大豆科学 Soybean Science
http://ddkx.haasep.cn
2021,40(2):177-185
DOI:10.11861/j.issn.1000-9841.2021.02.0177

大豆 PPR 基因家族生物信息学分析

彭诗巧,卢梦丹,杨梦琦,李佳雯,岳岩磊

(河南农业大学 生命科学学院,河南 郑州 450002)

摘　要:为明确大豆 PPR 基因家族在染色体上的分布、保守结构域、亚族种类、进化关系及表达特征等,采用生物信息学方法,基于 Pfam 的 PPR 种子序列模型筛选大豆基因组数据库,获得 631 个大豆 PPR 家族基因;利用 MEME、ExPASy、TBtools、FigTree 等工具对大豆 PPR 家族基因进行分类,分析各亚族的进化关系和保守结构域差异;筛选各亚族的代表基因,并进一步分析各亚族基因的保守率、等电点、UTR、CDS 及基因表达特异性。结果表明:大豆 PPR 基因家族分为 DYW、P、PLS、E/E + 亚族以及 1 个未知亚族,其中 DYW 亚族为第一大亚族,占总基因数目的 57.2%;各亚族基因在染色体上的分布是不均匀的,其内含子数目差异也较大,DYW 亚族基因的内含子数目较少,但 DYW 亚族结构域种类最多,具有在 C 端出现特有的 Motif7 和 Motif4 的显著特征;但大豆 PPR 家族的各亚族基因表达特异性比较相似,均表现为叶片中高表达,花、根和茎中低表达;而且各亚族中都有大量成员缺失 UTR;*Glyma.19G095500*、*Glyma.11G086900*、*Glyma.02G175900* 和 *Glyma.01G158100* 这 4 个基因具有独特的 Motif8 保守结构域,为新的亚族。

关键词:大豆;PPR 基因家族;生物信息学分析;进化;系统发育;组织特异性表达

Bioinformatic Analysis of PPR Gene Family in Soybean

PENG Shi-qiao, LU Meng-dan, YANG Meng-qi, LI Jia-wen, YUE Yan-lei

(College of Life Sciences, Henan Agricultural University, Zhengzhou 450002, China)

Abstract: In order to find out soybean PPR genes and figure out the chromosomal distribution, the conserved domains, the subgrouping, the phylogenetic relationship and the expression characteristics of these genes, this study found 631 PPR genes through screening the soybean reference genome with the PPR model in the Pfam database. Clustering analysis was further performed to divide them into different subfamily with the softwares of MEME, ExPASy, TBtools and FigTree. The phylogenetic relationship and conserved domains of these subfamilies were also analyzed. We selected the representative genes from these subfamilies, and further evaluated their isoelectric points, UTR, CDS, and gene expression patterns. The results showed that, soybean PPR gene family could be divided into five subfamilies, DYW-, P-, PLS-, E/E + and an unknown subfamily. Of them, DYW-subfamily was the biggest one, accounting for 57.2% of total soybean PPR genes. These subfamilies distributed unevenly in chromosomes, and their intron numbers varied greatly. DYW-subfamily contained fewer introns but contained more conserved motifs with two characteristical motifs of Motif7 and Motif4 locating in the C terminal. The representative genes of these subfamilies had a similar expression pattern with a high expression level in leaf but a low expression level in flower, root and stem. The UTR was missing in some members of these subfamilies. *Glyma.19G095500*, *Glyma.11G086900*, *Glyma.02G175900* and *Glyma.01G158100* were subgrouped into a new subgroup containing a unique motif of Motif8.

Keywords: Soybean; PPR gene family; Bioinformatic analysis; Evolution; Phylogenetic analysis; Tissue specificity expression

PPR(pentatricopeptide repeats)基因家族是一类含有 35 个氨基酸残基重复序列的蛋白,广泛存在于拟南芥、水稻等高等植物基因组中,在雄性不育、光合作用、胚胎发育等生物学反应中发挥重要调控作用。植物中最早发现的 PPR 基因是玉米的 *crp1* 基因[1],但对 PPR 基因家族的系统描述和研究是 Small 和 Peeters[2] 首先开始的,他们对拟南芥全基因组测序数据进行分析,在筛选拟南芥线粒体和叶绿体功能基因时发现了 PPR 基因家族。随后人们发现 PPR 基因家族几乎存在于所有的真核生物中,尤其是在高等植物中[3-6]。对比分析真核生物基因组,发现陆生植物包含的 PPR 基因最多,其次是海藻[2,7-8]。在已知全基因组测序的被子植物中,PPR 基因家族几乎都是最大的家族,占所有基因的 1% ~ 2%[6,9]。如模式植物拟南芥中有 441 个 PPR 基因[10];番茄中有 471 个 PPR 基因[5];卷柏基因组中有1 670多个 PPR 基因[11];通过对作物基因组进行分析预测,发现水稻中的 PPR 蛋白超过 491 个[9],谷子中的 PPR 基因也有 486 个[11]。

PPR 蛋白由 N 端信号序列、中间串联重复单元和 C 端结构域 3 个部分组成。N 端序列多变且多数具有线粒体或叶绿体定位序列,可以调节线粒体或

收稿日期:2020-10-27
基金项目:国家自然科学基金(32001573);河南农业大学青年英才项目(30500613)。
第一作者:彭诗巧(1997—),男,学士,专业为生物信息学。E-mail:15960753907@163.com。
通讯作者:岳岩磊(1984—),女,博士,副教授,主要从事大豆分子育种研究。E-mail:yueyanlei@henau.edu.cn。

者叶绿体基因表达水平。中间串联重复单元由 2 ~ 27 个串联重复的高度保守 PPR 结构域组成[2],根据该结构域特征 PPR 基因家族可分为两大亚族:P 亚族和 PLS 亚族[10]。C 端结构域的多样性也是 PPR 基因家族分类的主要依据[12]。PPR 基因家族在线粒体基因的翻译中起重要作用,如酵母的 *PET309* 和红面包菌的 *cya-5* 均作用于 COX1[13-14]。在植物中的研究表明番茄 70% PPR 蛋白具有 RNA 结合活性,拟南芥和玉米中也发现了同样的结果[6, 15-16]。除此,PPR 还参与胚胎发育、胁迫反应、生长等过程。如拟南芥的 *AtC401* 基因及矮牵牛的 *PnC401* 参与光周期调控植物生长过程[17-18]。玉米、水稻、矮牵牛及萝卜的部分育性基因 *Rf* 都属于 *PPR* 基因,这也表明了部分 *PPR* 与植物育性相关[7, 19-21]。除此,拟南芥 *Emb175* 调控植物发育过程[7, 22],著名的水稻 *BT* 也是 PPR 基因[21]。在大豆中,PPR 蛋白可能与种子储藏物质的积累有关,如邱红梅等[23]采用元分析方法,得到了 4 个调控含硫氨基酸含量的 PPR 蛋白。Song 等[24]发现 PPR 基因可与 miRNA 协同调节种子中干物质的积累,且该基因在花和子叶中高表达。PPR 蛋白还与大豆生育期相关,赵峰[25]发现 *GmZTL* 基因能影响开花,该基因含有1 段 PPR 重复区域,可影响下游基因的转录,进而使转基因植株表现出晚花。另外,大豆的 PPR 基因还与组织分化及器官发育有关,如大豆曲茎基因 *Glyma14g38760* 也属于 PPR 基因家族,该基因与拟南芥的 *LOJ* 基因功能相似,在顶端分生组织和边缘侧生分生组织分化中起重要的调控作用[26]。

大豆 PPR 家族基因在生物农艺性状定位、功能基因挖掘及差异基因表达等的研究中时而出现,但作为古四倍体的大豆,其 PPR 家族基因的数量、在染色体上的分布、结构域种类、基因结构等却尚待揭示,这无疑限制了进一步对 PPR 家族基因的系统研究。本研究基于 Pfam 的 PPR 种子序列模型筛选大豆基因组数据库,获得大豆 PPR 家族基因;并根据这些基因的蛋白和基因结构特征,利用 MEME、ExPASy、TBtools、FigTree 等工具,对大豆 PPR 家族基因进行分类,分析各亚族的进化关系和保守结构域差异;进一步通过筛选各亚族的代表基因,分析各亚族基因的保守率、等电点、UTR 和 CDS 及基因表达特异性。研究旨在为深入探讨大豆 PPR 基因家族成员的功能以及表达调控等相关研究奠定基础,为进一步克隆和鉴定大豆 PPR 基因提供依据。

1 材料与方法

1.1 材料

从植物基因组数据库(https://phytozome.jgi.doe.gov)中获取大豆全基因组蛋白序列,基于 Pfam 32.0 数据库(http://pfam.xfam.org/)的 PPR 种子序列的隐马尔可夫模型[27],在大豆全基因组蛋白序列中进行搜索与筛选,初步获得 PPR 家族候选序列;再以 NCBI 中的 Swiss-Prot 为目标数据库,对这些候选序列进行 BLAST 分析,去除基因序列长度显著低于其他序列的基因,同时去除亲缘关系较为远源的基因,最终获得高质量的大豆 PPR 基因家族相关序列。

1.2 方法

1.2.1 进化树分析 利用进化遗传分析软件 MEGA X 自带的 ClustalW 程序对上一步筛选获得的 PPR 家族相关蛋白序列进行多重比对分析;并构建最大似然树(ML),其中 Bootstrap 值设为1 000,其他值则为默认值[28];最后,利用 FigTree(http://tree.bio.ed.ac.uk/software/figtree/)软件对构建完成的 ML 树进行可视化修饰,并对分支区域进行可视化分区,其中,可视化分区主要根据 PPR 家族的 5 种亚族分类进行相应分区,最终获得代表 PPR 家族亚族分类与进化关系的最大似然树(maximum likehood tree)。

1.2.2 保守结构域分析 利用已经获得的高质量 PPR 家族基因序列,在 MEME(http://meme-suite.org)中进行保守结构域预估分析(保守结构域数量设置为 10 个),同时在 TBtools 软件中进行保守结构域可视化修饰[29],并依据可视化结果筛选出保守结构域明显清晰且具有各个亚族特色的基因,最终筛选出 PPR 家族各个亚族的代表基因序列。并对最终筛选的 PPR 基因家族代表序列再一次用 MEME(http://meme-suite.org)进行保守结构域预估,分析不同亚族的保守结构域分布差异、保守域特征和保守率情况。

1.2.3 代表序列等电点分析 利用 ExPASy 软件对大豆 PPR 基因家族各代表成员进行等电点与分子质量的测试[26],获得大豆 PPR 家族代表基因序列的等电点与分子质量状况。

1.2.4 代表序列结构分析 根据已知的大豆全基因组数据注释文件,将大豆 PPR 基因家族代表序列在软件 TBtools 中进行 UTR 和 CDS 可视化分析,并利用 TBtools 工具进行染色体定位预测,获得其染色体可视化分布图[29]。

1.2.5 代表序列基因表达特异性分析 从 Phytozome 数据库中获取大豆 PPR 基因家族代表基因在大豆不同部位中的表达含量值 FPKM(Fragments Per Kilobase of exon model per Million mapped fragments),并绘制成热图,分析大豆 PPR 基因家族代表序列的基因表达特异性。

2　结果与分析

2.1　大豆 PPR 基因家族进化分析

从 Phytozome 植物基因组数据库初步筛选获得 701 个 PPR 蛋白候选序列,去除冗余后最终获得 631 个高质量大豆 PPR 基因家族相关序列,进化树分析结果显示,大豆 PPR 基因家族可分为 5 个亚族:DYW 亚族、P 亚族、PLS 亚族、E/E + 亚族和 1 个未知亚族(图 1、表 1)。其中,DYW 亚族为第一大亚族,共有 361 个基因成员,占总基因数目的 57.2%;P 亚族为第二大亚族,共有 194 个基因成员(30.7%);PLS 亚族有 42 个基因成员(6.66%);E/E + 亚族为 30 个基因成员(4.75%);最后剩下的 4 个序列被划分为未知亚族,仅占总基因数目的 0.69%。同时,大豆 PPR 基因家族在基因组上的分布是不均匀的,PLS 亚族不存在于 4 号和 7 号染色体中,E/E + 亚族不存在于 7 号、10 号、11 号、14 号、16 号和 19 号染色体中。尤其值得注意的是,大豆 7 号染色体中完全缺乏大豆 PPR 家族基因(图 2)。

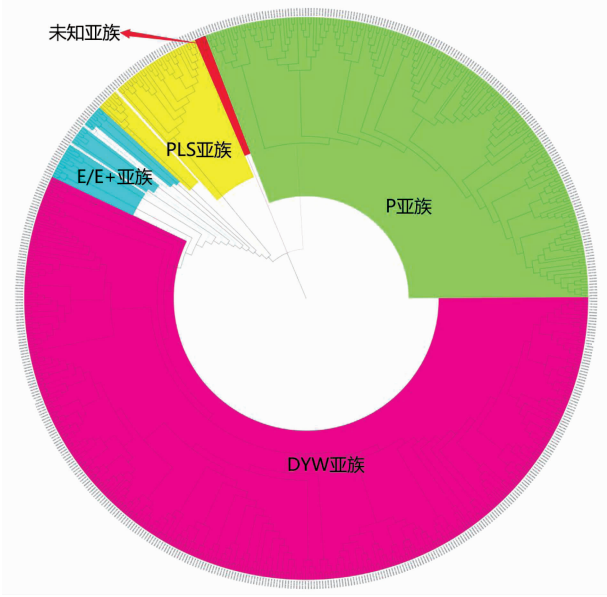

图 1　大豆 PPR 基因家族的最大似然树分析

Fig. 1　Phylogenetical analysis of maximum likelihood tree on soybean PPR gene family

表 1　大豆 PPR 基因家族分类信息

Table 1　Basic clustering of soybean PPR gene family

亚族 Subfamily	成员数量 No. of member	成员 Member
DYW	361	01G003400 01G004400 01G011300 01G011700 01G043500 01G048100 01G054900 01G056200 01G060400 01G099200 01G139700 01G147900 01G153700 01G159100 01G163800 01G173200 01G176700 01G180600 01G181400 01G201800 01G208900 01G228000 01G230500 01G234100 01G236000 01G237100 01G245300 02G006900 02G020100 02G043900 02G071600 02G077900 02G087100 02G094200 02G102000 02G113400 02G114500 02G118600 02G133100 02G136600 02G144100 02G160400 02G174500 02G179600 02G182700 02G201800 02G205900 02G217200 02G219000 02G223800 02G227300 02G250600 02G284700 02G285400 02G309700 03G000300 03G001300 03G022000 03G022100 03G026700 03G027800 03G081800 03G108300 03G108500 03G148000 03G161200 03G178400 03G184300 03G189000 03G205100 03G223800 03G227900 03G238000 03G239200 03G264700 04G007000 04G010200 04G039000 04G057300 04G062100 04G078800 04G127500 04G166500 04G186800 04G203700 04G243200 04G243300 04G255300 05G008800 05G026400 05G069000 05G076000 05G078900 05G079500 05G122700 05G125500 05G131900 05G132700 05G137900 05G157500 05G159700 05G182900 05G183900 05G228000 05G240300 05G244300 05G244400 06G006900 06G040000 06G057900 06G080400 06G080500 06G110000 06G119900 06G120100 06G122300 06G154600 06G161600 06G161900 06G179100 06G196000 06G206900 06G211900 06G222600 06G254600 06G289800 06G301600 06G311600 06G311700 06G322500 08G006700 08G034600 08G034900 08G077700 08G080300 08G086500 08G092900 08G096900 08G117400 08G123700 08G133500 08G140700 08G141600 08G160200 08G172500 08G208900 08G213800 08G241100 08G252400 08G254300 08G285300 08G291700 08G291800 08G294900 08G295900 08G302300 08G304400 08G349800 09G000400 09G006400 09G017100 09G043600 09G086600 09G088000 09G097800 09G126100 09G160300 09G168800 09G180500 09G200600 09G209700 09G227000 09G232200 09G235800 09G236700 09G237200 09G237800 09G251000 09G261900 09G272000 09G283000 09G286100 09G286300 10G008100 10G018100 10G074100 10G093800 10G094300 10G138900 10G145700 10G148000 10G190600 10G191100 10G230100 10G240000 10G247700 10G258200 10G260100 10G277600 10G284800 11G006200 11G007200 11G008800 11G013500 11G015000 11G033300 11G059600 11G061600 11G065300 11G069900 11G079700 11G081400 11G085600 11G104400 11G105600 11G121500 11G131400 11G136400 11G162700 11G190300 11G213300 11G254200 12G001100 12G004300 12G005600 12G009800 12G030500 12G046600 12G055700 12G060100 12G102800 12G115500 12G118200 12G146900

续表 1

亚族 Subfamily	成员数量 No. of member	成员 Member
		12G183600 12G184200 12G187900 12G188000 12G189100 12G239300 13G001700 13G008400 13G060800 13G062300 13G118700 13G119700 13G121800 13G135000 13G141400 13G150300 13G154700 13G169800 13G232000 13G239600 13G240000 13G260500 13G312600 13G313200 13G313300 13G332400 13G345000 13G346800 13G357000 14G003000 14G003900 14G017300 14G028500 14G066100 14G159200 14G184600 14G190500 14G194400 14G207100 14G224400 15G016600 15G059000 15G093800 15G103100 15G149500 15G193700 15G228600 15G254800 15G270800 15G272100 15G273200 16G022100 16G026000 16G034600 16G035600 16G049100 16G049800 16G115600 16G152800 16G161800 16G172800 16G204800 16G206000 16G209700 16G211900 16G218400 16G221300 17G024100 17G024900 17G057500 17G072100 17G099500 17G101900 17G117000 17G124200 17G124300 17G165800 17G175800 17G186700 17G209700 17G220100 17G262700 18G056000 18G085900 18G094700 18G126700 18G128100 18G134700 18G139800 18G151600 18G230200 18G241500 18G252800 18G259800 18G260300 18G261200 18G262200 18G263500 18G275000 18G277000 18G287500 18G288100 19G023400 19G025700 19G026400 19G090900 19G102300 19G141800 19G179200 19G202500 19G220900 20G002600 20G013300 20G044000 20G068100 20G094700 20G095100 20G095400 20G104500 20G112100 20G154600 20G155800 20G163300 20G199200 20G200100 U028000 U036700
P	194	01G016100 01G179200 01G228800 01G230600 01G233900 02G000600 02G009700 02G011200 02G086700 02G192000 02G192800 02G271600 02G276200 03G127100 03G136200 03G148400 03G156300 03G177000 03G190300 03G196000 03G261200 04G018000 04G054900 04G060500 04G097400 04G155800 04G159900 04G235800 05G024900 05G032300 05G060900 05G093700 05G105900 05G113300 05G115200 05G142300 05G173900 06G018300 06G099100 06G128900 06G144800 06G196100 06G198200 06G236700 06G246500 08G053900 08G072900 08G090700 08G098100 08G106500 08G131100 08G140300 08G174800 08G176700 08G186600 08G254000 08G283500 09G013300 09G013400 09G013500 09G013600 09G050500 09G058600 09G082100 09G089200 09G156000 09G175400 09G200200 09G218800 09G242500 09G256600 09G263600 09G279300 09G282100 10G000600 10G001500 10G011800 10G122100 10G162500 10G166300 10G166400 10G193100 10G265500 10G285200 11G001100 11G009000 11G011600 11G013600 11G013700 11G063000 11G078800 11G098900 11G103200 11G110600 11G217500 11G253700 11G256300 12G033700 12G048300 12G118300 12G135500 12G164700 12G191600 12G203800 13G034900 13G181600 13G197400 13G197700 13G221600 13G273600 13G297800 13G354300 13G364500 13G367800 13G370800 14G014900 14G039600 14G044600 14G108600 14G156100 14G184300 14G223300 15G017400 15G038300 15G073300 15G091700 15G118200 15G118300 15G131800 15G162500 15G164700 15G197000 16G001000 16G043600 16G052400 16G053700 16G066300 16G160700 16G197600 16G206600 17G006500 17G016400 17G049100 17G094300 17G099400 17G102300 17G149100 17G153900 17G160600 17G190200 17G197900 17G203100 17G205100 18G000700 18G039300 18G093100 18G108200 18G117600 18G143000 18G180200 18G197200 18G198500 18G207800 18G209700 18G228500 18G229500 18G276500 19G000100 19G010600 19G080500 19G088500 19G099200 19G108600 19G129900 19G138100 19G151700 19G177700 19G190700 20G009700 20G010100 20G014500 20G074200 20G092100 20G097200 20G103800 20G104200 20G109800 20G115000 20G125900 20G131400 20G158600 20G197300 20G222800 20G225100
PLS	42	01G057600 01G058200 01G073500 01G155000 01G176900 02G115700 03G093000 05G004200 05G072500 05G118000 05G182700 06G226800 08G061500 08G238900 08G254100 09G129400 09G157400 09G172300 09G268500 10G083900 11G007400 11G065700 11G111200 12G047100 12G201200 12G214200 13G206100 13G261100 13G287400 13G356700 14G073700 14G141400 15G092000 15G106500 15G236000 15G245700 16G177300 17G251300 18G221100 19G055700 20G010600 20G105600
E/E +	30	01G108100 01G132800 02G116700 03G035400 03G108100 04G175400 04G211700 05G030900 06G063100 06G189500 06G268800 06G284200 08G152900 08G255500 08G345700 09G244300 12G071400 13G227000 13G341300 15G264900 15G265000 17G146000 18G047500 18G249200 18G276600 18G278400 20G006300 20G023100 20G089500 20G220700
未知 Unknown	4	01G158100 02G175900 11G086900 19G095500

01G003400 是 *Glyma. 01G003400* 的缩写,其他同理。

01G003400 is briefed from *Glyma. 01G003400*. All others in the same way.

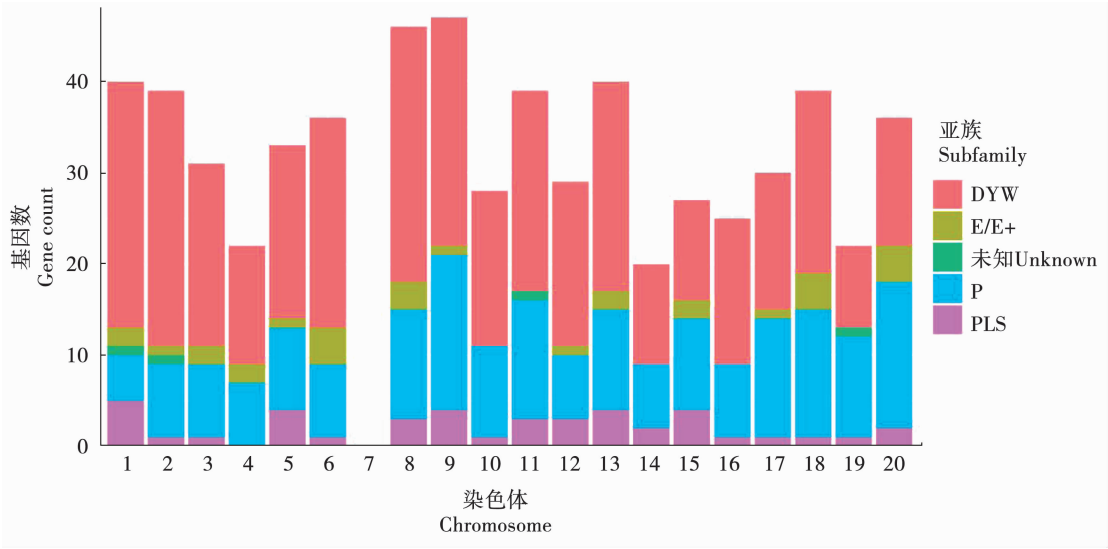

图 2　大豆 PPR 家族基因在染色体上的分布情况

Fig. 2　Genomic distribution of soybean PPR gene family members

2.2　大豆 PPR 基因家族代表序列保守性分析

根据进化树分析,结合各类群特点选取高质量具有代表性的基因 28 个,除未知亚族中包含 4 个基因外,其他亚族均包含 6 个基因(表 2)。对这些基因的保守区域、基因结构和表达特点的分析结果如图 3 所示,DYW 亚族的结构域种类最多,其显著特征是在 C 端出现了特有的 Motif 7 和 Motif 4 结构域,而其他亚族则没有。拟南芥中也报道 DYW 亚族是在 E/E + 族基础上多 1 个特有的 DYW 结构域[12,30]。除此,DYW 亚族还有 1 个显著的特征,Motif 5-Motif 3-Motif 6-Motif 2 结构域串联重复出现,以及 C 端具有特殊的 Motif 4 结构域。与 DYW 亚族相比,E/E + 亚族的保守结构域种类数量次之,但分布样式最为多样,即有以类似于 DYW 族保守结构域特点的 Motif 5-Motif 3-Motif 6 结构域串联方式分布出现的,也有类似于 PLS 亚族 Motif 1-Motif 2 串联结构域方式出现的,但其串联重复次数出现的较少。P 亚族较其他亚族来说,保守结构域种类比较少,多数为 Motif 1 和 Motif 2。从结构来看各结构域排列紧密,并且拥有特有的保守结构域分布模式——以 Motif 1-Motif 2 结构域形式串联紧密重复出现。PLS 亚族保守结构域分布与 P 亚族极其相似,大部分以 Motif 1-Motif 2 形式串联重复出现,但是其在 Motif 1-Motif 2 串联重复出现的时候也会出现 Motif 6 与 Motif 5 串联,整体来看 PLS 与 P 亚族不同,其保守结构域长度比 P 亚族和 DYW 亚族都较短,且串联重复次数较少,结构域的保守程度比 P 亚族和 DYW 族都弱。未知亚族仅有 4 个基因,*Glyma. 19G095500*、*Glyma. 11G086900*、*Glyma. 02G175900*、*Glyma. 01G158100*,其显著的特点是大部分都有 Motif 8 保守结构域分布,而其他亚族则没有发现该结构域。

表 2　大豆 PPR 基因家族代表序列基本信息

Table 2　Basic information of representative members of soybean PPR gene family

亚族 Sub family	基因编号 Gene ID	基因组位置 Genome location	蛋白长度 Length of protein/aa	外显子数目 Exon number	等电点 pI
DYW	*Glyma. 08G123700*	Chr08:9506498. . 9508788	721	1	8. 32
	Glyma. 08G294900	Chr08:41002768. . 41005138	611	1	6. 51
	Glyma. 08G295900	Chr08:41117707. . 41121522	631	2	8. 21
	Glyma. 09G200600	Chr09:42479152. . 42483030	676	1	7. 60
	Glyma. 09G209700	Chr09:43396577. . 43398265	562	1	6. 30
	Glyma. 09G283000	Chr09:49875720. . 49879432	618	1	6. 97
P	*Glyma. 03G136200*	Chr03:35205755. . 35212954	837	10	8. 85
	Glyma. 03G261200	Chr03:45476750. . 45478150	466	1	9. 18
	Glyma. 04G018000	Chr04:1403026. . 1407268	682	3	6. 50

续表 2

亚族 Sub family	基因编号 Gene ID	基因组位置 Genome location	蛋白长度 Length of protein/aa	外显子数目 Exon number	等电点 pI
	Glyma. 17G094300	Chr17:7374357. . 7383061	859	4	5. 56
	Glyma. 17G153900	Chr17:12919733. . 12922117	655	2	8. 39
	Glyma. 20G109800	Chr20:35213258. . 35219502	625	8	7. 03
PLS	*Glyma. 01G155000*	Chr01:49253634. . 49255352	423	1	8. 54
	Glyma. 01G176900	Chr01:51347518. . 51350161	457	3	8. 09
	Glyma. 05G072500	Chr05:8069457. . 8071713	728	2	8. 70
	Glyma. 05G182700	Chr05:37046146. . 37049433	503	2	9. 34
	Glyma. 11G065700	Chr11:4952513. . 4954342	395	1	6. 73
	Glyma. 11G111200	Chr11:8486920. . 8490026	703	2	8. 84
E/E +	*Glyma. 06G284200*	Chr06:47256370. . 47258568	732	1	6. 18
	Glyma. 09G244300	Chr09:46682979. . 46685027	682	1	9. 50
	Glyma. 18G047500	Chr18:4139387. . 4142265	234	2	9. 07
	Glyma. 18G249200	Chr18:53571953. . 53573976	492	3	9. 21
	Glyma. 18G276600	Chr18:55866582. . 55869929	749	2	4. 69
	Glyma. 20G023100	Chr20:2445807. . 2450475	817	4	7. 48
未知	*Glyma. 01G158100*	Chr01:49616252. . 49618590	507	2	7. 57
Unknown	*Glyma. 02G175900*	Chr02:28628623. . 28630688	334	5	5. 02
	Glyma. 11G086900	Chr11:6524266. . 6525385	322	3	5. 25
	Glyma. 19G095500	Chr19:33790733. . 33793088	540	6	6. 26

图 3 大豆 PPR 基因家族代表序列保守域分析

Fig. 3 Conserve motifs of representative members of soybean PPR gene family

2.3　大豆 PPR 基因家族代表序列基因结构分析

基因结构、UTR 和内含子的数量对基因表达调控具有重要意义,对大豆 PPR 家族基因结构的分析结果如图 4 所示,大豆 PPR 家族各个亚族内外显子和内含子数量分布不均匀,其中在 DYW 亚族中外显子数目最少,除 *Glyma. 08G295900* 之外,其他基因均含有 1 个外显子;P 亚族基因的内含子数量最多且序列最长,基因结构整体表现较为松散;未知亚族基因的内含子数目与 P 亚族相近,但内含子序列较短,基因结构较为紧凑;E/E + 亚族和 PLS 亚族基因的结构较为相似,内含子数目为 0 ~ 3 个(图 4)。值得注意的是,这些 PPR 家族基因都或多或少地缺乏 UTR 结构,除 P 亚族和未知亚族外,各亚族中缺少 UTR 结构的序列的含子数目较其它序列有减少趋势,该结果与前人报道的 PPR 家族基因来源于反转录转座的发生相符,早期 PPR 家族基因通过转录加工成为 RNA 后,反转录却生成了无内含子的双链 DNA,在基因组的复制中,形成了无内含子拷贝,这样的重复复制导致了 PPR 家族基因丢失大量内含子[10,31]。

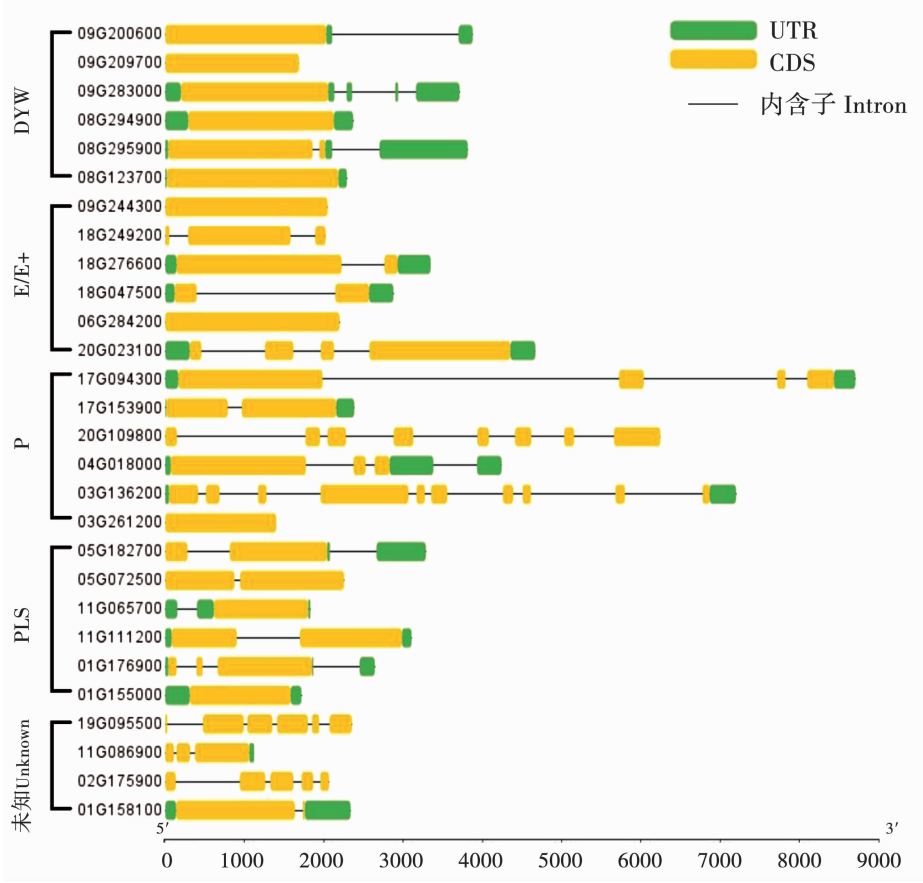

图 4　大豆 PPR 基因家族代表序列基因结构分析

Fig. 4　Gene structure of representative members of soybean PPR gene family

2.4　大豆 PPR 基因家族代表序列组织表达特异性分析

基因特异性表达分析是研究基因产物行使生物学功能方式的重要手段,大豆 PPR 基因家族代表序列的表达特性分析如图 5 所示,大豆 PPR 家族的各个亚族基因在花(包括 Flower open、Flower unopened 和 Flower)、根、根毛、茎及叶片中均有表达。其中,DYW、E/E + 和 P 亚族极为相似,都表现出叶片中高表达,根和花中低表达的趋势。而 PLS 亚族和未知亚族的基因表达情况表现出了个别基因的特异性,如 *Glyma. 19G095500* 和 *Glyma. 01G158100* 在花中高表达。较为一致的是 PPR 家族基因在根和茎中都表现为低表达。总体来看各个亚族在顶端分生组织以及种子中表达水平较高,显现出根中低表达、叶片中高表达、顶端分生组织次之的趋势,表明大豆 PPR 家族基因主要在叶片中发挥作用。

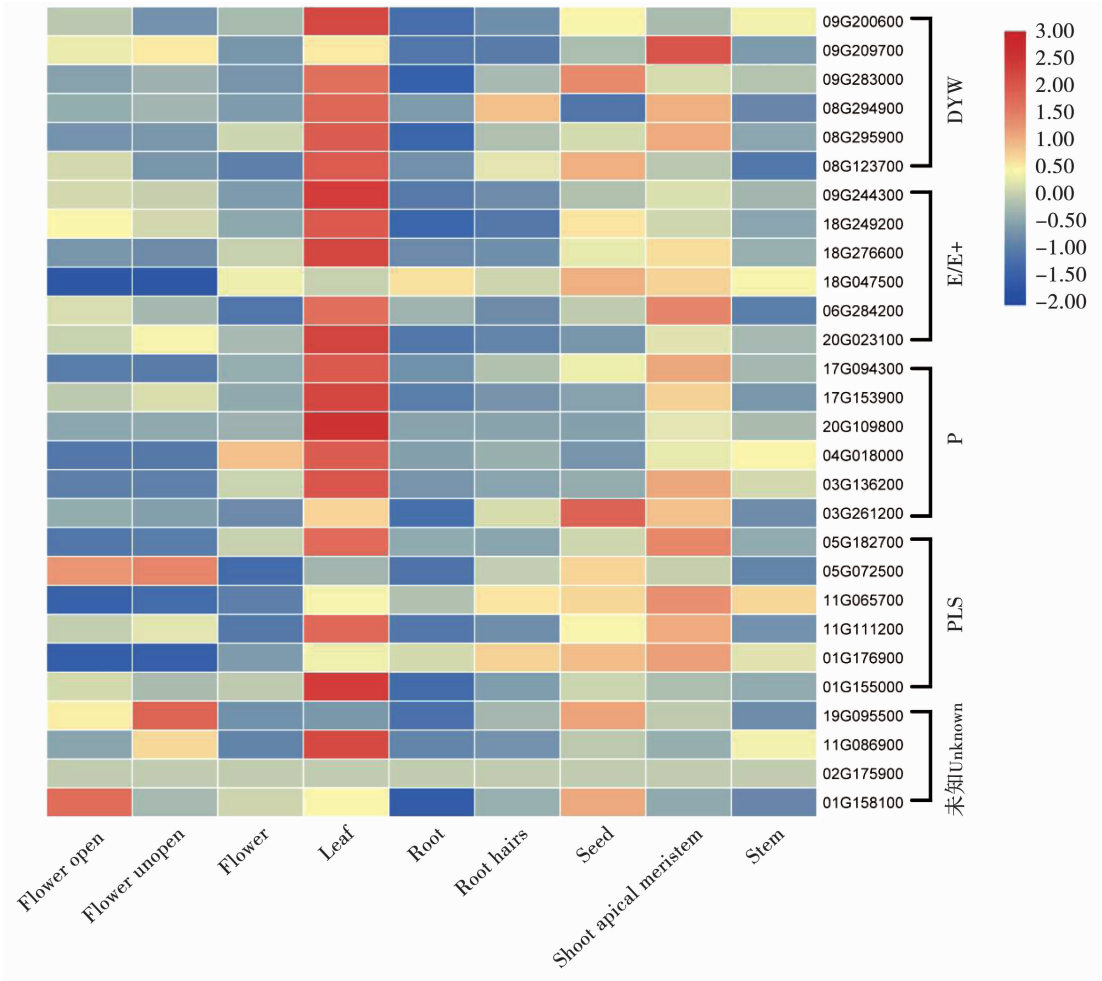

图 5 大豆 PPR 基因家族代表序列基因表达热图

Fig. 5 Expression heatmap of representative members of soybean PPR gene family

3 讨 论

随着各种植物基因组计划的完成,从基因组层面进行基因家族分类的研究、预估其保守结构域分布特点、进化特征、结构域功能特征等已成为生物学研究的重要手段。PPR 家族几乎是植物中最大的基因家族,占所有基因的 1% ~2%,作物中水稻和谷子等的 PPR 基因家族先后被揭示,其成员的功能与光合作用、呼吸作用及胚胎发育等重要的生命活动相关[9,11]。而大豆 PPR 家族基因功能虽有少量报道[23-26],但对 PPR 基因家族的整体分布情况及特征并不清楚。本研究从 701 个大豆 PPR 家族基因中去除冗余后,筛选出 631 个高质量的 PPR 蛋白,进行亚族分类建树,得到 5 个 PPR 亚族,进一步对 28 个亚族代表基因进行分析,得出保守结构域种类和分布,基因结构特征和组织表达特异性。本研究虽填补大豆 PPR 基因家族系统分析的空白,但用于结构域和基因结构分析的代表基因的数量还不足以阐释所有 PPR 家族基因的特征,该结果也是基于生物信息分析的初步预测结果,具体 PPR 家族基因的克隆和功能注释还有待完善。

拟南芥的 PPR 基因家族分为 P、PLS、E、E + 和 DYW 5 个亚族[8],与其类似,大豆 PPR 基因家族也分为 5 个亚族,除了 P、PLS、E/E + 和 DYW 4 个亚族外,还发现了 1 个未知的亚族。在 5 个亚族中 DYW 亚族是大豆 PPR 家族中占比最大的,P 亚族占比次之,PLS 与 E/E + 亚族占比相当。而对保守域预测的结果则显示,DYW 保守结构域最多样也最为保守,P 亚族次之,PLS 和 E/E + 亚族相当,这个特征与其他植物略有差异[3]。大豆 PPR 家族基因的 UTR 区域和内含子与拟南芥和番茄等植物相同[5],都存在大量缺失的情况,这与 PPR 庞大家族通过反转录转座而形成有关。

大豆是古四倍体作物,基因组结构复杂,本研究获得的 631 个大豆 PPR 基因成员,其数量远远多于已报道的拟南芥、水稻和谷子等作物。表明 PPR 基因在大豆生长发育过程中有更为广泛的作用,这与已报道的大豆 PPR 基因研究相符,如大豆 PPR 基因有与种子储藏物质积累功能相关的基因,也有与大豆生育期相关的基因,还有与组织分化及器官发育有关的基因[23-26]。

本研究中的 PPR 代表基因在大豆叶片中高表达,花、根、茎中低表达。该研究结果与前人报道的调控种子干物质积累的大豆 PPR 基因在花、子叶和发育中的种子中高表达有所不同[24]。PPR 基因数量之庞大,仅靠生物信息学预测的方法无法详尽地阐述 PPR 基因的功能,具体基因的功能还有待分子生物学的进一步揭示。

4 结　论

大豆 PPR 基因家族分为 DYW、P、PLS、E/E + 和 1 个未知亚族共 5 个亚族,其中 DYW 亚族为第一大亚族,占总基因数目的 57.2%;各亚族基因在染色体上的分布是不均匀的,其内含子数目也差异较大,DYW 亚族基因的内含子数目最少,但 DYW 亚族结构域种类最多,显著特征是在 C 端出现特有的 Motif 7 和 Motif 4;但大豆 PPR 家族的各亚族基因表达特异性比较相似,均表现为叶片中高表达,花、根和茎中低表达;大豆 PPR 基因家族各亚族都普遍缺乏 UTR 和内含子。除此,*Glyma. 19G095500*、*Glyma. 11G086900*、*Glyma. 02G175900* 和 *Glyma. 01G158100* 这 4 个基因具有独特的 Motif 8 保守结构域,为新的亚族。该研究为深入探讨 PPR 基因家族成员的功能以及表达调控机理等奠定了基础。

参考文献

[1] Barkan A, Walker M, Nolasco M, et al. A nuclear mutation in maize blocks the processing and translation of several chloroplast mRNAs and provides evidence for the differential translation of alternative mRNA forms[J]. EMBO Journal, 1994, 13(13): 3170-3181.

[2] Small I D, Peeters N. The PPR motif-A TPR-related motif prevalent in plant organellar proteins[J]. Trends in Biochemical Science, 2000, 25(2): 45-47.

[3] Hayes M L, Santibanez P I. A plant pentatricopeptide repeat protein with a DYW-deaminase domain is sufficient for catalyzing C-to-U RNA editing *in vitro*[J]. Journal of Biological Chemistry, 2020, 295(11): 3497-3505.

[4] Oldenkott B, Yang Y, Lesch E, et al. Plant-type pentatricopeptide repeat proteins with a DYW domain drive C-to-U RNA editing in *Escherichia coli*[J]. Communications Biology, 2019, 2(1): 85.

[5] 丁安明, 李凌, 屈旭, 等. 番茄 PPR 基因家族的鉴定与生物信息学分析[J]. 遗传, 2014, 36(1): 77-84. (Ding A M, Li L, Qu X, et al. Identification and bioinformatics analysis of PPR gene family in Tomato[J]. Hereditas, 2014, 36(1): 77-84.)

[6] O'Toole N, Hattori M, Andres C, et al. On the expansion of the pentatricopeptide repeat gene family in plants[J]. Molecular Biology and Evolution, 2008, 25(6): 1120-1128.

[7] 陆萍, 俞嘉宁. PPR 蛋白影响植物生长发育的研究进展[J]. 植物生理学报, 2013, 49(10): 989-999. (Lu P, Yu J N. Research Progress on the effect of PPR protein on plant growth and development[J]. Plant Physiology Journal, 2003, 49(10): 989-999.)

[8] 丁安明, 屈旭, 孙玉合. 植物 PPR 蛋白家族研究进展[J]. 中国农学通报, 2014, 30(9): 218-224. (Ding A M, Qu X, Sun H Y. Research progress of plant PPR protein[J]. Chinese Agricultural Science Bulletin, 2014, 30(9): 218-224.)

[9] Chen G L, Zou Y, Hu J H, et al. Genome-wide analysis of the rice PPR gene family and their expression profiles under different stress treatments[J]. BMC Genomics, 2018, 19(1): 720.

[10] Lurin C, Andrés C, Aubourg S, et al. Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis[J]. Plant Cell, 2004, 16(8): 2089-2103.

[11] Liu J M, Xu Z S, Lu P P, et al. Genome-wide investigation and expression analyses of the pentatricopeptide repeat protein gene family in foxtail millet[J]. BMC Genomics, 2016, 17(1): 840.

[12] Takenaka M, Verbitskiy D, Zehrmann A, et al. Reverse genetic screening identifies five E-class PPR proteins involved in RNA editing in mitochondria of *Arabidopsis thaliana*[J]. Journal of Biological Chemistry, 2010, 285(35): 27122-27129.

[13] Zamudio-Ochoa A, Camacho-Villasana Y, García-Guerrero A E, et al. The Pet309 pentatricopeptide repeat motifs mediate efficient binding to the mitochondrial COX1 transcript in yeast[J]. RNA Biology, 2014, 11(7): 953-967.

[14] Coffin J W, Dhillon R, Ritzel R G, et al. The *Neurospora crassa cya-5* nuclear gene encodes a protein with a region of homology to the *Saccharomyces cerevisiae* PET309 protein and is required in a post-transcriptional step for the expression of the mitochondrially encoded COXI protein[J]. Current Genetics, 1997, 32(4): 273-280.

[15] Tasaka M, Shikanai T, Kotera E. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts[J]. Nature, 2005, 433(7023): 326-330.

[16] Das A K. The structure of the tetratricopeptide repeats of protein phosphatase 5: Implications for TPR-mediated protein-protein interactions[J]. EMBO Journal, 1998, 17(5): 1192-1199.

[17] Oguchi T, Sage-Ono K, Kamada H, et al. Genomic structure of a novel *Arabidopsis* clock-controlled gene, *AtC401*, which encodes a pentatricopeptide repeat protein[J]. Gene, 2004, 330: 29-37.

[18] Oguchi T, Sage-Ono K, Kamada H, et al. Characterization of transcriptional oscillation of an *Arabidopsis* homolog of PnC401 related to photoperiodic induction of flowering in *Pharbitis nil*[J]. Plant and Cell Physiology, 2004, 45(2): 232-235.

[19] Alfonso A A, Bentolila S, Hanson M R. Evaluation of the fertility restoring ability of Rf-PPR592 in petunia[J]. Philippine Agricultural Scientist, 2003, 86(3): 303-315.

[20] Mora J R H, Rivals E, Mireau H, et al. Sequence analysis of two alleles reveals that intra-and intergenic recombination played a role in the evolution of the radish fertility restorer (Rfo)[J]. BMC Plant Biology, 2010, 10(1): 35-35.

[21] Kazama T, Nakamura T, Watanabe M, et al. Suppression mechanism of mitochondrial ORF79 accumulation by Rf1 protein in BT-type cytoplasmic male sterile rice[J]. Plant Journal, 2008, 55(4): 619-628.

[22] Cushing D A, Forsthoefel N R, Gestaut D R, et al. *Arabidopsis emb175* and other *ppr* knockout mutants reveal essential roles for pentatricopeptide repeat (PPR) proteins in plant embryogenesis[J]. Planta, 2005, 221(3): 424-436.

[23] 邱红梅, 历志, 于妍, 等. 基于元分析的大豆含硫氨基酸相关基因挖掘与信息学分析[J]. 中国油料作物学报, 2015, 37(2): 141-147. (Qiu H M, Li Z, Yu Y, et al. Mining and analysis of genes related to sulfur-containing amino acids in soybean based on Meta-QTL[J]. Chinese Journal of oil crops, 2015, 37(2): 141-147.)

[24] Song Q X, Liu Y F, Hu X Y, et al. Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing[J]. BMC Plant Biology, 2011, 11: 5.

[25] 赵峰. 大豆 *GmZTL* 基因的功能研究[D]. 贵州: 贵州大学, 2008: 42. (Zhao F. Research on the function of soybean gene *GmZTL*[D]. Guizhou: Guizhou University, 2008: 42.)

[26] 李丛丛. 大豆曲茎和扁茎性状基因的精细定位与候选基因分析[D]. 南京: 南京农业大学, 2011: 38. (Li C C. Fine mapping and candidate gene exploration of loci corresponding to brachytic and fasciation stem in soybean[D]. Nanjing: Nanjing Agricultural University, 2011: 38.)

[27] Finn R D, Mistry J, Schuster-Bockler B, et al. Pfam: Clans, web tools and services[J]. Nucleic Acids Research, 2006, 34: 247-251.

[28] Tamura K, Peterson D, Peterson N, et al. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods[J]. Molecular and Biology Evolution, 2011, 28(10): 2731-2739.

[29] Chen C, Chen H, Zhang Y, et al. TBtools: An integrative toolkit developed for interactive analyses of big biological data[J]. Molecular Plant, 2020, 13(8): 1194-1202.

[30] 宋学花. 植物 PPR 基因家族的鉴定与进化分析[D]. 广州: 华南理工大学, 2017. (Song X H. Identification and evolution of pentatricopeptide repeat gene family in plants[D]. Guangzhou: South China University of Technology, 2017.)

[31] Lecharny A, Boudet N, Gy I, et al. Introns in, introns out in plant gene families: A genomic approach of the dynamics of gene structure[J]. Journal of Structural and Functional Genomics, 2003, 3(1-4): 111-116.