Theor Appl Genet (2005) 111: 994–1012
DOI 10.1007/s00122-005-2011-y

ORIGINAL PAPER

R. R. Klein · P. E. Klein · J. E. Mullet · P. Minx
W. L. Rooney · K. F. Schertz

Fertility restorer locus *Rf1*of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12

Received: 8 December 2004 / Accepted: 17 March 2005 / Published online: 3 August 2005

Abstract With an aim to clone the sorghum fertility restorer gene *Rf1*, a high-resolution genetic and physical map of the locus was constructed. The *Rf1*locus was resolved to a 32-kb region spanning four open reading frames: a plasma membrane Ca^{2+}-ATPase, a cyclin D-1, an unknown protein, and a pentatricopeptide repeat (*PPR13*) gene family member. An ~19-kb region spanning the cyclin D-1 and unknown protein genes was completely conserved between sterile and fertile plants as was the sequence spanning the coding region of the Ca^{2+}-ATPase. In contrast, 19 sequence polymorphisms were located in an ~7-kb region spanning *PPR13*, and all markers cosegregated with the fertility restoration phenotype. *PPR13* was predicted to encode a mitochondrial-targeted protein containing a single exon with 14 PPR repeats, and the protein is classified as an E-type PPR subfamily member. To permit sequence-based comparison of the sorghum and rice genomes in the *Rf1* region, 0.53 Mb of sorghum chromosome 8 was sequenced and compared to the colinear region of rice chromosome 12. Genome comparison revealed a mosaic pattern of colinearity with an ~275-kb gene-poor region with little gene conservation and an adjacent, ~245-kb gene-rice region that is more highly conserved between rice and sorghum. Despite being located in a region of high gene conservation, sorghum *PPR13* was not located in a colinear position on rice chromosome 12. The present results suggest that sorghum *PPR13* represents a potential candidate for the sorghum *Rf1* gene, and its presence in the sorghum genome indicates a single gene transposition event subsequent to the divergence of rice and sorghum ancestors.

Keywords Sorghum · Fertility restorer · Map-based cloning · Pentatricopeptide repeat protein · Microsynteny

Communicated by E. Guiderdoni

R. R. Klein (✉) · K. F. Schertz
Southern Plains Agricultural Research Center, USDA–ARS,
College Station, TX 77845, USA
E-mail: rklein@tamu.edu
Tel.: +1-979-7774470
Fax: +1-979-2609333

P. E. Klein
Department of Horticulture and Institute for Plant Genomics and
Biotechnology, Texas A&M University, College Station,
TX 77843-2123, USA

J. E. Mullet
Department of Biochemistry and Biophysics and Institute for Plant
Genomics and Biotechnology, Texas A&M University,
College Station, TX 77843-2123, USA

P. Minx
Washington University Genome Sequencing Center, Washington
University School of Medicine, St. Louis, MO 63108, USA

W. L. Rooney
Department of Soil and Crop Sciences, Texas A&M University,
College Station, TX 77843-2474, USA

Introduction

Cytoplasmic male sterility (CMS) is a maternally inherited trait in which pollen development or normal anther dehiscence is impaired, but female fertility is normal (Pring et al. 1995). CMS is widely distributed in approximately 150 plant species (Laser and Lersten 1972) and is often associated with the expression of novel mitochondrial open reading frames (for review, see Hanson and Bentolila 2004). In sorghum (*Sorghum bicolor*) and other crop species such as maize (*Zea mays*) and rice (*Oryza sativa*), CMS is observed when a cytoplasm is transferred into a different nuclear background. In many CMS systems, male fertility can be restored by a series of fertility restorer (*Rf*) genes encoded in the nucleus. Thus, CMS/*Rf* systems appear to result from specific nuclear–mitochondrial interactions, and *Rf* genes block or compensate for specific mitochondrial dysfunctions that are phenotypically expressed during pollen development. The CMS/*Rf* systems are of great commercial importance for hybrid seed production, because they eliminate the need for hand emasculation. Apart from its commercial importance, CMS/*Rf* systems offer an excellent opportunity to investigate the intricate interaction of mitochondrial-encoded gene products and nuclear-encoded proteins targeted for this subcellular organelle.

There are a number of striking similarities in the proposed mechanisms of fertility restoration. In many instances, CMS restorer genes seem to function by preventing the expression of mitochondrial CMS-associated open reading frames (ORFs) or proteins. For example, in petunia, a CMS-associated abnormal protein termed PCF is greatly reduced by the action of the *Rf* gene (for review, see Hanson and Bentolila 2004). Similarly, fertility restoration has been shown to correlate with the transcript profile or the protein accumulation of the CMS-associated ORFs in *Brassica* (Brown 1999), Ogura cytoplasm in radish (Bonhomme et al. 1992; Gray et al. 1998), BT cytoplasm (or *ms-bo* cytoplasm) of rice (Iwabuchi et al. 1993), and A3-CMS cytoplasm of sorghum (Tang et al. 1996, 1999).

In order to elucidate the molecular mechanisms of fertility restoration, research groups have attempted to clone *Rf* genes from various plant species. Over the past several years, *Rf* genes have been cloned from petunia (Bentolila et al. 2002), radish (Brown et al. 2003; Desloire et al. 2003; Koizuka et al. 2003), and rice (Kazama and Toriyama 2003; Komori et al. 2004). Each of these *Rf* genes has been shown to encode a protein with a mitochondrial transit peptide and numerous pentatricopeptide repeat (PPR) motifs. Small and Peeters (2000) coined the term PPR for the major motif based on its similarity to the better-known tetratricopeptide repeat, or TPR motif. The PPR proteins contain 2 to at least 26, tandem copies of PPR motifs containing either 31, 35 or 36 amino acids, and represent a protein family of more than 450 members in*Arabidopsis* (Lurin et al. 2004). A recent study employing bioinformatics and functional genomics (Lurin et al. 2004) concluded that the family of PPR proteins play constitutive, often essential roles in mitochondria and chloroplasts, and they likely represent RNA-binding proteins involved in posttranscriptional processing. To date, all the cloned *Rf* genes, except *Rf2* of maize, are members of the PPR protein family. Maize *Rf2* was the first fertility restoration gene cloned, and it was identified as encoding for a mitochondrial aldehyde dehydrogenase (Cui et al. 1996; Liu et al. 2001).

The infrastructure for cereal genomics has progressed from low-density genetic maps to the nearly completed sequence of the 430-Mbp rice genome. In sorghum, genomic resources enabling analysis of sorghum trait loci, and gene expression have been developed. High-density sorghum genetic maps have been constructed (Bowers et al. 2003a; Menz et al. 2002) and integrated with an emerging BAC-based sorghum physical map (Draye et al. 2001; Klein et al. 2000, 2003). The DNA markers and sequence-based information are being used to align the sorghum genome map with the rice genome sequence and with maps of other grasses (Bowers et al. 2003b; Klein et al. 2003; Paterson et al. 2004b). In addition, a sorghum EST project has collected over 200,000 sequences from cDNA libraries derived from diverse tissues and treatments (http://www.fungen.org).

An obvious use of the genome resources of sorghum, and the genomic microcolinearity that exist between sorghum and rice, is to accelerate the positional cloning of genes of agronomic importance by chromosome walking within small colinear regions of the sorghum and rice genomes. The effectiveness of this approach relies not on overall genome colinearity but rather on local microcolinearity at sites where the genes targeted for positional cloning are located. Microcolinearity is apparent in sequence comparisons of the sorghum and rice genomes, but there are many small exceptions (Bennetzen 2000, 2002; Bennetzen and Ma 2003; Bennetzen and Ramakrishna 2002; Chen et al. 1998; Ilic et al. 2003; Klein et al. 2003; Li and Gill 2002; Ramakrishna et al. 2002; Song et al. 2002; Tikhonov et al. 1999). Comparative sequence analysis in the grasses indicates a relatively high frequency of small genetic rearrangements that include deletions, gene duplication, inversions, and translocations (for review, see Bennetzen 2000; Bennetzen and Ma 2003). Nevertheless, comparative sequence analysis between species such as sorghum and rice can provide useful tools for physical map closure, gene annotation, revealing details of local sequence evolution, and identifying conserved non-coding sequences that are critical for gene regulation (Hardison 2000; Kaplinsky et al. 2002).

In previous studies, we utilized the sorghum genomic resources to map the *Rf1* gene to the long arm of sorghum chromosome 8 (Kim et al. 2005a, b; Klein et al. 2001). Using a combination of AFLPs and SSRs, the *Rf1* locus was refined to a ~2-cM region of the sorghum genome. In the present study, we utilized a positional cloning approach in conjunction with microcolinear analysis of sorghum and rice to genetically and physically delimit the *Rf1*locus of sorghum. In conjunction with the construction of a high-resolution map of the *Rf1* trait locus, we also sequenced a 0.53-Mb region of the sorghum genome flanking the *Rf1* locus. This contiguous sorghum genome sequence permitted an examination of the syntentic relationship of the genomes of rice and sorghum. We present evidence of the potential identity of the *Rf1* gene, and we examine the genome organization of rice and sorghum in one of the longest contiguous stretches of sorghum genomic DNA sequenced to date.

Materials and methods

Plant materials and phenotypic classification

An F_2 population of 1,858 plants derived from the cross of ATx623 (*rf1, rf1*) and RTx432 (*Rf1, Rf1*) was created for segregation analysis. Inbred ATx623 has an A1-type of male sterile cytoplasm, whereas inbred RTx432 is a common R (restorer) line used in sorghum hybrids.

ATx623 and RTx432 have been noted to differ by a single major gene (designated *Rf1*) for fertility restoration although genes with minor effects appear to be involved (R.R. Klein, unpublished observation). Therefore, to map the major gene controlling fertility restoration in the A1 cytoplasm, only the phenotypic extremes of the F_2 population were used in linkage analysis. Seven hundred sixty-two F_2 gametes (241 sterile F_2s, 0% seed set; 140 homozygous fertile F_2s, 100% seed set) were utilized for linkage analysis. Phenotypic classification (percentage seed set) was conducted as previously described (Klein et al. 2001). Seeds from each F_2 fertile plant were progeny tested either in the field or under glasshouse conditions to differentiate homozygous (*Rf1, Rf1*) and heterozygous (*Rf1, rf1*) fertile F_2s.

BAC libraries and BAC contig development

Sorghum BAC libraries used for physical map construction were previously described (Klein et al. 2000, 2003). Briefly, three BAC libraries were constructed from inbred BTx623 (*Hin*dIII library, BAC prefix SB_BBe; *Bam*HI libraries, BAC prefixes SB_BBd, and SB_BBf) and a fourth from converted sorghum line IS3620C (*Hin*dIII, BAC prefix SB_IBa). The BAC contigs were assembled by a modified version of high-information content fingerprinting [(HICF) Luo et al. 2003; Klein et al. 2003] and DNA marker content mapping as detailed (Klein et al. 2000, 2003). For BAC-end sequencing, column-purified template was prepared with the Plasmid Midi kit (Qiagen,Valencia, Calif., USA) with the manufacturer's suggested procedural modifications for BAC DNA. The BAC ends were sequenced with BigDye Terminator mix, version 3.1, cycle sequencing kit (1:4 dilution) using appropriate forward and reverse primers. All other sequencing parameters were as previously detailed (Klein et al. 2003).

High-resolution mapping, chromosome walking, and molecular marker development

The closest marker at the inception of this project was AFLP marker *Xtxa2582*, mapping at a distance of ~0.92 cM from *Rf1*. To increase map resolution, new codominant genetic markers were discovered by searching for SSRs, indels, and SNPs by BAC sequence analysis. As the sorghum genome map has been constructed from two sorghum genotypes (BTx623, IS3620C); sequence analysis of overlapping IS3620C and BTx623 BAC clones revealed polymorphisms between the two genotypes. As inbred BTx623 has the same nuclear genome as ATx623 (sterile parent of *Rf1* mapping population), a significant percentage of polymorphisms (> 60%) detected by sequencing overlapping BTx623 and IS3620C BAC clones were also polymorphic in the parental lines of the *Rf1* mapping population. Random BAC sequences were obtained using either BAC DNA digested with two, six-base recognition restriction enzymes as previously detailed (Klein et al. 2003), or by sequencing subclones from small insert libraries prepared from sheared BAC DNA (see below).

Physical map closure (chromosome walking) in the region flanking the *Rf1* locus was accomplished by extending BAC contigs using PCR probes designed against BAC-end sequences, or by discovering new genetic markers in BACs that delimit a physical gap. The BAC-end PCR probes were subsequently used to screen BAC pools as previously detailed (Klein et al. 2000). The BAC clones identified through BAC-pool screenings were subjected to HICF to permit integration into existing BAC contigs. This process was repeated until a contiguous array of BAC clones was assembled across the *Rf1* locus.

High-resolution genetic mapping of the *Rf1* locus was performed with F_2 individuals that possessed a recombination event between two *Rf1* locus-delimiting markers, *Xtxp18* and *Xtxs560* (see Table 1; Klein et al. 2001). A total of 11 recombinant F_2 plants from 762 gametes were identified and used for high-resolution mapping. To assure the accuracy of the genetic map, the phenotypes of fertile recombinant F_2 plants were reexamined by repeated phenotyping of F_3 progeny in a glasshouse.

BAC subcloning and sequencing

DNA from sorghum BACs SB_IBa88M04, SB_IBa60H10, SB_BBf14I11, and SB_IBa75D09 was isolated from exponentially growing *Escherichia coli* cells using the Qiagen Large-Construct Preparation kit. Prior to column purification, DNA was treated with Plasmid Safe ATP-dependent DNase (Epicentre, 50 U, 16 h at 37°C) to hydrolyze linear dsDNA that included *E. coli* DNA. Purified BAC DNA was sheared physically to a size of 2–6 kb with a Hydroshear instrument (Genomic Instrumentation Services). Two different settings (speed codes 7 and 10) were routinely used. The ends of the sheared DNA were repaired with the DNA Termination End Repair kit (Lucigen, Middleton, Wis., USA), and the products were size separated on 1% agarose gels. The DNA fractions from 1.8 kb to 5 kb were excised from the gels and DNA recovered using the GENECLEAN Turbo Nucleic Acid Purification kit (Q-BIOgene, Irvine, Calif, USA). The DNA was ligated into a pSMARTHC vector using the CloneSmart Blunt Cloning kit (Lucigen). Ligated DNA was transformed with One Shot Top 10 chemically competent cells (Invitrogen, Valencia, Calif., USA) as per the manufacturer's instructions. Transformants were selected on Luria-Bertani broth agar with appropriate selective agents. Ninety-six clones were checked by sequence analysis for insert size and for *E. coli*contamination prior to large-scale shotgun sequencing.

Plasmid DNA minipreparations for large-scale shotgun sequencing were conducted as previously described

Table 1 Sorghum genetic markers discovered and used for linkage analysis

Sorghum Microsatellite (*Xtxp*), AFLP (*Xtxa*), STS (*Xtxs*)

		Map Position				
Marker	Marker name	Genetic (cM)	Physical (bp)	Motif and Length	Forward Primer (5'-3')	Reverse Primer (5'-3')
1	*Xtxp18*	-1.18	127,586	$(AG)_{20}$	ACTGTCTAGAACAAGCTGCG	TTGCTCTAGCTAGGCATTTC
2	*Xtxa2582*	-0.91	-[b]	-	*Eco*RI+GGA	*Mse*I+CGT
3	*Xtxa3431*	-0.91	182,446	-	*Eco*RI+CTG	*Mse*I+CCAAT
4	*Xtxp403*[a]	-0.52	274,706	$(TA)_{36}$	TACTGAACGATACTGGATTCTGT	TGATCTCTGTTCTTGTCGTTAAG
6	*Xtxa7179*[a]	-0.26	291,453	-	*Eco*RI+AGA	*Mse*I+CCCTAGGT
8	*Xtxp400*[a]	-0.13	338,533	$(GA)_8GT(GA)_4$	CTAAGAACCGACGCGTGTATAGT	GCATCTATCTTCACTCCGATTCT
9	*Xtxp408*	0	339,491	$(TTTCC)_2$	GCTGCCACCCACGATTCCGACT	GAGCCAAGGAGGGGTGTGCAGAC
28	*Xtxp407*	0	352,631	$(TA)_{13}GATATAGA(TA)_7$	TATGGCTCCAAATCAATATTACT	TTAGCTAGGCTACTAAAGTGATGA
29	*Xtxp406*	+0.13	371,387	$(AAAAC)_5$	GGCCTGAATCTCAGTGTTAAG	AGTTGCCTGCTTCGACACTT
32	*Xtxs560*	0.26	406,343	-	CTGCCTCTGCCGTGTTGGAT	GCTGCTCTTCTCTTGCCTTGCT
33	*Xtxp250*	+1.05	-	$(AAG)_{17}AAT(AAG)_4AAA(ACA)_9$	GCACATCCTCTAAAACTACTTAGT	GAACAGGACGATGTGATAGAT

Sorghum Indels (*Xtxi*), SNPs (*Xtxsn*)

		Map Position		
Marker	Marker name	Genetic (cM)	Physical (bp)	Altered Sequence[c]
5	*Xtxsn1*[a]	-0.39	281,851	GGCGATTA**TCT**CCGA **A**
7	*Xtxsn2*[a]	-0.26	330,868	GCTGGA**CTT**GGACGT **A**
10	*Xtxsn3*	0	340,138	TGGGT**GCC**GGATTC **A**
11	*Xtxsn*	0	345,326	GTCGT**CGT**CCTCCTCC **C**
12	*Xtxsn5*	0	345,781	CATTG**TGA**GTGCCCTT **A**
13	*Xtxsn6*	0	345,892	AACCCCA**ACA**CCACCACC **T**
14	*Xtxsn7*	0	346,131	AACAGA**GCG**GGGCCACTC **T**
15	*Xtxsn8*	0	346,320	TAGAGT**GTC**ATCTCCCTG **C**
16	*Xtxsn9*	0	346,590	ACACCA**TGT**CCAGCATTGG **A**

Table 1 (Contd.)

17	*Xtxsn10*	0	346 825	TGCTTTTCCAAATTTTA T
	Xtxsn11	0	347,643	CCCTCTCGGTACGCCA C
19	*Xtxi1*	0	349,763	CGAAGGAGCGCAGGG * * * AAA
20	*Xtxsn12*	0	349,997	CAAGATGAATCTTTT T
21	*Xtxi2*	0	350,070	TTTAAAAAAAA** CTTTTTG AA
22	*Xtxsn13*	0	350,258	ACTGTAGCATTTTCA T
23	*Xtxsn14*	0	350,607	GAATTTCAATAACCC C
24	*Xtxsn15*	0	350,815	GTGTCTGTAAGCCCAGTCTCAA C
25	*Xtxsn16*	0	350,993	CTTTCATGAAACAT T
26	*Xtxsn17*	0	352188	GCAAGGCCGATCCTA T
27	*Xtxsn18*	0	352,590	TTCTAATTTTTTTTT AA
30	*Xtxsn21*	+0.26	389,952	ACACATCCAAAAAAAA C
31	*Xtxi3*[a]	+0.26	405,906	A GGCCAGCCCTCCGTTCT TTTTTGATAAA**************** GA

[a] Not available

[b] Markers discovered by sequence analysis of overlapping BTx623 and IS3620C BAC clones

[c] Upper sequence is that of ATx623, whereas RTx432 is the lower sequence. Sequence polymorphisms depicted in *boldface*

(Klein et al. 2003). Sequencing reactions of both ends of each clone were carried out using the BigDye Terminator, version 3.1, Cycle Sequencing kit (Applied Biosystems, Foster City, Calif., USA). Sequencing products were analyzed on an ABI3700 capillary sequencer (Applied Biosystems). Base calling and quality assessment was conducted with the PHRED program (Ewing and Green 1998; Ewing et al. 1998). Sequence assembly utilized the PHRAP program (http://www.phrap.org), and assembled shotgun reads were viewed and edited with CONSED (Gordon et al. 1998). To close gaps, primer walking and sequencing of PCR products was conducted as detailed by Autofinisher. To resolve the sequence of regions that were difficult to read, the ABI Prism dGTP BigDye Terminator Cycle Sequencing kit was utilized as a 4:1 BigDye/dGTP mix. In addition, dimethyl sulfoxide [(DMSO) 5–10%] or betaine (10%) were added to both PCR and sequencing reactions to permit ‘‘read through’’ of difficult regions. To validate experimentally the assembly of shotgun reads by CONSED, BAC DNA was digested with a series of six-base recognition restriction enzymes, and the observed restriction pattern was compared to that predicted for the assembled shotgun reads. In each case, there was good agreement between the observed restriction pattern and that predicted in silico for each assembled BAC clone. All BAC clones have been advanced to phase III level sequence and deposited into GenBank. The GenBank accessions numbers are AY661656 (BAC SB_IBa88M04), AY661657 (BAC SB_IBa60H10), AY661658 (BAC SB_BBf14I11), and AY661659 (SB_IBa75D09).

The sequence of a 67-kb region flanking the *Rf1* locus in the genomes of RTx432 and ATx623 was obtained by directly sequencing PCR products. A series of sequential PCR products (that collectively span the *Rf1*locus) were amplified with*Pfu* Turbo DNA polymerase (50-µL reactions per Stratagene’s protocol guide), column purified to remove PCR primers (GENECLEAN Turbo Kit, Q-BIOgene), and sequenced from both ends with the appropriate PCR primers. All sequencing parameters and data analysis were as detailed above.

Sequence annotation and microcolinearity

Phase III level sequence of each BAC clone was initially submitted to the Rice Genome Automated Annotation System (http://ricegaas.dna.affrc.go.jp/), a system that integrates programs for prediction and analysis of protein-coding gene structure (Sakata et al. 2002). Subsequent to this initial examination, sorghum sequences were submitted to the prediction programs FGENESH (http://www.softberry.com/berry.phtml) with the monocot training set used for gene prediction, GENSCAN (http://genes.mit.edu/GENSCAN.html), and GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/). FGENESH was used to define the position of exons and introns in each predicted ORF (with the exception being gene*PPR13*, whose cDNA was sequenced). The criteria used to define a gene were (1) prediction as a gene from the three prediction programs listed above and (2) a sequence similarity to a putative, hypothetical, or known protein-coding sequence in the non-redundant database or in a protein database using BLASTX at an expected value greater than e-15. Alternatively, a predicted ORF that matched a monocot EST was defined as encoding for an unknown protein. Because EST libraries have been prepared from different species of sorghum (*S. bicolor*, *S. propinquum*), a homology cut-off of 97% sequence identity was required. If sorghum ESTs were hit at a homology ranging from 90% to 97%, the ORF was predicted to encode a gene if the ORF also identified other rice, maize, or sugarcane ESTs at greater than 90% homology over the same genomic region. These criteria were applied after eliminating ORFs homologous with repetitive DNA elements such as putative *gag-pol* gene, Kafirin cluster, and putative reverse transcriptase. After applying the above criteria, each predicted sorghum gene was examined for potential alignment to a specific rice chromosome. Predicted sorghum genes were matched with rice BAC sequences in the non-redundant or high-throughput genome sequence databases at an *e*-value of greater than e-12. Sequences from large protein families were examined closely to determine if orthologous rice and sorghum BAC sequences could be identified. If several rice BACs were identified with a given sorghum gene (all at a similar *e*-value), it was so noted in Table 3.

PipMaker was used to obtain a percent identity plot of the alignment of genomic regions of rice and sorghum (Schwartz et al. 2000). To prevent alignment of duplicated genes or exons, percent identity plots were generated using the chaining option and by searching a single DNA strand during sequence alignment. The results of this alignment were compared to the percent identity plots obtained with both of these options turned off. RepeatMasker (http://repeatmasker.org/) was used to identify repetitive elements using the Poaceae RepBase database (http://www.girinst.org). The results of this search were used to mask interspersed repeats in the 0.53 Mb of sorghum genome sequence during genome alignment with PipMaker. Annotation of sequences encoding repeat family members relied in part on RepeatMasker followed by BLAST analysis to classify the repeat family to which the element belonged. Long terminal repeats were identified by the AutoPredLTR function in RICEGAAS (http://ricegaas.dna.affrc.go.jp/). The miniature inverted-repeat transposable elements (MITEs) were identified using the FINDMITE program (Tu 2001) followed by confirmation by BLAST analysis against the The Institute for Genomic Research Gramineae Repeat Database (http://www.tigr.org/tdb/e2k1/plant.repeats/index.shtml). Repetitive elements that appear to be fragments of DNA transposable elements (e.g., highly fragmented colinearity to known transposons, degenerate terminal inverted repeats) were classified as transposon-like sequences.

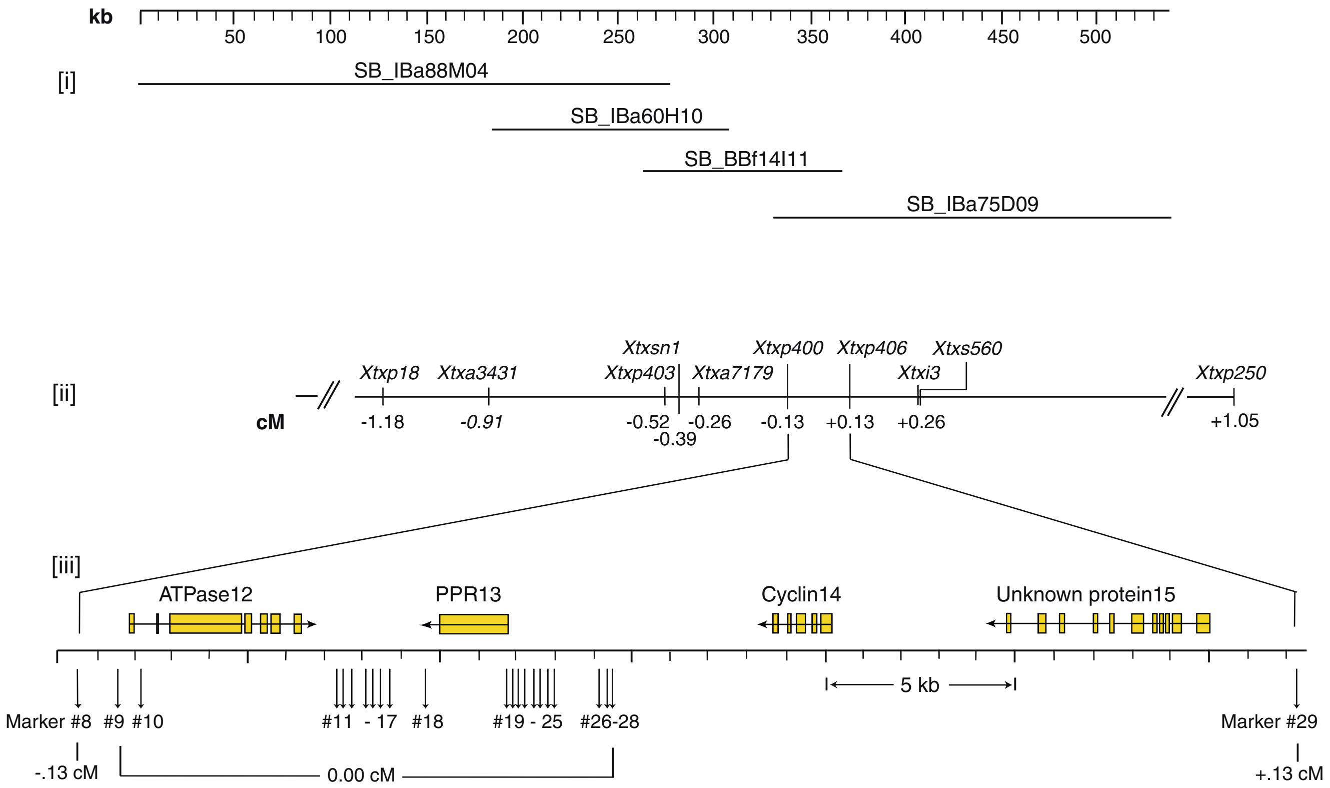


Fig. 1 A fine-scale, high-resolution genome map of the *Rf1* locus on sorghum linkage group 08. **i** Tiling path of BAC clones. **ii** Genetic map showing relative position of selected genetic markers in the sorghum pseudomolecule. Position of sequences corresponding to genetic markers was determined during BAC sequence analysis. **iii** Expanded view of the high-resolution map of the *Rf1* locus showing the position of candidate genes and genetic markers that delimit the locus. *Arrows* facing downward indicate the map position of markers used to refine the locus, and the numbers to which each *arrow points* correspond to marker number designations shown in Table 1

RNA extraction, cDNA synthesis, and 5′ and 3′ rapid amplification of cDNA ends

Total RNA was extracted from leaves of 6-day-old ATx623 or RTx432 sorghum seedlings. Tissue was pulverized in a mortar and pestle under liquid N_2, and RNA was extracted using the TRI REAGENT with the manufacturer's modifications to remove polysaccharide contamination (Molecular Research Center, Cincinnati, Ohio, USA). To remove DNA contamination, total RNA (25–75 µg) was treated with 2 U of TurboDNase for 30 min at 37°C (Ambion, Austin, Tex., USA). DNase-treated RNA was then purified using the MEGAclear Purification kit (Ambion). First-strand cDNA was synthesized with random hexamers using the TaqMan reverse transcription kit (Applied Biosystems). In each case, control reactions were run (plus RNA, minus reverse transcriptase) to reveal DNA contamination in RNA preparations. To sequence the cDNA of the pentatricopeptide repeat protein (*PPR13*), overlapping segments (700–1,500 bp in length) of the predicted ORF were PCR-amplified using PFU*Taq* polymerase (Stratagene, La Jolla, Calif., USA). Following PCR, the ends of the amplified fragments were filled by adding 1 U of *Taq* DNA polymerase (Promega, Madison, Wis., USA) and incubated at 72°C for 10 min. The PCR reactions were purified with the GENECLEAN Turbo Nucleic Acid Purification kit using the diluted salt wash solution to remove PCR primers. The PCR products were directly sequenced using the primers used in PCR amplification or were TA-cloned and sequenced as detailed above.

The 5′ and 3′ rapid amplifications of cDNA ends (RACE) were generated and sequenced to map the transcript ends of gene *PPR13*. Primers specific to gene *PPR13* were designed to permit PCR amplification of 3′ RACE products using the FirstChoice RLM-RACE kit (Ambion). Leaf total RNA from ATx623 and RTx432 was used as template for 3′ RACE amplification. For 5′ RACE, Invitrogen's 5′ RACE System was utilized. The presence of strong G/C-rich stretches in the 5′ end of the PPR cDNA required several modifications of first-strand cDNA synthesis and subsequent PCR amplification. DMSO (5%) was added to first strand synthesis, and the reaction temperature was elevated to 50°C. Poly(A) mRNA was used as template for first strand synthesis to increase the amount of target transcript and hence, increase the yield of synthesized *PPR13*cDNA. First-strand synthesis was primed with one of two PPR gene-specific primers (GSPs) and amplification of the target cDNA involved two sequential nested PCR reactions using additional GSPs (primer sequences available upon request). The PCR products from 3′ and 5′ RACE were directly sequenced as detailed above.

Alignment of cDNA sequences of *PPR13* from RTx432 and ATx623 was done with ClustalW (http://workbench.sdsc.edu/). Predicted proteins were analyzed for potential mitochondrial targeting pre-sequences with the programs Predotar, version 1.03 (http://genoplante-info.infobiogen.fr/predotar/predotar.html.), MitoProt II, version 1.0a4 (http://ihg.gsf.de/ihg/mitoprot.html), TargetP (http://www.cbs.dtu.dk/services/TargetP/), and PSORT (http://psort.nibb.ac.jp/). The lengths of mitochondrial pre-sequences were predicted with TargetP.

Results

Physical mapping of the *Rf1* locus

We have previously mapped the *Rf1* locus to a ~2-cM region of linkage group (LG) 8 of the high-density linkage map of sorghum (Kim et al. 2005a, b; Klein et al. 2001). Markers flanking the locus have been integrated in the physical map of sorghum, and BAC contigs containing each marker can be viewed at (http://sorgblast2.tamu.edu/index.html). FISH analysis (Kim et al. 2005b) of this region of sorghum LG-08 indicated that the *Rf1* gene was in a highly recombinant segment of euchromatin, which suggested that it was a realistic target for cloning using a map-based strategy. The physical size of the *Rf1* locus was also delimited using FISH analysis with BAC clones containing proximal linkage markers *Xtxp18* and *Xtxa606*. Estimates based on FISH analysis indicate a molecular size of the *Rf1*marker-delimited segment between 0.383 Mb and 0.418 Mb (Kim et al. 2005b).

Examination of the physical map in the region of the *Rf1* locus revealed a gap between BAC contigs containing the markers *Xtxp18* and *Xtxa606*. To close this physical gap, a series of approaches were used to identify gap-bridging BAC clones including (1) screening a recently created IS3620C BAC library with *Rf1*-linked molecular markers, (2) screening BAC pools with STS probes designed against BAC-end sequences, and (3) identifying AFLPs markers in sequences immediately adjacent to physical gaps. These techniques were designed to identify sequences that would permit PCR-based screening of BAC libraries for gap-bridging clones. The BAC clones identified in this manner were subjected to HICF and fingerprint contig analyses to permit integration into existing BAC contigs. The result was a contiguous array of BAC clones spanning the region of LG-08 harboring the *Rf1* locus. The resulting BAC minimum tiling path across this region of LG-08 is shown in Fig. 1i.

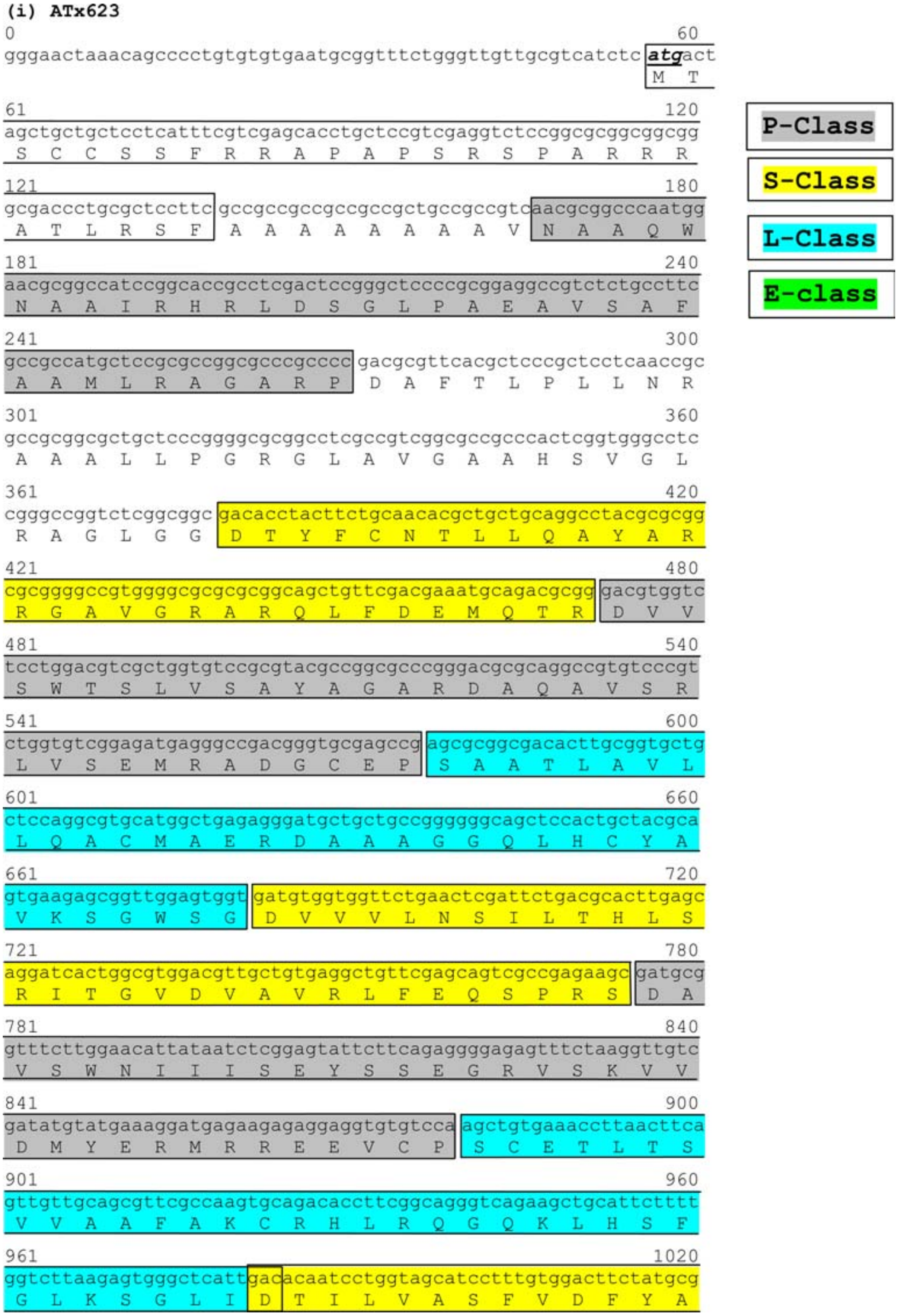
High-resolution genetic mapping of the *Rf1*gene

Having spanned the *Rf1*locus with a contiguous array of BAC clones, the resolution of the linkage map was increased utilizing new markers discovered through low-pass sequencing of sorghum BACs. A list of markers including SSRs, indels, and SNPs that were discovered in this manner are listed in Table 1. Breakpoint analysis with F_2recombinants placed these new genetic markers into the *Rf1* linkage map and narrowed the critical region in which the *Rf1* gene resides. Of the newly discovered genetic markers, two in particular, *Xtxp400* and *Xtxi3*, refined the *Rf1* locus to a 0.39-cM region of LG-08 (see Fig. 1ii). The BACs that contain *Xtxp400* and *Xtxi3* also provided new insight as to the physical location of the *Rf1* gene. Proximal marker *Xtxp400* was determined to reside on BACs SB_ BBf14I11, and SB_IBa75D09, and *Xtxi3* was landed only on BAC SB_IBa75D09. These results indicate that the *Rf1* gene is positioned physically between the overlap of BACs SB_BBf14I11 and SB_IBa75D09, and a defined physical position within BAC SB_IBa75D09.

Identification of candidate genes

Sequence analysis (sequenced to 6X minimum coverage) of BACs SB_BBf14I11 and SB_IBa75D09 revealed the physical location of markers *Xtxp400* and *Xtxi3* (see Table 1). The 67-kb region demarked by these two markers represented a total of three crossover events (*Xtxp400*, 1 F_2 recombinant; *Xtxi3*, 2 F_2recombinants). Breakpoint analysis of the three remaining F_2recombinants was used to refine further the location of the *Rf1* gene. Genetic markers within this 67-kb region were discovered by sequencing the genomes of RTx432 and ATx623 spanned by *Xtxp400* and *Xtxi3*. The results of breakpoint analysis are shown in Fig. 1iii. The locus is delimited by markers nos. 8 and 29, each exhibiting one recombination event and thereby providing the highest genetic resolution possible within the present mapping population. At this level of resolution, the *Rf1* locus spans 32 kb and four ORFs: a Ca^{2+}-ATPase (Ca^{2+}-ATPase12), a cyclin D-1 (cyclin14), an unknown protein (unknown protein15), and a PPR gene family member (*PPR13*). Marker no. 29 resides ~2,400 bp upstream of the first exon of unknown protein15 and represents a single recombinant event. No other sequence differences were observed between parental inbreds ATx623 and RTx432 in the adjoining ~19 kb. This 19-kb conserved region spans cyclin14, unknown protein15, and the intergenic region between these genes. A series of 17 sequence polymorphisms were located in an ~7-kb region spanning gene *PPR13*, with all markers cosegregating with the fertility restoration phenotype. Eight cosegregating sequence polymorphisms (markers nos. 19–25) reside in the ~1,200-bp 5′ region immediately flanking the *PPR13* ORF, and a cluster of three cosegregating SNPs reside slightly further upstream (markers nos. 26–28). Similarly, eight sequence polymorphisms reside in the 3′ region flanking the *PPR13* ORF (markers nos. 11–18). Markers in this ~2,700-bp region downstream of *PPR13* all cosegregated with the fertility restoration phenotype. The sequence of the Ca^{2+}-

Table 2 Nucleotide, amino acid sequences, and motifs of the sorghum *PPR13* cDNA



(i) ATx623

```
0                                                          60
gggaactaaacagcccctgtgtgtgaatgcggtttctgggttgttgcgtcatctcatgact
                                                        M  T

61                                                         120
agctgctgctcctcatttcgtcgagcacctgctccgtcgaggtctccggcgcggcggcgg
 S  C  C  S  S  F  R  R  A  P  A  P  S  R  S  P  A  R  R  R

121                                                        180
gcgaccctgcgctccttcgccgccgccgccgccgctgccgccgtcaacgcggcccaatgg
 A  T  L  R  S  F  A  A  A  A  A  A  A  A  V  N  A  A  Q  W

181                                                        240
aacgcggccatccggcaccgcctcgactccggctccccgcggaggccgtctctgccttc
 N  A  A  I  R  H  R  L  D  S  G  L  P  A  E  A  V  S  A  F

241                                                        300
gccgccatgctccgcgccggcgcccgccccgacgcgttcacgctcccgctcctcaaccgc
 A  A  M  L  R  A  G  A  R  P  D  A  F  T  L  P  L  L  N  R

301                                                        360
gccgcggcgctgctcccggggcgcggcctcgccgtcggcgccgcccactcggtgggcctc
 A  A  A  L  L  P  G  R  G  L  A  V  G  A  A  H  S  V  G  L

361                                                        420
cgggccggtctcggcggcgacacctacttctgcaacacgctgctgcaggcctacgcgcgg
 R  A  G  L  G  G  D  T  Y  F  C  N  T  L  L  Q  A  Y  A  R

421                                                        480
cgcggggccgtggggcgcgcgcggcagctgttcgacgaaatgcagacgcgggacgtggtc
 R  G  A  V  G  R  A  R  Q  L  F  D  E  M  Q  T  R  D  V  V

481                                                        540
tcctggacgtcgctggtgtccgcgtacgccggcgcccgggacgcgcaggccgtgtcccgt
 S  W  T  S  L  V  S  A  Y  A  G  A  R  D  A  Q  A  V  S  R

541                                                        600
ctggtgtcggagatgagggccgacgggtgcgagccgagcgcggcgacacttgcggtgctg
 L  V  S  E  M  R  A  D  G  C  E  P  S  A  A  T  L  A  V  L

601                                                        660
ctccaggcgtgcatggctgagagggatgctgctgccgggggcagctccactgctacgca
 L  Q  A  C  M  A  E  R  D  A  A  A  G  G  Q  L  H  C  Y  A

661                                                        720
gtgaagagcggttggagtggtgatgtggtggttctgaactcgattctgacgcacttgagc
 V  K  S  G  W  S  G  D  V  V  V  L  N  S  I  L  T  H  L  S

721                                                        780
aggatcactggcgtggacgttgctgtgaggctgttcgagcagtcgccgagaagcgatgcg
 R  I  T  G  V  D  V  A  V  R  L  F  E  Q  S  P  R  S  D  A

781                                                        840
gtttcttggaacattataatctcggagtattcttcagaggggagagtttctaaggttgtc
 V  S  W  N  I  I  I  S  E  Y  S  S  E  G  R  V  S  K  V  V

841                                                        900
gatatgtatgaaaggatgagaagagaggaggtgtgtccaagctgtgaaaccttaacttca
 D  M  Y  E  R  M  R  R  E  E  V  C  P  S  C  E  T  L  T  S

901                                                        960
gttgttgcagcgttcgccaagtgcagacaccttcggcagggtcagaagctgcattctttt
 V  V  A  A  F  A  K  C  R  H  L  R  Q  G  Q  K  L  H  S  F

961                                                       1020
ggtcttaagagtgggctcattgacacaatcctggtagcatcctttgtggacttctatgcg
 G  L  K  S  G  L  I  D  T  I  L  V  A  S  F  V  D  F  Y  A
```

P-Class

S-Class

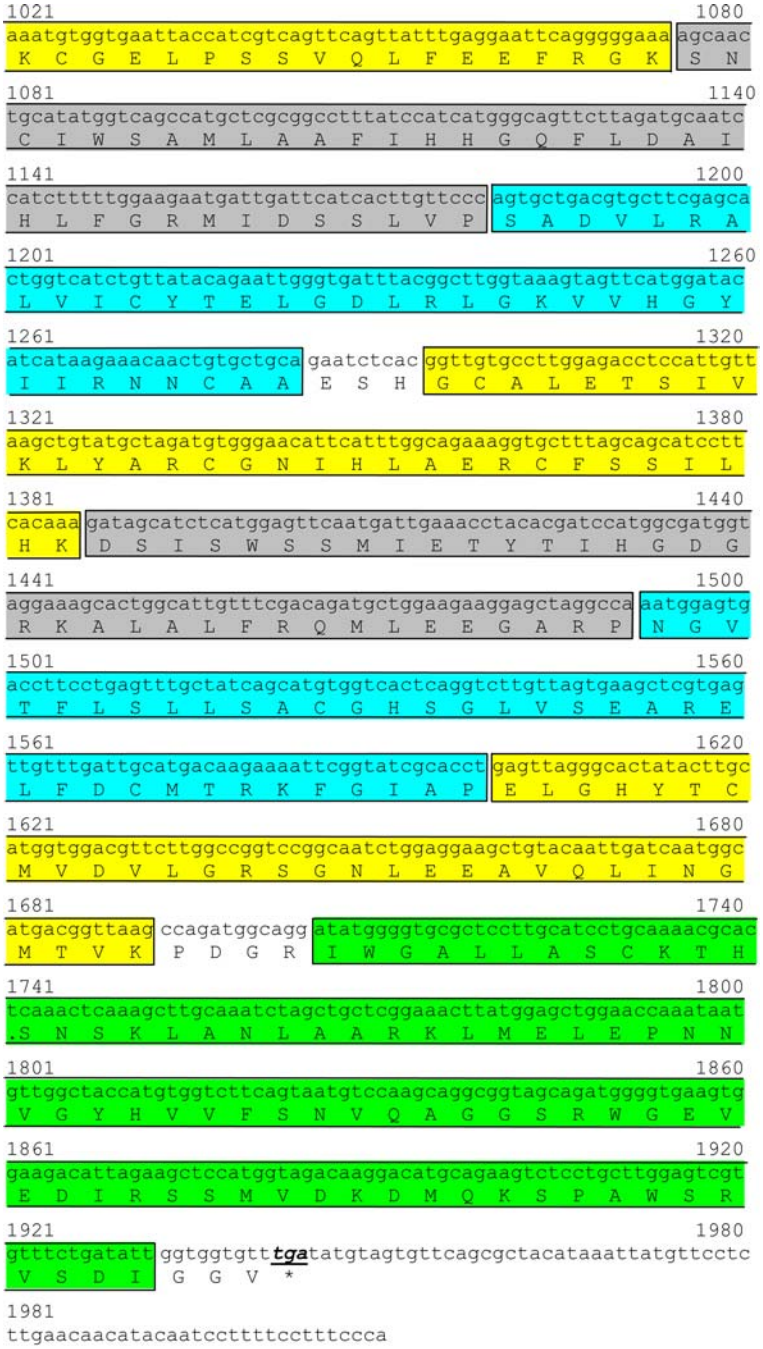
L-Class

E-class

ATPase12 gene was conserved across all introns and exons with two exceptions, a single base substitution in the first intron and a microsatellite upstream of the ATPase gene. Both of these polymorphisms cosegregated with the fertility restoration phenotype.

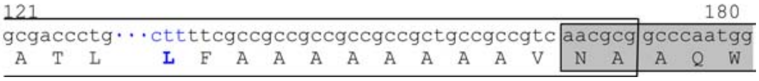
To gain further insight into the identity of the sorghum *Rf1* gene, ORFs within the *Rf1* locus were scanned *in silico* for organellar-targeting pre-sequences. CMS is maternally inherited, and a large volume of research indicates that CMS-*Rf* systems result from an interaction of the mitochondrial genome and nuclear-encoded proteins targeted for the mitochondria (Pring et al. 1995). Examining the four ORFs revealed that the PPR13 protein was likely targeted to the mitochondria (probabilities, MitoP, 0.9986; PSORT, 0.86; TargetP, 0.76; Predotar, 0.421). In contrast, PSORT*in silico*

Table 2 (Contd.)



```
1021                                                        1080
aaatgtggtgaattaccatcgtcagttcagttatttgaggaattcaggggaaa agcaac
 K  C  G  E  L  P  S  S  V  Q  L  F  E  E  F  R  G  K   S  N
1081                                                        1140
tgcatatggtcagccatgctcgcggcctttatccatcatgggcagttcttagatgcaatc
 C  I  W  S  A  M  L  A  A  F  I  H  H  G  Q  F  L  D  A  I
1141                                                        1200
catctttttggaagaatgattgattcatcacttgttccc agtgctgacgtgcttcgagca
 H  L  F  G  R  M  I  D  S  S  L  V  P   S  A  D  V  L  R  A
1201                                                        1260
ctggtcatctgttatacagaattgggtgatttacggcttggtaaagtagttcatggatac
 L  V  I  C  Y  T  E  L  G  D  L  R  L  G  K  V  V  H  G  Y
1261                                                        1320
atcataagaaacaactgtgctgca gaatctcac ggttgtgccttggagacctccattgtt
 I  I  R  N  N  C  A  A   E  S  H   G  C  A  L  E  T  S  I  V
1321                                                        1380
aagctgtatgctagatgtgggaacattcatttggcagaaaggtgctttagcagcatcctt
 K  L  Y  A  R  C  G  N  I  H  L  A  E  R  C  F  S  S  I  L
1381                                                        1440
cacaaa gatagcatctcatggagttcaatgattgaaacctacacgatccatggcgatggt
 H  K   D  S  I  S  W  S  S  M  I  E  T  Y  T  I  H  G  D  G
1441                                                        1500
aggaaagcactggcattgtttcgacagatgctggaagaaggagctaggcca aatggagtg
 R  K  A  L  A  L  F  R  Q  M  L  E  E  G  A  R  P   N  G  V
1501                                                        1560
accttcctgagtttgctatcagcatgtggtcactcaggtcttgttagtgaagctcgtgag
 T  F  L  S  L  L  S  A  C  G  H  S  G  L  V  S  E  A  R  E
1561                                                        1620
ttgtttgattgcatgacaagaaaattcggtatcgcacct gagttagggcactatacttgc
 L  F  D  C  M  T  R  K  F  G  I  A  P   E  L  G  H  Y  T  C
1621                                                        1680
atggtggacgttcttggccggtccggcaatctggaggaagctgtacaattgatcaatggc
 M  V  D  V  L  G  R  S  G  N  L  E  E  A  V  Q  L  I  N  G
1681                                                        1740
atgacggttaag ccagatggcagg atatggggtgcgctccttgcatcctgcaaaacgcac
 M  T  V  K   P  D  G  R   I  W  G  A  L  L  A  S  C  K  T  H
1741                                                        1800
tcaaactcaaagcttgcaaatctagctgctcggaaacttatggagctggaaccaaataat
.S  N  S  K  L  A  N  L  A  A  R  K  L  M  E  L  E  P  N  N
1801                                                        1860
gttggctaccatgtggtcttcagtaatgtccaagcaggcggtagcagatggggtgaagtg
 V  G  Y  H  V  V  F  S  N  V  Q  A  G  G  S  R  W  G  E  V
1861                                                        1920
gaagacattagaagctccatggtagacaaggacatgcagaagtctcctgcttggagtcgt
 E  D  I  R  S  S  M  V  D  K  D  M  Q  K  S  P  A  W  S  R
1921                                                        1980
gtttctgatatt ggtggtgtttgatatgtagtgttcagcgctacataaattatgttcctc
 V  S  D  I   G  G  V  *
1981
ttgaacaacatacaatccttttcctttccca
```

(ii) RTx432



```
121                                                       180
gcgaccctg···cttttcgccgccgccgccgccgctgccgccgtc aacgcg gcccaatgg
 A  T  L     L  F  A  A  A  A  A  A  A  A  V   N  A   A  Q  W
```

(i) Predicted sequence of the gene encoded by *sterile line* ATx623. The mitochondrial targeting pre-sequence predicted by TargetP for ATx623 is indicated by the *unshaded box*, whereas predicted start and stop codons are *underlined*. *Shaded boxes* indicate different classes of PPR domains. (ii) Divergence of protein-coding region of RTx432 *PPR13* gene. *Colored text*indicates amino acid and nucleotide sequences that diverge between *sterile* and *fertile lines*. The mitochondrial targeting pre-sequence predicted by TargetP for the RTx432 protein is indicated by the *unshaded box*

predictions indicate that Ca^{2+}-ATPase12 is targeted to the plasma membrane ($P = 0.60$) or the chloroplast thylakoids ($P = 0.68$), but is not predicted to target the mitochondria ($P < 0.14$). The BLASTX analysis of Ca^{2+}-ATPase12 revealed high homology with type IIB, plasma membrane-type Ca^{2+}-ATPase (e-value of 0.00).

Similarly, unknown protein15 may contain a pre-sequence targeting the chloroplast thylakoids (probabilities, TargetP, 0.561: PSORT 0.979; Predotar, 0.41), but a mitochondrial pre-sequence was not predicted ($P < 0.14$). Finally, cyclin14 protein was not predicted to contain any targeting pre-sequence ($P < 0.1$ in all programs examined).

Characteristic of*PPR13*cDNA

Based on the results of marker segregation analysis and pre-sequence targeting predictions, *PPR13*represents a potential *Rf1* candidate gene and therefore warranted further characterization. To characterize the *PPR13* transcript, first-strand cDNA was prepared from

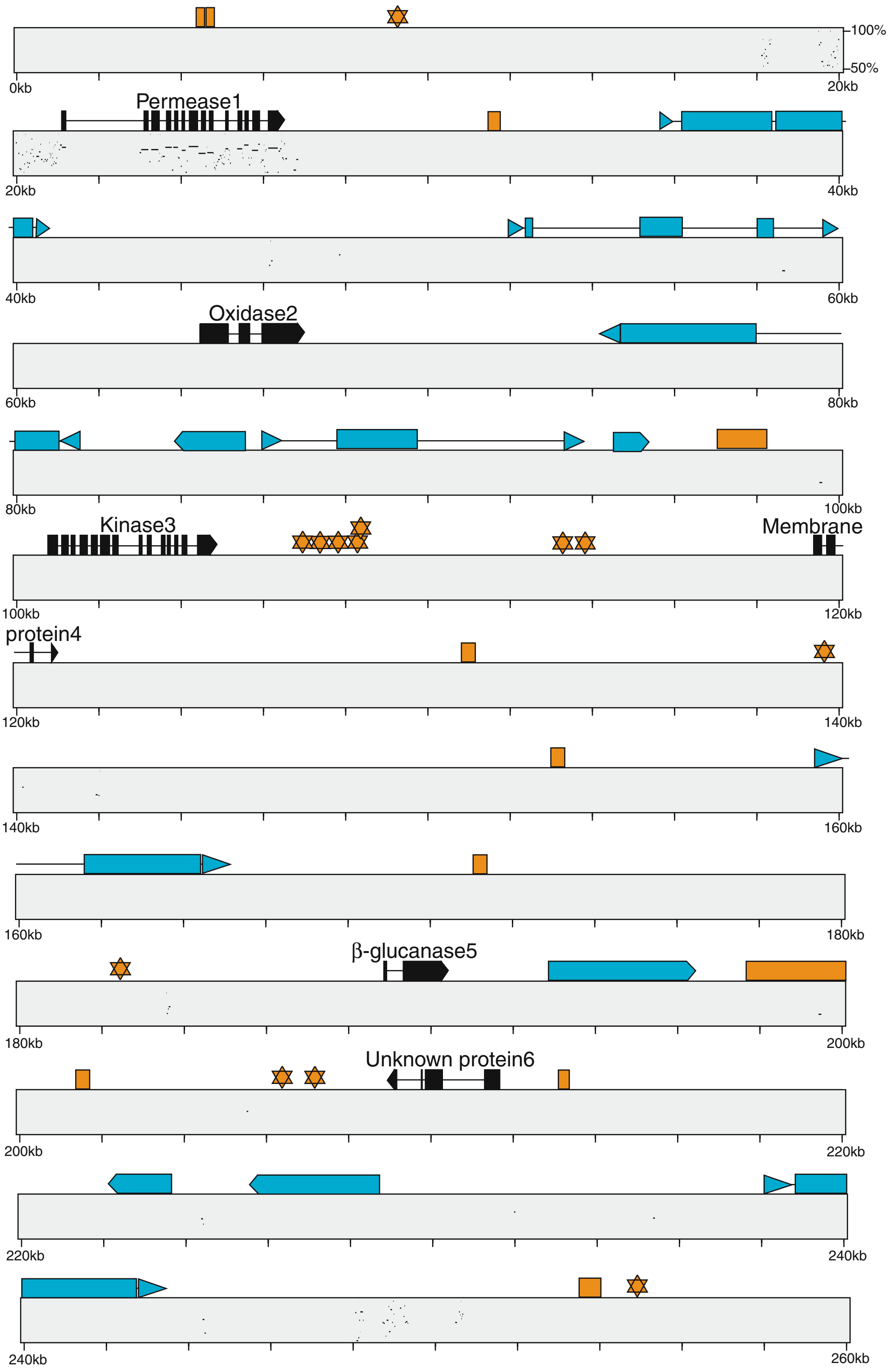
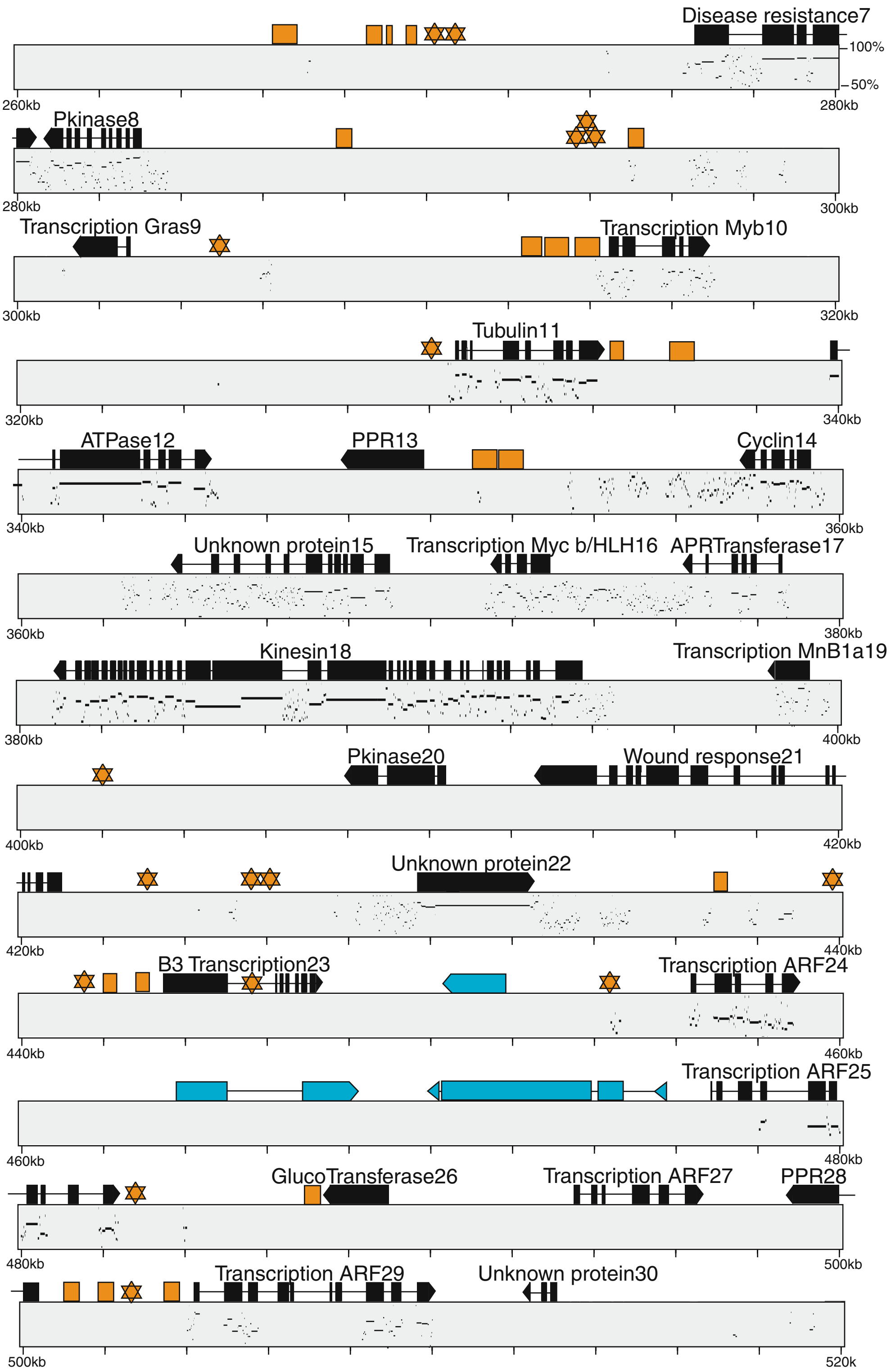


Fig. 2 Gene and repetitive element map of 0.53-Mb segment of sorghum chromosome 8 and percent identity plot with rice chromosome 12. Transcription orientation of predicted genes is indicated by *arrowheads*. Exons are indicated as *black rectangles* and introns by *thin lines* between them. Genes are those listed in Table 3. Retrotransposable elements are depicted in blue with LTRs delineating retroelements shown as *arrowheads*. Positions of DNA transposon-like sequences are indicated by *orange rectangles*, whereas *stars* show the position of predicted miniature inverted-repeat transposable elements. Percent identity plot showing the alignment of sorghum and rice sequences is shown in the *light gray boxes*

Fig. 2 (Contd.)



DNase-treated total leaf RNA, and subsequently PCR-amplified with GSPs, and the ends of the *PPR13* transcript were obtained by 5′ and 3′ RACE. The PCR amplicons from both ATx623 and RTx432 PPR transcripts were single products of the expected size. Control reactions (plus template, minus reverse transcriptase) did not yield a detectable PCR product, indicating that the amplification from RNA preparations was not due to DNA contamination. A single amplicon was also obtained from 5′ and 3′ RACE (data not shown). The full-length transcript and predicted ORF of *PPR13* are shown in Table 2. The 5′ untranslated region appears to extend 55 bp upstream of the predicted start codon, whereas the 3′ RACE product placed the poly(A) site

67 bp downstream of the predicted termination codon. The *PPR13* gene of ATx623 encodes a 2,010-bp full-length transcript that contains a single exon encoding 628 amino acids, with a predicted molecular weight of 67.96 kDa (Table 2). The cDNA of RTx432 differs by a 3-bp deletion and an adjacent, non-synonymous single-base substitution that results in a two-amino acids difference between the predicted proteins of sterile and restorer lines (see Table 2). The two-amino acid differences between the RTx432 and ATx623 alleles do not reside in a PPR motif; however, the amino acid differences are positioned at a consensus pre-sequence cleavage site (I. Small, personal communication). As such, the TargetP-predicted cleavage site in the PPR13 protein differs between sterile and fertile lines, and unique N termini are predicted to arise after cleavage of the transit peptide pre-sequence in sterile and fertile lines (compare unshaded boxes, Table 2i vs ii).

Examination of the deduced amino acid sequence of PPR13 revealed 14 PPR motifs, of which 13 appear in tandem (Table 2, colored boxes). Based on a recent study by Lurin et al. (2004), sorghum PPR13 is classified as an E-subfamily PPR protein. Of the different known classes of PPR motifs, PPR13 contains five S-class, five P-class, and four L-class motifs (a degenerate L-class motif resides in the region between amino acids D73 and G108). The C terminus of PPR13 (amino acids I550–I625) also contains an E-type motif; a motif only found in the C-terminal domains of members of the PPR family in*Arabidopsis* (Lurin et al. 2004).

Comparisons of orthologous sorghum and rice genomic regions

To permit a detailed comparison of the genomes of rice and sorghum spanning the *Rf1* locus, the sorghum BAC tiling path shown in Fig. 1 was sequenced (phase III). The prediction of genes present in this region of the sorghum genome is shown in Table 3. Thirty genes were predicted to be encoded in this 0.53-Mb region of LG-08, averaging one gene for every ~18 kb. Categories of genes included predicted proteins involved in plant defense (disease resistance and wound-responsive factor), photosynthesis/respiration/carbohydrate metabolism (permease, β-glucanase, phosphoribosyltransferase, and adenine glucosyltransferase), regulatory or signaling functions (kinases and transcription factors), growth-related processes (cyclin, tubulin, and kinesin), and a number of hypothetical or proteins of unknown function. Nearly all of the predicted genes were homologous with a grass EST, although a significant portion showed only 90–95% homology with sorghum ESTs in the databases. As these predicted genes also show greater than 90% homology with ESTs from other grasses, it appears that they may represent members of multigene families, and the EST for this predicted gene family member has not been sequenced. Several notable exceptions to the discovery of homologous ESTs were a series of transcription factors and sorghum gene *PPR13* (see Table 3). The lack of *PPR13*homologues in the public EST databases is consistent with the low expression level of the P–L–S class of PPR genes (Lurin et al. 2004), and the high threshold cycle values displayed by *PPR13* during real-time PCR assays (R.R. Klein, unpublished).

The distribution of genes and repetitive elements across this genomic region is shown in Fig. 2. The distribution of genes was not uniform, as the first ~270 kb of the pseudomolecule encoded only six predicted genes, averaging one gene every ~45 kb. In contrast, gene density was considerably greater in the region spanning 275–515 kb of the pseudomolecule (one gene every ~10 kb). Examination of the region for transposable elements revealed LTR and LINE retrotransposons, degenerate enhancer/suppressor-mutator DNA elements, and a series of MITEs. As predicted, the distribution of the transposable elements showed a greater number of LTR and LINE retrotransposons in the gene-poor region of the pseudomolecule (first ~270 kb), whereas the distribution of MITEs was indicative of gene distribution with the vast majority residing in gene-rich regions.

Sequence-based alignment of the rice and sorghum genomes has revealed macrocolinearity between sorghum chromosome 8 and rice chromosome 12 (Paterson et al. 2004a). A plot of the percent identity between sorghum LG-08 and the colinear BACs of rice chromosome 12 (~95.1–99 cM, BACs OJ1260_B01, OSJNBB0049H14, OJ1573_A06, OSJNBa0044E20, OSJNBa0010M16) are shown in the lower panels of Fig. 2. With few exceptions, regions of microcolinearity between rice and sorghum reflected gene-coding sequences and the regions immediately flanking genes, whereas gaps reflect missing genic sequences and intergenic spaces. Genomic regions that were gene-rich (regions 270–515 kb) generally showed conserved gene content, gene order, and transcription orientation. In the gene-poor region from 1 kb to 270 kb, microcolinearity was interrupted. Sorghum genes nos. 1–6 revealed a mosaic pattern of genome alignment displaying homology with genes from four rice chromosomes: chromosomes 01, 04, 09, and 12 (Table 3). In the gene-rich region from 270 kb to 515 kb of sorghum, homologues on rice chromosome 12 were observed for a great majority of the predicted genes. In a 125-kb region spanning genes nos. 7–19, gene order, gene content, and transcription orientation were highly conserved with two notable exceptions; homologues to sorghum *PPR13* and Gras transcription factor9 were not detected on rice chromosome 12 BACs (Fig. 2, ~301.5–302.7 kb). Closer examination reveals that a 10-kb region of the sorghum genome spanning *PPR13* was devoid of homologous sequences with rice chromosome 12. Rice sequences showing homology with *PPR13* however, were detected on chromosome 4 (BLASTN analysis, Table 3). Closer examination of the BLAST results showed that tandem genes located on rice chromosome 4 encode for PPR proteins. Percent identity plots of sorghum *PPR13* and these rice chromosome 4 homologues (OSJNBa0065J03.21 and OSJNBb0032D24.15) revealed that

Table 3 Predicted genes in a 528-kb region of sorghum chromosome 8 and alignment with homologous sequences of rice

Gene	EST access. no.[a,b] (species)	*e*-value (percent ident.)[b]	Protein access. no.[b]	Homology [a,b] (species)	BLASTX[a,b] (*e*-value)	BLASTN[b] rice BAC/PAC	Rice Chr[b] Chr No.: cM	Accession No.[b]	BLAST[b] (*e*-value)
1	CD230109 (*Sb*)	e-124 (100%)	NP_180966.1	Xanthine/uracil permease (*Ta*)	2e-30	OJ1260_B01	12:95.1	AL772427.4	6e-54
2	CD231935 (*Sb*)	0 (98%)	XP_472075	Putative oxidase-like protein (*Os*)	0.0	OSJNBa0034E2[c]	4:30.8	AL662967.2	e-113
3	CA276132 (*So*)	2e-86 (98%)	XP_476111	Protein kinase family protein (*Os*)	0.0	OSJNBa0014K08	1:127.3	AP003376.3	8e-53
4	–	–	BAD29095.1	Putative membrane protein (*Os*)	1e-28	OSJNBb0085I16	9:34.9	AP005700.2	1e-12
5	CN127710 (*Sb*)	0 (100%)	NP_914607.1	Beta-1,3-glucanase precursor (*Os*)	e-106	P0432C03	1:166.9	AP004031.3	5e-39
6	CA484679.1 (*Ta*)	e-135 (96%)	NP_921170.1	Unknown protein (*Os*)	6e-33	–	–	–	–
7	BE359692.1 (*Sb*)	0 (100%)	Q9SZA7	Disease resistance protein (*Ta*)	e-161	OSJNBb0049H14	12:95.1	AL844880.3	0.0
8	BE594366.1 (*Sb*)	e-176 (97%)	BAD18005.1	Serine/threonine protein kinase (*Os*)	3e-61	OSJNBb0049H14	12:95.1	AL844880.3	9e-54
9	–	–	NP_910364.1	Transcription factor, Gras (*Os*)	4e-33	P0425F02	6:6.0	AP001168.1	3e-27
10	BG411322.1 (*Sb*)	e-148 (99%)	AAP50967.1	Transcription factor, Myb-related (*Os*)	2e-30	OSJNBb0049H1[c]	12:95.1	AL844880.3	8e-22
11	BF657456.1 (*Sb*)	e-128 (98%)	BAC84426.1	Tubulin-related protein (*Os*)	3e-44	OSJNBb0049H14	12:95.1	AL844880.3	1e-69
12	BI098936.1 (*Sb*)	0 (100%)	AAM44081.1	Calcium ATPase, type IIB (*Mt*)	0.0	OSJNBb0049H14	12:95.1	AL844880.3	0.0
13	–	–	NP_195239.1	PPR repeat-containing protein (*At*)	3e-86	OSJNBa0065J03 OSJNBb0032D24	4:19.94:19.9	AL731615.4 AL662995.2	5e-353e-30
14	CN124680.1 (*Sb*)	e-168 (99%)	NP_177178.1	Cyclin delta-1 (*At*)	5e-16	OJ1573_A06	12:95–99	AL954852.4	2e-60
15	BG411381.1 (*Sb*)	0 (98%)	NP_567993.1	Unknown protein (*At*)	2e-40	OJ1573_A06	12:95–99	AL954852.4	2e-76
16	–	–	BAC83262.1	Transcription factor, Myc b/HLH (*Os*)	3e-23	OJ1573_A06	12:95–99	AL954852.4	2e-19
17	AW747763.1 (*Sb*)	0 (99%)	BAB08003.1	Putative adenine phosphoribosyltransferase (*Hv*)	4e-44	OSJNBa0010M16	12:95.1	AL831797.4	5e-16
18	CF036321.1 (*Zm*)	0 (94%)	NP_188535.3	Kinesin motor protein (*At*)	7e-93	OSJNBa0044E20	12:95.1	AL954829.5	e-145
19	CK144497.1 (*Zm*)	8e-54 (93%)	P38564	Transcription factor, MNB1a (*Zm*)	4e-21	OSJNBa0044E20	12:95.1	AL954829.5	2e-58
20	CA196723.1 (*So*)	e-149 (94%)	NP_919804.1	Cell-wall protein kinase-like (*Os*)	9e-91	–	–	–	–
21	CD232990.1 (*So*)	0 (99%)	XP_481914.1	Putative wound-responsive (*Os*)	3e-69	P0421H07	1:122.3	AP003245.4	5e-20
22	CA175400.1 (*So*)	0 (96%)	AAP50958.1	Unknown protein (*Os*)	0	OSJNBa0044E20	12:95.1	AL954829.5	0.0
23	CA264834.1 (*So*)	e-126 (97%)	XP_463192.1	Transcription factor, B3 (*Os*)	4e-43	OSJNBb0089B03	4:30.8	AL607000	9e-12
24	–	–	XP_463202.1	Transcription factor, ARF (*Os*)	4e-34	–	–	–	–
25	–	–	AAO34491.1	Transcription factor, ARF (*Os*)	3e-19	OSJNBa0044E20	12:95.1	AL954829.5	2e-18
26	CA279755.1 (*So*)	0 (94%)	NP_916449.1	Flavonol glucosyltransferase, (*Os*)	e-121	B1060H01	1:126.5	AP003560.2	7e-63
27	–	–	AAO34502.1	Transcription factor, ARF (*Os*)	1e-20	OSJNBa0010M16 OSJNBa0044E20	12:95.112:95.1	AL831797.4 AL954829.5	2e-272e-18
28	BG268102.1 (*Zm*)	e-146 (90%)	NP_921705.1	PPR-repeat containing protein (*Os*)	3e-85	–	–	–	–
29	CA230819.1 (*So*)	0 (93%)	AAO34502.1	Transcription factor, ARF (*Os*)	5e-18	OSJNBa0010M16 OSJNBa0044E20	12:95.112:95.1	AL831797.4 AL954829.5	1e-363e-25
30	CF628680.1 (*Zm*)	2e-44 (95%)	–	–	–	OSJNBa0010M16	12:95.1	AL831797.4	3e-20

[a] *Sb Sorghum bicolor*, *So Saccharum officinarum*, *Zm Zea mays*, *Ta Triticum aestivum*, *At Arabidopsis thaliana*, *Os Oryza sativa*, *Mt Medicago truncatula*, *Hv Hordeum vulgare*, *ARF* Auxin-responsive factor, *PPR* pentatricopeptide repeat

[b] *Dashes* indicate homology values below threshold score

[c] Multiple BACs hit at similar *e*-values

conservation was confined to the protein-coding region, whereas homology between the 5′ and 3′ sequences was not apparent (data not shown). The percent nucleotide identity between sorghum *PPR13* and rice predicted genes OSJNBa0065J03.21 and OSJNBb0032D24.15 was 73% (total length, 1,230 bp; gaps, 59 bp) and 71% (total length 1,127 bp; gaps, 77 bp), respectively.

Examination of percent identity plots between rice and sorghum revealed several additional interesting features. As stated previously, microcolinearity was largely confined to gene coding regions, and sequences of exons displayed the highest percent identities between rice and sorghum. Several intergenic regions of 3 to 4 kb in size did show a detectable level of microcolinearity (~354–358, 373–376 kb). Present evidence does not predict that protein-coding sequences reside in these regions, and the significance of this homology is unclear. Additionally, Auxin-responsive transcription factors (ARF) appear to have duplicated and diverged in this region of rice and sorghum. In a ~56-kb region of the sorghum pseudomolecule, four ARF transcription factors are predicted, with each showing varying levels of homology with rice chromosome 12 (see Table 3; Fig. 2). The ARF transcription factor27 appears to have duplicated after the divergence of rice and sorghum ancestors as a homologue was not found in the same relative position of rice. Finally, unknown protein30 may reside in a colinear position in rice and sorghum, but the transcription orientation appears opposite in the two species. Whereas the transcription orientation of most colinear genes was conserved, this hypothetical protein may represent an exception to this assertion.

Discussion

The completed sequence and structure of the rice genome will create an unprecedented opportunity to accelerate genome research for all grass species through comparative genome alignment. To realize fully the investment in the rice genome-sequencing project, genomic resources are needed for other grass species. The sorghum research community has generated an extensive set of genetic, physical, cytogenetic, and comparative map resources (for review, see Mullet et al. 2001). Fortunately, the rice genome exhibits substantial colinearity with the genome of sorghum and other grasses. This suggests that alignment of the rice genome sequence to high-resolution integrated genome maps of related grass species can accelerate the isolation and analysis of genes of agronomic importance. Microcolinearity between rice and sorghum have been shown by the sequencing of a limited number of genomic segments from orthologous loci (for review, see Bennetzen and Ma 2003). Although microcolinearity between rice and other grasses was observed, a relatively high frequency of small genetic rearrangements had occurred including gene insertion, inversions, and duplications. These small rearrangements will make approaches like chromosome walking in surrogate genomes (such as rice) fairly risky, as was observed with cloning the barley stem rust resistance gene *Rpg1* (Brueggeman et al. 2002; Han et al. 1999; Kilian et al. 1997). Nevertheless, comparative sequence analysis between grasses including rice, sorghum, and maize can provide a wealth of information on speciation and genome evolution, along with practical applications such as more precise gene annotation and the identification of *cis* regulatory elements.

Rf1 candidate gene

In the present study, we have tentatively identified the sorghum *Rf1* gene, one of the major loci controlling fertility restoration in A1 cytoplasm of sorghum. The cloning and characterization of *Rf1* will provide needed insight into the mechanism of action of fertility restoration in sorghum, and permit a detailed examination of this nuclear–mitochondrial interaction. As the A1 male sterile cytoplasm remains the primary CMS system used for commercial hybrid seed production in the United States, the development of a rapid, robust molecular screen for the *Rf1* gene is a major target for the hybrid seed industry. The discovery of a series of polymorphisms in the *Rf1* locus that cosegregate with fertility restoration provides the necessary information to develop molecular phenotyping tools to classify sorghum germplasm without the need for time-consuming test crosses.

The *Rf1* gene was tentatively identified by high-resolution genetic and physical mapping and by exclusion of other potential candidate genes by sequence analysis. Conceptual translation of the potential target sequence showed amino acid homology to the PPR protein family. Three other genes reside in the region of zero recombination, although either these genes do not contain a mitochondrial targeting pre-sequence or their sequence is conserved completely between fertile and sterile lines. Over the past several years, fertility restorer genes have been cloned from several plant species including petunia (Bentolila et al. 2002), radish (Brown et al. 2003; Desloire et al. 2003; Koizuka et al. 2003), and rice (Kazama and Toriyama 2003; Komori et al. 2004). Each of these *Rf* genes has been shown to encode a protein with a mitochondrial targeting peptide pre-sequence and a series of PPR repeat motifs. The candidate *Rf1* gene of sorghum identified herein would add to a growing list of fertility restorer genes that are members of the PPR gene family. This fact does not preclude fertility restoration being conditioned by non-PPR proteins that are targeted to cytoplasmic-inherited organelles (e.g., mitochondria and plastid). Maize *Rf2*, the first fertility restoration gene cloned, has been identified as encoding for a mitochondrial aldehyde dehydrogenase (Cui et al. 1996; Liu et al. 2001). Hence, further studies will be required to prove definitively the identity of the sorghum *Rf1* gene.

The PPR genes are found in all eukaryotes analyzed but with a great discrepancy between numbers in plants and non-plant organisms. Lurin et al. (2004) identified hundreds of PPR genes in the rice and*Arabidopsis*genomes, whereas the human genome encoded for only six putative PPR proteins. The limited numbers of biochemical studies (Lurin et al. 2004; Meierhoff et al. 2003; Nakamura et al. 2003; Tsuchiya et al. 2004) indicate that PPR proteins may represent RNA-binding proteins involved in posttranscriptional processes in organelles. A significant portion of PPR proteins in*Arabidopsis*are predicted to contain a targeting signal peptide at the N terminus, although the percentage of organellar-targeting varied with the particular PPR subfamilies (Lurin et al. 2004). Evidence that these proteins are involved in organelle biogenesis comes largely from mutant studies in which organellar transcript processing and translation are correlated with a mutant PPR protein (Meierhoff et al. 2003; Nakamura et al. 2003). Positional cloning of CMS restorer genes has provided further evidence for a role of PPR proteins in organelle gene expression.

Lurin et al. (2004) recently characterized PPR proteins into a series of discrete subfamilies on the basis of the nature of PPR motif. Using a bioinformatics approach, two new PPR motifs of 31 and 35 to 36 amino acids were defined. These two motifs, PPR-like S (for short) and PPR-like L (for long) motifs share the same α-helix secondary structure that is predicted for the P-type classical motif. The PPR subfamilies are classified as those containing P-type, L-type, S-type, and P-L-S type motifs with the latter showing a triple motif (P-L-S) repeated along the protein. Examination of sorghum PPR13 clearly revealed the triple motif of P-L-S repeated along the protein sequence. Thirteen motifs form a single unbroken tandem array that spans from amino acid D108 to amino acid K545. Assuming the presence of a fourteenth degenerate L-type motif from D73 to G108, an unbroken string of tandem motifs would exist from N37 to K545. In addition to the PPR motifs, sorghum PPR13 contains a C-terminal motif that has only been observed in PPR proteins from*Arabidopsis*(Lurin et al. 2004). The E motif is related to neither the PPR motifs nor any other polypeptide motif characterized to date. Whether the terminal motif has a catalytic function is presently unknown, and future site-directed mutational analysis of this motif in PPR13 may elucidate the functional significance of this C-terminal domain. Nevertheless, the recent identification of an E-type PPR protein in*Arabidposis*as a specificity factor in RNA editing (Allen et al. 2004) lends credence to a possible role of PPR13 in posttranscriptional processes within sorghum mitochondria.

Examination of an ~7-kb region of the *Rf1* locus revealed 19 sequence polymorphisms between *rf1* and *Rf1* lines within either the coding region of *PPR13* or the 5′ or 3′ flanking sequences. This polymorphic region is bordered by gene and intergenic sequences that are conserved in fertile and sterile lines. Examination of the *PPR13* coding sequences of fertile and sterile plants revealed a nonsynonymous substitution (S27 to L26) adjacent to an amino acid insertion (R26) in the protein of sterile plants (Fig. 2). A series of SNPs and a small indel also exist in sequences immediately 5′ (and 3′) of the *PPR13* transcript. Views of gene regulation in plants have posited that regulatory elements are located within several kilobases of transcript sequences, although promoter elements may exist more distant from the presumed promoter region. Functional mutations in *Rf* of petunia (Bentolila et al. 2002) and *Rf1* (Kazama and Toriyama 2003; Komori et al. 2004) of rice relate to a large deletion in the promoter region and to the synthesis of a truncated *Rf* protein in non-restorer genotypes, respectively. In contrast, the CMS restorer in Kosena radish (Koizuka et al. 2003) was associated with amino acid substitutions in PPR motifs of the restorer allele. In the case of sorghum *Rf1*, the presence of PPR transcripts in both sterile and restorer leaf tissue indicates that complete disruption of transcription of *PPR13* in sterile plants has not occurred. Whether any of the mutations found 5′ of the *PPR13* gene affect the temporal or tissue-specific gene expression of *PPR13*, and whether the functional mutation has been exposed, is presently under detailed examination. In vitro-synthesized PPR13 polypeptides will allow determination of whether the observed coding region differences between the *Rf1* and *rf1* alleles affect the uptake of the protein and pre-sequence processing by mitochondria. Nevertheless, the allelic diversity that exists between *PPR13* in fertile and sterile plants, set within a highly conserved regional sequence background, provides an interesting opportunity to examine gene evolution and allelic diversity in a series of sterile and restorer sorghum germplasm accessions.

Genome organization and microcolinearity

This study also provides an opportunity for detailed examination of a 0.53-Mb region of the sorghum genome spanning the *Rf1* locus. The analysis was enriched by its comparison to the orthologous region in rice, thereby providing insight into modes of local sequence evolution over the past 60 million years. Comparative DNA sequence analysis revealed a mosaic pattern of colinearity with an extended 270-kb gene-poor region that showed little gene conservation between rice and sorghum. In stark contrast, an adjoining gene-rich region extended over ~245 kb with high gene conservation between rice chromosome 12 and sorghum chromosome 8. As expected, intergenic regions, transposable elements, and other features (SSRs) showed little sequence conservation between rice and sorghum. *PPR13* was localized to the gene-rich region of sorghum, and the genes immediately flanking *PPR13* are found on contiguous rice BACs, with gene order and transcription orientation preserved between these two species. However, an orthologous PPR gene was not present in the colinear region of rice chromosome 12, suggesting that sorghum *PPR13* moved as a single gene to its current location in the genome after the divergence of sorghum and rice ancestors. At present, the mechanism of single-gene translocation is unknown, and it cannot be ascertained if the transposition occurred before or after sorghum diverged from maize. Hence, comparative sequence analysis of the orthologous maize region will permit a determination of whether the transposition occurred after maize and sorghum diverged, some 15–20 million years ago.

Small-scale gene rearrangements/events including gene duplication have been found in numerous eukaryotic genome comparisons and are believed to represent major forces in speciation and genome evolution. Evidence of PPR gene duplication has been observed as a clustering of PPR genes within the loci of radish *Rfo* (Brown et al. 2003), petunia *Rf* (Bentolila et al. 2002), and rice *Rf-1* (Komori et al. 2004). Members of each cluster show a high degree of similarity, indicating gene duplication occurred in relatively recent evolutionary time. In most of these cases, all but one member of the cluster is nonfunctional, either lacking a functional promoter or exhibiting a frame shift mutation. In stark contrast, there exists a single PPR gene in the *Rf1* locus of sorghum. The region of the rice genome showing highest homology to sorghum *PPR13* spans a region of chromosome 4 that encodes for tandem PPR genes. It is unclear whether either of these rice genes is functional or encodes a fertility restorer protein. Nevertheless, why sorghum *PPR13* does not reside within a PPR gene cluster is a matter of speculation. Knowing whether the *PPR13* transposition occurred in recent evolutionary time (e.g., after sorghum and maize ancestors diverged) may in part explain why PPR gene duplication has not yet occurred in the *Rf1* locus.

It was originally postulated that the cloning of genes in species such as sorghum would utilize surrogate model species such as rice. It has become apparent from the present study and that of the barley *Rpg1* gene (Brueggeman et al. 2002; Han et al. 1999; Kilian et al. 1997) that the value of model species such as rice for positional cloning depends on the extent of microcolinearity within the targeted region. Within the genomic region examined within this study, there existed a region of poor gene conservation between rice and sorghum that resided adjacent to a region of high gene conservation. Even in the region of higher microcolinearity, exceptions existed due either to gene transposition, gene creation, or gene removal. Hence, whereas the draft sequence of the rice genome is invaluable for BAC contig building and physical map closure, the ultimate challenge for scientists working with related grass species is to develop the genomic resources that complement resources already developed for rice. By developing these resources, the draft sequence of rice should hasten the task of cloning in targeted species, especially when a narrow colinear region can be identified. Other valuable information is also extracted from sequence colinearity, especially for specific cases where genes are missing from one species but present in another. These unique features of a genome are central to speciation and will provide valuable insight into how species such as sorghum have adapted to specific environmental niches.

Acknowledgements The authors wish to thank Ms. Julie McCollum (USDA–ARS) and Tracie Miner (Washington University) for expert technical assistance. Special thanks to Professor Ian Small (Unité de Recherches en Gènomique Végétale INRA CNRS, Evry, France) for critical guidance and thoughtful insight concerning PPR genes. The research is supported by the United States Department of Agriculture's Agricultural Research Service (R.R.K.). This work was also supported in part by National Science Foundation Plant Genome Grant DBI-0077713 (J.E.M. and P.E.K) and DBI-0321578 (P.E.K., R.R.K. and J.E.M.).

References

Allen E, Xie Z, Gustafson AM, Sung G-H, Spatafora JW, Carrington JC (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in*Arabidopsis thaliana*. Nature Genet 36:1282–1290

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12:1021–1029

Bennetzen JL (2002) Opening the door to comparative plant biology. Science 296:60–63

Bennetzen JL, Ma J (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. Curr Opin Plant Biol 6:128–133

Bennetzen JL, Ramakrishna W (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol 48:821–827

Bentolila S, Alfonso AA, Hanson MR (2002) A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci USA 99:10887–10892

Bonhomme S, Budar F, Lancelin D, Small I, Defrance MC, Pelletier G (1992) Sequence and transcript analysis of the Nco2. Mol Gen Genet 235:340–348

Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lennington J, Li ZK, Lin YR, Liu SC, Luo LJ, Marler BS, Ming R, Mitchell SE, Qiang D, Reischmann K, Schulze SR, Skinner DN, Wang YW, Kresovich S, Schertz KF, Paterson AH (2003a) A high-density genetic recombination map of sequence-tagged sites for *Sorghum*, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics 165:367–386

Bowers JE, Chapman BA, Rong JK, Paterson AH (2003b) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Brown GG (1999) Unique aspects of cytoplasmic male sterility and fertility restoration in *Brassica napus*. J Hered 90:351–356

Brown GG, Formanova N, Jin H, Wargachuk R, Dendy C, Patil P, Laforest M, Zhang JF, Cheung WY, Landry BS (2003) The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. Plant J 35:262–272

Brueggeman R, Rostoks N, Kudrna D, Kilian A, Han F, Chen J, Druka A, Steffenson B, Kleinhofs A (2002) The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. Proc Natl Acad Sci USA 99:9328–9333

Chen M, SanMiguel P, Bennetzen JL (1998) Sequence organization and conservation in *sh2/a1*-homologous regions of sorghum and rice. Genetics 148:435–443

Cui X, Wise RP, Schnable PS (1996) The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. Science 272:1334–1336

Desloire S, Gherbi H, Laloui W, Marhadour S, Clouet V, Cattolico L, Falentin C, Giancola S, Renard M, Budar F, Small I, Caboche M, Delourme R, Bendahmane A (2003) Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family. EMBO Rep 4:588–594

Draye X, Lin YR, Qian XY, Bowers JE, Burow GB, Morrell PL, Peterson DG, Presting GG, Ren SX, Wing RA, Paterson AH (2001) Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. Plant Physiol 125:1325–1341

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. Genome Res 8:175–185

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Littlejohn TG, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger G (1998) Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res 26:865–878

Han F, Kilian A, Chen JP, Kudrna D, Steffenson B, Yamamoto K, Matsumoto T, Sasaki T, Kleinhofs A (1999) Sequence analysis of a rice BAC covering the syntenous barley *Rpg1*region. Genome 42:1071–1076

Hanson MR, Bentolila S (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell 16(Suppl):S154–S169

Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16:369–372

Ilic K, SanMiguel PJ, Bennetzen JL (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci USA 100:12265–12270

Iwabuchi M, Kyozuka J, Shimamoto K (1993) Processing followed by complete editing of an altered mitochondrial atp6 RNA restores fertility of cytoplasmic male sterile rice. EMBO J 12:1437–1446

Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M (2002) Utility and distribution of conserved noncoding sequences in the grasses. Proc Natl Acad Sci USA 99:6147–6151

Kazama T, Toriyama K (2003) A pentatricopeptide repeat-containing gene that promotes the processing of aberrant *atp6*RNA of cytoplasmic male-sterile rice. FEBS Lett 544:99–102

Kilian A, Chen J, Han F, Steffenson B, Kleinhofs A (1997) Towards map-based cloning of the barley stem rust resistance genes *Rpg1* and *rpg4* using rice as an intergenomic cloning vehicle. Plant Mol Biol 35:187–195

Kim JS, Klein PE, Klein RE, Price HJ, Mullet JE, Stelly DM (2005a) Chromosome identification and nomenclature of *Sorghum bicolor*. Genetics 169:1169–1173

Kim JS, Klein PE, Klein RR, Price HJ, Mullet JE, Stelly DM (2005b) Molecular cytogenetic maps of sorghum linkage groups 2 and 8. Genetics 169:955–965

Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong J, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, Mullet JE (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res 10:789–807

Klein RR, Klein PE, Chhabra AK, Dong J, Pammi S, Childs KL, Mullet JE, Rooney WL, Schertz KF (2001) Molecular mapping of the *rf1* gene for pollen fertility restoration in sorghum (*Sorghum bicolor* L.). Theor Appl Genet 102:1206–1212

Klein PE, Klein RR, Vrebalov J, Mullet JE (2003) Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1 reveals extensive conservation of gene order and one major chromosomal rearrangement. Plant J 34:605–621

Koizuka N, Imai R, Fujimoto H, Hayakawa T, Kimura Y, Kohno-Murase J, Sakai T, Kawasaki S, Imamura J (2003) Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish. Plant J 34:407–415

Komori T, Ohta S, Murai N, Takakura Y, Kuraya Y, Suzuki S, Hiei Y, Imaseki H, Nitta N (2004) Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa*L.). Plant J 37:315–325

Laser KD, Lersten NR (1972) Anatomy and cytology of microsporogenesis in cytoplasmic male sterile angiosperms. Bot Rev 38:425–454

Li W, Gill BS (2002) The colinearity of the *Sh2/A1*orthologous region in rice, sorghum and maize is interrupted and accompanied by genome expansion in the *Triticeae*. Genetics 160:1153–1162

Liu F, Cui X, Horner HT, Weiner H, Schnable PS (2001) Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. Plant Cell 13:1063–1078

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82:378–389

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

Meierhoff K, Felder S, Nakamura T, Bechtold N, Schuster G (2003) HCF152, an*Arabidopsis*RNA binding pentatricopeptide repeat protein involved in the processing of chloroplast*psbB-it psbT-it psbH-it petB-it petD*RNAs. Plant Cell 15:1480–1495

Menz MA, Klein RR, Mullet JE, Obert JA, Unruh NC, Klein PE (2002) A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Mol Biol 48:483–499

Mullet JE, Klein RR, Klein PE (2001) *Sorghum bicolor*—an important species for comparative grass genomics and a source of beneficial genes for agriculture. Curr Opin Plant Biol 5:118–121

Nakamura T, Meierhoff K, Westhoff P, Schuster G (2003) RNA-binding properties of HCF152, an *Arabidopsis*PPR protein involved in the processing of chloroplast RNA. Eur J Biochem 270:4070–4081

Paterson AH, Bowers JE, Chapman BA (2004a) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA 101:9903–9908

Paterson AH, Bowers JE, Chapman BA, Peterson DG, Rong JK, Wicker TM (2004b) Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. Curr Opin Biotechnol 15:120–125

Pring DR, Van TH, Schertz KF (1995) Cytoplasmic male sterility and organelle DNAs of sorghum. In: Levings CS III, Vasil IK (eds) Advances in cellular and molecular biology of plants. Kluwer Academic, Dordrecht, pp 461–495

Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics 162:1389–1400

Sakata K, Nagamura Y, Numa H, Antonio BA, Nagasaki H, Idonuma A, Watanabe W, Shimizu Y, Horiuchi I, Matsumoto T, Sasaki T, Higo K (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. Nucleic Acids Res 30:98–102

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10:577–586

Small ID, Peeters N (2000) The PPR motif—a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25:46–47

Song R, Llaca V, Messing J (2002) Mosaic organization of orthologous sequences in grass genomes. Genome Res 12:1549–1555

Tang HV, Pring DR, Shaw LC, Salazar RA, Muza FR, Yan B, Schertz KF (1996) Transcript processing internal to a mitochondrial open reading frame is correlated with fertility restoration in male-sterile sorghum. Plant J 10:123–133

Tang HV, Chen W, Pring DR (1999) Mitochondrial *orf107* transcription, editing, and nucleolytic cleavage conferred by the gene *Rf3* are expressed in sorghum pollen. Sex Plant Reprod 12:53–59

Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous*adh* regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409–7414

Tsuchiya N, Fukuda H, Nakashima K, Nagao M, Sugimura T, Nakagama H (2004) LRP130, a single-stranded DNA/RNA-binding protein, localizes at the outer nuclear and endoplasmic reticulum membrane, and interacts with mRNA in vivo. Biochem Biophys Res Commun 317:736–743

Tu Z (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc Natl Acad Sci USA 98:1699–1704