Pentatricopeptide Repeat Proteins in Cotton

Zongfu Han[1], Mingzhou Song[2], and Jinfa Zhang[1]

(1) Plant and Environmental Sciences, New Mexico State University, Las Cruces, NM, (2) Computer Science, New Mexico State University, Las Cruces, NM

Abstract

Pentatricopeptide repeat (PPR) proteins are encoded by genes from one of the largest families in higher plants. PPR proteins are translocated to mitochondria or chloroplasts and play a broad role including RNA processing, fertility restoration in cytoplasmic male sterility, embryogenesis, and plant development. In this study, four recently sequenced cotton (Gossypium) genomes were analyzed to identify PPR protein-coding genes. A phylogenetic tree for each species is constructed and compared among the four species. Homologous and homeologous PPR genes are further identified for identification of sequence variations and evolutionary analysis. Candidate PPR genes for fertility restoration of cytoplasmic male sterility in cotton will be identified and analyzed.

Objectives

To identify and map the PPR proteins two cultivated tetraploid cotton species (*G. hirsutum* and *G. barbadense*) and their two ancestral diploid species (*G. arboreum* and *G. raimondii*).

To characterize the candidate PPR encoding *Rf* genes in cotton.

Materials & Methods

Identification of PPR proteins: **The PPR seed protein sequence alignment, named PF01535 (http://pfam.xfam.org), was used as query by searching against the predicted protein sequences in cotton genome sequence database (Table 1) using the Hmmer3.1 program with default parameters. Then, the predicted proteins were further queried with P domain HMM model (Lurin et al., 2004) with e-value <-10. To identify P or PLS subfamily members, all candidate PPR genes were queried with L1, L2, S, E, E+ and DYW domains using HMM models with e-value <-10.**

Subcellular Predication: **TargetP 1.1 (http://www.cbs.dtu.dk) with default parameters was used.**

Phylogenetic analysis and sequence alignment: **All the PPR sequences in this study were aligned using the ClustalX version 2.1. FastTree was used to estimate the maximum-likelihood phylogeny. Trees were visualized through the Figtree version 1.4.2.**

Chromosomal mapping: **The chromosome location information of PPR genes was searched from the cotton genome database. MapChart 2.30 software was performed to generate the chromosomal distribution image of all candidate PPR genes in *G. arboreum*, *G. hirsutum*, *G. barbadense* and *G. raimondii*. Then the homologous and homeologous genes were linked with straight lines manually.**

Prediction of the candidate PPR encoding *Rf* genes: **Using the linkage maps of CMS fertility restorer genes (Wang et al., 2009; Wu et al., 2014), candidate PPR genes were located in a target region carrying markers associated with *Rf* genes.**

Table 1 The genome sources in this study

Species	Genome	Cultivar	Database_name	Source	Date	Publication
G. arboreum	A2	Shixiya 1	*Gossypium arboreum* (A2) Genome BGI Assembly v2.0 & Annotation v1.0	CottonGen	2014-05-31	Li et. al.,. Nature Genetics. 46, 567–572. 2014
G. raimondii	D5	CMD 10	*Gossypium raimondii* (D5) genome JGI assembly v2.0 (annot v2.1)	CottonGen	2013-02-18	Wang et. al., Nature Genetics. 44, 1098–1103. 2012
G. hirsutum	AD1	TM-1	*Gossypium hirsutum* (AD1) Genome NAU-NBI Assembly v1.1 & Annotation v1.1	CottonGen	2015-04-20	Zhang et. al., Nature Biotechnology. 33, 531–537. 2015
G. barbadense	AD2	Xinhai-21	*Gossypium_barbadense* cv. Xinhai-21 genome	CHGC	2016-04-01	Liu et al., Scientific reports, 5. 2015

Results

Table 2 The total number of PPR genes in four Gossypium species

	Total	P subfamily	PLS subfamily	PLS class E	class E+	class DYW	class
G. arboreum	513	253	260	31	142	5	82
G. raimondii	558	297	261	20	134	6	101
G. hirsutum	990	490	500	45	254	7	194
G. barbadense	1054	556	498	45	261	10	182

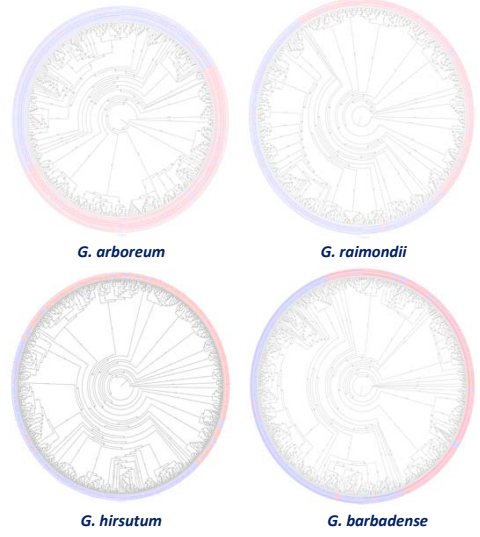

Fig 1. Phylogenetic trees of PPR genes in four Gossypium species. Blue color denotes P subfamily members; Red color denotes PLS subfamily members.

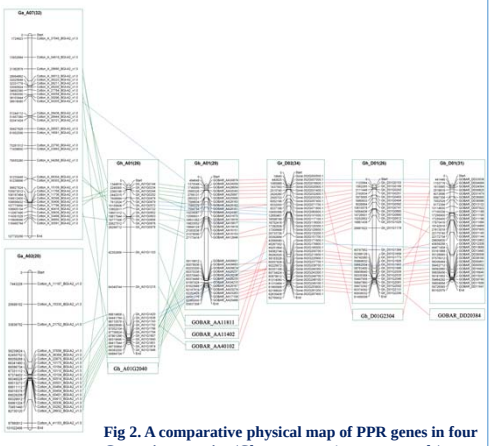

Fig 2. A comparative physical map of PPR genes in four Gossypium species (Chromosome 1 as an example).

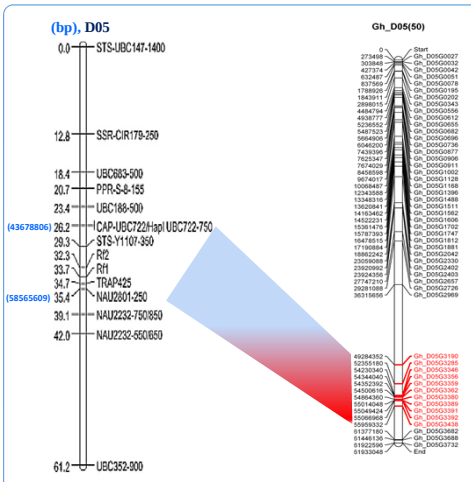

Fig 3. PPR proteins in the target region derived from a linkage map of *Rf* genes (Wang et al., 2009; Wu et al., 2014).

Conclusions

- **A total of 513, 558, 990 and 1054 PPR genes were identified in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively.**
  - The allotetraploid species also have similar numbers of PPR genes, which are almost sum of these from the two diploid progenitors.
  - The P subfamily contained roughly half of the identified PPR genes in the four species.
  - PPR genes are not distributed randomly on chromosomes.
  - 295 (57.5%), 345 (61.8%), 600 (60.6%), 544 (51.6%) PPR proteins were predicted to be targeted to mitochondria or chloroplast in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively.
- **Phylogenetic trees indicated clustering of the PPR subfamily members.**
  - A total of 14 (5.5%), 15 (5.1%), 47 (9.6%) and 46 (8.3%) P subfamily members in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively, were clustered to the PLS subfamily.
  - A total of 12 (4.6%), 9 (3.4%), 21 (4.2%) and 18 (3.6%) PLS subfamily members *in G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively, were clustered to the P subfamily.
  - Genetic variations might have happened in the C-terminal motifs during natural or artificial selection in cotton.
- **The physical maps showed that most of the PPR genes in *G. raimondii*, *G. hirsutum* and *G. barbadense* had good collinear relationships. However, PPR genes on a single chromosome in *G. hirsutum* usually correspond to homologous PPR genes on several chromosomes in *G. arboreum*.**
- **Eleven PPR proteins were located in the target region of *Rf* genes. A further analysis indicated that four of them were targeted to chloroplasts.**

References

Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Ret ML, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, and Small I(2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. The Plant Cell 16: 2089-2103.

Sykes T, Yates S, Nagy I, Asp T,, Small I, and Studer B(2016) In-silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.). Genome Biology and Evolution, 2016: evw047.

Wang F, Yue B, Hu J, Stewart JM and Zhang JF (2009) A target region amplified polymorphism marker for fertility restorer gene and chromosomal localization of and in cotton. Crop Science 49: 1602-1608.

Wu J, Cao X, Guo L, Qi T, Wang H, Tang H, Zhang J, and Xing C (2014) Development of a candidate gene marker for Rf 1 based on a PPR gene in cytoplasmic male sterile CMS-D2 Upland cotton. Molecular Breeding 34: 231-240.

Acknowledgements

The graduate students in the cotton lab of NMSU for their help.