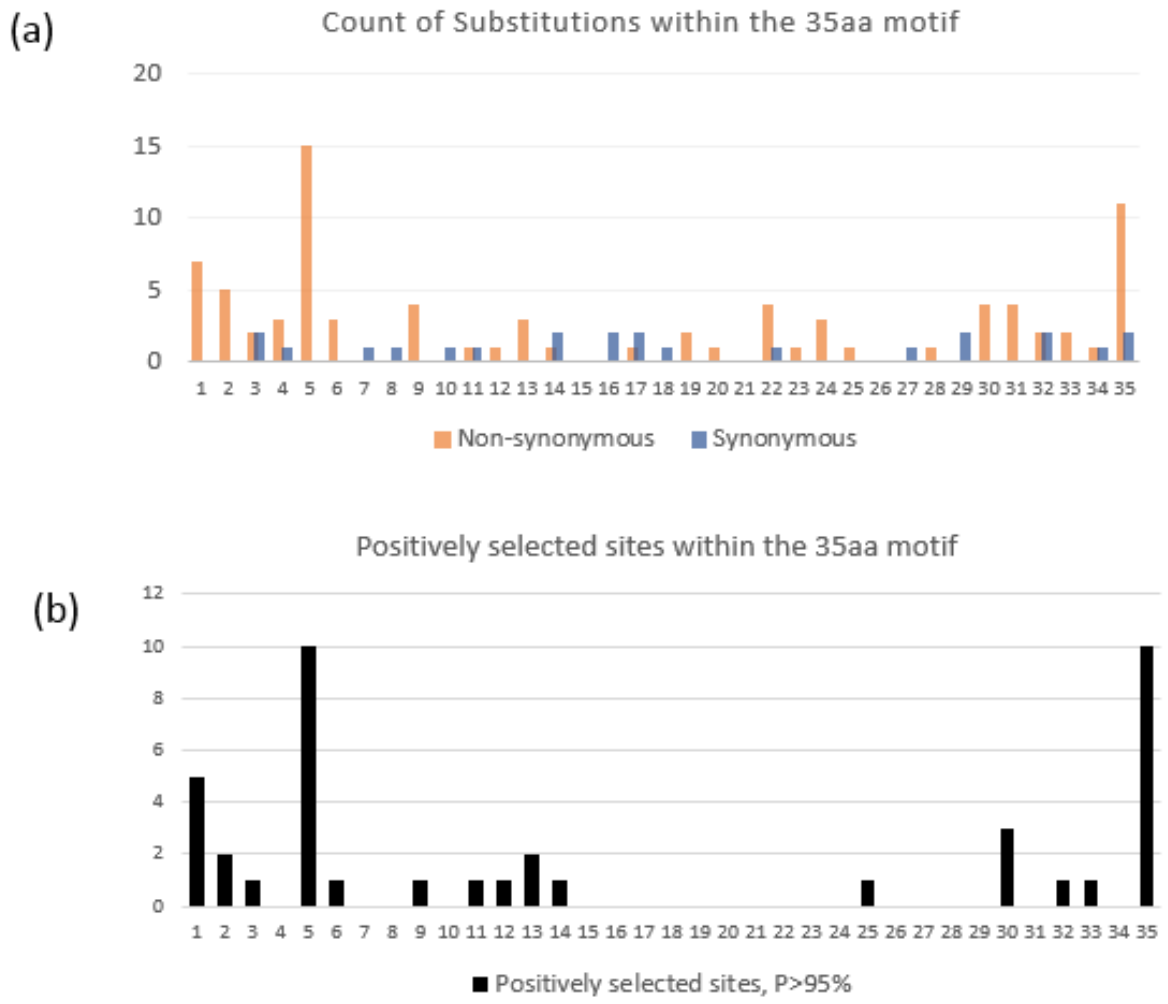

Figure 5. High diversity in PPRCode showing the 5th and 35th amino acid in each repeat, grouped by similar RFL sequences. The colors were added to aid in visualizing the amino acid differences.

Group	# Members	-1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	-	AD	SD	NS	ND	SD	ND	NN	TV	ND	TD	HD	ST	SD	GS	ND	NN	NS	NN	NE
2	2	-	AD	ND	SS	ND	SN	ND	SN	TV	ND	TD	HD	ST	SD	GS	ND	NN	NS	NN	NE
3	1	-	AN	ND	SS	ND	SN	ND	SN	TV	ND	TD	HD	ST	SD	GS	ND	NN	NS	NN	NE
4	7	-	AN	ND	SS	SD	SN	ND	SN	TV	ND	TD	HD	ST	SD	GS	ND	NN	NS	NN	NE
5	1	-	AD	ND	SS	ND	SN	ND	ND	NV	NN	TD	HD	ST	SD	GS	NN	ND	NC	SD	NE
6	7	-	AD	ND	NS	TD	SN	ND	NN	TD	NN	RD	ND	SD	SN	CS	SD	ND	NC	SD	NE
7	3	LL	AD	SD	NS	ND	SD	ND	ND	TD	NH	MD	HD	GD	NN	GS	NN	SD	NC	SD	NE
8	29	-	AD	SD	NS	ND	SD	ND	ND	TD	NH	MD	HD	GD	NN	GS	NN	SD	NC	SD	NE
9	1	-	AN	ND	SS	ND	SD	ND	ND	TD	NH	MD	HD	GD	NN	GS	NN	SD	NC	SD	NE
10	1	-	AN	ND	SS	SD	SN	ND	NN	TV	ND	TD	HD	GD	SD	GS	SD	NN	NC	SN	NE
11	10	-	AN	ND	SS	SD	SN	ND	NN	TD	ND	TD	HD	SD	SD	GS	SD	NN	NC	SN	NE
12	1	-	AN	ND	SS	SD	SN	ND	NN	TD	ND	TD	HD	-	SD	GS	SD	NN	NC	SN	NE
13	1	-	AD	ND	-	NE	NN	ND	-	ND	ND	RD	ND	GD	SN	GS	SD	NN	NC	SN	NE
14	3	-	AD	ND	-	SD	NN	ND	NN	TV	ND	RD	ND	GD	SN	GS	ND	NN	NC	SN	NE
15	1	-	AD	ND	-	SD	NN	ND	NN	TV	ND	RD	ND	GD	SN	GS	NN	ND	NC	SD	NE
16	1	-	AD	ND	-	SN	-	ND	NN	TV	ND	RD	ND	GD	SN	GS	ND	NN	NC	SN	NE
17	2	-	AN	ND	SS	SD	SN	ND	SN	TD	ND	RD	ND	ST	SN	GS	ND	SN	NC	SN	NE
18	1	-	AN	ND	SS	ND	NR	AD	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NN	SN	NE
19	2	-	AN	ND	SS	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NN	SN	NE
20	1	-	AN	ND	IS	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NS	SN	NE
21	33	-	AN	ND	SS	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NS	SN	NE
22	1	-	AN	NV	-	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NS	SN	NE
23	1	-	RN	ND	SS	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NS	SN	NE
24	1	-	AN	ND	SS	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CS	SD	NN	NS	SI	
25	7	-	AD	SD	TS	ND	SN	ND	NN	ND	NN	RD	HD	GT	SN	GS	ND	SD	SC	SN	NE
26	1	-	AD	SD	TN	-	-	ND	NN	ND	NN	RD	HD	GT	SN	GS	ND	SD	SC	SN	NE
27	7	-	AN	ND	-	ND	NN	ND	NN	TD	SN	RD	HD	ST	SD	GS	ND	NN	NC	SN	NE
28	1	-	AN	NA	ND	NN	ND	SD	ND	ND	RD	SD	SK	NN	NS	SN	NE				
29	1	-	AN	ND	SS	ND	NN	ND	SD	ND	ND	RD	ND	SD	SN	CQ					
30	1	-	NS	ND	SA	NH	MD	HD	GD	NN	GS	NN	SD	NC	SD	NE					
31	1	-	RN	ND	SS	SE	NT	TD	ND	TD	HD	SD	SD	GS	SD	NE					
	131																				