GENOME-WIDE ANALYSIS OF PENTATRICOPEPTIDE REPEAT PROTEINS IN *GOSSYPIUM* SPECIES

Zongfu Han
Cotton Research Center
Shandong Academy of Agricultural Sciences
Jinan, China
Jianyong Wu
Chaozhu Xing
Xihua Li
Jiwen Yu
Institute of Cotton Research
Chinese Academy of Agricultural Sciences
Anyang, China
Mingzhou Song
Jinfa Zhang
New Mexico State University
Las Cruces, NM

Abstract

Pentatricopeptide repeat (PPR) proteins are encoded by genes from one of the largest gene families in higher plants. Most PPR genes are localized in mitochondria and chloroplasts functioning in regulations of plant growth and development, fertility restoration for cytoplasmic male sterility (CMS), RNA editing and stress defense. In this study, four recently sequenced cotton (*Gossypium*) genomes were analyzed to identify PPR protein-coding genes by an *in-silico* analysis. A phylogenetic tree for each species was constructed and compared among the four species. Homologous and homeologous PPR genes were further identified for identification of sequence variations and evolutionary analysis. Candidate PPR genes for ovule and fiber development and fertility restoration of CMS in cotton were further identified and analyzed.

Introduction

The PPR family genes are particularly numerous in land plants, with 450 PPRs identified in Arabidopsis and 477 in rice (Fujii and Small, 2011; Schmitz-Linneweber and Small, 2008). The PPR proteins usually have 2–27 repeating motifs consisting of 35 amino acids arranged in clusters (Small and Peeters, 2000; Lurin et al., 2004) and most of them lack introns in structure. The PPR gene family is divided into the P subfamily and the PLS subfamily based on the P motif variation. The PLS subfamily can be further divided into 4 subclasses by three distinctive C-terminal motifs, i.e., E, E+ and DYW.

As the PPR proteins are closely related to the function of mitochondria and chloroplasts, they play important roles in organelle function in plants (Colcombet et al., 2013). PPR genes can regulate the expression of organelle genes throughout the process of transcription and translation and are involved in plant growth and development, fertility restoration for cytoplasmic male sterility (CMS), and stress defense (Fan et al., 2008). Fertility restoration for CMS crops represents a major challenge in heterosis utilization. PPR genes can restore pollen fertility by inhibiting the expression of mitochondrial CMS genes (O'Toole et al., 2008). Most of the reported restorer genes including *Rfo* in radish (Brown et al., 2003), *Rf-1* in rice (Akagi et al., 2004), and *Rf-PPR592* in petunia (Alfonso and Hanson, 2003) belong to the PPR family. Previous studies indicate that P subfamily is specifically linked to fertility restoration of CMS, while PLS subfamily is more likey involved in RNA editing in both mitochondria and chloroplasts.

Cotton is the world's leading fiber crop as the source of the most important natural textile fiber (Zhang et al., 2014). Two most grown tetraploid species (i.e., *Gossypium hirsutum* and *G. barbadense*) for fiber production and their ancestral diploid species (i.e., *G. arboreum* and *G. raimondii*) have been sequenced, providing an opportunity to study cotton gene families. The role of the PPR gene family has been extensively studied in Arabidopsis, rice, corn, carrots and petunias. However, very few studies have examined this gene family in cotton and the functions of PPRs in *Gossypium* remain to be determined. In this study, we first performed an *in-silico* analysis of the four *Gossypium* genome sequences to identify PPR genes. We then focused on gene expressions to identify candidate PPR genes for ovule and fiber development and fertility restoration of CMS in *G. hirsutum*. The results of this study will contribute to the understanding of the distribution and functions of PPR genes in cotton.

Materials and Methods

Materials

We performed three sets of transcriptome analysis by RNA-seq with following genotypes and tissues: 1) Floral buds about 3 mm in length (representing the stage of male meiosis) of *G. harknessii* cytoplasmic- based CMS line (i.e. A) and its isogenic restorer line with the restorer gene *Rf1* (i.e. R) and maintainer line (i.e. B); 2) Ovules at 0 and 3 DPA (day post-anthesis) and fibers at 10 DPA in two backcross inbred lines (BILs; a long fiber line NMGA-062 & a short fiber line NMGA-105); 3) Ovules of -3 and 0 DPA from Xuzhou 142 and its fibreless and fuzzless mutant *fl*. All of the genotypes were planted under normal field conditions. The tissues were harvested with three biological replicates and immediately frozen in liquid nitrogen and stored at -76°C before use.

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 database (Table 1) using the Hmmer3.1 program with default parameters. Then, the proteins were further queried with P domain HMM model (Lurin et al., 2004) with e-value <-10. All the predicted PPR proteins were analyzed on a standalone PfamScan pipeline and proteins with only one P motif were double checked with TPRpred tools (https://toolkit.tuebingen.mpg.de/tprpred). The sequences contained less than 2 P motifs were excluded. 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. TargetP 1.1was used to predict the subcellular location of the PPR proteins.

Table 1. The genome sources in this study

Species	Genome	Cultivar	Database_name	Source	Publication
G. arboreum	A2	Shixiya 1	*Gossypium arboreum* (A2) Genome BGI Assembly v2.0 & Annotation v1.0	CottonGen	Li et al. (2014)
G. raimondii	D5	CMD 10	*Gossypium raimondii* (D5) genome JGI assembly v2.0 (annot v2.1)	CottonGen	Wang et al. (2012)
G. hirsutum	AD1	TM-1	*Gossypium hirsutum* (AD1) Genome NAU-NBI Assembly v1.1 & Annotation v1.1	CottonGen	Zhang et al. (2015)
G. barbadense	AD2	Xinhai-21	*Gossypium_barbadens*e cv. Xinhai-21 genome	CHGC	Liu et al. (2015)

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 (http://tree.bio.ed.ac.uk/software/figtree/).

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 the four species, i.e., *G. arboreum*, *G. hirsutum*, *G. barbadense* and *G. raimondii*. Then the homologous and homeologous genes were linked with straight lines manually.

Prediction of Candidate PPR-Encoding *Rf* Genes

By 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.

Results and Discussion

A total of 513, 558, 990 and 1054 PPR genes were identified in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively (Table 2). Two diploids contained a similar number of genes and so did tetraploids. The

number of genes in tetraploid species was almost the sum of these from the two diploid progenitors. The P subfamily contained roughly half of the identified PPR genes in the four species. In *G. raimondii* and *G. barbadense*, the P subfamily was slightly larger, representing 53.2% and 52.8%, respectively. However, in *G. arboreum* and *G. hirsutum*, the P subfamily was slightly smaller, representing 49.3% and 49.5%, respectively. Of the 4 subclasses in the PLS subfamily, the E class contained almost half of the PLS subfamily genes, followed by the DYW subclass (representing 31.5%, 38.7%, 38.8% and 36.5% in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively). The E+ class contained the least number of PPR genes. Using TargetP program, 295 (57.5%), 345 (61.8%), 600 (60.6%), 544 (51.6%) PPR proteins were predicted to be targeted to mitochondria or choloroplast in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively.

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

As shown in Figure 1, most of the subfamily members clustered together. 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. This indicated that genetic variations might have occurred in the C-terminal motifs during natural and/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 except in *G. arboreum*. As shown in Figure 2, PPR genes on a single chromosome in *G. hirsutum* usually corresponded to homologs on several chromosomes in *G. arboreum*.

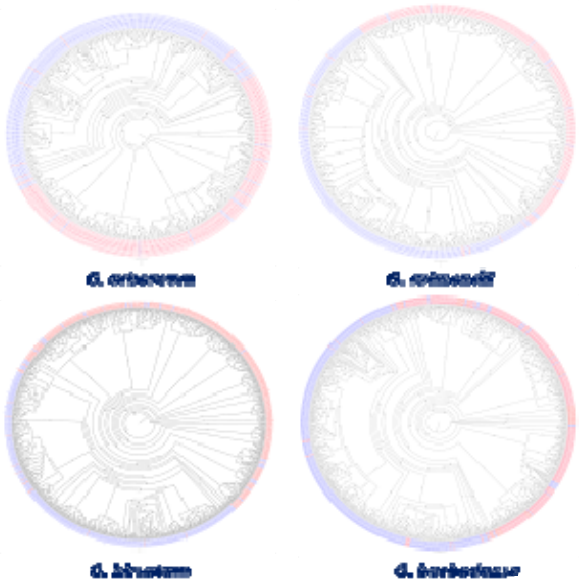

Figure 1. Phylogenetic trees of PPR genes in four Gossypium species. Blue color denotes P subfamily members; Red color denotes PLS subfamily members.

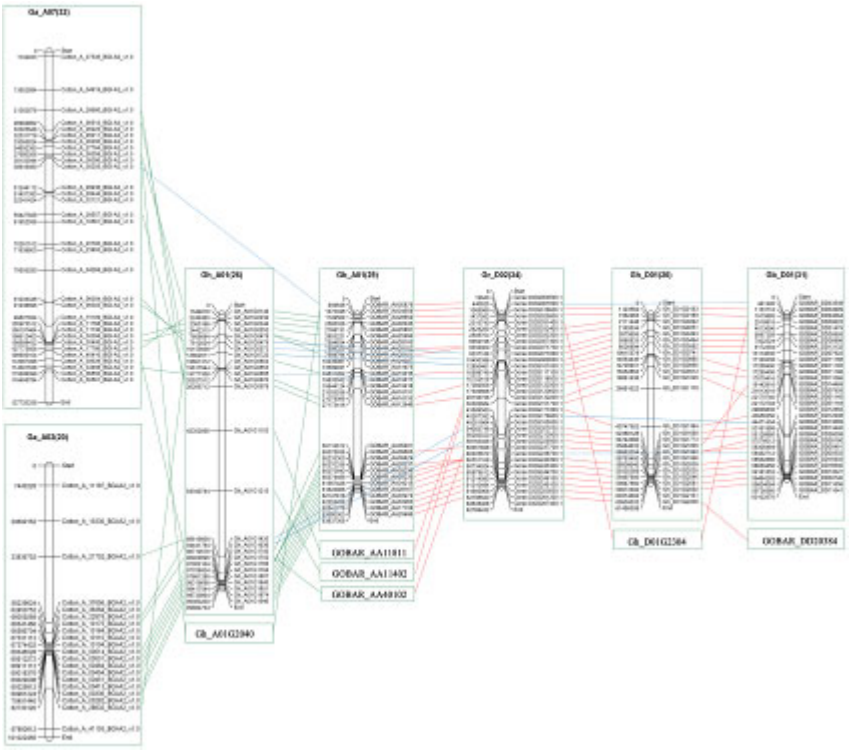

Figure 2. A comparative physical map of PPR genes in four *Gossypium* species (Chromosome 1 as an example).

Previous studies reported that *Rf1* and *Rf2* of CMS in *G. hirsutum* are located on D05 chromosome (Wang et al., 2009; Wu et al., 2014). By using the Blast search program, we determined the physical positions of the two flanked SSR markers for the two restorer genes, which were UBC722 at 43678806 bp and NAU2801 at 58565609 bp on the physical map (Figure 3). Eleven PPR proteins were located in this target region of *Rf* genes. A further analysis indicated that four of them were targeted to chloroplasts.

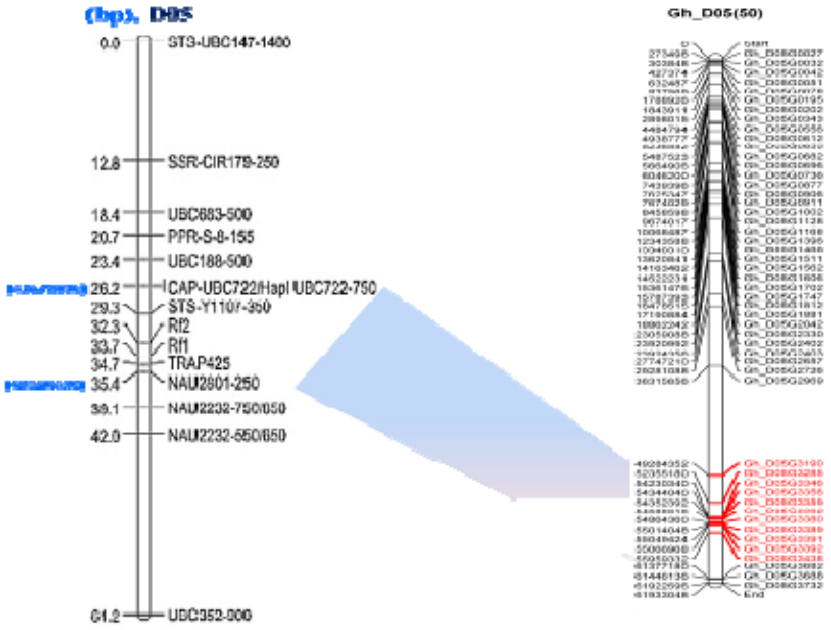

Figure 3. PPR proteins in the target region derived from a linkage map of *Rf* genes (Wang et al., 2009; Wu et al., 2014).

A high throughput RNA-seq indicated that 820 out of 990 PPR genes were expressed in flowering buds of the near-isogenic A, B and R lines (Figure 4A). The R line was significantly different from the A line due to its restorer gene. We identified 6 differentially expressed genes (DEGs) between A and R lines, including two pairs of homeologous genes (Gh_A06G0542 and Gh_D06G0610; Gh_A11G0734 and Gh_D11G0852). Three of them were up-regulated in R line, all of which were from the P subfamily, including one gene which was located onto the target region of the *Rf* genes on the linkage map. In addition, we identified 11 DEGs between R and B lines, and all of them were down-regulated in R line. Gh_A02G1102 was the only common differentially expressed gene between R and B (and A). We also evaluated the sequence variation between R and B, A and TM-1 to find unique sequences in R line. The results indicated that 211 genes carried 272 unique SNPs, and most of them (83.5%, 227/272) were in exons, causing 110 nonsynonymous mutations. A further analysis indicated that two PPR genes, Gh_D05G3190 and Gh_D05G3356, which were in the target region of the two *Rf* genes, carried unique SNPs in R line. However, both of them were not differentially expressed between R and B (or A). Among the DEGs in Table 3, two carried unique SNPs, and both were down-regulated in R line compared to B line. The SNPs in Gh_A02G0212 might cause an early stop codon mutation (Table 4).

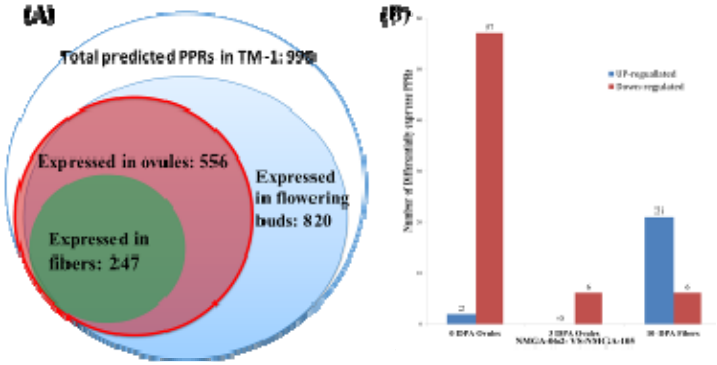

Figure 4. An expression analysis of PPR genes by transcriptome sequencing. (A) The number of PPR genes expressed in flowering buds at meiosis (3 mm in length), ovules (0 and 3 DPA) and fibers (10 DPA); (B) Differentially expressed PPRs (|log2FC| ≥1) between a long fiber and a short fiber BIL in ovules at 0 and 3 DPA and fibers at 10 DPA.

Table 3. Differentially expressed PPRs between A (or B) and R lines.

Gene_id	Chr.	Length (aa)	Subclass	Subcellular location*	log_2FC(R/B)	Regulation	log_2FC(R/A)	Regulation
Gh_A02G1102	A02	700	PLS	C	-1.02	down	-1.34	down
Gh_A06G0542	A06	613	P	_	0.17	-	1.37	up
Gh_A11G0734	A11	637	E	M	-0.12	-	-1.17	down
Gh_D05G3392	D05	532	P	S	0.77	-	1.47	up
Gh_D06G0610	D06	613	P	_	-0.09	-	1.14	up
Gh_D11G0852	D11	637	E	M	-0.84	-	-1.66	down
Gh_A02G0212	A02	702	DYW	C	-1.06	down	-0.06	-
Gh_A03G1158	A03	342	P	S	-1.04	down	-0.59	-
Gh_A04G0297	A04	400	P	_	1.03	down	0.71	-
Gh_A04G1184	A04	516	P	M	-1.12	down	-0.96	-
Gh_A07G1125	A07	855	E	_	-1.96	down	-0.74	-
Gh_A08G0790	A08	918	P	C	-1.12	down	-0.79	-
Gh_A10G2338	scaffold2697_A10	796	E	_	-1.01	down	-0.66	-
Gh_A11G3103	scaffold2759_A11	959	E	_	-1.26	down	-0.69	-
Gh_D05G0032	D05	613	E	_	-1.20	down	-0.56	-
Gh_D12G1745	D12	895	E	_	-1.14	down	-0.71	-

* C: Chloroplast; M: Mitochondrion; S: Secretory pathway; and _: Any other location

A total of 556 and 247 PPRs were expressed in ovules (0 and 3 DPA) and fibers (10 DPA), respectively (Figure 4 A), among which 59, 6, and 27 genes were differentially expressed in 0 and 3 DPA ovules and 10 DPA fibers between long and short fiber BILs, respectively (Figure 4 B). Most of the DEGs (57/59) in 0 DPA ovules and all of the DEGs in 3 DPA ovules were down-regulated in the long fiber BIL. On the contrary, 77.8% (21/27) DEGs between the two BILs in 10 DPA fibers were up-regulated in the long fiber line, indicating an enhancement of gene expression during fiber development might be responsible for the fiber length trait. Only 2 PPR genes were differentially expressed between Xuzhou 142 and its fibreless and fuzzless mutant, and both were up-regulated in Xuzhou 142 (Table 5). Gh_A02G1425 was differentially expressed both in -3 and 0 DPA ovules, and the decreased abundance of transcripts in *fl* led to a higher log2ratio (2-fold higher than -3 DPA) in 0 DPA ovules. Gh_A12G2511 was only differentially expressed in 0 DPA ovules with a log2ratio of 1.2 between Xuzhou 142 and *fl*.

Table 4. Unique SNPs in R line for differentially expressed PPR genes between A/B and R lines.

Gene	Expression variation	Chr.	Position (bp)	TM-1	B	A	R	Annotation	MUT_type
Gh_A02G0212	R/B ↓	A02	2402632	A	A	A	T	exonic	stopgain
Gh_A02G0212	R/B ↓	A02	2402742	C	C	C	T	exonic	nonsynonymous
Gh_A02G0212	R/B ↓	A02	2403007	G	G	G	A	exonic	synonymous
Gh_A10G2338	R/B ↓	scaffold2697_A10	164529	A	A	A	G	exonic	.

Table 5. Differentially expressed PPRs between Xuzhou 142 and its *fl* mutant in ovules.

Stage	Gene	*fl*__FPKM	Xuzhou142__FPKM	log_2FC(Xuzhou142/*fl*)	Regulation
-3 DPA	Gh_A02G1425	4.26	12.65	1.57	Up
0 DPA	Gh_A02G1425	1.34	11.91	3.15	Up
	Gh_A12G2511	8.35	19.31	1.21	Up

Summary

In this study, 513, 558, 990 and 1054 PPR genes were identified in *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, respectively. Phylogenetic trees indicated clustering of the PPR subfamily members. Most of the PPR genes in *G. raimondii*, *G. hirsutum* and *G. barbadense* had a good collinear relationship except in *G. arboreum*. Eleven PPR proteins were located in the previous published target region of *Rf* genes for CMS-D2 and CMS-D8 systems. A high throughput RNA-seq indicated that 820 PPRs were expressed in flowering buds. The number of differentially expressed genes between R and B (and A) lines was 11 (and 6), two of which carried SNPs. A total of 556 and 247 PPRs were expressed in ovules (0 and 3 DPA) and fibers (10 DPA), respectively, among which 59, 6, and 27 genes were differentially expressed in 0 and 3 DPA ovules and 10 DPA fibers between long and short fiber BILs, respectively. Only 2 PPR genes were differentially expressed between Xuzhou 142 and its fibreless and fuzzless mutant, and both of them were up-regulated in Xuzhou 142. This study provided an essential piece of information on PPR gene family in four *Gossypium* species, and some differentially expressed PPRs related to ovule and fiber development and fertility restoration of cytoplasmic male sterility in G. *hirsutum*. The results of this study will contribute to the understanding of the distribution and functions of PPR genes in cotton including their associations with agronomic traits in cotton.

References

Akagi, H., A. Nakamura, Y. Yokozeki-Misono, A. Inagaki, H. Takahashi, K. Mori, 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.

Alfonso, A. A., S. Bentolila, and M. R. Hanson. 2003. Evaluation of the fertility restoring ability of *Rf-PPR592* in Petunia. Philippine Agric. Sci. 86:303-315.

Brown, G. G., N. Formanová, H. Jin, R. Wargachuk, C. Dendy, P. Patil, et al. 2003. The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. The Plant Journal 35:262–272.

Colcombet, J., M. Lopez-Obando, L. Heurtevin, C. Bernard, K. Martin, R. Berthomé, et al. 2013. Systematic study of subcellular localization of Arabidopsis PPR proteins confirms a massive targeting to organelles. RNA Biol 10:1557–1575.

Fan, M., L. P. Jin, Q. C. Liu, and D. Y. Qu. 2008. Cloning of *SoDIPPR* gene of pentatricopeptide repeat (PPR) protein family in potato and analysis of expression characteristics under drought conditions. Sci. Agric. Sin. 41:2249–2257.

Fujii, S., and I. Small. 2011. The evolution of RNA editing and pentatricopeptide repeat genes. New Phytol. 191:37–47.

Lurin, C., C. Andrés, S. Aubourg, M. Bellaoui, F. Bitton, C. Bruyère, et al. 2004. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. The Plant Cell 16:2089–2103.

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

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

Small, I. D., and N. Peeters. 2000. The PPR motif–a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25:45–47.

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

Wang, F., B. Yue, J. Hu, J. M. Stewart, and Jinfa Zhang. 2009. A target region amplified polymorphism marker for fertility restorer gene and chromosomal localization of and in cotton. Crop Sci. 49:1602–1608.

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

Zhang, J., R. G. Percy, and J. C. McCarty Jr. 2014. Introgression genetics and breeding between Upland and Pima cotton: a review. Euphytica 198:1–12.