Journal of
Data Mining in Genomics & Proteomics

Kaur et al., J Data Mining Genomics Proteomics 2016, 7:3
DOI: 10.4172/2153-0602.1000203

Research Article Open Access

Insights into PPR Gene Family in *Cajanus cajan* and Other Legume Species

Parampreet Kaur, Mohit Verma, Pavan K Chaduvula, Swati Saxena, Nikita Baliyan, Alim Junaid, Ajay K Mahato, Nagendra Kumar Singh and Kishor Gaikwad*

National Research Centre on Plant Biotechnology, Pusa, New Delhi, India

Abstract

PPR proteins comprises of several hundred members among land plants and govern a fascinating array of functions in organeller genomes that ranges from participation in stabilization of organeller transcripts, RNA editing to fertility restoration of CMS lines. Despite the availability of genome sequences of several legume species, comprehensive cataloguing of members of PPR gene family has not been carried out. In the current study, we identified 523, 830, 534, 816, 441 and 677 PPR proteins in *Cajanus*, *Glycine, Phaseolus, Medicago, Vigna* and *Cicer* genomes, respectively and their complete *in silico* categorization was undertaken to classify them into various sub-classes and their localization prediction. Chromosomal coordinates of 271 *Cajanus* PPR genes were predicted and their homologues were identified in 5 other legumes revealing extensive genome conservation. PPR genes of all 6 legume species were further probed to identify restorer of fertility-like PPRs (RFLs) on the basis of protein clustering and followed by homology searches to already known Rf-PPR genes. Seventy RFL PPR genes (P sub-class) were identified and were scrutinized by phylogenetic analysis which revealed extended similarity and common features shared by these RFLs across the species. Some of these RFL PPRs were present as small clusters in *Glycine, Phaseolus, Vigna* and *Cicer* genomes. This study has generated a knowledge base about PPR gene family in legumes and opens several avenues for future investigations into their molecular functions, evolutionary relationships and their potential in identifying markers to enable cloning of Rf genes.

Keywords: PPR protein; Legumes; Restorer of fertility like-PPR (RFL); Synteny; P sub-class; Mitochondrion

Introduction

PPR motifs containing proteins were first discovered from the genome of *Arabidopsis thaliana* [1,2] and later reported in other sequenced eukaryotes. PPR proteins have gained importance in context of their role in various RNA processing events such as RNA stabilization, splicing, editing, cleavage and transcriptional activation [3]. Though PPRs are encoded by nuclear genome, they are mostly targeted to either mitochondria or plastids for their functions [4] and thus play an important role in organeller gene regulation. By using classical genetic screens, number of PPR mutants have been characterized with varied phenotypes ranging from those showing photosynthetic defect [5] to restricted growth [6], defective seed and embryo development [7], aberrant leaf growth [8] and restoration of pollen fertility [9]; implying the role of PPRs as sequence specific RNA binding proteins in organelles. Other reports also suggest important role of PPR and these includes, abnormal splicing of chloroplast targeted PPR encoding *Rpl2* gene in rice resulted in mutant with white stripe leaf (WSL mutant) characterized by enhanced sensitivity to abiotic stresses and chlorotic striations during its early development [10], *Rf1A* in rice functions in atp6 mRNA editing [11], *RPF2* affects mitochondrial nad9 and cox3 mRNAs in arabidopsis [12] and so on. Non plant organisms have very few PPRs whereas great expansion of this gene family via retrotransposition has been observed in plants [13]. Their number in a particular species could range from less than 30 in eukaryotes (*Chlamydomonas reinhardtii*) [14] to 1882 members in *T. aestivum* [15].

PPR proteins are categorized into different sub-classes and sub-groups on the basis of the sequence content and arrangement of peptide repeat motifs that constitutes their structural and functional divergence [16]. It is the sequence variability within repeats that provides specificity to the action of different members of this protein family. The two major sub-classes are denoted as P and PLS. Classical PPRs or P class PPRs are defined as those containing degenerate 35 amino acid peptide motif present in multiple tandem repeats and this sub-class constitutes half of the PPR family in any plant species. PPR motif is known to form two anti-parallel α-helices that interact to produce a helix-turn-helix motif, series of which forms a superhelix with central groove for interaction with RNA [17]. Many P class proteins have special appendages present at C-terminal domain (PRORP, SMR, LAGLIDADG etc.) that confers functional specificity to proteins due to presence of variable motifs. Proteins with LAGLIDADG motif are involved in catalytic processes due to its similarity with group-1 intron maturases [18] and those with SMR domain are related to MutS2 family which participate in transcription or repair of chloroplast DNA [19]. PRORP (proteinaceous RNaseP) sub-class possess metallonuclease domain which are involved in processing of mitochondrial tRNA, for example arabidopsis PRORP3 protein [20]. The classical P motif when interspersed by L motifs (36 amino acids) and S motifs (31 amino acids) in triplets constitute PLS sub-class, wherein this ordered association could have variable number of S motif repeats [21]. PLS-PPRs also possess additional C terminal domains designated as E (extended), E^+ (slightly longer than E domain) and DYW (characterised by Asp-Tyr-Trp triplet at terminating end). Thus, a PLS protein will terminate with either a PPR motif or a non-PPR motif i.e., E motif, EE^+ motif or EE^+DYW motif sequence. The members of these three sub-groups are mainly involved in RNA editing in chloroplast and mitochondria [22].

***Corresponding author:** Kishor Gaikwad, National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi, India, Tel: 011-25841787/25842789; Fax: +911125843984; E-mail: kish2012@nrcpb.org

Received June 30, 2016; **Accepted** July 11, 2016; **Published** July 18, 2016

Citation: Kaur P, Verma M, Chaduvula PK, Saxena S, Baliyan N, et al. (2016) Insights into PPR Gene Family in *Cajanus cajan* and Other Legume Species. J Data Mining Genomics Proteomics 7: 203. doi:10.4172/2153-0602.1000203

Copyright: © 2016 Kaur P, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Due to certain limitations of demarking PPR motifs by classical system of categorization, PPR motifs has been redefined by Cheng et al. [15] whereby 10 PPR motif variants have been described and used to annotate PPR sequences in 109 genomes. Newly identified motif variants in PLS sub-class includes P1 and P2 motifs that differ from classical P motif in first helix; S2 (32 amino acid), SS (31 amino acid), E1 (34 amino acid) and E2 (34 amino acid) motif. This revision of PPR classification has provided a clearer picture of PPR structure and thus will provide new insights into their role in molecular and structural evolution.

Most of these nuclear encoded PPRs carry N-terminal mitochondrial or chloroplast targeting sequence and are important contributors of various organellar post-transcriptional processes by virtue of their sequence specific RNA binding activity. Considering the array of functions governed by PPR proteins, identification and characterization of homologous and species specific PPR proteins in other plant species is critical for understanding the dynamics of nuclear cytoplasmic interactions.

Cytoplasmic male sterility system is widely exploited/ phenomenon for hybrid seed production and has been extensively studied at molecular and biochemical level in various crops. Fertility restorers are an important component of hybrid breeding that suppresses the male sterility in plants bearing defective mitochondrial transcripts. These Rf genes belong to several protein families, out of which majority have been found to encode for PPR proteins [23]. Few examples of PPR containing genes include *Rf1* gene in Petunia [24], *Rf1* in rice [25, 26] and *Rfk* and *rfo* in radish [27-29]. On the basis of their homology within PPR family and also with known CMS restorers PPRs from related plant species, restorers of fertility-like PPRs (RFL) can be identified. RFLs generally constitutes around 10-30 members/plant genome. *In silico* based approaches to identify RFLs on the basis of phylogenetic analysis and orthologous clustering has been used to identify candidate genes for fertility restoration in perennial ryegrass by Sykes et al. [30].

With approximately 20000 species, legumes are placed second to grasses in term of their economic contribution to world agricultural system. Approximately 33% of human nitrogen requirement is fulfilled by grain legumes as they contain twice the amount of proteins in comparison to cereals [31] and legumes are the single chief dietary source of proteins in many developing countries. They are unique in their capacity for symbiotic nitrogen fixation and thus enhance soil fertility along with serving as an important source of fodder, forage, secondary metabolites and industrial and edible oils. Legumes are divided in three sub-families and the important species fall under two papilionoid clades i.e., phaseoloid clade and galegoid clade [32]. Considering the importance of legumes, in depth analysis of their genome structure is important to improve their yield and quality using various genetics and genomics approaches.

Though recently, a documentation of PPR proteins in 109 genomes has been done [15] and includes few legume species, the goal of the current study is to expand the knowledge base on these proteins in legumes. The members of Phaseoleae, *Cicer*eae and Trifolieae tribe i.e., *Cajanus cajan, Glycine max, Vigna radiata, Phaseolus vulgaris, Cicer arietinum* and *Medicago truncatula* were selected to provide an understanding of the PPR gene family in legumes. As the draft genome sequence is available for all these species, insights onto PPR family in legumes will be provided by i) identifying PPR encoding genes, ii) classifying and categorizing them on the basis of the domain structure, iii) mapping them onto genome, iv) studying their evolutionary relationship among legumes and v) isolation of potential RFLs that could serve as candidate Rf genes.

Materials and Methods

PPR gene identification and classification in *Cajanus cajan* and other legumes

Genome sequencing data of 80.4 Gb from Illumina sequencing platform of cultivar Asha (unpublished data) was used as a seed sequences to search against nr database for PPR hits followed by gene prediction using FGENESH [33] and domain identification using Interproscan (version 5) [34]. Further, the predicted protein sequences were used as query against Uniprot (http://www.uniprot.org/) PPR database of *Glycine max*. The search was based on BLASTx with an e value of $1e^{-3}$ to identify putative PPR proteins. Simultaneously, already available draft genome sequence data of Asha (ICPL87119) was also used for PPR identification. Annotated data of 454-FLX sequencing chemistry of Asha [35] was searched to identify PPR genes directly whereas protein sequence data submitted by [36] was used as query in BLASTp search against arabidopsis and rice PPR dataset from Uniprot database (http://www.uniprot.org/) followed by confirmation by domain prediction using Interproscan. To remove redundancy between three datasets, reciprocal BLAST was done among the putative PPRs to identify unique PPRs. Among PPRs found in common between either datasets, the longest ones were retained to compute the actual number of putative PPRs in *Cajanus* and subsequent downstream analysis.

PPR proteins present in other legume species viz. *Glycine max*, *Phaseolus vulgaris* and *Medicago truncatula* were downloaded from Uniprot database (http://www.uniprot.org/). PPR proteins in *Vigna radiata* and *Cicer arietinum* were identified from Legume Information System (www.legumeinfo.org). Unique PPRs for each legume were sorted out to eliminate alternative splicing products.

Domain architecture of PPR proteins was described as per new classification system [15] using PPR browser website (http://plantppr.genomics.cn:8080/plantppr/nav.do?flag=group).

Subcellular localization prediction

TargetP v. 1.01 (http://www.cbs.dtu.dk/services/TargetP/) and Predotar v. 1.03 (https://urgi.versailles.inra.fr/predotar/predotar.html) were used to predict the organelle targeting domains of PPR proteins. In case of ambiguity between the results of two predicting software, the prediction with the better confidence was retained.

Genome organization and chromosome distribution of PPRs in *Cajanus cajan* and Glycine max

The chromosomal location for *Cajanus* and *Glycine* PPRs was obtained through BLASTp searches against whole genome sequence information using LIS database (http://legumeinfo.org/) and soybase (http://soybase.org/, Wm82.a2.v1), respectively. Their physical distribution on chromosomes was drawn using ArkMAP (http://www.bioinformatics.roslin.ed.ac.uk/arkmap/help/) on the basis of their coordinates in their respective genomes and the position of each gene on the chromosome was represented in base pairs.

Comparative genome analysis

Homologues of *Cajanus cajan* PPRs were identified from *Glycine max*, *Phaseolus vulgaris*, *Medicago truncatula*, *Vigna radiata* and *Cicer arietinum* by BLASTp search using LIS database. Chromosomal coordinates i.e., chromosome number and position were used to depict the homologues using Circos program (circos.ca/).

Prediction of restorers of fertility like (RFLs) PPR genes

All predicted PPR protein sequences from six legume species were analyzed for putative RFL genes. CD-hit [37,38] was used to cluster all proteins at different identity percents. PPRs clustered with ≥3PPRs/cluster and at 60% identity were selected for alignment with already well characterised restorer of fertility (Rf) genes that encodes for PPR proteins. Rf protein sequence of 5 species i.e., *Brassica napus* (ACJ70132.1), *Zea mays* (ACN24620.1), *Oryza sativa* (AB110016.2), *Petunia hybrida* (AY1027.1) and *Raphanus sativus* (DQ445625.1) were downloaded from NCBI. Using Mega7, these Rf protein sequences were then aligned individually with the PPR proteins clusters of individual species followed by construction of phylogenetic tree using iTOL (http://itol.embl.de/), PPRs showing homology with atleast two of these 5 known Rfs were designated as putative RFL genes. To provide authenticity to this approach, PPRs of all legumes were simultaneously subjected to online version of OrthoMCL (http://www.orthomcl.org/orthomcl/) to cluster the proteins into orthologous clusters.

Results

PPR genes in *Cajanus cajan* and other legumes

PPR genes in *Cajanus* were identified using three different data sets as described in Figure 1 and in totality 523 putative PPRs were identified (Supplementary file 1). The predicted numbers of PPRs in current study are based on homology searches and use of annotated data sets, thus explaining the less number of predicted PPRs to those already reported by Cheng et al. [15]. Similarly, a total of 677, 830, 534, 816 and 441 PPR proteins were identified in other legumes i.e., *Cicer arietinum*, *Glycine max*, *Phaseolus vulgaris, Medicago truncatula* and *Vigna radiata*, respectively. It was observed that though the genome size of *Medicago* (257 Mb) is considerably smaller than that of *Cicer* (738 Mb) but higher frequency of PPRs were predicted in *Medicago*. Interestingly, despite large variation in genome sizes between *Glycine* (1115 Mb) and *Medicago* (257 Mb), there was little variation in the number of PPR genes predicted.

Domain architecture, classification and organelle targeting of PPR protein

PPR proteins in legumes were classified into two sub-classes (Figure 2) i.e., P and PLS, based on the presence and arrangement of different motifs. For 1 and 5 of predicted PPR protein sequences in *Cajanus* and *Medicago,* respectively no PPR domains were predicted and hence these could not be classified and were not included in further analysis. For rest of the legume species, all the predicted proteins were classified. In *Cajanus, Glycine* and *Medicago* 51.1%, 50.2% and 62.5% of the predicted PPRs, respectively were classified in P sub-class while for other 3 legumes less than half of the PPRs were categorized as P sub-class i.e., *Cicer* (45.05%), *Phaseolus* (49.8%), *Vigna* (48.5%). In all 6 species, small proportion of PPRs within P sub-class (2-6% of P sub-class) was observed to possess C-terminal motifs i.e., SMR, PRORP and LAGLIDADG. PLS sub-class was further sub categorized into PLS, E1, E2 and DYW and majority of the proteins were found to posess DYW editing motif (Figure 2b) except in *Cicer*. None of the PPR was categorized into E^+ sub-group that is known to constitute proteins with a degenerate or truncated DYW domain [15]. A small proportion of sequences were identified with E1 motif present as a C terminal domain in all legumes (Figure 2b).

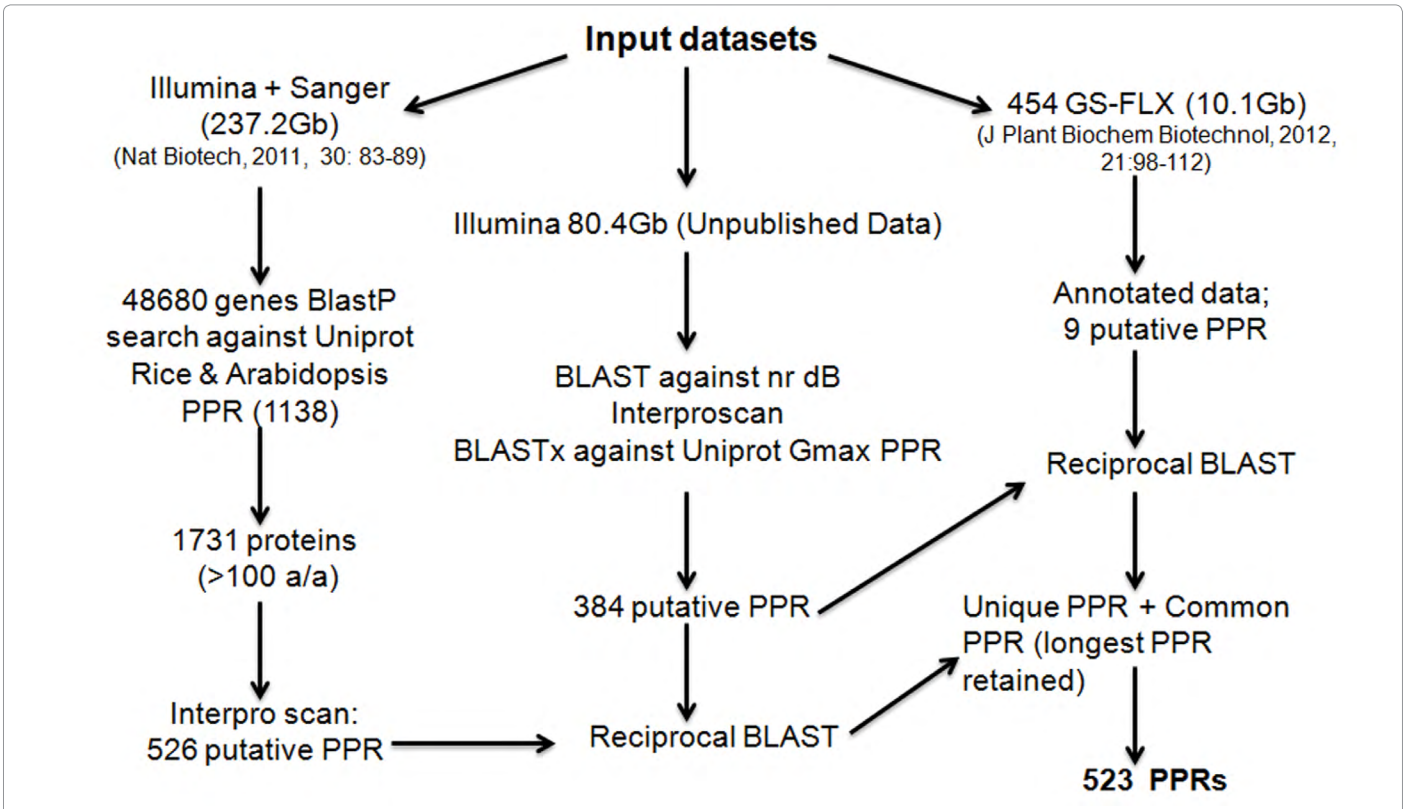

Figure 1: Outline of strategy used for mining PPR genes in *Cajanus cajan*. PPR mining was done using three data sets of cultivar Asha, in which Illumina + Sanger data (237.2Gb) and 454GS-FLX data (10.1Gb) were already publically available while 80.4Gb Illumina data is yet unpublished. BLAST against nr dB (non redundant database) was performed at $1e^{-3}$ while BLASTp, BLASTx and reciprocal BLAST searches were conducted at an e value of $1e^{-5}$.

N-terminal PPR protein sequences of all 6 species were characterized to predict their sub cellular localization. Majority of the PPRs (Figure 3) were classified and were found to be targeted to mitochondria in all legumes. For *Cajanus*, *Glycine*, *Medicago* and *Vigna,* <30% of total PPRs were not predicted with any targeting signal whereas for *Cicer* and *Phaseolus*, 42% of the PPRs lack to possess localization signal. Out of the PPRs with predicted localization in all legumes, more than 80% of proteins were found to be targeted to mitochondria and chloroplast except in *Cicer*, where only 65% of the sequences with predicted sub cellular localization were targeted to mitochondria and chloroplasts.

Despite the differences between number of PPRs predicted in each legume, approximately same percent of PPRs were found to be targeted to chloroplast and mitochondria, except for *Cicer*, in which mitochondrial targeting PPRs were less as compared to other legumes (Table 1).

Genome organization and chromosome distribution of PPRs in *Cajanus cajan* and *Glycine max*

Chromosomal location i.e., chromosome number and position of the predicted PPRs of *Cajanus cajan* and *Glycine max*, is shown in Figure 4 and Figure 5, respectively and depicts their random distribution across the genome. No correlation was observed between the PPR sub-class and subcellular localization with respect to mapping in either *Cajanus* or *Glycine* and some chromosomes possessed dense distribution of PPR genes whereas on other chromosomes, genes were sparsely distributed.

In *Cajanus,* 271 out of 523 were mapped onto its eleven chromosomes. Maximum number of PPRs i.e., 41 were mapped onto chromosome 3. Highest number of PPRs mapped/Mb of chromosome i.e., gene density was found for chromosome 6 where 37 PPRs mapped onto 23.79 Mb of chromosome giving gene density of 1.55 PPRs mapped/Mb while lowest gene density i.e., 0.57 PPRs mapped/Mb was observed for chromosome 10. PPRs were designated using the following convention: Name of the species 'Cc' (*Cajanus cajan*) followed by chromosome number, PPR number on chromosome and lastly the sub-class of particular PPR.

Similarly for *Glycine max*, 827 PPRs were found to be mapped on its 20 chromosomes while only 2 PPRs mapped to the scaffolds. Gene density ranging from 1-1.2 PPRs mapped/Mb was observed for 5 chromosomes i.e., 8, 9, 11, 13 and 16 while for rest of the chromosomes it was less than 1PPR mapped/Mb. For ease in handling, the naming of the *Glycine* PPR was changed and limited to only its unique identifier number as obtained from Uniprot database. For instance, name of the PPR protein 'tr_K7K128_K7K128_SOYBN' (as given in Uniprot database) was changed to 'K7K128'.

Conserved genome synteny among legumes

The level of synteny between *Cajanus* and other 5 legumes was assessed and shown in Figure 6 by comparison of physical map positions of 271 *Cajanus* PPRs with corresponding map positions in other legume genomes i.e., *Glycine, Phaseolus, Medicago, Vigna* and Cicer. It was observed that maximum homologues of *Cajanus* PPRs were obtained with *Glycine* and *Phaseolus* genomes where except for 1, hits were identified for all *Cajanus* PPRs. In *Cajanus* versus *Medicago, Cicer* and *Vigna* synteny, BLAST hits were obtained for 264 (97.41%), 251 (92.61%) and 220 (81.18%) sequences, respectively. This high level of homology shared across legumes supports their close evolutionary relationship. Majority of the *Cajanus* PPRs were found to map on chromosome 7 (41 PPRs), chromosome 17 (21 PPRs), chromosome 6

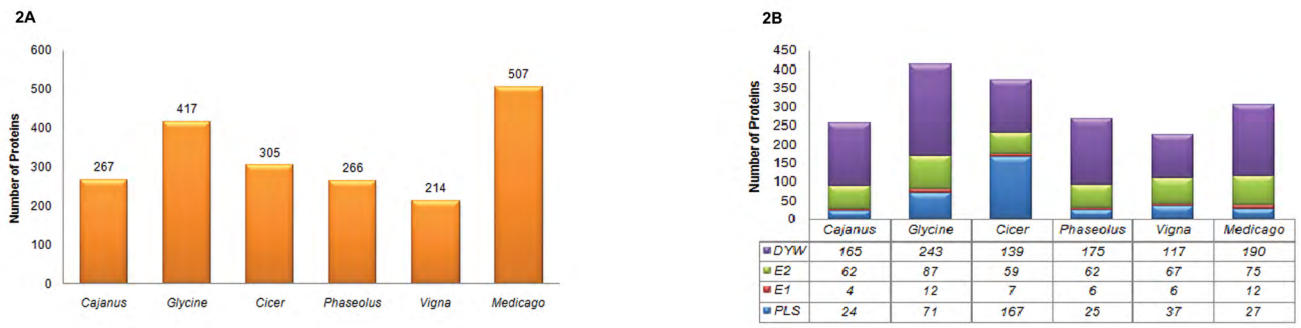

Figure 2: Classification of PPRs in P sub-class (a) and PLS sub-class (b) in different legume species using PPR browser website. y axis represents number of proteins of a particular sub-class in different legume species represented on x axis. Numbers indicated on top of each bar in Figure 2 (a) represents number of P sub-class proteins in a particular species. PLS sub-class was further classified on the basis of C-terminal motifs i.e., DYW, E1, E2 and PLS as shown in Figure 2 (b) where, total height of each bar corresponds to total number of proteins in PLS sub-class of each species while different colours in each bar reflects proteins with different C-terminal motifs, their numbers represented in the table below y axis.

Species	Cp genome (Kb)	Mt genome (Kb)	PPRs with sub cellular localization	% PPR targeted to Mt	% PPR targeted to Cp
Cajanus cajan	152.2	545.7	365	60.0	28.7
Glycine max	152.2	392.0	601	59.9	28.1
Phaseolus vulgaris	150.2	-	307	59.9	29.6
Vigna radiata	151.2	401.2	309	59.5	26.8
Cicer arietinum	125.3	-	388	42.2	23.4
Medicago truncatula	124.0	271.6	578	54.8	28.0

Table 1 PPR targeted to organelles in each legume in reference to genome size of chloroplast (Cp) and mitochondria (Mt).

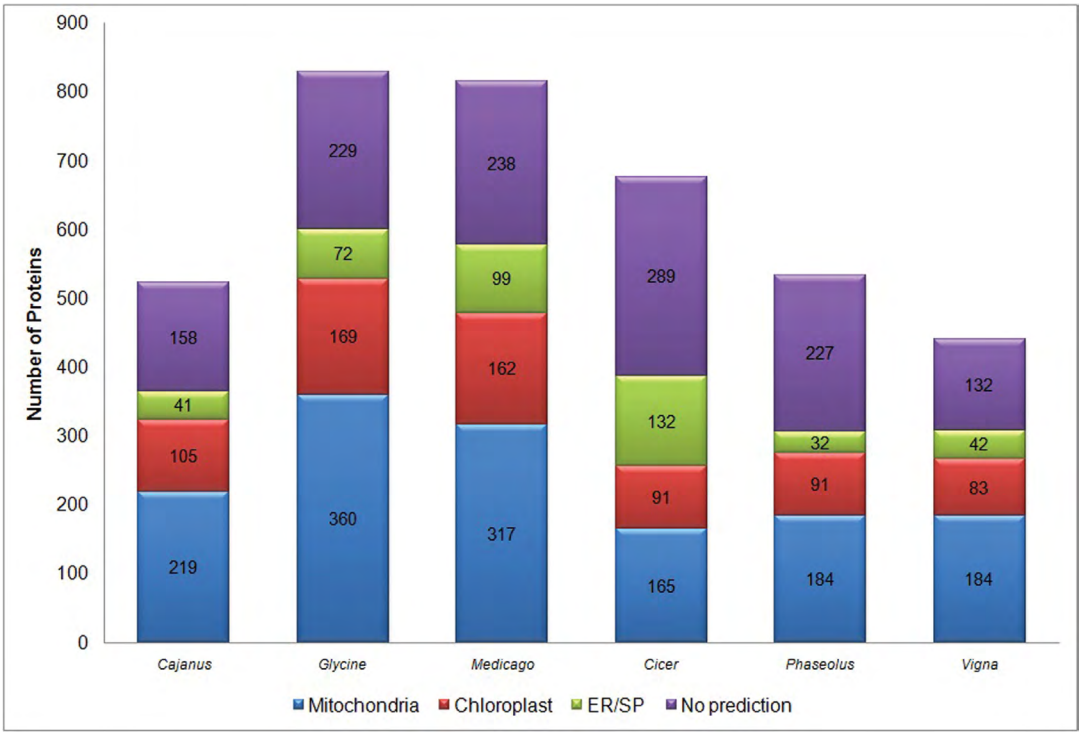

Figure 3: Classification of PPRs on the basis of their sub cellular localization in different legumes using TargetP v 1.01 and Predotar v 1.03. Total height of each bar corresponds to total number of predicted PPR proteins (Y axis) in each legume species (X axis). Different colors in each bar reflects number of proteins with different sub cellular targeting peptides i.e., mitochondrial, chloroplast and endoplasmic reticulum (ER) or secretory pathway (SP). Number of proteins in which no organelle targeting peptides were predicted are represented as no predictions.

Figure 4: Distribution of 271 Cajanus cajan (Cc) PPRs on its 11 chromosomes drawn using ArkMAP. Each chromosome is designated as Cc followed by chromosome number. Length of each chromosome corresponds to number of base pairs as represented on the axis drawn to the left. PPRs are designated on the right side of chromosomes using the convention: name of the species 'Cc' followed by chromosome number, PPR number on chromosome and lastly the sub-class of particular PPR.

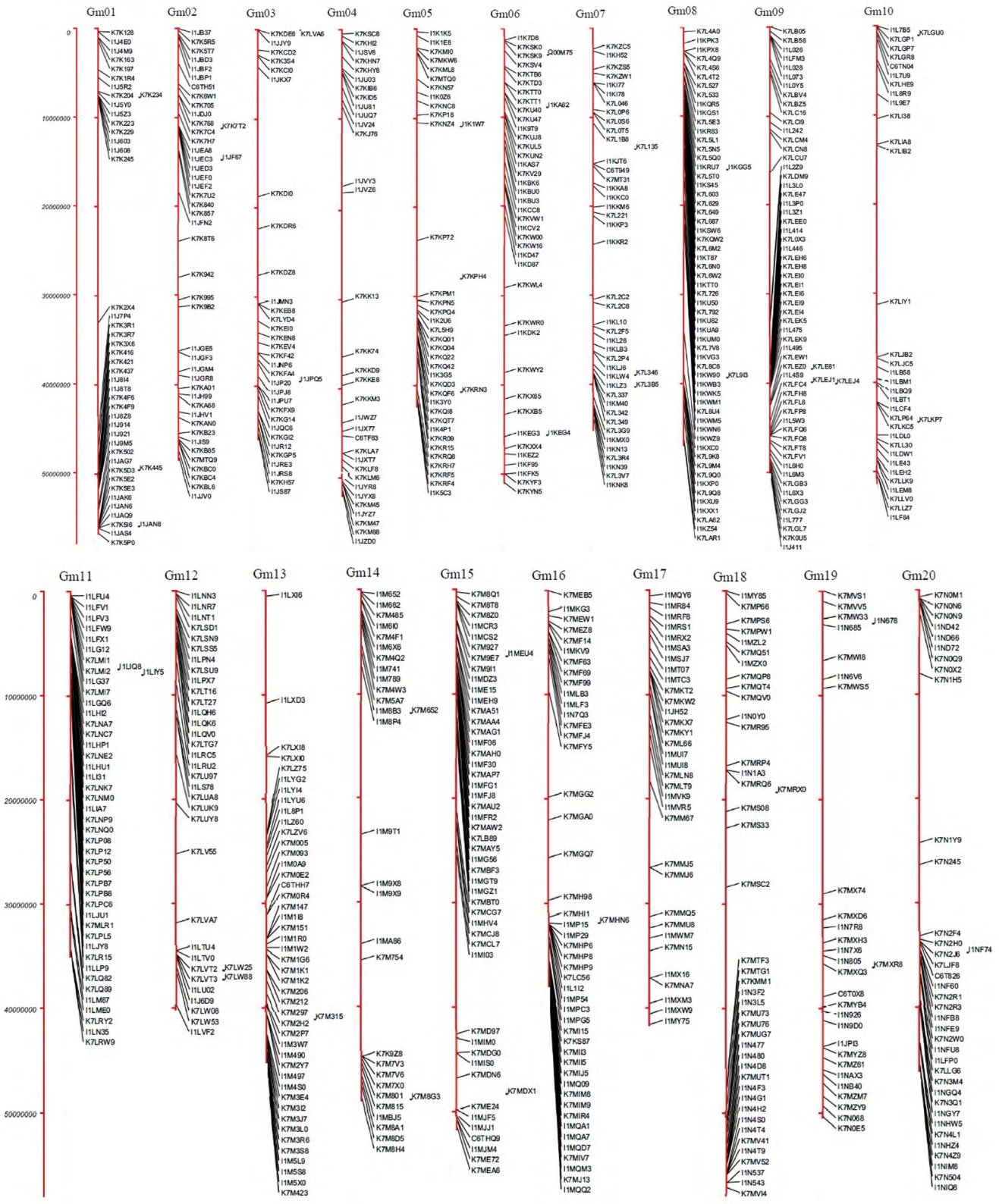

Figure 5: Distribution of 827 *Glycine max* (Gm) PPRs across its genome (5a-chromosome 1-10, 5b-chromosome 11-20) drawn using ArkMAP. Each chromosome is designated as Gm followed by chromosome number. Scale drawn on the left is in base pairs. PPRs are designated on the right side of chromosomes using only its unique identifier number (Uniprot database number). For instance, PPR protein designated as 'K7K128' instead of 'tr_K7K128_K7K128_SOYBN' (Uniprot database).

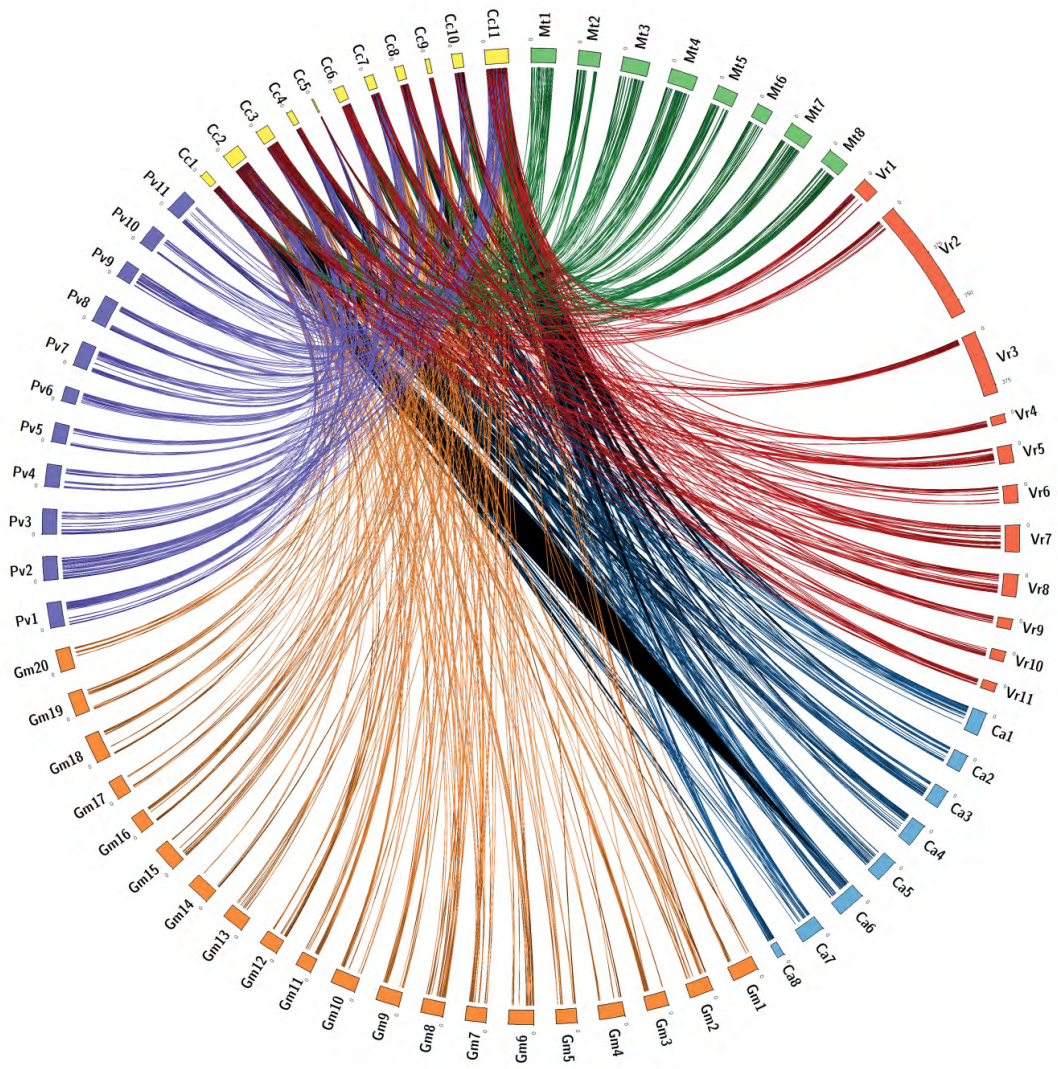

Figure 6: Circos diagram illustrating syntenic relationship of 271 mapped Cajanus cajan (Cc) PPRs with Medicago truncatula (Mt), Vigna radiata (Vr), Cicer arietinum (Ca), Glycine max (Gm) and Phaseolus vulgaris (Pv). Different species are reprersented in different color bars and number of bars correspond to chromosome number in each species. The connecting lines originating from chromosomes of Cajanus to those of other species join the syntenic regions between them.

(46 PPRs), chromosome 9 (29 PPRs) and chromosome 8 (33 PPRs) of *Medicago, Glycine, Cicer, Phaseolus* and *Vigna*, respectively.

PPRs that mapped onto a particular chromosome of *Cajanus*, identified their homologues scattered onto different *Glycine* chromosomes. For instance, 40 PPRs mapped onto *Cajanus* chromosome 11, identified their homologues that were distributed across the entire *Glycine* genome except for chromosome 4 and 19. Similar trend was observed between *Cajanus* and other legume species. Further it was observed that 14% of the *Cajanus* PPRs were found to map to the same genomic regions in one or the other target legumes while other 86% of the genes were getting mapped uniquely. Some of the *Cajanus* PPRs showed homology with small clusters of genes across all the five legumes, for e.g., a group of 5 genes mapped on *Cajanus* chromosome 1 showed homology with chromosome 7, 8, 3, 18 and 4 of *Medicago*, *Phaseolus*, *Cicer*, *Glycine* and *Vigna,* respectively. Maximum number of PPRs i.e., 41 were mapped onto *Cajanus* chromosome 3. Out of which, 15 PPRs found their homologues both onto chromosome 1 and chromosome 7 of *Phaseolus* and *Medicago*, respectively. While 10, 13 and 11 PPRs of *Cajanus* chromosome 3 were mapped onto chromosome 3 of *Glycine, Cicer* and *Vigna*, respectively. This is a representative of conserved synteny that exists across legumes. Certain other small groups of genes also displayed homology in small clusters in one or more than one legume species. No large syntenic blocks were observed between *Cajanus* and *Glycine* or with other target legumes though a high level of shared synteny was observed to cover all linkage groups of 5 legume species.

Prediction of restorers of fertility like (RFLs) PPR genes

RFL PPRs could be predicted by cross species comparison of PPR proteins with Rf-PPR genes on the basis of extended sequence similarity shared between them. Our analysis revealed that same set

of PPR genes of *Glycine, Cajanus, Medicago* and *Vigna*, respectively showed homology with Rf genes of *Brassica*, *Zea*, *Oryza*, *Petunia* and *Raphanus*. Similarly, common set of PPRs from *Cicer* was identified to be homologous to Rf gene of *Brassica*, *Petunia* and *Raphanus* while different set of PPRs displayed homology with rice and maize Rf gene. Those PPR genes that were found to be in common in terms of similarity with atleast two of the Rf genes, were selected. In totality, 70 PPR genes (8-*Cajanus*, 8-*Glycine*, 6-*Vigna*, 16-*Medicago*, 13-*Phaseolus* and 19-*Cicer*) were found to be candidate RFLs, on the basis of their homology with known Rf genes from 5 different species. All 70 genes were found to belong to P sub-class, which encode for fertility restorer genes reported so far.

Phylogenetic analysis (Figure 7) among the 70 P sub-class RFL PPR genes, revealed that with exception of 1 RFL from *Cicer* and *Medicago* each, all *Cicer* proteins were present in two sub-clusters while *Medicago* RFLs were either present as outliers or in few small sub-clusters. Six RFLs from *Phaseolus* were found to be present in one clade along with *Vigna* RFLs while 7 other *Phaseolus* RFLs were present as a cluster along with *Cajanus* RFLs. Eight *Glycine* RFLs formed separate clade with 4 *Cajanus* RFLs. Three seperate clusters constituted all candidate RFLs of *Glycine*, *Cajanus*, *Vigna* and *Phaseolus*.

Out of 3821 proteins subjected for orthologous clustering using OrthoMCL, 89.66% were assigned to 397 orthologous clusters whereas no orthologous cluster was assigned to 396 proteins. Thirty four clusters represented single protein sequences, representing species specific clusters whereas the largest cluster identified consisted of 268 proteins from all 6 legume species. A total of 249 clusters represents proteins comprising all legumes and this indicated high level of similarity between PPR proteins from different species. *Cajanus* PPRs (469 proteins) were represented in 330 clusters and a plot of number of proteins with respect to number of species represented in these clusters showed a linear relationship with presence of few outliers (Figure 8). Largest cluster that comprised of 268 PPRs formed an outlier and represented 56 of the candidate RFL genes identified. Fourteen other

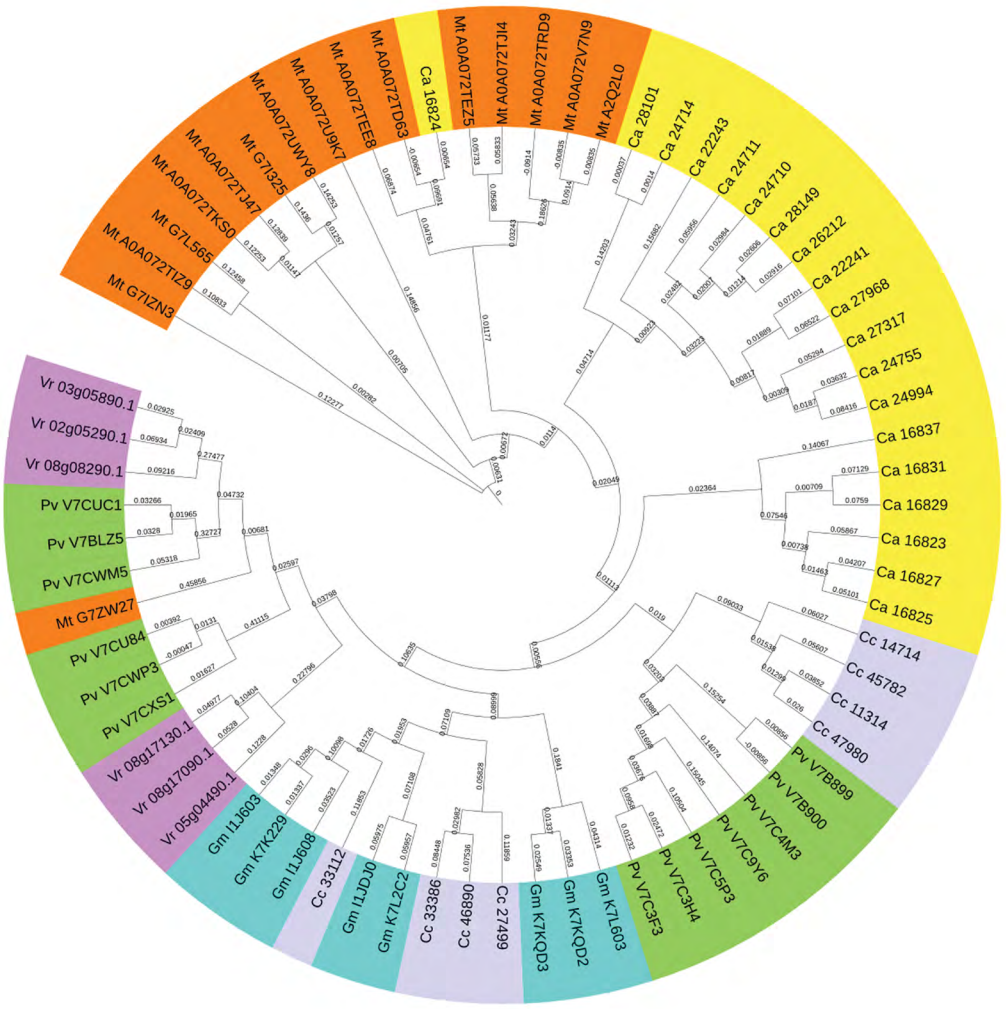

Figure 7: Phylogenetic tree of 70 RFLs predicted from 6 legume species. Amino acid sequences were aligned using Clustal W (Mega7), NJ tree was constructed and designed using iTOL. PPRs from different species are represented in different colors and legume species are represented as *Cajanus cajan* (Cc), *Medicago truncatula* (Mt), *Vigna radiata* (Vr), *Cicer arietinum* (Ca), *Glycine max* (Gm) and *Phaseolus vulgaris* (Pv).

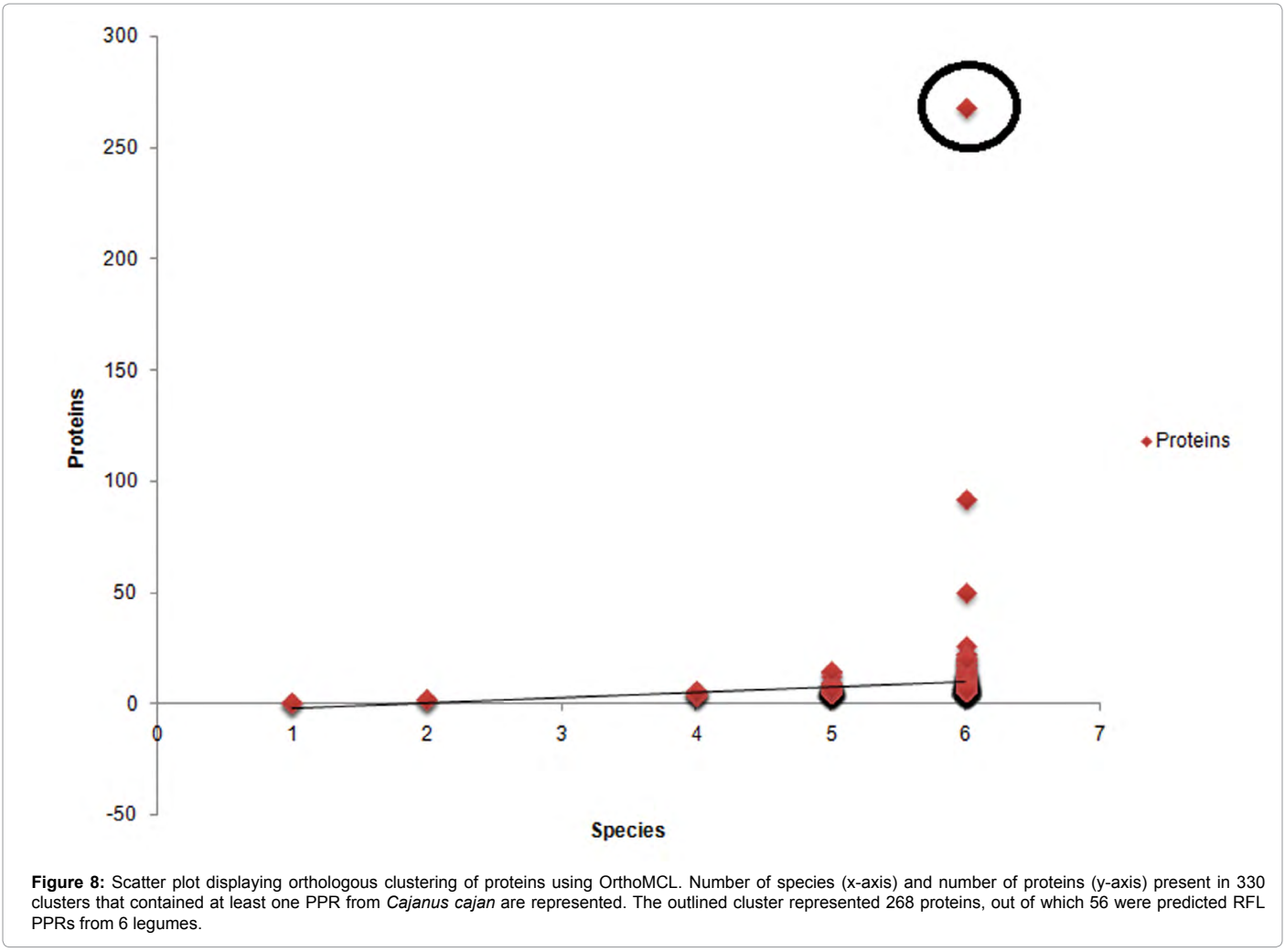

Figure 8: Scatter plot displaying orthologous clustering of proteins using OrthoMCL. Number of species (x-axis) and number of proteins (y-axis) present in 330 clusters that contained at least one PPR from *Cajanus cajan* are represented. The outlined cluster represented 268 proteins, out of which 56 were predicted RFL PPRs from 6 legumes.

RFLs (6-*Vigna*, 7-*Phaseolus*, 1-*Medicago*) were present in six other minor clusters.

Further, the genomic coordinates i.e., chromosome number and position of these 70 RFLs on their respective genomes and other 5 legumes were identified. The mapping of these RFLs in legumes proved to be advantageous in identifying regions that correspond to high RFL density. Four such genomic regions i.e., one each on genomes of *Glycine*, *Vigna*, *Cicer* and *Phaseolus* were identified (Table 2). Except 1 RFL of *Cicer,* all other RFLs identified in these clusters were also found to be present in the same groups in their phylogenetic analysis.

Discussion

Nuclear genome encoded PPR protein family is widely associated with processing of mitochondrial and chloroplast transcripts. Considering the wealth of genome sequence information available for various legume species, *Cajanus cajan* along with 5 other species i.e., *Glycine max, Cicer arietinum, Phaseolus vulgaris, Medicago truncatula* and *Vigna radiata* were selected for genome wide analysis of PPR gene family using various bioinformatics tools. Owing to the important role governed by this gene family, the RFLs i.e., restorers of fertility like genes were narrowed down from PPR genes which could potentially serve as candidate Rf genes in legumes.

PPR gene family in legumes

The number of proteins identified in all 6 legumes were well in range and as already described for other land plants. More than 80% of the documented *Arabidopsis* and rice PPRs are known to form orthologous pairs, indicating a remarkable conservation in terms of sequence and functioning [13,39]. The number of PPRs identified for *Arabidopsis*, rice and *Glycine* from Uniprot database were used to scan the *Cajanus* genome sequence for presence of PPR genes. No direct correlation was observed between the genome size and number of members of PPR gene family in legumes and is in agreement with a similar study undertaken for members of AP2/ERF transcription factor superfamily where number of genes predicted in *Cicer*, *Cajanus, Phaseolus, Medicago* and *Lotus* did not show any relation with the genome size of the legume [40]. Categorization of <50% of PPRs as P sub-class members in *Cicer, Phaseolus* and *Vigna* could be attributed to the lack of availability of complete sequence data, presence of gaps and sequencing errors. The numbers representing PPRs for any representative species may change in future with the availability of their completely finished genome sequence data.

As all the proteins required for organeller functioning cannot be encoded by their own genomes, the rest are encoded by nuclear genomes. These include genes for respiratory pathway, photosynthesis, mRNA maturation etc. To possess an organelle localization feature, N

Species	RFL Gene	Chromosome	Start (bp)	End (bp)	Putative region of high RFL density	Number of RFLs
Glycine max	Gm_K7K229	Gm01	7774549	7777396	226Kb	4RFLS
	Gm_I1J603	Gm01	7790857	7792847		
	Gm_I1JDJ0	Gm01	7995614	8000555		
	Gm_I1J608	Gm01	7995614	8000555		
Phaseolus vulgaris	Pv_V7C5P3	Pv04	42980045	42981604	148Kb	3RFLs
	Pv_V7C3H4	Pv04	43033197	43034818		
	Pv_V7C3F3	Pv04	43128239	43128922		
Vigna radiata	Vr_08g17090.1	Vr08	38093705	38095304	41.6Kb	2RFIs
	Vr_08g17130.1	Vr08	38133690	38135303		
Cicer arietinum	Ca_16837	Ca8	11204750	11206138	142.7Kb	7RFLs
	Ca_16831	Ca8	11268414	11269907		
	Ca_16829	Ca8	11289683	11291083		
	Ca_16827	Ca8	11308003	11309391		
	Ca_16825	Ca8	11319550	11320938		
	Ca_16824	Ca8	11335760	11336293		
	Ca_16823	Ca8	11346389	11347480		

Table 2: Species specific genomic regions in legumes predicted with high density of RFL genes.

terminal of PPR protein is either merged with 40-50 amino acid long mitochondrial targeting peptide or a chloroplast targeting peptide of upto 60 amino acid long [4]. Except for *Cicer* and *Phaseolus*, number of PPRs predicted as untargeted proteins, is equivalent to the false negative results (~20-30%) obtained by Predotar and TargetP [41,42].

Presence of more number of PPRs with mitochondrial targeting signal could be related to the larger mitochondrial genome of land plants (200-2000 Kbp) harboring low gene densities. Remarkable increase of mitochondrial genome size in plants is reported to occur in union with proportional expansion of PPR gene family and range of post transcriptional activities, for eg., RNA editing required in higher plants necessitating the diversification of PPR protein functioning [43]. Expansion of PPR gene family in land plants is hypothesised to be in proportion with the editing of organelle transcripts [44]. Mitochondrial and chloroplast genome of *Arabidopsis* contain 525 and 34 editing sites, respectively and possess 225 PLS proteins [45] whereas >800 PLS proteins are identified in *Selaginella* with 2150 and 1041 editing sites in mitochondria [46] and chloroplast [44], respectively. This further implies the importance of PPRs in organelle communication. In the current study, no relation was observed between organelle genome size and number of members of PLS sub-class, but in future with decoding of all editing sites in organelle genomes of different legumes, number of PPRs in PLS sub-class could be related to number of editing sites in a genome.

Further the sub cellular targeting of PPR proteins was found to be independent of their sub-class or of C-terminal domains they possess, as reported in other studies [47], though a high proportion of members of both sub-class in all 6 legume species were predicted to be targeted to chloroplast or mitochondria, which reflects their basic necessary feature in organelle functioning.

In chromosomal mapping of members of homeobox genes, more genes were located onto scaffolds of *Cajanus, Cicer* and *Lotus* as compared to that in *Medicago* and *Glycine*, wherein except few all genes are mapped onto distinct chromosomes and was attributed to the availability of incomplete genome sequence data of these 3 legume species [48]. Similarly, current study revealed that chromosomal localization of PPR genes across the *Cajanus* and *Glycine* genome is characterized by their uneven distribution where approximately half of the *Cajanus* genes were located on unanchored scaffolds while only 2 genes were located onto *Glycine* scaffolds and rest were assigned to 20 differernt chromosomes. Comparison of a collinear region between arabidopsis and *Brassica rapa* with respect to PPR genes also demonstrated their random distribution [47].

Synteny studies

Comparative information from the well characterized species has often been used to accelerate genetic and genomic studies in less charcaterized orphan species for varied purposes viz. candidate gene identifications for important traits. Similarly, identification of conserved regions across legumes will assist in the detailed analysis of legume genome evolution. Most of the genes in papilionoid legume species are likely to be found within syntenic regions (ranging from 100s of Kb to Mb) to any other given papilionoid species, so that an orthologue of a gene with known phenotype is most likely to be found in a similar genomic region in closely related species [49]. A similar trend is visible from the fact that except 1, homologues were obtained for all *Cajanus* PPRs in *Glycine* and *Phaseolus* genome and majority of *Cajanus* PPRs were mapped across other 3 legume species as well, this reflects high level of synteny conservation across species that could also be utilized as a resource to identify syntenic regions in other species. Individual *Cajanus* chromosomes are known to be syntenic to two or more than two *Glycine* chromosome [36] and was also in accordance with the current study where clusters of homologous PPRs were observed between various *Glycine* and *Cajanus* chromosomes. Similarly, chromosome 1 of *Phaseolus* is known to exhibit synteny with chromosome 3 of *Cajanus* with respect to genes such as those governing determinacy in *Cajanus* [50]. It was observed that homologues of 12 out of 41 PPRs mapped onto chromosome 3 of *Cajanus* (7.3 Mb), were identified as a cluster on chromosome 1 of *Phaseolus* in a region spanning 12.93 Mb.

Studies between *Arabidopsis* and rice PPR proteins also observed exceptionally high degree of interspecies individual protein conservation [39]. Legume Tentative Orthologous Genes (TOG) markers have been used to study evolution across pigeon pea [51], common bean [52] and other legumes [32,53]. In a study, 128 out of 377 TOGs that mapped onto *L. ervoides* genome found their orthologue both in *Medicago* and *Cicer* genome, thus reflecting a high level of conservation of synteny among the species and serves as a resource to identify syntenic regions in other species [54]. Remarkably high levels of collinearity were observed

in 0.5 Mb region surrounding *Rhg1* and *Rhg4* SCN resistance loci of *Glycine* and its corresponding region in *Medicago* in terms of perfect conservation of gene order and orientation [55]. Key TF orthologs for nodulation and floral meristem development were identified between *Medicago* and *Lotus*, which allow direct genome comparison to predict orthologs, for eg., LjNIN and PsSym35 [56]. Further, a clear existence of one-to-two relationship between the *Phaseolus* and *Glycine* genomes has already been demonstrated [57].

Conserved synteny among legume species is often disrupted by chromosomal rearrangements defined in terms of translocations or inversions [53] as targeted search for synteny between *Glycine* and *Medicago* with respect to Rpg1 displayed limited synteny [58]. Lack of synteny for PPR genes even in otherwise collinear segments of arabidopsis and *Brassica rapa* genome has also been reported [47].

RFLs prediction in legumes

Rf protein superfamily is known to constitute atleast 51 different families [23] and the number is more likely to expand with the availability of finished genome sequence data of other plant species in future. Till date, only few Rf genes have been cloned, majority of which are known to encode for PPR proteins. Almost all plant genomes contain 10-30 Rf like proteins [59,60] that share significant sequence similarity with the Rf-PPRs from other plant species and thus cross species comparison of PPR proteins with known Rf genes can be used to identify a subset of Rf proteins known as RFL PPRs [39,47,61,62]. A microsynteny analysis was conducted between *Arabidopsis* and radish and was used to clone the PPR encoding *Rfo* locus in radish [28]. Utilizing this facet of RFL PPR proteins, RFLs could be identified in any plant. These genes generally belong to the P sub-class of PPR gene family with the exception of *Rf1* gene in *Sorghum bicolor* that belongs to PLS sub class and possess domain for RNA editing [63].

Though RFL PPR represents a small group of PPR proteins, they possess certain features that are distinct. The first distinctive feature is reflected by the observation that upon mapping of RFL PPRs of a particular legume on the genome of other legumes, RFL PPRs formed species specific paralogous groups and displayed limited but significant inter species orthology which is in contrast to non RFL PPR proteins [39]. Their second distinctive feature is clustering in non-conserved genomic locations in comparison to random distribution behaviour of PPRs on the genome [47,61]. These regions where RFL genes are clustered together could be considered as candidate regions to identify Rf genes in these species. For eg., 26 Rf like PPRs were identified in *Arabidopsis* in two clusters on chromosome 1 [28,60], chromosome 10 of rice possess 9 PPRs, out of which 3 were Rf PPRs [62]. Small cluster of RFL PPRs were identified on genome of *Glycine*, *Vigna*, *Cicer* and *Phaseolus* and allowed us to narrow down the list to 15 potential Rf genes in 4 legumes (Table 2). Twenty five candidate genes for fertility restoration in CMS perennial reyegrass has been predicted based on homology with known Rf genes and DNA sequence clustering; efficacy of both approaches depending upon the type and quality of input data [30]. Similarly, prediction of RFL PPRs and identification of genomic regions where these are present as clusters on the respective legume genomes is based up on the *in silico* analysis of draft sequence assemblies. Therefore a complete repertoire of PPR genes and subsequently RFLs could not be predicted efficiently as incomplete genome assemblies do not reveal all clusters that would have otherwise formed, also observed in case of barley [63,64]. However the 70 RFls deduced in this study provides an handle for investigating the typical conserved features and subsequent functioning across legumes.

The organelle genomes of flowering plants are now thoroughly evolved and yet they retain their basic functions. These circular genomes are also very dynamic and are frequently involved in structural changes, leading to disturbances in terms of altered transcripts and generation of new orfs. It is probable that these conserved PPR proteins help in minimizing the abnormal manifestations and hence the high degree of conservation seen across genomes. Utilizing the genome sequence information from 6 legume species, an *in silico* study was conducted to provide a catalogue of PPR genes and to identify potential candidate Rf genes. Analysis of synteny between *Cajanus* PPRs and 5 other species revealed a high level of similarity that exists between legumes indicating its evolutionary lineage and conservedness of functionality. To date, PPR genes have been documented in other plant species but this forms the first comprehensive study on the PPR gene family in legumes revealing a repertoire of knowledge that can be further investigated to reveal details about their structure, evolutionary relationships and functional analysis.

Acknowledgement

This work was supported by Indian Council of Agricultural Research- Network Projects on Transgenics in Crops (ICAR- NPTC), New Delhi, India.

References

1. Aubourg S, Boudet N, Kreis M, Lecharny A (2000) In *Arabidopsis thaliana,* 1% of the genome codes for a novel protein family unique to plants. Plant Mol Biol 42: 603-613.
2. Small ID, Peeters N (2000) The PPR motif-A TPR related motif prevalent in plant organellar proteins. Trends Biochem Sci 25: 46-47.
3. Barkan A, Small I (2014) Pentatricopeptide Repeat Proteins in Plants. Annu Rev Plant Biol 65: 415-442.
4. Colcombet J, Lopez-Obando M, Heurtevin L, Bernard C, Martin K, et al. (2013) Systematic study of subcellular localization of *Arabidopsis* PPR proteins confirms a massive targeting to organelles. RNA Biology 10: 1557-1575.
5. Yamazaki H, Tasaka M, Shikanai T (2004) PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in Arabidopsis. Plant J 38: 152-163.
6. Zhu Q, Dugardeyn J, Zhang C, Takenaka M, Kuhn K, et al. (2012) SLO2, a mitochondrial pentatricopeptide repeat protein affecting several RNA editing sites, is required for energy metabolism. Plant J 71: 836-849.
7. Gutierrez-Marcos JF, Pra MD, Giulini A, Costa LM, Gavazzi G, et al. (2007) *Empty pericarp4* encodes a mitochondrion-targeted pentatricopeptide repeat protein necessary for seed development and plant growth in maize. Plant Cell 19: 196-210.
8. Petricka J, Clay N, Nelson T (2008) Vein patterning screens and the defectively organized tributaries mutants in *Arabidopsis thaliana*. Plant J 56: 251-263.
9. Akagi H, Nakamura A, Yokozeki-Misono Y, Inagaki A, Takahashi 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.
10. Tan J, Tan Z, Wu F, Sheng P, Heng Y, et al. (2014) A novel chloroplast-localized pentatricopeptide repeat protein involved in splicing affects chloroplast development and abiotic stress response in rice. Mol Plant 7: 1329-1349.
11. Rudinger M, Volkmar U, Lenz H, Groth-Malonek M, Knoop V (2012) Nuclear DYW-type PPR gene families diversify with increasing RNA editing frequencies in liverwort and moss mitochondria. J Mol Evol 74: 37-51.
12. Jonietz C, Forner J, Holzle A, Thuss S, Binder S (2010) RNA PROCESSING FACTOR 2 is required for 5'end processing of nad9 and cox3 mRNAs in mitochondria of *Arabidopsis thaliana*. Plant Cell 22: 443-453.
13. Lurin C, Andres C, Auborg S, Bellaoui M, Bitton F, et al. (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16: 2089-2103.
14. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

15. Cheng S, Gutmann B, Zhong X, Ye Y, Fisher MF, et al. (2016) Redifining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. The Plant J 85: 532-547.

16. Manna S (2015) An overview of pentatricopeptide repeat proteins and their applications. Biochemie 113: 93-99.

17. Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci 13: 663-670.

18. de Longevialle AF, Hendrickson L, Taylor NL, Delannoy E, Lurin C, et al. (2008) The pentatricopeptide repeat gene OTP51 with two LAGLIDADG motifs is required for the cis-splicing of plastid ycf3 intron 2 in *Arabidopsis thaliana*. Plant J 56: 157-168.

19. Nakamura T, Yagi Y, Kobayashi K (2012) Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. Plant Cell Physiol 53: 1171-1179.

20. Brillante N, Gobringer M, Lindenhofer D, Toth U, Rossmanith W, et al. (2016) Substrate recognition and cleavage-site selection by a single-subunit protein-only RNase P. Nucl Acid Res 44: 2323-2336.

21. Rivals E, Bruyere C, Nioche T, Lecharny A (2006) Formation of the Arabidopsis pentatricopeptide repeat family. Plant Physiol 141: 825-839.

22. Schallenberg-Ruedinger M, Lenz H, Polsakiewicz M, Gott JM, Knoop V (2013) A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes. RNA Biol 10: 1549-1556.

23. Kotchoni SO, Jimenez-Lopez JC, Gachomo EW, Seufferheld MJ (2010) A new and unified nomenclature for male fertility restorer (RF) proteins in higher plants. PLOS One 5: e15906.

24. 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.

25. Kazama T, Toriyama K (2014) A fertility restorer gene, *Rf4*, widely used for hybrid rice breeding encodes a pentatricopeptide repeat protein. Rice 7: 1-5.

26. Komori T, Ohta S, Murai N, Takakura Y, Kuraya Y, et al. (2004) Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa L.*). Plant J 37: 315-325.

27. Brown G, Formanova N, Jin H, Wargachuk R, Dendy C, 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.

28. Desloire S, Gherbi H, Laloui W, Marhadour S, Clouet V, et al. (2003) Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family. EMBO Reports 4: 588-594.

29. Koizuka N, Imai R, Fujimoto H, Hayakawa T, Kimura Y, et al. (2003) Genetic characterization of pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish. Plant J 34: 407-415.

30. Sykes T, Yates S, Nagy I, Asp T, Small I, et al. (2016) *In silico* identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.). Genome Biol Evol.

31. Zhu H, Choi HK, Cook DR, Shoemaker RC (2005) Bridging model and crop legumes through comparative genomics. Plant Physiol 137: 1189-1196.

32. Choi HK, Mun JH, Kim DJ, Zhu H, Baek JM, et al. (2004) Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci USA 101: 15289-15294.

33. Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7: 10.1-10.12.

34. Jones P, Binns D, Chang HY, Fraser M, Li W, et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236-1240.

35. Singh NK, Gupta DK, Jayaswal PK, Mahato AK, Dutta S, et al. (2012) The first draft of the pigeonpea genome sequence. J Plant Biochem Biotec 21: 98-112.

36. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, et al. (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotech 30: 83-92.

37. Li WZ, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.

38. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics 28: 3150-3152.

39. O'Toole N, Hattori M, Andres C, Iida K, Lurin C, et al. (2008) On the expansion of the pentatricopeptide repeat gene family in plants. Mol Biol Evol 25: 1120-1128.

40. Agarwal G, Garg V, Kudapa H, Doddamani D, Pazhamala LT, et al. (2016) Genome-wide dissection of AP2/ERF and HSP90 gene families in five legumes and expression profiles in chickpea and pigeonpea. Plant Biotech J.

41. Emanuelson O, Nielson H, Brunak S, Von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016.

42. Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4: 1581-1590.

43. Gutmann B, Gobert A, Giege P (2012) Mitochondrial genome evolution and the emergence of PPR proteins. Advances in Botanical Research (Volume 63), Elseiver.

44. Fujii S, Small I (2011) The evolution of RNA editing and pentatricopeptide repeat genes. New Phytologist 191: 37-47.

45. Chateigner-Boutin AL, Small I (2010) Plant RNA editing. RNA Biol 7: 213-219.

46. Hecht J, Grewe F, Knoop V (2011) Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genome Biol Evol 3: 344-358.

47. 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: 130-143.

48. Bhattacharjee A, Ghangal R, Garg R, Jain M (2015) Genome-wide analysis of homeobox gene family in legumes: Identification, Gene duplication and Expression profiling. PLOS One 10: e0119198.

49. Cannon SB, May GD, Jackson SA (2009) Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. Plant Physiol 151: 970-977.

50. Mir RR, Kudapa H, Srikanth S, Saxena RK, Sharma A, et al. (2014) Candidate gene analysis for determinancy in pigeonpea (*Cajanus* spp.). Theor Appl Genet 127: 2663-2678.

51. Kassa MT, Penmetsa RV, Carrasquilla-Garcia N, Sarma BK, Datta A, et al. (2012) Genetic patterns of domestication in pigeon pea (*Cajanus cajan* (L.) Mill sp.) and wild *Cajanus* relatives. PLOS One 7: e39563.

52. Blair MW, Cortes AJ, Penmetsa RV, Farmer A, Carrasquilla-Garcia N, et al. (2013) A high-throughput SNP marker system for parental polymorphism screening and diversity analysis in common bean (*Phaseolus vulgaris L.*). Theor Appl Genet 126: 535-548.

53. Choi HK, Kim D, Uhm T, Limpens E, Lim H, et al. (2004) A sequence-based genetic map of *Medicago truncatula* and comparison of marker collinearity with *M. sativa*. Genetics 166: 1463-1502.

54. Gujaria-Verma N, Vail SL, Carrasquilla-Garcia N, Penmetsa RV, Cook DR, et al. (2014) Genetic mapping of legume orthologs reveals high conservation of synteny betweeen lentil species and the sequenced genomes of Medicago and chickpea. Frontiers Plant Sci 5.

55. Mudge J, Cannon SB, Kalo P, Oldroyd GED, Roe BA, et al. (2005) Highly syntenic regions in the genomes of soybean, *Medicago truncatula* and *Arabidopsis thaliana*. BMC Plant Biol 5: 15-31.

56. Libault M, Joshi T, Benedito VA, Xu D, Udvardi MK, et al. (2009) Legume transcription factor genes: what makes legumes so special? Plant Physiol 151: 991-1001.

57. Lee JM, Grant D, Vallejos CE, Shoemaker RC (2001) Genome organization in dicots. II. Arabidopsis as a ‚bridging species' to resolve evolution events among legumes. Theor Appl Genet 103: 765-773.

58. Cannon S, Makarevich G, Savage E, Denny R, Mudge J, et al. (2005) A phylogenetic and structural comparison of homologous Rpg1 R-gene-containing regions in soybean and *Medicago truncatula*. Plant and Animal genomes XIII: San Diego, CA.

59. Fujii S, Bond CS, Small I (2011) Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. Proc Natl Acad Sci USA 108: 1723-1728.

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

61. Chen L, Liu YG (2014) Male sterility and fertility restoration in crops. Annu Rev Plant Biol 65: 579-606.

62. Klein RR, Klein PE, Mullet JE, Minx P, Rooney WL, et al. (2005) Fertility restorer locus Rf1 of sorghum (Sorghum bicolour L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12. Theor Appl Genet 111: 994-1012.

63. Tsai IJ, Otto TD, Berriman M (2010) Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol 11: R41.

64. Ui H, Sameri M, Pourkheirandish M, Chang MC, Shimada H, et al. (2015) High-resolution genetic mapping and physical mapconstruction for the fertility restorer Rfm1 locus in barley. Theor Appl Genet 128: 283-290.