Journal of Cotton Research

RESEARCH Open Access

Genome-wide analysis of *Rf-PPR-like* (*RFL*) genes and a new InDel marker development for *Rf1* gene in cytoplasmic male sterile CMS-D2 Upland cotton

ZHANG Bingbing, ZHANG Xuexian, GUO Liping, QI Tingxiang, WANG Hailin, TANG Huini, QIAO Xiuqin, SHAHZAD Kashif, XING Chaozhu* and WU Jianyong*

Abstract

Background: Cytoplasmic male sterility in flowering plants is a convenient way to use heterosis via hybrid breeding and may be restored by nuclear restorer-of-fertility (*Rf*) genes. In most cases, *Rf* genes encoded pentatricopeptide repeat (PPR) proteins and several *Rf* genes are present in clusters of similar *Rf-PPR-like* (*RFL*) genes. However, the *Rf* genes in cotton were not fully characterized until now.

Results: In total, 35 *RFL* genes were identified in *G. hirsutum*, 16 in *G. arboreum*, and 24 in *G. raimondii*. Additionally, four RFL-rich regions were identified; the RFL-rich region in Gh_D05 is the probable location of *Rf-PPR* genes in cotton and will be studied further in the future. Furthermore, an insertion sequence was identified in the promoter sequence of *Gh_D05G3392* gene in the restorer line, as compared with the CMS-D2 line and maintainer lines. An InDel-R marker was then developed and could be used to distinguish the restorer line carrying *Rf1* from other genotypes without the *Rf1* allele.

Conclusion: In this study, genome-wide identification and analysis of *RFL* genes have identified the candidate *Rf-PPR* genes for CMS in *Gossypium*. The identification and analysis of *RFL* genes and sequence variation analysis will be useful for cloning *Rf* genes in the future and also for three-line hybrid breeding in cotton.

Keywords: Upland cotton, CMS, *Rf-PPR-like* gene, Restorer gene, InDel marker

Introduction

Cotton is an important fiber crop worldwide. Improving cotton yield and quality is becoming critical to meet industrial demands. Hybrid breeding is an important strategy to increase yield and quality by efficiently exploiting heterosis and has been applied to many important crops, including rice, maize, and cotton (Huang et al. 2016). In China, more than 90% of cotton hybrids are produced by artificial emasculation and pollination (Yu et al. 2016). It is time-consuming, labor-intensive, and costly and the purity of hybrid seeds cannot be guaranteed, representing an important limiting factor for hybrid seed production. One of the major challenges is the absence of a pollination control strategy that could efficiently produce hybrid seed on a commercial level. In other crops, cytoplasmic male sterility (CMS) is an indispensable resource for commercial hybrid seed production (Schnable and Wise 1998; Hanson and Bentolila 2004; Chase 2007; Pelletier and Budar 2006).

CMS is a maternally inherited trait in flowering plants that cannot produce functional pollen (Hanson and Bentolila 2004). The CMS trait is caused by the rearrangement of the mitochondrial genome and several CMS genes have been identified in many crops (Schnable and Wise 1998; Hanson and Bentolila 2004; Chase 2007). The products of CMS genes destroy the normal function of mitochondria and cause a deficiency in the energy supply required for pollen development, resulting in aborted pollen (Schnable and Wise 1998). The CMS phenotypes could be restored

* Correspondence: chaozhuxing@126.com; dr.wujianyong@live.cn
State Key Laboratory of Cotton Biology/Institute of Cotton Research of Chinese Academy of Agricultural Sciences, 38 Huanghe Dadao, Anyang 455000, Henan, China

by the fertility restorer (*Rf*) genes from the nuclear genome. Previous studies have indicated that the *Rf* genes identified in petunia (Bentolila et al. 2002), radish (Brown et al. 2003; Desloire et al. 2003), rice (Tan et al. 2004, 2008; Fujii et al. 2014; Igarashi et al. 2016), and sorghum (Klein et al. 2005) belong to a pentatricopeptide repeat (PPR) gene family. Exceptions are the maize *Rf2*, which encodes an aldehyde dehydrogenase that may be involved in the production of the plant hormone indole-3-acetyl acetate (Cui et al. 1996; Liu and Schnable 2002), and the *Rf2* gene in rice for Lead-type CMS that encodes a protein containing a glycine-rich domain (Itabashi et al. 2010). Additionally, three *PPR* genes cosegregated with the *Rf3* gene of S type CMS in maize (Xu et al. 2009), and the *Rf5* gene in rice encodes a PPR protein interacting with a glycine-rich domain protein (GRP) which restores fertility in Hong-Lian CMS lines (Hu et al. 2012). These studies indicated that PPR genes have important relationships with the *Rf* genes in plants.

In cotton, two main CMS systems, CMS-D2–2 and CMS-D8, have been developed by transferring exotic cytoplasm from *Gossypium harknessii* Brandegee (D2) and *G. trilobum* (DC.) Skovst. (D8) into the Upland cotton (*G. hirsutum*, AD1) nuclear background (Meyer 1975; Yin et al. 2006; Zhang et al. 2007; Wang et al. 2010; Wu et al. 2011). So far, no studies have reported the cloning of cotton *Rf* genes, with most studies focusing on genetic mapping and the development of related markers. Previous studies have indicated that the *Rf1* gene from *G. harknessii* (D2) can restore the fertility of both CMS-D2 and CMS-D8, whereas the *Rf2* gene from *G. trilobum* only restores male fertility to CMS-D8 (Zhang and Stewart 2001a, 2001b). Additionally, the *Rf1* and *Rf2* genes in cotton function sporophytically and gametophytically, respectively. These two restorer genes are not allelic but tightly linked in 0.93 cM (Yin et al. 2006; Wang et al. 2009; Wu et al. 2011, 2014). Yin et al. (2006) identified that the marker NAU4047 is closely linked to *Rf1* (within 0.2 cM) and delimited the *Rf1* gene to a 100-kb region. Furthermore, the *Rf1* gene is located on the Gh_D05 chromosome, with genetic mapping indicating that the nearest SSR markers to *Rf1* are BNL3535 (within 0.049 cM) and NAU3652 on the other side (within 0.078 cM). An *Rf1*-specific CAPS marker was developed based on a candidate *PPR* gene and could ensure the purity of restorer lines (Wang et al. 2007, 2009; Wu et al. 2014). Wang et al. (2007) constructed a linkage map with nine markers flanking the *Rf2* gene including a PPR-AFLP marker. A whole-genome resequence was completed for the restorer N (*Rf1Rf1*) and maintainer N (*rf1rf1*) lines that indicated that most of the InDels were distributed near the region containing the *Rf1* gene in Gh_D05. Furthermore, an InDel-1891 marker was developed for fine mapping of the *Rf1* gene (Wu et al. 2017).

The PPR gene family constitute a large family of RNA-binding proteins in plants and the members are involved in many cellular functions and biological processes in organelles, including gene expression, RNA stabilization, RNA cleavage, and RNA editing (Schmitzlinneweber and Small 2008; Prikryl et al. 2010). Previous studies indicated that all cloned *Rf-PPR* genes might have a common ancient ancestor and that *Rf-CMS* genes have coexisted during the evolutionary process (Geddy and Brown 2007; Fujii et al. 2011; Joanna et al. 2016; Sykes et al. 2017). For example, *Rf1a* and *Rf1b* genes in rice share 70% identity between their protein sequences (Wang et al. 2006) while in radish the Rf3 protein shows 85% similarity with the Rf0 protein (Wang et al. 2013). Additionally, several studies indicated that *Rf-PPR* genes are targeted to mitochondria where they prevent the accumulation of the CMS-specific gene product (Bentolila et al. 2002; Wang et al. 2006; Kazama et al. 2008). Furthermore, these *Rf-PPR* genes are presented in clusters of similar *Rf-PPR-like* (*RFL*) genes in almost all cases (Bentolila et al. 2002; Wang et al. 2006; Kazama et al. 2008; Uyttewaal et al. 2008; Barr and Fishman 2010). *RFL* genes at the same genomic region are most likely to be active restorer genes and several *PPR-Rf* genes present within the RFL-rich region such as the rice *Rf1* and *Rf4* genes presented in the RFL-rich region of rice chromosome 10 (Wang et al. 2006; Fujii et al. 2011; Luo et al. 2013). Additionally, the *Rf5* gene in rice was mapped to a 200-kb region on chromosome 8 that contains three *RFL* genes, one of which, *Os08g01870,* was located within 15 kb of the marker and cosegregated with the *Rf* gene (Hu et al. 2012; Huang et al. 2016). In maize, the *Rf8* locus was mapped to an RFL cluster on chromosome 2 (Meyer et al. 2011). The only *PPR-Rf* gene identified in sorghum was found to be located outside of the RFL-rich regions, however, occurs on chromosome 8. This gene most likely encodes a PPR protein belonging to the PLS (P-L-S motifs) subfamily that is involved in RNA editing events, indicating that the mechanism of fertility restoration in sorghum may be unique (Klein et al. 2005; Schmitzlinneweber and Small 2008; Dahan and Mireau 2013). This allowed us to further explore the candidate *Rf* genes in cotton by identifying the RFL-rich region that shows a similar pattern to other species.

In cotton, we have characterized the DYW (Asp-Tyr-Trp tripeptide in C terminal domain) deaminase domain-containing PPR genes belonging to PLS subfamily and have determined that these genes may not directly function in the occurrence of CMS or in fertility restoration, while P (common PPR motif) subfamily genes might have a critical role in the fertility restoration process (Zhang et al. 2017). However, no results have been reported regarding the identification and analysis of *RFL* genes in cotton until now. Here, to identify the candidate *Rf-PPR* genes for CMS in cotton, a genome-wide identification and analysis

of *RFL* genes were completed in *Gossypium*. The *RFL* genes identified and analyzed in our study will be useful for cloning the *Rf* genes and for three-line cotton hybrid breeding in the future.

Materials and methods

Cotton genome and RNA-seq resources

The genome sequence and annotation information of three *Gossypium* species (*G. raimondii, G. arboreum,* and *G. hirsutum*) were downloaded from Cottongen (https://www.cottongen.org). The raw sequence data of a 3 mm floral bud transcriptome from three-line hybrid cotton (CMS-D2 line A, maintainer line B, and restoration line R) could be found in the National Center for Biotechnology Information (NCBI) under accession number SRX3421007.

Identification and chromosomal mapping *RFL* genes in *Gossypium*

To precisely identify the RFL genes in *Gossypium*, BLAST (http://www.ncbi.nlm.nih.gov/Tools/) was used to search sequences in three cotton genomes. The sequence of *Rf-PPR592* from *Petunia hybrida* identified previously was used for searches against the whole genome database of the three cotton species. Hits with an estimated E-value under $1e^{-100}$ were set as threshold (Fujii et al. 2011). The number of PPR domains in the protein structure was further validated using SMART software (http://smart.embl-heidelberg.de).

The physical location data of *RFL* genes were retrieved from genome sequence data of three cotton species. Mapping of these *RFL* genes was then performed using Mapchart software (Voorrips 2002).

Subcellular location analysis

The signal peptide prediction program Target P (http://www.cbs.dtu.dk/services/TargetP/) was used to predict the subcellular location of RFL proteins.

Quantitative (q) RT-PCR validation of DEG expression

The CMS-D2 three-line hybrid cotton system was obtained from the Institute of Cotton Research, Chinese Academy of Agricultural Science (ICR, CAAS). The three lines were planted under normal production conditions. Samples were collected as described previously (Wu et al. 2011; Suzuki et al. 2013); floral buds approximately 3 mm in length (corresponding roughly to the meiosis stage) were collected with three independent biological replicates. All collected floral buds were cut above ovaries and immediately frozen in liquid nitrogen and stored at − 80 °C. Total RNAs were extracted from floral buds and reverse transcribed to cDNA using a PrimeScript RT reagent kit (Takara, Dalian) following the manufacturer's guidelines. For qRT-PCR, reactions were performed in 20-μL volumes containing 1 μL diluted cDNA, 10 μL 2× SYBR Green Mix (Takara), 7 μL water and 1 μL each of forward primer and reverse primer. The amplifications were carried out as follows: 94 °C for 30 s, then 40 cycles of 94 °C for 5 s, 55 °C for 15 s, and 72 °C for 25 s. The cotton *histone 3* (*GhHIS3*) was used as a reference gene for normalization. All the primers were listed in Additional file 1: Table S1.

Promoter sequence analysis and InDel marker development

Total genomic DNA from the three lines was extracted from leaves using the CTAB method (Paterson et al. 1993), respectively. Additionally, gene-specific primers were designed by using Primer Premier 5.0 software (http://www.premierbiosoft.com) to amplify the promoter sequence of *Gh_D05G3392* gene in the A, B and R lines. A 20-μL mixture consisting of 1× reaction buffer, 2.0 $mmol{\cdot}L^{-1}$ $MgCl_2$, 0.2 $mmol{\cdot}L^{-1}$ dNTPs, 0.5 $mmol{\cdot}L^{-1}$ of each primer, 1 U Taq DNA polymerase (Takara, Japan), and 50 ng DNA template was used. The PCR procedure was as follows: 35 cycles of 94 °C for 30 s, then 58 °C for 30 s, and 72 °C for 60 s. The PCR mixture was separated and purified by TaKaRa DNA Fragment Purification Kit. Then the DNA fragment was ligated into the pEASY-T1 vector (TransGen, Beijing), following the manufacture's guidelines. Then five clones were selected in every sample for sequencing. The MEGA7.0 was used for sequence alignment.

The *cis*-acting element identification in the promoter region was completed by using plant *cis*-acting regulatory DNA elements (https://www.dna.affrc.go.jp/htdocs/PLACE/).

An InDel-R marker was then developed and the primer pair (forward: 5′- GAAAGTTGGACAACAATGAGAAGTC-3′; reverse: 5′- CCAATTTCTAATAAAGAAAAGAAAGAG-3′) were designed for applications. A 20-μL mixture consisting of 1× reaction buffer, 2.0 $mmol{\cdot}L^{-1}$ $MgCl_2$, 0.2 $mmol{\cdot}L^{-1}$ dNTPs, 0.5 $mmol{\cdot}L^{-1}$ of each primer, 1 U Taq DNA polymerase (Takara, Japan), and 50 ng DNA template was used. PCR was performed as follows: 30 cycles of 94 °C for 30 s, then 56 °C for 30 s, and 72 °C for 10 s. The PCR products were then separated using 3.0% agarose gel electrophoresis.

Results

Genome-wide identification and chromosomal distribution of *RFL* genes in *Gossypium*

To identify potential *RFL* genes in the *G. hirsutum*, *G. arboreum,* and *G. raimondii* protein databases, the sequence of *Rf-PPR592* from *P. hybrida* was used for BLAST searching against the three cotton genomes, as per the previous study by Fujii et al. (2011). Hits with an estimated E-value under $1e^{-100}$ were collected (Fujii et al. 2011). In total, 75 *RFL* genes were identified, of which

35 were obtained from *G. hirsutum*, 16 from *G. arboreum*, and 24 from *G. raimondii*. Analysis of the 75 predicted cotton RFL proteins, which identified by homology to the known restorer genes *Rf-PPR592* from *P. hybrida*, revealed that these proteins also belonged to the P subfamily. Further analysis indicated that the number of PPR motifs in the proteins ranged from 9 to 20 (Table 1).

The 35 *RFL* genes which identified from *G. hirsutum* were found to be located on 15 chromosomes, with 17 and 18 genes distributed to the A and D sub-genomes, respectively (Fig. 1), with the *Gh_A04G1306* and *Gh_A04G1307* genes localized to scaffold756_A04. Additionally, six and five genes were located on chromosome 5 and 10 in the D sub-genome, respectively. Chromosomes 1, 5, 6, 7, 12, and 13 in the A sub-genome and chromosomes 1, 4, 6, 7, and 12 in the D sub-genome were the exceptions and did not contain any *RFL* genes. Previously, the rice *Rf1* (Wang et al. 2006) and *Rf4* (Luo et al. 2013) genes were found to occur in the RFL-rich region of rice chromosome 10. In our study, four RFL-rich regions were identified, including three *RFL* genes in Gh_A04, four *RFL* genes in Gh_A10, six *RFL* genes in Gh_D05, and five *RFL* genes in Gh_D10. The *RFL* genes in these regions will be studied further.

Expression patterns of *RFL* genes and qPCR validation

Additionally, because of the tissue and time-specific expression of *RFL* genes (Prasad et al. 2003; Tomohiko and Kinya 2014), transcriptomic data from 3 mm floral buds of three-line hybrid cotton (CMS-D2 line (A), maintainer line (B), and restoration line (R)) were used to identify candidate *Rf-PPR* genes within the RFL-rich region (Fig. 2) (Additional file 2: Table S2). Interestingly, three genes (*Gh_D05G3356*, *Gh_D05G3389,* and *Gh_D05G3392*) in Gh_D05 were up-regulated in the R line as comparing with the A and B lines. To verify the expression profiles of the *RFL* genes, three genes (*Gh_D05G3356*, *Gh_D05G3389,* and *Gh_D05G3392*) were selected for qPCR analysis using the 3 mm floral buds from the A, B, and R lines. Their gene expression patterns were similar to the RNA-seq data and indicated that all three genes were up-regulated in the R line as comparing with the A and B lines. This suggests that these genes might play critical roles in fertility restoration.

Sequence variation of DEGs on Chr_05

Furthermore, the above transcriptomic data were further used to identify single nucleotide polymorphism (SNPs) in the three differentially expressed *RFL* genes (*Gh_D05G3356*, *Gh_D05G3389,* and *Gh_D05G3392*) on Chr_05. In total, 37 SNP loci were identified between the sequences from the R line and that from the non-restoring genome A and B lines (Additional file 3: Table S3). The results implied that these SNPs might be linked to the fertility restoring gene on Chr_05. In addition, promoter sequence analysis of *Gh_D05G3392* gene among the A, B, and R lines was also conducted. Consistent with the coding region between the R line and the A and B lines, a high level of polymorphisms was observed in the promoter region (Fig. 3). Multiple alignments indicated that several SNP loci and seven InDels specifically exist between the restoration R line and the non-restoring genome A and B lines. Furthermore, there was a 12 nt insertion "TAGAAGACTGGA" in the restorer line as comparing with the A and B lines.

A search for *cis*-acting elements in the promoter region of *Gh_D05G3392* gene was completed by using plant *cis*-acting regulatory DNA elements (https://www.dna.affrc.go.jp/htdocs/PLACE/). Except for the core promoter element "TATA" box, we also found other motifs associated with light responsiveness (GA-motif (AAGGAAGA) and I-box (GATATGG)) and a TCA-element (CCATCTTT) involved in salicylic acid responsiveness. Furthermore, five copies of the pollen specific motifs POLLEN1LELAT52 (AGAAA) (Filichkin and Nonogaki 2004) were also identified, which indicated that transcriptional activation of *Gh_D05G3392* gene might be controlled by the pollen specific *cis*-regulatory elements.

An InDel-R marker was then developed for this insertion sequence that was verified as a co-dominant marker in the three lines. A total of 24 randomly selected individual BC_5F_2 plants were checked using this InDel-R marker. As shown in Fig. 4, the InDel-R marker could be used to distinguish the restorer line carrying *Rf1* from other genotypes without the *Rf1* allele. The result showed three different PCR band models in which a single PCR band of nearly 149 base pairs (bp) represented plants homozygous for the *Rf* gene allele N(*Rf1Rf1*) and a single PCR band of nearly 137 bp represented plants lacking the *Rf* gene allele (*rf1rf1*). Plants containing both PCR bands were considered heterozygous at the *Rf* gene locus N(*Rf1rf1*). These results indicated that this InDel-R marker could be used in the marker-assisted breeding of fertility restoration lines carrying the *Rf1* gene.

Discussion

Previous studies have indicated that most *Rf* genes came from the same small clade of *PPR* genes, with many similarities and are usually presented as clusters of similar *Rf-PPR-like* (*RFL*) genes in many plants (Bentolila et al. 2002; Kazama et al. 2008; Uyttewaal et al. 2008; Barr and Fishman 2010; Fujii et al. 2011). The importance of the *Rf* gene in the CMS/*Rf* system of cotton resulted in many studies aiming to identify molecular markers linked to the *Rf* gene; there have been no reports regarding cloning of the *Rf* gene until now. In this study, we performed genome-wide identification and analysis of *RFL* genes in

Table 1 Characteristics of *RFL* genes and predicted properties of RFL proteins in three *Gossypium* species

Gene ID	Chromosome number	Location	Intron	Length /aa	Domain number	Subcellular location
RFL genes in *G. hirsutum*						
Gh_A02G0346	A02	4061961-4063892(–)	0	643	12	Chloroplast
Gh_A03G0085	A03	1310395-1312206(+)	0	603	12	Chloroplast
Gh_A04G0298	A04	6938035-6939906(+)	1	538	14	Signal peptide
Gh_A04G0299	A04	6958902-6961157(+)	1	740	17	Chloroplast
Gh_A04G0308	A04	7067031-7068902(+)	0	623	12	Signal peptide
Gh_A04G1306	scaffold756_A04	11713-13662(+)	0	649	15	_
Gh_A04G1307	scaffold756_A04	42960-44756(+)	0	598	13	Mitochondrial
Gh_A08G1858	A08	99291747-99293615(–)	1	538	14	Mitochondrial
Gh_A08G1886	A08	99594679-99596448(+)	1	557	14	Signal peptide
Gh_A09G0071	A09	1545143-1547526(–)	1	443	12	_
Gh_A09G0099	A09	2441359-2443299(+)	0	646	14	Chloroplast
Gh_A09G1959	A09	72667483-72669399(–)	0	638	10	_
Gh_A10G1153	A10	58877471-58931193(–)	2	867	20	Signal peptide
Gh_A10G1192	A10	62428148-62429707(+)	0	519	13	Mitochondrial
Gh_A10G1204	A10	62755821-62757800(–)	2	485	12	_
Gh_A10G1206	A10	62766997-62769312(–)	1	726	18	_
Gh_A11G1174	A11	14344990-14346882(+)	1	558	12	Mitochondrial
Gh_D02G0409	D02	5265341-5267272(–)	0	643	12	Chloroplast
Gh_D03G1566	D03	44987228-44989030(–)	0	600	12	Mitochondrial
Gh_D05G3346	D05	54230340-54232205(–)	0	621	12	Chloroplast
Gh_D05G3356	D05	54344042-54346277(–)	1	602	13	_
Gh_D05G3362	D05	54500616-54502460(–)	0	614	14	_
Gh_D05G3380	D05	54864359-54866280(+)	1	557	13	_
Gh_D05G3389	D05	55014049-55028124(–)	6	749	14	_
Gh_D05G3392	D05	55066970-55068568(–)	0	532	13	Signal peptide
Gh_D08G2249	D08	62092399-62094342(+)	0	647	15	Mitochondrial
Gh_D09G0096	D09	2499241-2501169(+)	0	642	14	Chloroplast
Gh_D09G2163	D09	48772802-48774718(–)	0	638	10	_
Gh_D10G1292	D10	23933234-23935420(+)	0	728	18	_
Gh_D10G1294	D10	23946799-23948358(+)	0	519	14	_
Gh_D10G1307	D10	24224714-24226273(–)	0	519	13	Mitochondrial
Gh_D10G1342	D10	25844049-25845929(+)	0	626	13	Mitochondrial
Gh_D10G1344	D10	25875652–25882818(+)	6	906	18	_
Gh_D11G1331	D11	12830305-12832155(+)	0	616	12	Chloroplast
Gh_D13G0526	D13	6880247-6881917(+)	0	556	14	_
RFL genes in *G. arboreum*						
Cotton_A_08373	Ca13	57501074-57503005(+)	0	643	13	Chloroplast
Cotton_A_14708	Ca10	102602256-102604196(–)	0	646	14	Chloroplast
Cotton_A_14743	Ca10	103472578-103474212(+)	0	544	12	_
Cotton_A_16847	Ca7	9513817-9515736(–)	1	573	13	_
Cotton_A_18522	Ca7	61596824-61598692(–)	0	622	14	Chloroplast
Cotton_A_23070	Ca8	78350187-78351653(+)	0	488	13	Signal peptide
Cotton_A_23084	Ca8	78693577-78695556(–)	0	659	15	Mitochondrial

Table 1 Characteristics of *RFL* genes and predicted properties of RFL proteins in three *Gossypium* species *(Continued)*

Gene ID	Chromosome number	Location	Intron	Length /aa	Domain number	Subcellular location
Cotton_A_24432	Ca8	85527702-85529582(−)	0	626	13	Mitochondrial
Cotton_A_24724	Ca10	34691224-34693117(−)	1	630	10	Mitochondrial
Cotton_A_26557	Ca3	23320169-23321491(−)	0	440	10	Signal peptide
Cotton_A_26837	Ca5	136263118-136264365(+)	0	415	11	_
Cotton_A_29292	Ca6	90346627-90348504(−)	0	625	13	Chloroplast
Cotton_A_29299	Ca6	90193385-90195133(−)	0	582	14	_
Cotton_A_29300	Ca6	90149515-90151278(−)	0	587	12	Mitochondrial
Cotton_A_30591	Ca4	116495782-116497152(+)	0	456	11	Mitochondrial
Cotton_A_33520	Ca6	91413647-91415518(+)	0	623	12	Mitochondrial
RFL genes in *G. raimondii*						
Cotton_D_gene_10000174	scaffold587	10218-12381(+)	1	698	15	Chloroplast
Cotton_D_gene_10000410	scaffold520	19372–20826(−)	0	484	13	Signal peptide
Cotton_D_gene_10000446	scaffold516	21891-23345(+)	0	484	13	Signal peptide
Cotton_D_gene_10000448	scaffold516	47453-49639(+)	0	540	11	Chloroplast
Cotton_D_gene_10000451	scaffold516	90985-92439(+)	0	484	13	Signal peptide
Cotton_D_gene_10000822	scaffold512	92947-94929(+)	0	660	13	Chloroplast
Cotton_D_gene_10000826	scaffold512	134156-135944(+)	1	564	10	Chloroplast
Cotton_D_gene_10002529	scaffold461	131009-132463(−)	0	484	13	Signal peptide
Cotton_D_gene_10003142	Chr9	41720023-41721747(+)	1	520	9	Signal peptide
Cotton_D_gene_10003980	scaffold288	253425-255650(−)	0	623	14	Mitochondrial
Cotton_D_gene_10003981	scaffold288	259126-260817(−)	1	479	11	_
Cotton_D_gene_10004373	scaffold326	137744-139303(−)	1	519	13	Mitochondrial
Cotton_D_gene_10005258	Chr5	742735-745217(+)	0	641	12	Chloroplast
Cotton_D_gene_10007940	Chr4	2510485-2512428(+)	0	647	15	Mitochondrial
Cotton_D_gene_10009676	Chr7	41367207-41369072(−)	0	621	12	Chloroplast
Cotton_D_gene_10009740	Chr3	1451913-1453283(+)	0	456	10	Mitochondrial
Cotton_D_gene_10013435	scaffold333	1235526-1237394(−)	0	622	12	Chloroplast
Cotton_D_gene_10013437	scaffold333	1247144-1249009(−)	0	621	12	Mitochondrial
Cotton_D_gene_10014531	scaffold324	965465-968260(+)	1	579	13	_
Cotton_D_gene_10014534	scaffold324	998554-1000283(+)	1	541	13	_
Cotton_D_gene_10021157	Chr6	45498703-45500927(−)	0	638	10	_
Cotton_D_gene_10026507	Chr11	23161755-23163734(+)	0	659	15	Mitochondrial
Cotton_D_gene_10027032	Chr6	2797685-2799676(+)	1	555	12	_
Cotton_D_gene_10027066	Chr6	2028765-2030705(−)	0	646	14	Mitochondrial

G. hirsutum, *G. arboreum*, and *G. raimondii* to identify candidate *Rf* genes for CMS in cotton.

The *RFL* genes in *Gossypium*

In the draft genome sequence of cotton, a total of 35 *RFL* genes were identified from *G. hirsutum*; this is in contrast to previous studies that have suggested the presence of around 10–30 *RFL* genes per plant genome (Andrés et al. 2007; Fujii et al. 2011; Joanna et al. 2016; Sykes et al. 2017). This difference may be associated with the polyploidization of Upland cotton that has resulted in whole genome duplication (WGD). Additionally, 16 and 24 *RFL* genes were identified from *G. arboreum* and *G. raimondii*, respectively. Gene structure analysis revealed that *RFL* genes only contain the PPR domain and that these genes belong to the P subfamily.

Identification of an RFL-rich region

Previous studies have indicated that *Rf-PPR* genes are targeted to mitochondria where they prevent the accumulation of the CMS-specific gene products (Bentolila et al. 2002; Kazama et al. 2008; Uyttewaal et al. 2008;

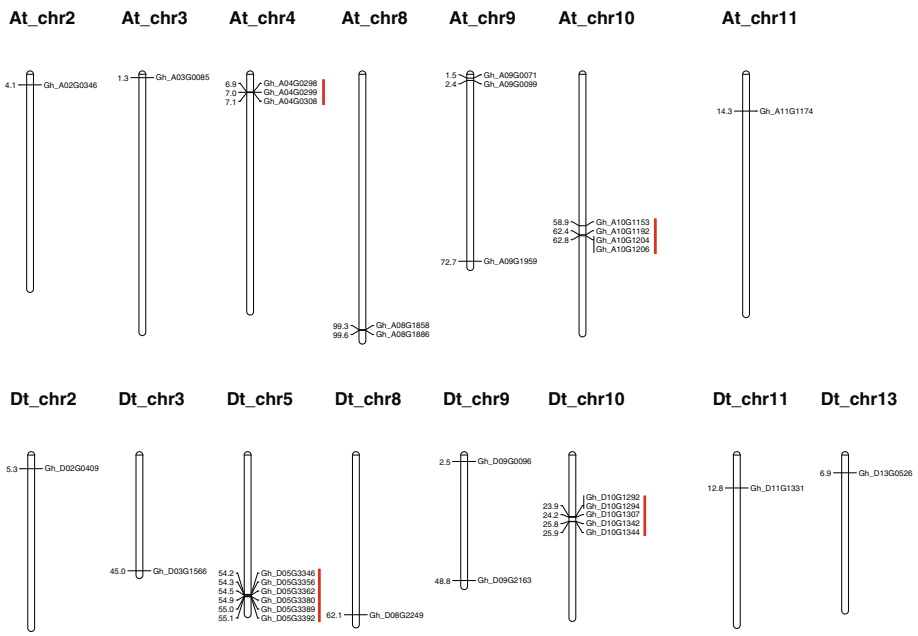

Fig. 1 The putative chromosome location of *RFL* genes on *G. hirsutum*. The scale represents megabases (Mb). The red column represents the RFL-rich region

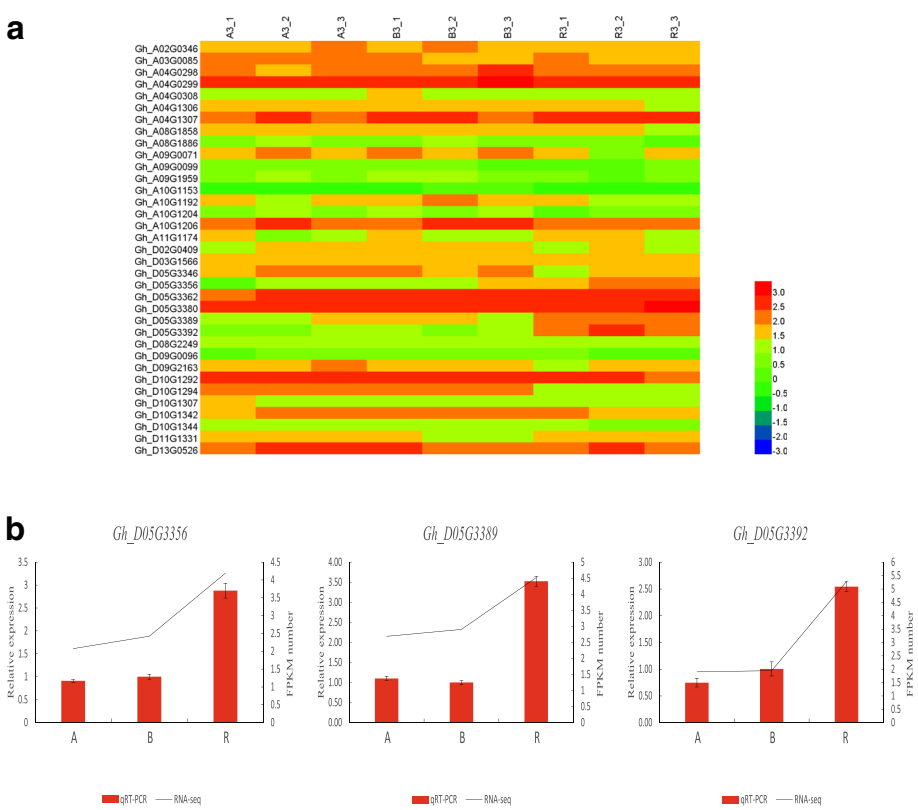

Fig. 2 qRT-PCR analysis of *RFL* gene expression compared with the RNA-seq data from three-line hybrid cotton lines (CMS-D2 line A, maintainer line B, and restoration line R). **a** The RNA-seq data of *RFL* genes in A, B and R lines. **b** The qRT-PCR analysis of three differentially expressed *RFL* genes. The red columns represented the relative expression levels of the genes; the black lines represented the FPKM number. A: CMS-D2 line, B: maintainer line, R: restorer line

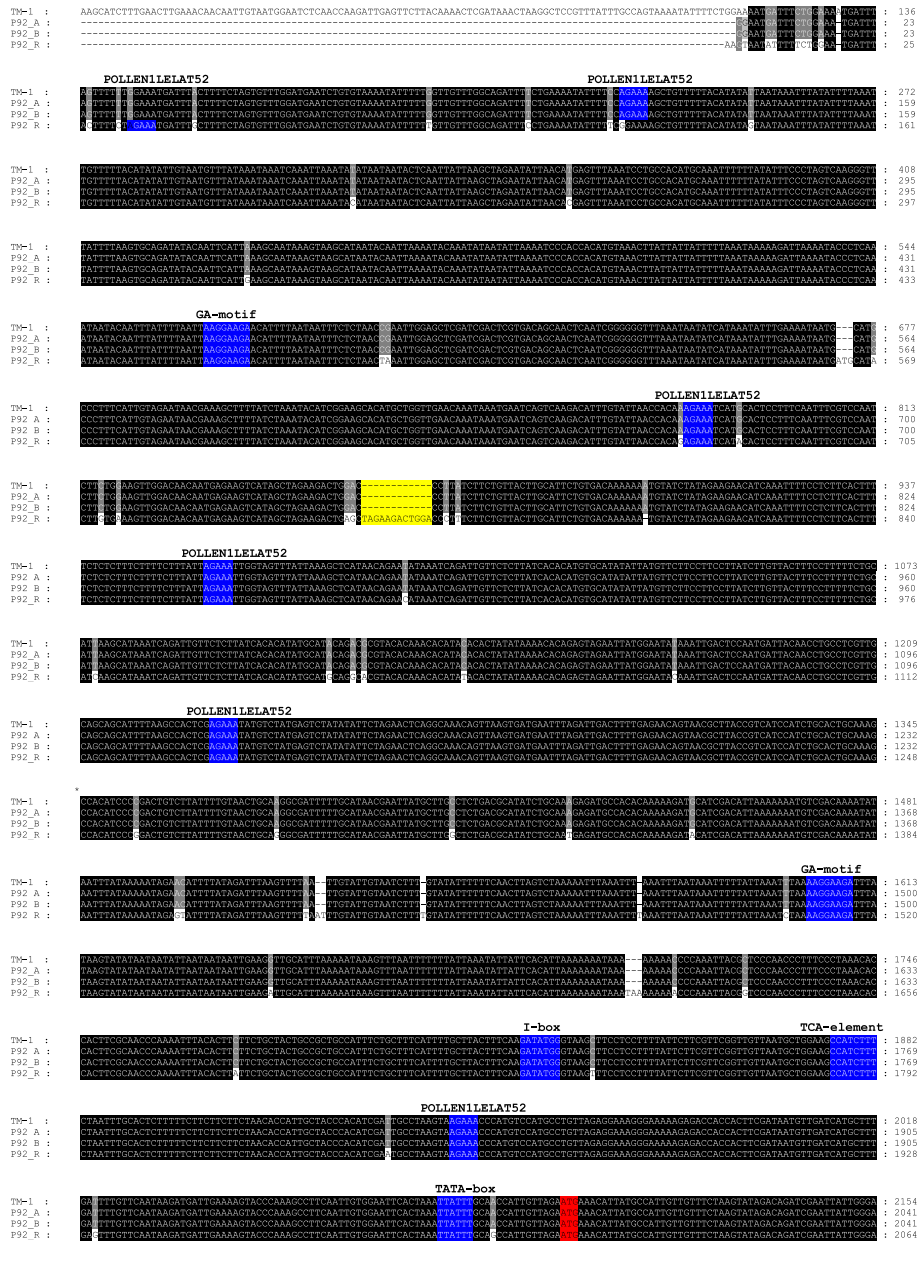

Fig. 3 Multiple alignments for promoter sequence of *Gh_D05G3392* gene in CMS-D2 line A, maintainer line B, and restoration line R. The red color represent the translation start site. The yellow color represent the insertion in the restorer R line. The blue color represent specific *cis*-regulatory elements in the promoter region

Barr and Fishman 2010; Fujii et al. 2011). *RFL* genes in the same genomic region are most likely active restorer genes, with several *PPR-Rf* genes presenting within the RFL-rich region, such as the rice *Rf1* and *Rf4* genes in the RFL-rich region of rice chromosome 10 (Wang et al. 2006; Fujii et al. 2011; Luo et al. 2013; Huang et al. 2016; Sykes et al. 2017). Additionally, the *Rf6* gene in rice was mapped to a 200-kb region on chromosome 8 that contains three *RFL* genes. Of these, *Os08g01870* was located within 15 kb of the marker and cosegregated with the *Rf* gene (Hu et al. 2012; Huang et al. 2016). The only identified *PPR-Rf* gene in sorghum is, however, located outside the RFL-rich regions on chromosome 8. This gene most likely encodes a PPR protein belonging to the PLS subfamily that is involved in RNA editing events, indicating that the mechanism of fertility restoration in

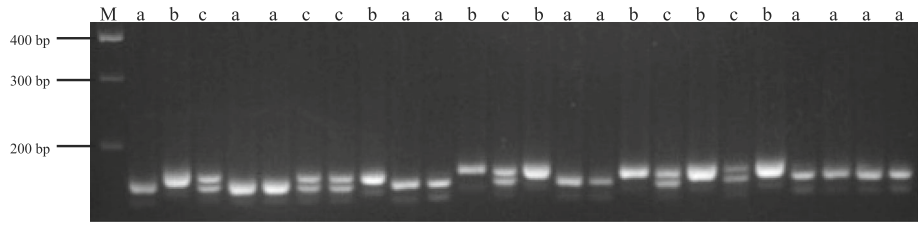

Fig. 4 BC_5F_2 plants were screened with InDel-R. M: DNA marker III, a plants lacking the restorer gene *Rf1*, b *Rf1* homozygous plants *Rf1*, c *Rf1* heterozygous plants

sorghum may be unique (Klein et al. 2005; Schmitzlinneweber and Small 2008; Dahan and Mireau 2013). This allowed us to further refine the candidate *Rf* genes in cotton by identifying the RFL-rich region common to other species. Previous studies indicated that *Rf1* and *Rf2* in cotton functioned sporophytically and gametophytically, respectively, and that the two *Rf* genes are not allelic but are tightly linked in 0.93 cM (Wang et al. 2007; Wang et al. 2009; Wu et al. 2011). Furthermore, the *Rf1* gene is located on chromosome Gh_D05 and genetic mapping has indicated that the nearest SSR marker to *Rf1* was BNL3535 (within 0.049 cM) and NAU3652 on the other side (within 0.078 cM) (Wang et al. 2007; Wu et al. 2014). In this study, four RFL-rich regions were identified in four chromosomes with six *RFL* genes found to cluster in the Gh_D05 chromosome near the *Rf* region. Contrary to our expectations, six *RFL* genes were not targeted to the mitochondria based on the TargetP software prediction. This may be because some *RFL* genes were overlooked because of assembly errors and gaps in the draft genome or because of repetitive features in the RFL-rich genomic regions. For example, most of the InDels were distributed near the region of the *Rf1* gene on chromosome Gh_D05 in cotton (Wu et al. 2017). In barley, an *RFL* gene was identified on an unordered contig from the chromosome 6HS containing a recently mapped *Rf* locus that could not be associated with an RFL cluster (Tsai et al. 2010; Ui et al. 2014).

Furthermore, a *Rf1*-specific CAPS marker was developed based on a SNP occurring within a PPR gene and an InDel-1891 marker was developed for fine mapping of the *Rf1* gene (Wu et al. 2014; Wu et al. 2017). The application of these markers could ensure the purity of restorer lines in cotton. In this study, three genes (*Gh_D05G3356*, *Gh_D05G3389*, and *Gh_D05G3392*) were up regulated in the R line as compared with the A and B lines. In total, 37 SNP loci in these three genes were identified between the R line and the A and B lines. Furthermore, a 12 nt insertion "TAGAAGACT GGA" was identified in the promoter region of *Gh_D05G3392* in the restorer R line as comparing with the A and B lines. An InDel-R marker was then developed for this insertion sequence that could be used to distinguish the restorer line carrying *Rf1* from other genotypes without the *Rf1* allele. The results implied that these SNPs and InDels might be used for fine mapping of the *Rf1* gene in cotton.

Conclusion

In our study, we tried to identify candidate *Rf-PPR* genes for CMS in cotton via genome-wide identification and analysis of *RFL* genes in *G. hirsutum*, *G. arboreum*, and *G. raimondii*. Furthermore, four RFL-rich regions were identified. Within one of these regions on Gh_D05, expression of three *RFL* genes was up-regulated in the R line as comparing with the A and B lines. Sequence variation analyses indicated that several SNPs and InDels exist in the R line as comparing with the non-restoring genome A and B lines, providing excellent sites for marker development and further mapping approaches. An InDel-R marker was then developed that could be used to distinguish the restorer line carrying *Rf1* from other genotypes without the *Rf1* allele. These results will not only be useful for guiding future identification and cloning of *Rf* genes responsible for CMS but will also be useful in heterosis in cotton.

Additional files

Additional file 1: Table S1. Information of primers for qRT-PCR and promoter analysis in A, B, and R lines. (XLSX 8 kb)

Additional file 2: Table S2. Expression level of *RFL* genes in *G. hirsutum* in different tissues and A, B, and R lines. (XLSX 11 kb)

Additional file 3: Table S3. SNP information of *RFL* genes in *G. hirsutum* in A, B, and R lines. (XLSX 16032 kb)

Abbreviations

A: CMS line; B: Maintainer line; CMS: Cytoplasmic male sterility; GRP: Glycine-rich domain protein; L motif: Long PPR motif; P motif: Common PPR motif; PCR: Polymerase chain reaction; PPR: Pentatricopeptide repeats; R: Restorer-of-fertility line; *Rf* gene: Restorer-of-fertility gene; S motif: Short PPR motif

Acknowledgements

The authors are grateful for Professor Liu F providing the materials of *G. harknessii*. The authors are also grateful for Doctor Liu GY and Zhang M, Li X, Feng JJ and the whole group of Professor Yu JW for analyzing the RNA-seq data, figures and helpful comments on the manuscript.

Funding
This research was financed by National Key Research and Development Program of China (2016YFD0101400) and Foundation of State Key Laboratory of Cotton Biology (CB2018C06).

Availability of data and materials
The raw sequence data of transcriptome in this study could be found in the National Center for Biotechnology Information (NCBI) under accession number SRX3421007.

Authors' contributions
Xing CZ, Wu JY conceived and designed the research. Zhang BB, Zhang XX performed the experiments. Guo LP, Qi TX and Wang HL prepared the materials. Tang HN, Qiao XQ and Shahzad K helped field investigation. Zhang BB wrote the paper. Xing CZ and Wu JY revised the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Received: 10 July 2018 Accepted: 12 October 2018
Published online: 29 October 2018

References

Andrés C, Lurin C, Small ID. The multifarious roles of PPR proteins in plant mitochondrial gene expression. Physiol Plant. 2007;129(1):14–22. https://doi.org/10.1111/j.1399-3054.2006.00766.x.

Barr CM, Fishman L. The nuclear component of a cytonuclear hybrid incompatibility in mimulus maps to a cluster of pentatricopeptide repeat genes. Genetics. 2010;184(2):455–65. https://doi.org/10.1534/genetics.109.108175.

Bentolila S, Alfonso AA, Hanson MR. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci. 2002;99(16):10887–92. https://doi.org/10.1073/pnas.102301599.

Brown GG, Formanová N, Jin H, et al. The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. Plant J. 2003;35(2):262–72. https://doi.org/10.1046/j.1365-313X.2003.01799.x.

Chase CD. Cytoplasmic male sterility: a window to the world of plant mitochondrial–nuclear interactions. Trends Genet. 2007;23(2):81. https://doi.org/10.1016/j.tig.2006.12.004.

Cui X, Wise RP, Schnable PS. The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. Science. 1996;272(5266):1334–6. https://doi.org/10.1126/science.272.5266.1334.

Dahan J, Mireau H. The *Rf* and *Rf-like* PPR in higher plants, a fast-evolving subclass of PPR genes. RNA Biol. 2013;10(9):1469–76. https://doi.org/10.4161/rna.25568.

Desloire S, Gherbi H, Laloui W, et al. Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family. EMBO J. 2003;4(6):588–94. https://doi.org/10.1038/sj.embor.embor848.

Filichkin SA, Nonogaki H. A novel endo-β-mannanase gene in tomato LeMAN5 is associated with anther and pollen development. Plant Physiol. 2004;134(3):1080–7. https://doi.org/10.1104/pp.103.035998.

Fujii S, Bond CS, Small ID. Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. Proc Natl Acad Sci. 2011;108(4):1723–8. https://doi.org/10.1073/pnas.1007667108.

Fujii S, Kazama T, Ito Y, et al. A candidate factor that interacts with RF2, a restorer of fertility of Lead rice-type cytoplasmic male sterility in rice. Rice. 2014;7(1):21. https://doi.org/10.1186/s12284-014-0021-6.

Geddy R, Brown GG. Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. BMC Genomics. 2007;8(1):130. https://doi.org/10.1186/1471-2164-8-130.

Hanson MR, Bentolila S. Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell. 2004;16(Suppl:S1):54–69. https://doi.org/10.1105/tpc.015966.

Hu J, Wang K, Huang W, et al. The rice pentatricopeptide repeat protein RF5 restores fertility in Hong-Lian cytoplasmic male-sterile lines via a complex with the glycine-rich protein GRP162. Plant Cell. 2012;24(1):109–22. https://doi.org/10.1105/tpc.111.093211.

Huang X, Yang S, Gong J, et al. Genomic architecture of heterosis for yield traits in rice. Nature. 2016;537(7622):629–33. https://doi.org/10.1038/nature19760.

Igarashi K, Kazama T, Toriyama K. A gene encoding pentatricopeptide repeat protein partially restores fertility in RT98-type cytoplasmic male sterile rice. Plant Cell Physiol. 2016;57(10):2187–93. https://doi.org/10.1093/pcp/pcw135.

Itabashi E, Iwata N, Fujii S, et al. The fertility restorer gene, *Rf2*, for Lead rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein. Plant J. 2010;65(3):359–67. https://doi.org/10.1111/j.1365-313X.2010.04427.x.

Joanna M, Stone JD, Ian S. Evolutionary plasticity of restorer-of-fertility-like proteins in rice. Sci Rep. 2016;6:35152. https://doi.org/10.1038/srep35152.

Kazama T, Nakamura T, Watanabe M, et al. Suppression mechanism of mitochondrial ORF79 accumulation by Rf1 protein in BT-type cytoplasmic male sterile rice. Plant J. 2008;55(4):619–28. https://doi.org/10.1111/j.1365-313X.2008.03529.x.

Klein RR, Klein PE, Mullet JE, et al. Fertility restorer locus *Rf1* of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12. Theor Appl Genet. 2005;111(6):994–1012. https://doi.org/10.1007/s00122-005-2011-y.

Liu F, Schnable PS. Functional specialization of maize mitochondrial aldehyde dehydrogenases. Plant Physiol. 2002;130(4):1657–74. https://doi.org/10.1104/pp.012336.

Luo D, Xu H, Liu Z, et al. A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. Nat Genet. 2013;45(5):573. https://doi.org/10.1038/ng.2570.

Meyer J, Pei D, Wise RP. *Rf*8-mediated T-*urf*13 transcript accumulation coincides with a pentatricopeptide repeat cluster on maize chromosome 2L. Plant Genome. 2011;4(3):283–99. https://doi.org/10.3835/plantgenome2011.05.0017.

Meyer VG. Male sterility from *Gossypium harknessii*. J Hered. 1975;62(1). https://doi.org/10.1093/oxfordjournals.jhered.a108566.

Paterson AH, Brubaker CL, Wendel JF. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Report. 1993;11(2):122–7. https://doi.org/10.1007/BF02670470.

Pelletier G, Budar F. The molecular biology of cytoplasmically inherited male sterility and prospects for its engineering. Curr Opin Biotechnol. 2006;18(2):121–5. https://doi.org/10.1016/j.copbio.2006.12.002.

Prasad K, Kushalappa K, Vijayraghavan U. Mechanism underlying regulated expression of *RFL*, a conserved transcription factor, in the developing rice inflorescence. Mech Dev. 2003;120(4):491–502. https://doi.org/10.1016/S0925-4773(02)00457-4.

Prikryl J, Rojas M, Schuster G, Barkan A. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. Proc Natl Acad Sci. 2010;108(1):415–20. https://doi.org/10.1073/pnas.1012076108.

Schmitzlinneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 2008;13(12):663–70. https://doi.org/10.1016/j.tplants.2008.10.001.

Schnable PS, Wise RP. The molecular basis of cytoplasmic male sterility and fertility restoration. Trends Plant Sci. 1998;3(5):175–80. https://doi.org/10.1016/S1360-1385(98)01235-7.

Suzuki H, Rodriguez-Uribe L, Xu J, Zhang J. Transcriptome analysis of cytoplasmic male sterility and restoration in CMS-D8 cotton. Plant Cell Rep. 2013;32(10):1531–42. https://doi.org/10.1007/s00299-013-1465-7.

Sykes T, Yates S, Nagy I, et al. In-silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.). Genome Biol Evol. 2017;9(2):351–62. https://doi.org/10.1093/gbe/evw047.

Tan XL, Tan YL, Zhao YH, et al. Identification of the *Rf* gene conferring fertility restoration of the CMS Dian-type 1 in rice by using simple sequence repeat markers and advanced inbred lines of restorer and maintainer. Plant Breed. 2004;123(4):338–41. https://doi.org/10.1111/j.1439-0523.2004.01004.x.

Tan YP, Li SQ, Wang L, et al. Genetic analysis of fertility-restorer genes in rice. Biol Plant. 2008;52(3):469–74. https://doi.org/10.1007/s10535-008-0092-6.

Tomohiko K, Kinya T. A fertility restorer gene, *Rf4*, widely used for hybrid rice breeding encodes a pentatricopeptide repeat protein. Rice. 2014;7(1):28. https://doi.org/10.1186/s12284-014-0028-z.

Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 2010;11(4):1–9. https://doi.org/10.1186/gb-2010-11-4-r41.

Ui H, Sameri M, Pourkheirandish M, et al. High-resolution genetic mapping and physical map construction for the fertility restorer *Rfm1* locus in barley. Theor Appl Genet. 2014;128(2):283–90. https://doi.org/10.1007/s00122-014-2428-2.

Uyttewaal M, Arnal N, Quadrado M, et al. Characterization of *Raphanus sativus* pentatricopeptide repeat proteins encoded by the fertility restorer locus for Ogura cytoplasmic male sterility. Plant Cell. 2008;20(12):3331–45. https://doi.org/10.1105/tpc.107.057208.

Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered. 2002;93(1):77–8. https://doi.org/10.1093/jhered/93.1.77.

Wang F, Stewart JM, Zhang JF. Molecular markers linked to the *Rf2* fertility restorer gene in cotton. Genome. 2007;50(9):818–24. https://doi.org/10.1139/G07-061.

Wang F, Yue B, Hu JG, et al. A target region amplified polymorphism marker for fertility restorer gene *Rf1* and chromosomal localization of *Rf1* and *Rf2* in cotton. Crop Sci. 2009;49(5):1602–8. https://doi.org/10.2135/cropsci2008.09.0531.

Wang Z, Zou Y, Li X, et al. Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. Plant Cell. 2006;18(3):676–87. https://doi.org/10.1105/tpc.105.038240.

Wang ZW, De WC, Wang C, et al. Heterozygous alleles restore male fertility to cytoplasmic male-sterile radish (*Raphanus sativus* L.): a case of overdominance. J Exp Bot. 2013;64(7):2041–8. https://doi.org/10.1093/jxb/ert065.

Wu JY, Cao XX, Guo LP, et al. Development of a candidate gene marker for Rf_1 based on a *PPR* gene in cytoplasmic male sterile CMS-D2 upland cotton. Mol Breed. 2014;34(1):231–40. https://doi.org/10.1007/s11032-014-0032-4.

Wu JY, Gong YC, Cui MH, et al. Molecular characterization of cytoplasmic male sterility conditioned by *Gossypium harknessii* cytoplasm (CMS-D2) in upland cotton. Euphytica. 2011;181(1):17–29. https://doi.org/10.1007/s10681-011-0357-6.

Wu JY, Zhang M, Zhang XX, et al. Development of InDel markers for the restorer gene *Rf1* and assessment of their utility for marker-assisted selection in cotton. Euphytica. 2017;213(11):251. https://doi.org/10.1007/s10681-017-2043-9.

Xu XB, Liu ZX, Zhang DF, et al. Isolation and analysis of rice Rf1-orthologus PPR genes co-segregating with *Rf3* in maize. Plant Mol Biol Report. 2009;27(4):511. https://doi.org/10.1007/s11105-009-0105-4.

Yin JM, Guo WZ, Yang LM. Physical mapping of the Rf_1 fertility-restoring gene to a 100 kb region in cotton. Theor Appl Genet. 2006;112(7):1318–25. https://doi.org/10.1007/s00122-006-0234-1.

Yu SX, Fan SL, Wang HT, et al. Progresses in research on cotton high yield breeding in China. Sci Agric Sin. 2016;49(18):3465–76. https://doi.org/10.3864/j.issn.0578-1752.2016.18.001.

Zhang B, Liu G, Li X, et al. A genome-wide identification and analysis of the DYW-deaminase genes in the pentatricopeptide repeat gene family in cotton (*Gossypium* spp.). PLoS One. 2017;12(3):e0174201. https://doi.org/10.1371/journal.pone.0174201.

Zhang JF, Stewart JM. CMS-D8 restoration in cotton is conditioned by one dominant gene. Crop Sci. 2001a;41(2):283–8. https://doi.org/10.2135/cropsci2001.412283x.

Zhang JF, Stewart JM. Inheritance and genetic relationships of the D8 and D2-2 restorer genes for cotton cytoplasmic male sterility. Crop Sci. 2001b;41(2):289–94. https://doi.org/10.2135/cropsci2001.412289x.

Zhang JF, Turley RB, Stewart JM. Comparative analysis of gene expression between CMS-D8 restored plants and normal non-restoring fertile plants in cotton by differential display. Plant Cell Rep. 2007;27(3):553–61. https://doi.org/10.1007/s00299-007-0492-7.