POLISH JOURNAL OF NATURAL SCIENCES
Abbrev.: Pol. J. Natur. Sc., Vol 35(4): 395–418, Y. 2020

COMPUTATIONAL ANALYSIS FOR CHARACTERIZATION AND EVALUATION OF PENTATRICOPEPTIDE REPEAT-CONTAINING PROTEIN (PPR) IN *ARABIDOPSIS THALIANA*

***Abbas Saidi*[1], *Zahra Hajibarat*[2]**

[1] ORCID: 0000-0001-6721-5389
[2] ORCID: 0000-0002-0322-9404
[1,2] Department of Plant Sciences and Biotechnology, Faculty of Life Sciences and Biotechnology
Shahid Beheshti University, Tehran

Key words: pentatricopeptide repeat protein, Tajima, RNA-binding, purifying selection.

Abstract

Pentatricopeptide repeat (PPR) proteins are a great group of RNA-binding proteins which critical role plays in different range of biological stages. However, further investigation is necessary for deeper insight to their roles. Here, a total of 41 sequences of *PPR* genes were identified and characterized in *Arabidopsis*. A comprehensive analysis of *PPR* gene was performed containing chromosomal distribution, phylogenetic relationships, conserved motifs, and detection of transcription factor binding sites (TFBs). Analysis of TFBs illustrated that several transcription factors binding sites (TFBs) namely MYB, bZIP, WRKY, Homeodomain, and AP2 act as basic TFBs linked to abiotic stress responses as well as different growth stages. Our findings revealed a positive correlation between PPR promoter regions of genes and other genes. Expression analysis revealed that PCMP-H52 is induced under iron deficiency and shift low to high light stresses. PCMP-H52 was highly up-regulated in senescence stage in *Arabidopsis*. Our results can provide a comprehensive insight into the expression analysis of *PPRs* and their roles in optimizing biological structure and representing varied roles in *PPR* genes.

Introduction

The pentatricopeptide repeat (PPR) proteins are one of the greatest protein families in terrestrial plants. This family has more than 400 members in *Arabidopsis thaliana*, rice, and foxtail millet (Liu et al. 2016). Most researchers have suggested that the PPR proteins are engaged in post

Addres: Abbas Saidi, Shahid Beheshti University, G.C, Tehran, Iran, phone: +98(21) 29901964, e-mail: abbas.saidi@gmail.com

transcriptional control of gene expression in organelles such as plastid and mitochondria (LURIN et al. 2004). The family members are detected by the arrays tandem of PPR motifs, approximately 35 amino acids, which contain highly degenerate units from 2 to 30 motifs. The PPR family in plants can be divided into the P and PLS subfamilies related to the PPR signature motifs. P class PPR proteins consist of 35 amino acids and lack additional domains whereas, PLS class PPR proteins possess three different types of PPR array repeats of P, short (S), and long (L). Many of PLS subfamily members also encompass C-terminal domains subdivided further into a four subsets: PLS, E, E+ and DYW (RIVALS et al. 2006). Studies have shown that PPR proteins are identified to be localized in the mitochondrial or chloroplast intracellular space, whereas few PPR proteins have been detected to inhabit in other cellular sections such as cytosol and/or nucleus. PPR proteins play a particularly significant role in RNA metabolism such as RNA cleavage, splicing, translation, RNA stability, and RNA editing. Several previous studies have shown that PPR proteins mediate some of the various functions in the plant biological and physiological stages. In addition, PPR proteins have key role in response to plant growth and development as well as biotic and abiotic stresses. For example, PPR40 is one of the PPR proteins that provide a signaling link between mitochondrial electron transport. Mutation of *PPR40* resulted in increased accumulation of reactive oxygen species (ROS) which enhanced toxicity in cell.

In *Arabidopsis*, *LPA66* is encoded in the chloroplast and is necessary for conversion of amino acids and mutation in *LPA66* causing a defect at the RNA transcription level (CAI et al. 2009). HAMMANI et al. (2011) revealed that the Organelle Transcript Processing 87 (*OTP87*) gene encoded a PPR protein which was indicated at the editing of nad7 and atp1 transcripts in *Arabidopsis*. The MLT1 is another pentatricopeptide repeat engaged in the translation of mitochondrial nad7 mRNA in *Arabidopsis*. This protein is a localized membrane-bound mitochondrial protein, indicated its function in nad7 mature mRNA translation. In rice, ASL3 encodes a novel PPR protein with 10 tandem repeats, having an essential role in chloroplast development and seedling growth. Recently, molecular evidences have revealed that PPRs play a vital role in organelle biogenesis and function and, subsequently, on growth, development, and various biotic and abiotic stresses (BARKAN and SMALL 2014). In rice, *WSL5* is important for chloroplast ribosome biogenesis under cold stress. Knock-out of *wsl5* resulted in inability to assemble functional ribosomes due to the abnormal splicing of *rpl2* and *rps12*. Consequently, the absence of RPL2 and RPS12 proteins prevent formation of functional ribosomes (LIU et al. 2018).

Empty pericarp12 (EMP12), a PPR protein, is implicated in the splicing of three nad2 introns and seed development in maize. Mutation in Emp12 severely arrests embryo and endosperm development, leading to embryo lethality in maize (Sun et al. 2019). PPR40 is implicated to increase seed and seedling development of the plants under salt stress, whereas, *ppr40* causes an increased accumulation of ROS, enhanced sensitivity to abiotic stresses, and less intense growth retardation. PPR protein SVR7 is localized in *Arabidopsis* chloroplast and is implicated in RNA processing and plastid gene expression. Further, *svr7* mutants have been demonstrated to aggregate under higher levels of ROS and reveal sensitivity to H_2O_2 with reduced photosynthetic activity (Lv et al. 2014). In the present study, comprehensive analysis of *PPR* genes including phylogenetic tree, chromosomal distribution, genes structure, transcription factor binding sites (TFBs), and their gene expression were performed. Our results can provide an understanding on the molecular mechanisms of the *PPR* genes in response to developmental stages and environmental stresses in *Arabidopsis.*

Material and Methods

Phylogenetic Analysis *PPR* Genes of *A. thaliana* and Its Structure Analysis

Gene sequences of 41 PPR proteins of *A. thaliana* was retrieved from the *Arabidopsis* Information Resource (TAIR). All of the gene sequences were confirmed against NCBI and Plant Genome and System Biology (PGSB) databases. Alignment of the sequences of the *PPR* genes was performed using CLUSTALW program with MEGA software version 6. Phylogenetic tree was constructed in the NJ method and diagrams of phylogenetic trees were drawn with MEGA6 software with bootstrap analysis of 1,000 replicates.

Chromosomal Locations and Analysis of TFBS

Chromosome map of *A. thaliana PPR* genes were constructed by Chromosome Map Tools available at TAIR (https://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp). All of *PPR* genes were analyzed to identify their cellular status using CELLO database (http://cello.life.nctu.edu.tw/) (Yu et al. 2006). Promoter regions of 41 PPR proteins were analyzed using PlantPAN (http://plantpan2.itps. ncku.edu.tw/) for the detection of transcription factor binding sites (TFBS) in *PPR* gene promoters. Pfam program was used to find out the PPR proteins domain in the predicted sequences.

Analysis of Gene Ontology and Gene Characterization

The list of *PPR* genes were subjected to GO analysis using the Classification Super Viewer web-based tool (http://bar.utoronto.ca/ntools/cgibin/ntools_classification_superviewer.cgi) available from http://bioinfo.cau.edu.cn/. The tool generated an overview of functional classification of a list of AGI IDs based on the GO database. This database was used to identify the biological processes, molecular functions, and cellular component.

Expression Study of *PPR* Genes

To evaluation the *PPR* gene expression, microarray expression data were taken (ZIMMERMANN et al. 2008) from *Arabidopsis thaliana* database using Affymatrix *Arabidopsis* ATH1 Genome Array. In addition, the genes up and/or down regulated by 1.5 folds were considered as differentially expressed genes (DEG) and these DEG find out using "Perturbations" tool under biotic and abiotic stresses. Differentially expressed genes (DEG) were utilized to generate gene expression heatmap using compendium--wide analysis in genevestigator program. The "red" and "green" colors reflect up and down-regulation of genes, respectively. 'Development' tool was used to identify *PPR* gene expression study using microarray AT-AFFY-ATH1-0 dataset in 10 developmental stage (seed germination, seedling young rosette, developed rosette, bolting, young flower, developed flower, flowers and siliques, senescence).

PPR Proteins Sequence Identification and Domain Analysis

Peptide length, molecular weight, and PI were calculated using the ProParam tool (https://web.expasy.org/protparam/). Characterization of *PPR* genes were performed from PGSB PlantsDB. Sub-cellular localization was predicated using Plant-Mploc server (http://csbio.sjtu. edu.cn/bioinf/plant multi) (HALL 2002). MEME program (http://alternate.meme--suite.org/tools/ meme) and the Pfam tool were utilized to detect the conserved motifs and domains of PPR proteins, respectively. Motifs functions were determined using the hmmscan (https:// www.ebi.ac.uk/Tools /hmmer /search/hmmscan) tool. Then, detected PPR sequences were aligned using Muscle, and identity residue was calculated. GSDS program was utilized to analyze the exon-intron structures of *PPR* genes.

Results and Discussion

Characteristics of 41 *PPR* genes of *A. thaliana* taken from TAIR database are illustrated in table 1 and their locations on each chromosome have also been represented. In this analysis, the names of the PPR proteins, their corresponded protein ID, and the 5'-upstream promoters of each *PPR* gene were surveyed and chromosomal location of the related *PPR* genes have been depicted (Table 1).

Chromosomal Organization of *PPR* Genes and Phylogenetic Analysis

To characterize the genomic distribution of the *PPR* genes on the *A. thaliana* genome, we extracted their chromosomal locations from the PGSB database and found their positions. Total of 41 *PPR* genes were mapped irregularly to the five chromosomes. While chromosome 3 and 5 each were found to possess 10 *PPR* genes, five *PPR* genes were observed to be located in chromosome 4 (Fig. 1). Chromosome 2 was detected to possess the least number of *PPR* genes (4) while the highest number of *PPR* genes (11) were found on chromosome 1. Gene clusters are orders of functionally related genes on a chromosome and clustering of some *PPR* genes were evident on all of the chromosomes, indicating valuable information about their evolution (XING et al. 2018). Accordingly, based on the Holub's criterion, we found 11 *PPR* gene clusters containing a total of eight genes. The clustering patterns of these sequences were compared with their placement on each chromosome, confirming the presence of the sequences of the same chromosomal origin in the same phylogeny (LIU et al. 2016). Only one gene cluster was found on each of the chromosomes 2 and no cluster was found on chromosome 4. Chromosome 5 has the highest number (5) of gene clusters, whereas the least number (only 1) of gene cluster was identified on chromosome 2. The phylogenetic tree was constructed on the nucleotide sequences of 41 *Arabidopsis* PPR proteins using the NJ method. The tree was classified into two distinct subfamilies (P and PLS subfamily) – Figure 2. However, some of PPR members of the PLS subfamily were clustered with the P subfamily, which is consistent with the results from the *Arabidopsis* phylogenetic analysis in which some of the PPR proteins possessed the PLS structure, but were clustered into the P subfamily. Our results agreed with XING et al. (2018) where some of the PPR proteins were classified in PLS subfamily and were clustered with the P subfamily (XING et al. 2018).

Table 1

Description of *Arabidopsis PPR* genes and their cell position

No.	Specification	Gene name	Promoter regions	Localization	Gene size	Chromosome number	Protein length (aa)	Strand	Type of domain
1	*AT1G09410*	PCMP-H18	3035400-3037687	mitochondrial	2118	1	705	+strand	PLS-type
2	*AT3G11460*	PCMP-H21	3608250-3610121	mitochondrial	1872	3	623	+ strand	PLS-type
3	*AT5G52630*	PCMP-E49	21350375-21352333	chloroplast	1767	5	588	+ strand	P-type
4	*AT1G20230*	PCMP-H69	7009568-7012107	mitochondrial	2539	1	759	+ strand	PLS-type
5	*AT2G37310*	PCMP-E5	15664800-15667115	mitochondrial	2280	2	759	-strand	P-type
6	*AT3G12770*	PCMP-E51	4056953-4059284	chloroplast, mitochondrial	2203	3	733	-strand	PLS-type
7	*AT1G03100*	PCMP-E55	743885-746701	mitochondrial	2595	1	864	-strand	PLS-type
8	*AT3G02010*	PCMP-E56	337906-340442	mitochondrial	2478	3	825	-strand	PLS-type
9	*AT1G56690*	PCMP-E57	21253686-21256048	mitochondrial	2115	1	704	+ strand	P-type
10	*AT5G61400*	PCMP-E35	24681550-24683514	mitochondrial	1965	5	654	+ strand	P-type
11	*AT5G37570*	PCMP-E103	14923911-14926333	chloroplast	1653	5	550	-strand	PLS-type
12	*AT1G77010*	PCMP-H13	28942710-28944827	chloroplast, mitochondrial	2088	1	695	+ strand	PLS-type
13	*AT3G15930*	PCMP-H3	5387444-5389690	chloroplast	2247	3	748	+ strand	PLS-type
14	*AT1G26900*	PCMP-H38	9319643-9321512	mitochondrial	1719	1	572	-strand	PLS-type
15	*AT1G31430*	PCMP-E105	11253912-11255745	chloroplast	1802	1	599	-strand	P-type
16	*AT1G32415*	PCMP-E25	11695596-11697964	mitochondrial	2286	1	761	+ strand	P-type
17	*AT1G33350*	PCMP-E101	12089249-12091743	mitochondrial	2285	1	760	-strand	P-type
18	*AT5G52850*	PCMP-E29	21414935-21417616	chloroplast, mitochondrial	2682	5	893	-strand	PLS-type
19	*AT3G25970*	PCMP-H56.1	9500016-9502253	chloroplast	2106	3	701	+ strand	PLS-type
20	***AT2G17140***	**PCMP-H52**	**7462809-7466898**	mitochondrial	3816	2	1271	+ strand	PLS-type
21	*AT5G43790*	PCMP-H43	17591929-17593666	chloroplast	1475	5	490	-strand	PLS-type
22	*AT3G14730*	PCMP-H36	4949178-4951346	mitochondrial	1962	3	653	-strand	PLS-type

23	*AT2G39620*	PCMP-H84	16518890-16521544	mitochondrial	2511	2	836	-strand	PLS-type
24	*AT4G20770*	PCMP-H22	11130762-11133086	mitochondrial	2325	4	774	-strand	PLS-type
25	*AT2G46050*	PCMP-H5	18939262-18941034	mitochondrial	1773	2	590	+ strand	PLS-type
26	*AT3G08820*	PCMP-H52.1	2676990-2679265	chloroplast	2062	3	686	-strand	PLS-type
27	*AT3G02330*	PCMP-H58	473774-476662	mitochondrial	2712	3	903	-strand	PLS-type
28	*AT5G47460*	PCMP-E37	19252463-19254336	chloroplast	1731	5	576	-strand	P-type
29	*AT1G68930*	PCMP-E54	25918066-25921034	mitochondrial	2232	1	743	+ strand	P-type
30	*AT4G14050*	PCMP-H31	8103521-8105637	mitochondrial	1877	4	624	-strand	PLS-type
31	*AT4G37170*	PCMP-E46	17498411-17500655	mitochondrial	2076	4	691	-strand	P-type
32	*AT3G22690*	PCMP-E30	8021229-8024594	chloroplast, mitochondrial	3306	3	1101	-strand	P-type
33	*AT5G50390*	PCMP-E31	20520600-20523212	chloroplast, mitochondrial	2192	5	729	-strand	P-type
34	*AT4G14820*	PCMP-E39	8507268-8510113	chloroplast	2245	4	747	-strand	P-type
35	*AT5G48910*	PCMP-E45	19832785-19835148	chloroplast	2063	5	686	-strand	PLS-type
36	*AT5G08305*	*AT2G17140*	2669944-2671782	chloroplast	1632	5	543	-strand	PLS-type
37	*AT4G38010*	At1g03100	17859338-17861407	mitochondrial	1680	4	559	-strand	PLS-type
38	*AT5G16420*	At5g61400	5367971-5370242	mitochondrial	1608	5	535	+ strand	P-type
39	*AT1G09220*	**PCMP-E33**	2977792-2979466	mitochondrial	1675	1	557	-strand	P-type
40	*AT4G18840*	PCMP-E90	10336566-10340378	mitochondrial	1638	4	545	-strand	P-type
41	*AT3G21470*	At5g16420	7563503-7565160	chloroplast	1572	3	523	+ strand	P-type

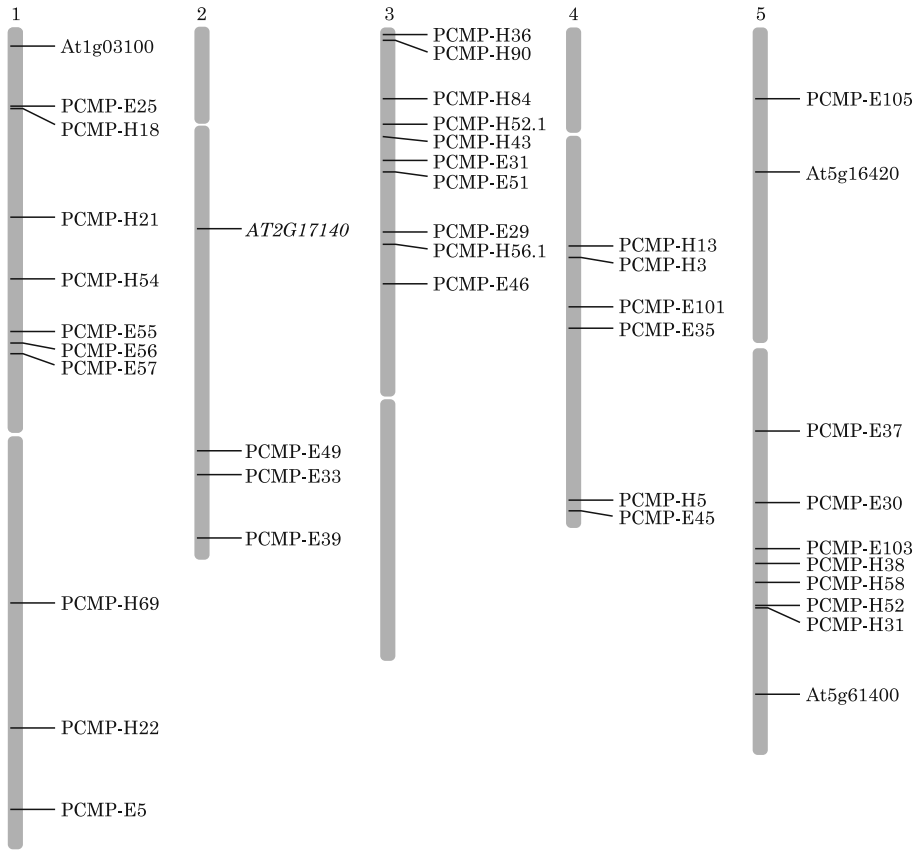

Fig. 1. Chromosomal distribution of 41 different *Arabidopsis PPR* genes

Phylogenetically, 41 *PPR* genes were divided into two different clusters with bootstrap values of 0 to 100 (Fig. 2). Cluster 1 (PLS subfamily) contained nine sequences from chromosome 1, while cluster 2 (P subfamily) had two sequences belonging to the chromosome 1. Cluster 1 contained four sequences (PCMP-H5, PCMP-H52, PCMP-E5, and PCMP-H84) from chromosome 2 while cluster 2 was formed of only one sequence (*AT2G17140*) from chromosome 2. Cluster 2, as the smallest clade, included 15 sequences; of which one (PCMP-E56) was from chromosomes 3, four sequences (3100, E39, E46, E90) from chromosome 4, and seven sequences (PCMP-E29, *AT5G61400*, PCMP-E103, PCMP-E35, PCMP-E31, and PCMP-E49*)* from chromosome 5. Cluster 1, as the largest clade, consisted three nucleotide sequences from chromosome 5 and the remaining nine sequences belonged to chromosome 1, two sequences were belonged to chromosome 4, eight sequences belonged to chromosome 3, and four sequences belonged to chromosome 2.

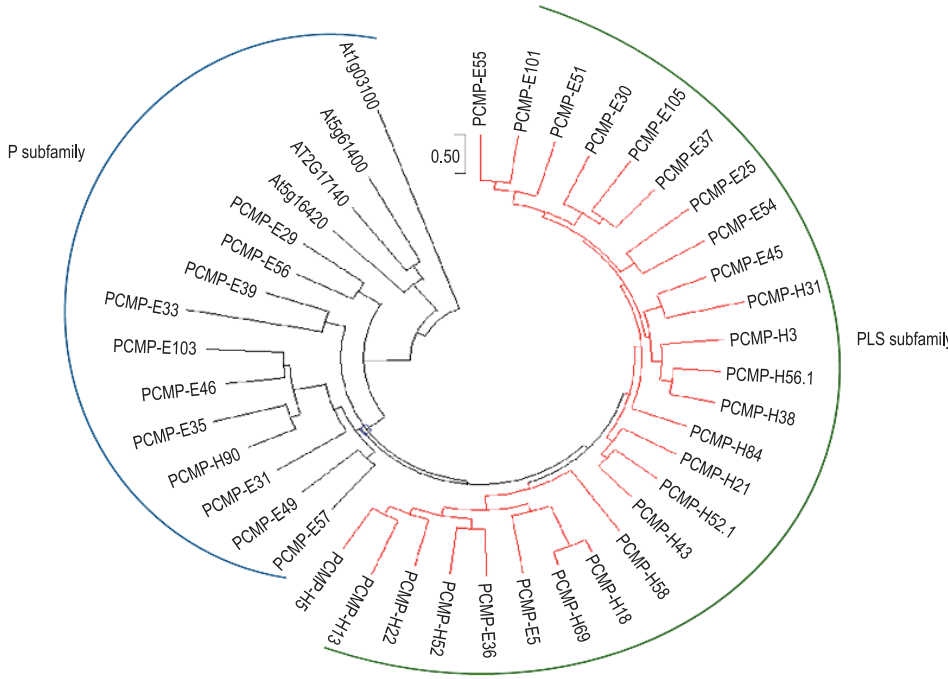

Fig. 2. Molecular Phylogenetic analysis of *PPR* genes from *A. thaliana* by Maximum Likelihood method. The green demonstrated the P class and the black demonstrated the PLS class. The tree was produced in MEGA7.0 with the NJ method and 1000 bootstrap replicates

Analysis of the TFBs

TF families (MBY, bZIP, WRKY, Homeodomain, and AP2) have been identified on the promoter regions of both strands, and were mostly located in the upstream region of 1000 bp. Description of the first five most frequently occurring TFBs of the total detected elements is provided in Table 2. In the following section the importance of surveyed TFBs in the *PPR* genes upstream promoter regions have been investigated, carefully considering their correlation with regulatory patterns and their roles.

Mby. MYB transcription factors play a key role in controlling various processes like metabolism, development, differentiation, defense and responses to biotic and abiotic stresses (AMBAWAT et al. 2013, SAIDI et al. 2020a). Microarray expression study showed at *MYB60* gene is up-regulated upon exposure pathogen and abiotic stress (RASHEED et al. 2016). Five MYB members, *MYB15*, *MYB20, MYB44, MYB52,* and *MYB96* genes are implicated in drought, ABA, salt or/and cold responses. In another study *OsMYB48-1* has been over-expressed in transgenic rice which resulted in lower rate of water loss and improved drought tolerance in comparison to non-transgenic line under drought stress (XIONG et al. 2014).

Table 2

Summary of the transcription factor binding sites (TFBS) detected in the promoter regions of *PPR* genes in *A. thaliana*

TF/ Motifs related to	Specification	AT1G03100	AT1G09220.1	AT1G09410.1	AT1G20230.1	AT1G31430.1	AT1G32415.1	AT1G33350.1	AT1G56690.1	AT1G68930	AT1G77010.1	AT2G17140	AT2G37310.1	AT2G39620.1	AT2G46050	AT3G02010.1	AT3G02330.1	AT3G08820.1	AT3G11460.1	AT3G12770	AT3G14730.1	AT3G15930.1	AT3G21470	AT3G22690	AT3G25970.1	AT4G14050.1	AT4G14820.1	AT4G18840	AT4G20770.1	AT4G37170.1	AT4G38010.1	AT5G08305	AT5G16420.1	AT5G37570.1	AT5G43790	AT5G47460	AT5G48910.1	AT5G50390.1	AT5G52630	AT5G52850.1	AT5G61400
Hormone response	EIN3; EIL	2	4	2	6	3	5	6	4	2	6	3	2	5	2	6	4	3	2	6	3	5	5	6	1	0	1	5	6	5	6	0	3	3	2	1	3	1	0	1	4
	AP2	4	16	15	15	13	16	32	20	13	30	28	49	30	25	34	25	21	22	25	30	25	6	24	31	33	58	30	37	18	13	21	15	17	6	26	40	26	14	17	14
	BES1	0	0	0	0	0	0	3	0	0	0	2	2	2	2	3	0	2	2	0	0	2	3	2	0	0	1	2	0	2	3	0	2	0	0	0	1	0	0	0	0
Tissue-specific	TCR	3	1	2	5	1	2	5	3	4	3	4	1	1	5	1	4	3	2	4	5	4	2	3	6	3	5	4	3	3	3	0	5	3	6	0	0	2	4	2	3
	SBP	17	3	3	4	21	6	20	3	4	20	22	22	6	3	18	19	21	23	3	20	4	21	23	3	18	22	3	12	3	35	3	2	12	3	19	14	4	15	4	2
	WOX	1	3	4	2	0	2	1	2	1	1	1	3	3	3	0	3	3	3	3	4	5	2	4	5	2	4	3	4	1	4	2	5	4	3	3	4	4	3	0	3
	AT-Hook	12	7	11	8	6	6	8	8	13	4	23	5	12	6	5	6	5	4	9	9	10	5	11	6	12	12	10	11	6	8	3	4	9	17	9	14	14	6	15	7
Light response	GATA	13	4	16	13	17	16	18	16	14	9	15	5	17	11	13	12	18	10	17	16	11	9	18	17	12	17	14	14	8	14	17	15	15	14	17	11	12	14	12	17
Cell cycle	E2F	2	2	0	2	1	2	2	4	0	1	1	2	0	0	1	2	1	1	1	2	0	1	1	1	1	2	1	1	1	1	0	2	1	0	1	1	1	1	1	1
Basic transcription	NF-YB	2	3	2	2	1	2	1	2	1	1	1	2	3	2	2	2	2	1	3	2	1	2	2	1	1	2	1	1	1	1	1	8	5	1	1	1	1	2	1	1
	TBP	4	5	2	9	1	0	8	3	8	7	5	7	5	2	3	3	1	4	1	6	4	4	5	7	7	8	7	2	5	5	8	5	3	4	8	6	7	2	3	7

Other binding	TCP	2	9	3	5	3	3	4	3	4	2	4	4	3	0	9	4	6	3	15	4	3	2	8	4	5	9	11	9	2	2	3	5	4	4	4	6	3	9	4	6
	WRKY	26	26	6	14	37	20	28	3	32	24	35	31	29	20	10	16	31	23	19	15	31	19	19	3	13	34	26	27	19	2	34	23	12	13	26	24	12	2	16	24
	MYB	22	22	55	39	19	15	14	16	57	41	55	39	38	47	39	39	12	40	44	51	46	37	46	41	35	41	47	50	41	19	26	43	58	31	36	45	50	47	43	40
	ZF-HD	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	GRAS	0	0	0	6	2	1	2	1	1	2	1	1	0	0	0	0	2	2	0	0	0	1	1	0	1	1	0	0	0	1	0	0	0	0	2	0	0	0	0	0
	bZIP	8	19	30	17	24	20	35	20	28	34	42	41	40	35	42	25	34	33	39	23	38	32	47	27	14	46	32	27	23	18	25	30	9	36	28	10	21	31	29	31
	Dof	10	16	12	6	15	16	14	10	8	10	15	10	9	12	15	15	12	10	6	14	7	12	11	8	14	15	15	12	15	10	10	12	10	5	3	9	15	13	14	12
	bHLH	8	12	6	8	7	9	40	10	17	12	13	39	40	35	40	10	16	40	15	9	10	38	12	8	7	10	14	8	34	14	10	10	10	10	12	10	9	8	16	11
	NAC	9	9	18	17	10	12	18	15	7	6	10	5	4	9	17	10	9	7	7	6	7	7	12	12	9	10	8	9	5	6	9	11	8	4	5	8	7	7	8	6
Stress response	CDS	0	0	0	1	2	2	2	1	1	1	1	2	1	1	2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	HSF	3	2	1	1	2	2	2	2	2	1	1	0	0	0	0	0	0	3	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	C2H2	7	7	8	8	9	7	7	9	9	9	8	9	7	5	10	7	5	5	17	8	4	7	14	8	16	13	5	4	3	6	8	7	6	21	5	11	16	1	9	6
	Homeo-domain	12	18	16	26	11	14	13	17	20	16	22	14	16	14	21	20	16	19	20	29	28	13	22	25	11	26	25	15	14	16	17	14	17	15	20	28	12	23	13	15

bZIP. In plants, bZIPs are master regulators of several processes including seed formation, pathogen defense, light signaling, abiotic and biotic stress responses (RAHAIE et al. 2011, SAIDI et al. 2020b). Some bZIP are activated by salt stress, and act as salt stress sensor in *Arabidopsis*. In addition, stress-inducible expression through participating transfer from the endoplasmic reticulum to the nucleus and subsequently up-regulation of salt stress genes. Beside abiotic stress control, nine of bZIP TFs increased expression during the course of *Ustilago maydis* infection in maize where same expression profile was observed by *Colletotrichum graminicola* infection (WEI et al. 2012).

WRKY. Members of this family of WRKY contain at least a conserved class of DNA-binding region for abiotic and biotic stress management in rice (ROSS et al. 2007). The upregulation of some members of WRKY TFs have been reported to be positive regulators of drought tolerance. Also, under salt and pathogen infection, most WRKY were significantly upregulated (JIANG et al. 2009, SAIDI and HAJIBARAT 2019).

Homeodomain. Homeodomain encoded by homeobox genes contain a specific DNA sequence that provides instruction for making a string of 60 protein building blocks. This TFBs was increased against pathogen infection (COEGO et al. 2005)

AP2. AP2 has a role in controlling seed mass, seed development, and development of the ovule and seed coat (OHTO et al. 2005). AP2, as a novel role, is incorporated in the floral homeotic gene APETALA2 during *Arabidopsis* fruit development (RIPOLl et al. 2011).

In the present study, five TFBs were identified among 41 *PPR* gene promoters with the greatest number of TFBs detected in MYB and least number of TFBs identified in AP2 (Table 2). Also, the maximum and minimum number of TFBs was observed on PCMP-H52 and PCMP-E55, respectively. Although the correlation between TFBs and genes response under stress conditions need more experimental and systematic analysis of most of PPR proteins, these results only showed the stress-responsive nature of *PPR* genes (CHEN et al. 2018). In silico analysis of transcription factor binding site has demonstrated that the availability of a TFBs bZIP in PCMP-H52 is far higher as compared to PCMP-E55 promoter region. Pervious study is shown that TFBs "bZIP" has been identified to be implicated in drought, senescence, and pathogen defense response (JAKOBY et al. 2002, CHEN et al. 2018) and it could be expected that PCMP-H52 might be highly responsive to drought stress. Further experimentation is required with the high and low TFBs possessing PPR promoters to unravel their biological significance (DAS et al. 2019).

Gene Expression of *PPR* Genes in Developmental Stages

To investigate the expression profile of the PPR proteins in *Arabidopsis* development, we analyzed microarray dataset in genevestigator (Fig. 3).

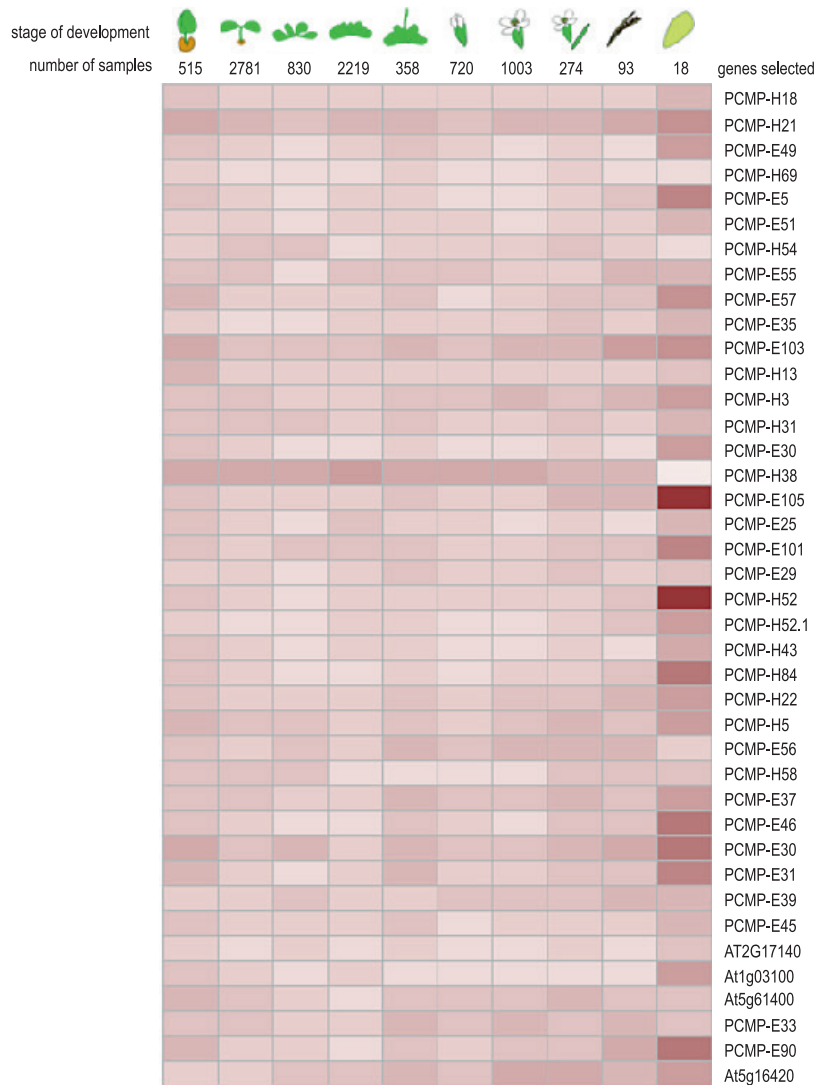

Fig. 3. Expression profiles of *PPR* genes at different developmental stages of *A. thaliana*

PCMP-H38 was a highly expressed in all stages of development except senescence, where, PCMP-E105 and PCMP-H52 were highly expressed in senescence stage. PCMP-H21 was slightly up-regulated in all developmental stages in *Arabidopsis* (Fig. 3). In the course of seed germination stage, all

genes showed almost similar level of expression. Expression of all genes were maximal during senescence and seed germination stage whereas, expression of these genes was minimal during the young rosette and flower development stages.

Gene Expression of *PPR* Genes in Abiotic and Biotic Stresses

Expression analysis of *A. thaliana PPR* genes in response to abiotic stress was surveyed by genevestigator (Fig. 4). Based on our results, a positive correlation was obtained between PPR up-stream promoter regions of genes and presence of TFBs. Based on the available microarray data, it has been observed that PCMP-H52 and PCMP-E105 were highly up-regulated in response to iron deficiency and shift low to high light stresses. Whereas, these genes were down-regulated in RNA labeling and EMS mutation (Fig. 5). Additionally, the gene expression data showed that the gene expression of most *PPR* genes were up-regulated in response to a shift low to high light stress. But, majority of *PPR* genes were down-regulated in response to RNA labeling and EMS mutation. Furthermore, some *PPR* genes were up-or down-regulated in response to cordycepin stress whereas, others showed no response to cordycepin stress.

Plants utilize complex signaling pathways containing stress-related TFBs and regulate their compatibility to changing stresses. To identify linked TFBs to abiotic stresses, the 5' up-stream sequences of *PPR* genes of *A. thaliana* were surveyed using PlantPAN. In this study, PCMP-H52 and PCMP-E105 were expressed under iron deficiency and shift low to high light stresses perhaps due to the presence of both MYB and bZIP. MYB has been shown to possess regulatory effects on cell fate, hormonal action, response to environmental factors, as well as in the control of Fe transport, and tissue partitioning under iron deficiency (WANG et al. 2018).

Microarray data also showed up-regulation of most genes during iron deficiency which can be related to the presence of bZIP. These results are in agreement with the finding that bZIP possesses indirect roles in Fe-response in *Arabidopsis* (SINCLAIR et al. 2018). Cordycepin, a transcription inhibitor, causes premature termination of protein synthesis. PCMP-H58 and PCMP-E33 were up-regulated after cordycepin; this can be due to the presence of MYB and bZIP in their promoter regions. Whereas, other genes were down-regulated under cordycepin stress. BZIP present in PCMP-E35, PCMP-E5, and PCMP-E51 play an important role in glucose-ABA interaction network, regulating mRNA decay in cordycepin stress (MATIOLLI et al. 2011). Most of *PPR* genes are highly expressed under a shift low to high light stress condition which can be due to presence of MYB TFBs, which act as a clock-controlled element (YANG et al. 2018).

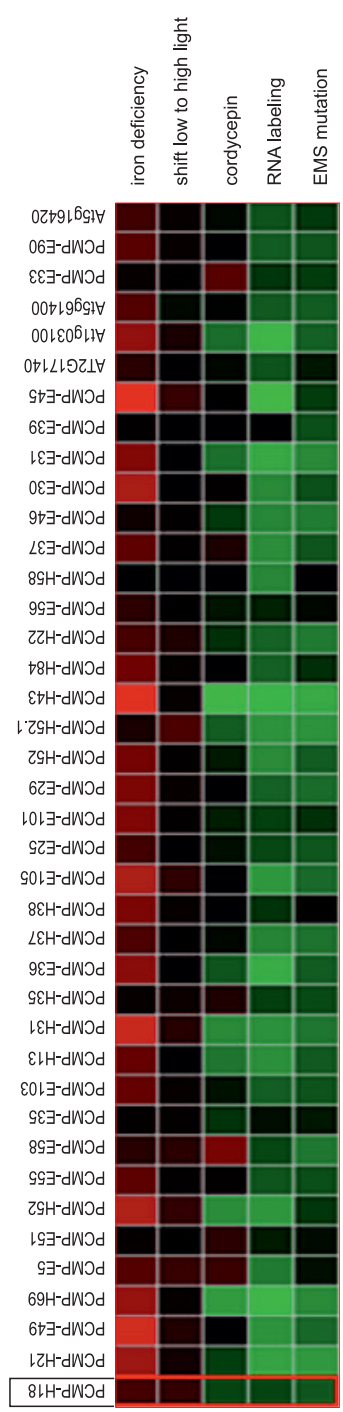

Fig. 4. Heatmap representation of expression analysis in response to biotic stresses of *A. thaliana PPR* genes based on microarray data present at Genevestigator

Structure and Characteristic Analysis of *PPR* Genes

The characteristics of the *PPR* genes were analyzed in detail. The length of protein sequences of *PPR* genes ranged from 490 (PCMP-H43) to 1271 (PCMP-H52) amino acids and their gene size ranged from 1475 (PCMP-H43) to 3816 (PCMP-H52) KDa (Table 1).

To examine the structural diversity of *PPR* genes, exon-intron distribution and conserved motifs were analyzed according to maize and Brachypodium (SAIDI and HAJIBARAT 2018, SUN et al. 2019). Gene structure analysis revealed that the number of introns in the *PPR* genes of the three clusters ranged from 1 to 2. Most of the PPR members were classified in the same subfamily. Both PCMP-H43 and PCMP-E35 genes each contained two exons. The PCMP-H43 and PCMP-E35 possess 1 intron (Fig. 5).

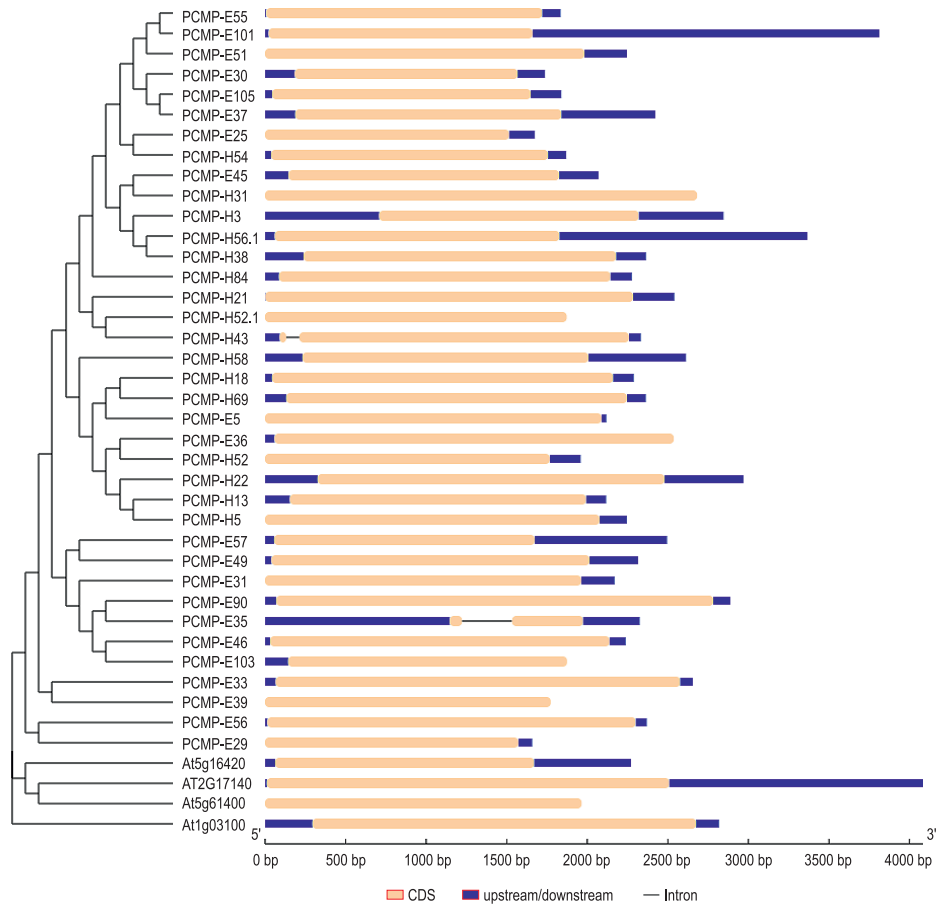

Fig 5. Distributions of the conserved motifs detected by GSDS and displayed in different colored boxes

Multiple Sequence Alignment, Conserved Motifs, and Domain Architectures in PPR Proteins

To obtain more understanding into the structure characteristics of the PPR proteins and conserved motif analysis, their amino acid sequences were submitted to MEME program. As shown in Figure 6, three motifs were identified in the PPR family members which were explored to encode functional domains when subjected to Pfam. Motif 3 was annotated as DYW family of nucleic acid deaminases while, motifs 1 and 2 were assigned by the Pfam as PPR domain. Highly similarity motifs are expected to have similar functions. DYW family belonged to PLS group having three motifs sequences namely, motif 1, 2, and 3. PPR proteins relevant to P group contained two motifs (1 and 2) (Fig. 5, Fig. 7).

The sequence alignment was performed among the reduced amino acid sequences of the 41 templates by Muscle, and identity residue was calculated (Fig. 6). The three general domain such as E, E^+, and DYW were identified as the dedicated motifs in the domain PLS-type proteins. According to other findings, the PLS subfamily have four subclasses: 1) proteins that do not possess none of the three motifs, 2) proteins have only E motif, 3) proteins with both E and E^+ motifs, and 4) proteins with the E, E^+, and DYW motifs (Fig. 8). Whereas, P-type proteins lack E, E^+ and DYW. Table 1 and Table 2 supplies details of the numbers of *PPR* genes in each subgroup and characterization of type motifs. The largest difference is the number of *PPR* genes in the PLS/P subgroups, with 24 *PPR* genes in PLS-type but only 17 in P-type (Table 1, Fig. 1). Previous studies have shown that these three motifs were located in C-terminal ends of *PPR* proteins and which were only present in PPR proteins and PCMPs and not in any other proteins of *Arabidopsis* (LURIN et al. 2004). AUBOURG et al. (2000) concluded that both E and E^+ motifs were extremely degenerate, but DYW motif was highly conserved in the amino acid sequence (Fig. 6).

As described earlier, the DYW domain is a typical feature of the PLS subfamilies. Hence, the domain architectures in PLS subfamilies were analyzed. Results showed that most of the PPR proteins included the DYW domain. However, some members of the PLS groups do not possess a DYW domain, required for site-specific editing factors in chloroplast. The DYW domain, one of the PLS family, is the candidate domain for cytidine deaminase, a highly evolutionarily domain correlated with RNA editing. While, P-type is a type of classical PPR proteins with p-motifs, having a vital role in RNA-interacting with other protein and RNAs ae well as the role played by combinatorial motifs (Fig. 6).

PLS-subfamily
E
E+
DYW

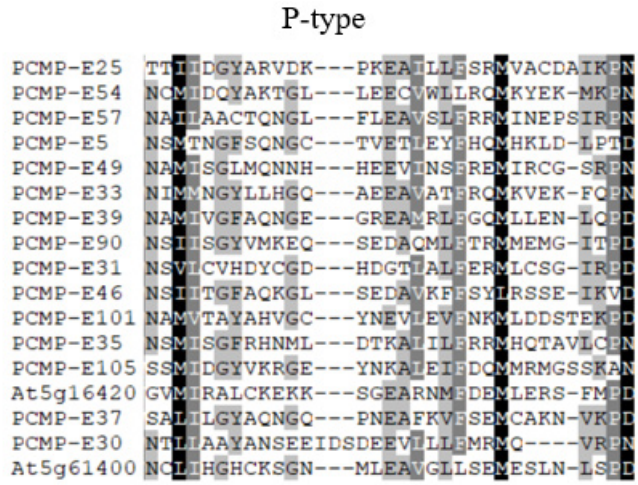

Fig. 6. Sequence alignment of the PPR motifs in *A. thaliana*. The amino acid sequences of the three domains are shown in boxes

Gene Ontology Annotation and Subcellular Localization Prediction

GO analysis was performed the using BAR database, suggesting the putative participation of PPR in various biological processes, molecular functions and cellular components (Fig. 8). All 41 PPR proteins were grouped into seven separate categories of biological processes. Our finding suggested that a most of PPR were probably associated to metabolic and cellular, followed by cell organization and biogenesis processes within mitochondria and chloroplasts including RNA editing, RNA splicing, RNA stability and translation, response to abiotic or biotic stimulus, and response to stress. PPR proteins were indicated to participate in molecular function such as hydrolase activity and other binding. The second most frequently annotated molecular function was nucleic acid and protein binding function, which is in agreement with the role of PPR10 protein in interaction with protein/RNA using two binding sites (BARKAN et al. 2012). Cellular component prediction suggested that PPR proteins were localized in mitochondria (60%), chloroplasts (30%), ribosome (6%), and cytosol (4%). It has been reported that PPR proteins are targeted approximately 54% in mitochondria and 28% in chloroplast (CHEN et al. 2018). The GO analysis results predicted that the PPR proteins incorporated in different biological and molecular processes, seed development, and fertility and can provide useful information for further gene functions studies in *Arabidopsis*. Our results have revealed that great PPR proteins are localized in mitochondria and a few in chloroplast. It has been reported that the majority of PPR proteins are indicated to target both mitochondria and chloroplast (SMALL and PEETERS 2000).

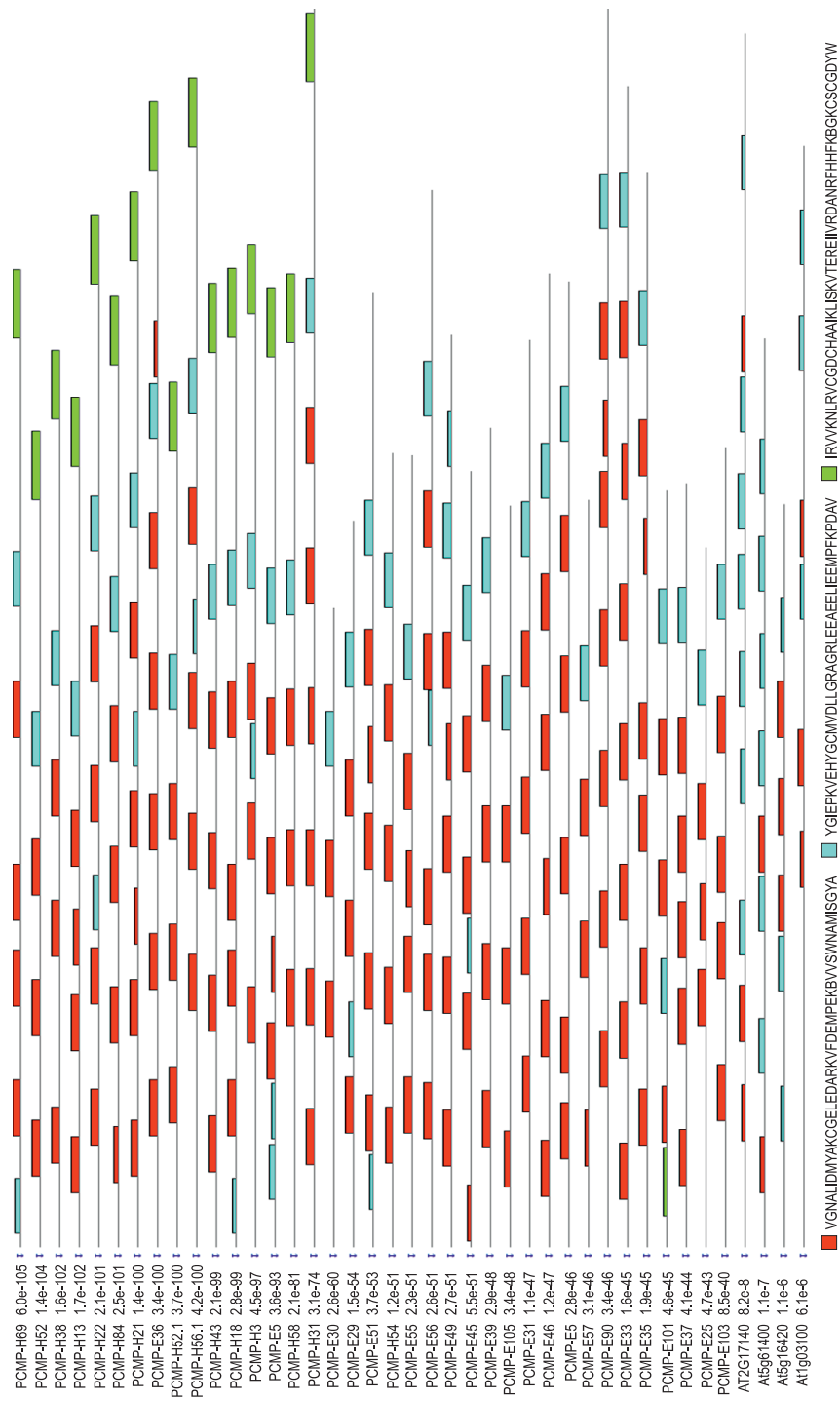

Fig. 7. Conserved motifs detected by MEME and displayed in different colored boxes

Biological process			
2.54	*0.289*	**7.254e-08**	unknown biological processes (Input set freq.: 0.65; 0.25)
1.85	*1.301*	0.318	electron transport or energy pathways (Input set freq.: 0.02; 0.01)
0.53	*0.163*	**5.105e-03**	other metabolic processes (Input set freq.: 0.21; 0.41)
0.51	*0.157*	**3.078e-03**	other cellular processes (Input set freq.: 0.21; 0.42)
0.27	*0.196*	0.092	cell organization and biogenesis (Input set freq.: 0.02; 0.08)
0.21	*0.152*	**0.041**	response to abiotic or biotic stimulus (Input set freq.: 0.02; 0.11)
0.19	*0.136*	**0.023**	response to stress (Input set freq.: 0.02; 0.12)
Molecular function			
1.6	*0.291*	**0.012**	other binding (Input set freq.: 0.41; 0.25)
1.48	*0.55*	0.095	hydrolase activity (Input set freq.: 0.17; 0.11)
1.14	*0.427*	0.152	DNA or RNA binding (Input set freq.: 0.17; 0.14)
1.04	*0.276*	0.137	unknown molecular functions (Input set freq.: 0.26; 0.25)
0.51	*0.256*	0.090	protein binding (Input set freq.: 0.07; 0.14)
Cellular component			
6.9	*0.437*	**2.930e-26**	mitochondria (Input set freq.: 0.85; 0.12)
1.37	*0.954*	0.356	ribosome (Input set freq.: 0.02; 0.01)
0.99	*0.398*	0.174	chloroplast (Input set freq.: 0.14; 0.14)
0.65	*0.402*	0.224	unknown cellular components (Input set freq.: 0.04; 0.07)
0.63	*0.232*	0.069	other intracellular components (Input set freq.: 0.14; 0.23)
0.33	*0.236*	0.148	cytosol (Input set freq.: 0.02; 0.07)
0.17	*0.099*	**1.771e-04**	other cytoplasmic components (Input set freq.: 0.04; 0.27)
0.1	*0.073*	**1.506e-04**	other membranes (Input set freq.: 0.02; 0.24)

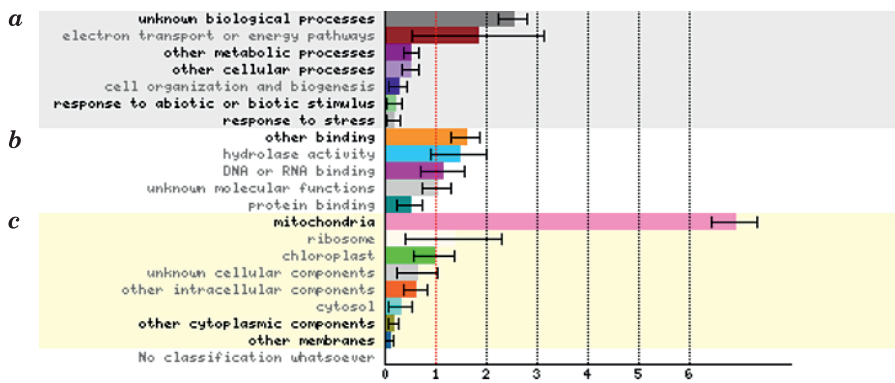

Fig. 8. GO classification of the PPR proteins. The BAR database defined the GO under three main categories: (*a*) biological processes; (*b*) molecular functions and (*c*) cellular component

Conclusion

In the current study, computational analysis of 41 PPR proteins of *Arabidopsis* revealed the presence of various types of TFBs. Consequently, analysis of PPR proteins on phylogeny, transcription factor binding sites (TFBs), chromosomal location, stress associated TFBs, expression profiles in different tissues and biotic stresses, and analysis of conserved motifs were performed based on bioinformatics. Our findings identified five

important TFBs such as MYB, bZIP, WRKY, Homeodomain, and AP2 which were involved in abiotic and biotic stress. According to analysis of TFBs, PCMP-H52 was found to be different from other *PPR* genes. The PCMP-H52 revealed a significant role in different abiotic stresses, possibly due to having multiple TFBs in its promoter region. Microarray data indicated that PCMP-H52 gene was up-regulated during senescence stage and iron deficiency and a shift low to high light stresses. Also, analysis of PCMP-H52 gene expression revealed that this gene can be considered for resistance to abiotic stresses in plant breeding programs.

Accepted for print 5.10.2020

References

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 Molecular Biology, 42: 603–613.

AMBAWAT S., SHARMA P., YADAV N.R., YADAV R.C. 2013. *MYB transcription factor genes as regulators for plant responses: an overview*. Physiology and Molecular Biology of Plants, 19: 307–321.

BARKAN A., ROJAS M., FUJII S., YAP A., CHONG Y.S., BOND C.S. SMALL I. 2012. *A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins*. PLoS genetics 8.

BARKAN A., SMALL I. 2014. *Pentatricopeptide repeat proteins in plants*. Annual Review of Plant Biology, 65: 415–442.

CAI W., JI. D., PENG L., GUO J., MA J., ZOU M., LU C., ZHANG L. 2009. *LPA66 is required for editing psbF chloroplast transcripts in Arabidopsis*. Plant Physiology, 150: 1260–71.

CHEN G., ZOU Y., HU ., DING Y. 2018. *Genome-wide analysis of the rice PPR gene family and their expression profiles under different stress treatments*. BMC Genomics, 19: 720.

COEGO A., RAMIREZ V., GIL M.J, FLORS V., MAUCH-MANI B., VERA P. 2005. *An Arabidopsis homeodomain transcription factor, OVEREXPRESSOR OF CATIONIC PEROXIDASE 3, mediates resistance to infection by necrotrophic pathogens*. The Plant Cell., 17: 2123–37.

DAS A., PRAMANIK K., SHARMA R., GANTAIT S., BANERJEE J. 2019. *In-silico study of biotic and abiotic stress-related transcription factor binding sites in the promoter regions of rice germin--like protein genes*. PloS one 14.

Hall J. 2002. *Cellular mechanisms for heavy metal detoxification and tolerance*. Journal of Experimental Botany, 53: 1–11.

HAMMANI K., DES FRANCS-SMALL C.C., TAKENAKA M., TANZ S.K., OKUDA K., SHIKANAI T., BRENNICKE A., SMALL I. 2011. *The pentatricopeptide repeat protein OTP87 is essential for RNA editing of nad7 and atp1 transcripts in Arabidopsis mitochondria*. Journal of Biological Chemistry, 286: 21361–71.

JIANG Y., DEYHOLOS M.K. 2009. *Functional characterization of Arabidopsis NaCl-inducible WRKY25 and WRKY33 transcription factors in abiotic stresses*. Plant Molecular Biology, 69: 91–105.

JAKOBY M., WEISSHAAR B., DRÖGE-LASER W., VICENTE-CARBAJOSA J., TIEDEMANN J, KROJ T., PARCY F. 2002. *bZIP transcription factors in Arabidopsis*. Trends in Plant Science, 7: 106–11.

LIU J.M., XU Z.S., LU P.P., L.I. W.W., CHEN M., GUO C.H. MA Y.Z. 2016. *Genome-wide investigation and expression analyses of the pentatricopeptide repeat protein gene family in foxtail millet*. BMC Genomics, 17: 840.

LIU X., LAN J., HUANG Y., CAO P., ZHOU C., REN Y., HE N., LIU S., TIAN Y., NGUYEN T., JIANG L. 2018. *WSL5, a pentatricopeptide repeat protein, is essential for chloroplast biogenesis in rice*

under cold stress. Journal of Experimental Botany, 69: 3949–61.

Lv H.X., Huang C., Guo G.Q., Yang Z.N. 2014. *Roles of the nuclear-encoded chloroplast SMR domain-containing PPR protein SVR7 in photosynthesis and oxidative stress tolerance in Arabidopsis*. Journal of Plant Biology, 57: 291–301.

Lurin C., Andrés C., Aubourg S., Bellaoui M., Bitton F., Bruyère C., Caboche M., Debast C., Gualberto J., Hoffmann B., Lecharny A. 2004. *Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis*. The Plant Cell, 16: 2089–2103.

Matiolli C.C., Tomaz J.P., Duarte G.T., Prado F.M., Del Bem L.E., Silveira A.B., Gauer L., Corrêa L.G., Drumond R.D., Viana A.J., Di Mascio P. 2011. *The Arabidopsis bZIP gene AtbZIP63 is a sensitive integrator of transient abscisic acid and glucose signals*. Plant Physiology, 157: 692–705.

Ohto M.A., Fischer R.L., Goldberg R.B., Nakamura K., Harada J.J.2005. *Control of seed mass by APETALA2*. Proceedings of the National Academy of Sciences, 102: 3123–3128.

Rahaie M., Gomarian M., Alizadeh H., Malboobi M.A., Naghavi M.R. 2011. *The expression analysis of transcription factors under long term salt stress in tolerant and susceptible wheat genotypes using reverse northern blot technique*. Iranian Journal of Crop Sciences, 13: 580–95.

Rasheed S., Bashir K., Matsui A., Tanaka M., Seki M. 2016. *Transcriptomic analysis of soil-grown Arabidopsis thaliana roots and shoots in response to a drought stress*. Frontiers in Plant Science, 7: 180.

Ripoll J.J., Roeder A.H., Ditta G.S., Yanofsky M.F. 2011. *A novel role for the floral homeotic gene APETALA2 during Arabidopsis fruit development*. Development, 138: 5167–76.

Rivals E., Bruyere C., Toffano-Nioche C., Lecharny A. 2006. *Formation of the Arabidopsis pentatricopeptide repeat family*. Plant Physiology, 141: 825–39.

Ross C.A., Liu Y., Shen Q.J. 2007. *The WRKY gene family in rice (Oryza sativa)*. Journal of Integrative Plant Biology, 49: 827–42.

Saidi A., Hajibarat Z. 2018. *In silico Analysis of floral MADS-BOX gene in Brachypodium distachyon*. Bionature, 22: 366–375.

Saidi A., Hajibarat Z. 2019. *Characterization of cis-elements in hormonal stress-responsive genes in Oryza sativa*. Asia-Pacific Journal of Molecular Biology and Biotechnology, 27: 95–102.

Saidi A., Hajibarat Z., Hajibarat Z. 2020a. *Transcriptome analysis of Phytophthora infestans and Colletotrichum coccodes in tomato to reveal resistance mechanisms*. Asia-Pacific Journal of Molecular Biology and Biotechnology, 28: 39–51

Saidi A., Hajibarat Z., Hajibarat Z. 2020b. *Identification of responsive genes and analysis of genes with bacterial-inducible cis-regulatory elements in the promoter regions in Oryza sativa* L. Acta Agriculturae Slovenica, 116: 115–123.

Small I .D., Peeters N. 2000. *The PPR motif – A TPR-related motif prevalent in plant organellar proteins*. Trends in Biochemical Sciences, 25: 46–47.

Sinclair S.A., Senger T., Talke I.N., Cobbett C.S., Haydon M.J., Kraemer U. 2018. *Systemic upregulation of MTP2-and HMA2-mediated Zn partitioning to the shoot supplements local Zn deficiency responses*. The Plant Cell, 30: 2463–79.

Sun F., Xiu Z., Jiang R., Liu Y., Zhang X., Yang Y.Z, Li X., Zhang X., Wang Y., Tan B.C. 2019. *The mitochondrial pentatricopeptide repeat protein EMP12 is involved in the splicing of three nad2 introns and seed development in maize*. Journal of Experimental Botany, 70: 963–72.

Wang F.P., Wang X.F., Zhang J., Ma F., Hao Y.J. 2018. *MdMYB58 modulates Fe homeostasis by directly binding to the MdMATE43 promoter in plants*. Plant and Cell Physiology, 59: 2476–89.

Wei K.A., Chen J.U., Wang Y., Chen Y., Chen S., Lin Y., Pan S., Zhong X., Xie D.2012. *Genome--wide analysis of bZIP-encoding genes in maize*. DNA Research, 19: 463–76.

Xing H., Fu X., Yang C., Tang X., Guo L., Li C., Xu C., Luo K.2018. *Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses*. Scientific Reports, 8: 1–9.

Xiong H., Li J., Liu P., Duan J., Zhao Y., Guo X., Li Y., Zhang H., Ali J., Li Z. 2014. *Overexpression of OsMYB48-1, a novel MYB-related transcription factor, enhances drought and salinity tolerance in rice*. PLoS One 9.

YANG B., SONG Z., LI C., JIANG J., ZHOU Y., WANG R., WANG Q., NI C., LIANG Q., CHEN H., FAN L.M. 2018. *RSM1, an Arabidopsis MYB protein, interacts with HY5/HYH to modulate seed germination and seedling development in response to abscisic acid and salinity*. PLoS genetics14: e1007839.

YU C.S., CHEN Y.C., LU C.H., HWANG J.K. 2006. *Prediction of protein subcellular localization. Proteins. Structure.* Function and Bioinformatics, 64: 643-51.

ZIMMERMANN P., LAULE O., SCHMITZ J., HRUZ T., BLEULER S., GRUISSEM W. 2008. *Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases*. Molecular Plant, 1: 851–7.

Supplement 1

Table Appx. 1

The sequences and the Pfam annotations of conserved motifs in PPR proteins

No. motifs	Specification	–
1	VGNALIDMYAKCGELEDARKVFDEMPEKBVVSWNAMISGYA	PPR
2	YGIEPKVEHYGCMVDLLGRAGRLEEAEELIEEMPFKPDAV	PPR_1
3	IRVVKNLRVCGDCHAAIKLISKVTEREIIVRDANRFHHFKBGKCSCGDYW	DYW family of nucleic acid deaminases