Theor Appl Genet (2006) 112: 808–817
DOI 10.1007/s00122-005-0182-1

ORIGINAL PAPER

Swarup K. Parida · K. Anand Raj Kumar · Vivek Dalal
Nagendra K. Singh · Trilochan Mohapatra

Unigene derived microsatellite markers for the cereal genomes

Received: 22 July 2005 / Accepted: 30 November 2005 / Published online: 21 January 2006

Abstract Unigene derived microsatellite (UGMS) markers have the advantage of assaying variation in the expressed component of the genome with unique identity and positions. We characterized the microsatellite motifs present in the unigenes of five cereal species namely, rice, wheat, maize, *Sorghum* and barley and compared with those in *Arabidopsis*. The overall UGMS frequency in the five cereal species was 1/7.6 kb. The maximum UGMS frequency was in rice (1/3.6 kb) and the lowest in wheat (1/10.6 kb). GC-rich trinucleotide repeat motifs coding for alanine followed by arginine and the dinucleotide repeat motif GA were found to be abundant UGMS classes across all the five cereal species. Primers could be designed for 95% (wheat and barley) to 97% (rice) of the identified microsatellites. The proportion and frequency of occurrence of long hypervariable class I (≥20 nucleotides) and potentially variable class II (12–20 nucleotides) UGMS across five cereal species were characterized. The class I UGMS markers were physically mapped in silico on to the finished rice genome and bin-mapped in wheat. Comparative mapping based on class I UGMS markers in rice and wheat revealed syntenic relationships between the two genomes. High degree of conservation and cross-transferability of the class I UGMS markers were evident among the five cereal species, which was validated experimentally. The class I UGMS-conserved orthologous set (COS) markers identified in this study would be useful for understanding the evolution of genes and genomes in cereals.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-005-0182-1 and is accessible for authorized users.

Communicated by F. Salamini

S. K. Parida · K. Anand Raj Kumar · V. Dalal
N. K. Singh · T. Mohapatra (✉)
National Research Centre on Plant Biotechnology,
Indian Agricultural Research Institute,
110012 New Delhi, India
E-mail: tm@nrcpb.org
Fax: +91-11-25843984

Introduction

Microsatellites or simple sequence repeats (SSRs) are tandemly arranged repeats of 1–6 nucleotide long DNA motifs that frequently exhibit variation in the number of repeats at a locus. Microsatellite markers are reliable, co-dominant, multi-allelic, chromosome-specific and highly informative genetic markers well established for genetic analysis in crop plants. These are amenable to high-throughput genotyping and thus suitable for construction of high-density linkage maps, gene mapping and marker-assisted selection. The microsatellite linkage maps have been developed in many plant species including major cereals namely, wheat (Roder et al. 1998), rice (McCouch et al. 2002) and maize (Sharopova et al. 2002).

In the past, the advantages of microsatellite markers were partially offset by difficulties in the marker development, as laborious iterations of genomic DNA library screening with SSR probes and sequencing of a large number of SSR positive clones (Panaud et al. 1996; Chen et al. 1997) were involved. Availability of complete genome sequence in rice has eliminated the above steps and enabled the development of a large number of SSR markers in silico (McCouch et al. 2002). Another source of SSR markers is the expressed sequence tag (EST) databases, which are continuously growing in size in most crop species (http://www.ncbi.nlm.nih.gov). These EST databases can be mined for microsatellite motifs that would serve as locus-specific markers. Development of EST-based microsatellite markers thus would involve considerably lower cost and effort. Besides, such markers being derived from the conserved expressed component of the genome are expected to show greater cross-transferability between species and genera (Varshney

et al. 2005). The unavailability of genomic and cDNA sequences, thus would not limit the development of robust microsatellite markers and various marker-based applications in related species and genera.

A major disadvantage of the EST-derived microsatellites is the sequence redundancy that yields multiple sets of markers at the same locus. However, more recently the random EST sequences are being assembled into unique gene sequences called unigenes (http://www.ncbi.nlm.nih.gov) that circumvents the problem of redundancy in EST databases. For example, 41,600 EST sequences in barley have been assembled into just 1,240 unigenes. The unigene-based microsatellite markers (UGMS markers) would therefore, have the advantages of unique identity and positions in the transcribed regions of the genome. With the availability of large unigene databases, it is now possible to systematically search for microsatellites in the unigenes. The UGMS markers can be used for accurately assaying functional diversity in the natural populations and the available germplasm collections as well as for comparative mapping and evolutionary studies as anchor markers. In the present study, unigene databases of five major cereal species namely, rice, wheat, maize, *Sorghum* and barley were mined for the presence of microsatellites. The nature of these microsatellites was analysed and compared with that based on the EST sequences in the respective crops and in *Arabidopsis thaliana*. Primers flanking these microsatellite motifs were designed for genotyping applications. Physical and comparative mapping of the UGMS markers with longer microsatellite repeat motifs (≥20 nucleotides) was carried out in rice and wheat. In silico analysis of conservation and cross-transferability of these markers was also carried out to study their utility in comparative mapping of genes and genomes.

Materials and methods

Accessing unigenes and mining of microsatellites

Unigenes comprising 48.8 Mb in rice (*Oryza sativa*), 34 Mb in *Arabidopsis*, 17.7 Mb in wheat (*Triticum aestivum*), 13.3 Mb in maize (*Zea mays*), 9.5 Mb in barley (*Hordeum vulgare*) and 5.8 Mb in *Sorghum* (*Sorghum bicolor*) were acquired in bulk through a FASTA sequence retrieval system from GenBank using Batch Entrez (http://www.ncbi.nlm.nih.gov) along with their constituent EST sequences. These sequences were mined for microsatellites using a program *MISA* (MIcroSAtellite) written in the Perl 5 scripting language that locates microsatellite patterns in FASTA formatted sequence files and reports the GenBank ID, microsatellite motifs (monomers to hexamers), number of repeats and sequence coordinates for each microsatellite. The *MISA* script is available from the IPK website (http://www.pgrc.ipk-gatersleben.de/misa), which is capable of identifying both perfect and compound microsatellites. The compound microsatellites were categorized into non-interrupting [e.g., (CTC)5(CT)17] and interrupting [maximum 100 nucleotides interrupting two microsatellites, e.g., [(CTC)5AGTCTACTGA(CT)17] groups. While classifying the microsatellites into different repeat categories, sequence complementarity was considered, e.g., repeat motifs AG, GA, TC and CT were put in the same class. Microsatellites were classified into class I (≥20 nucleotides) and class II (12–20 nucleotides) based on the length of the microsatellite motifs (Temnykh et al. 2000).

Primer design

The microsatellite (excluding monomers) containing unigenes were used for designing primers employing the microsatellite primer discovery tool (http://www.hornbill.cspp.latrobe.edu.au) in a batch module manner. Optimum and maximum primer sizes were 22 and 24 nucleotides, respectively, and the other options were left on default mode. The software is capable of designing both forward and reverse primers and detecting their T_m, GC (%) and pair complementarity values at 3′ and 5′ ends.

Physical mapping of UGMS markers in rice and wheat

Class I microsatellite containing unigenes of rice were subjected to a local BLASTN search (http://www.192.168.1.12) against the International Rice Genome Sequencing Project (IRGSP) pseudomolecule databases of RGP (http://www.rgp.dna.affrc.go.jp/IRGSP/Build3) to display the physical location (bp) of each microsatellite marker in each of the 12 chromosomes of rice. Optimum settings of the BLAST search options were used for analysis, since the default search parameters being highly stringent were picking up only short sequence stretches of high homology. Therefore, advanced BLAST search parameters with optimized value of open gap cost ($G = 5$), extended gap cost ($E = 1$), penalty for a mismatch ($q = 1$), reward for a match ($r = 1$) and low complexity filter options were used for analysis (Singh et al. 2004). Only top hits of the BLASTN search results ($e = 0$) were plotted on specific locations in the rice chromosomes. Class I microsatellite containing unigenes of wheat were BLAST searched against GrainGenes databases (http://www.wheat.pw.usda.gov/GG2/blast.html) of wheat EST sequences, bin-mapped earlier using deletion stocks of *cv. Chinese spring*. Optimized BLASTN search parameters (homology filter of expectation value ≤ 1e-15 and bit score 100) were used to find the bin-mapped wheat EST sequences. The matching microsatellite markers were then plotted on specific bins of the wheat chromosomes. For comparative mapping of wheat markers in rice, the microsatellite containing wheat unigenes mapped to homoeologous chromosome groups 1 and 3 (that carried

more UGMS markers than other groups) were used. These were analysed using a local BLASTN search tool (http://www.192.168.1.12) against the rice IRGSP pseudomolecule databases of RGP (http://www.rgp.dna.affrc.go.jp/IRGSP/Build3). For comparative mapping of rice UGMS markers in wheat, the class I microsatellite containing unigenes of rice were BLAST searched against GrainGenes databases (http://www.wheat.pw.usda.gov/GG2/blast.html) of wheat. Optimized BLAST search parameters as mentioned above for wheat physical mapping were used for this analysis.

Assessment of conservation and cross-transferability

To examine the conservation and cross-transferability of microsatellite repeat motifs, the class I microsatellite containing unigenes of a species were BLAST searched against another using the local BLASTN search tool (http://www.192.168.1.12) and the matches with *e* values of ≤ 1e-15, ≤ 1e-141 and 0 were extracted to Excel sheets using the Perl script "mblastex" (developed by V. Dalal and N.K. Singh, unpublished). The putative functions of these conserved motifs and physically mapped markers in rice and wheat were determined using BLASTX search tool (http://www.genome.ad.jp) against the nr-protein databases. The BLASTX output was annotated and extracted to Excel sheets using the Perl script "annotate" (developed by V. Dalal and N.K. Singh, unpublished). Optimized BLAST search parameters with a cutoff bit score 100 (≤ 1e-15) were considered for this analysis.

Results and discussion

Frequency of microsatellites

The EST databases of the five cereal species and *Arabidopsis* were highly redundant with regard to the size and the number of sequences analysed, and the frequency of the microsatellites contained in these sequences (Table 1). The overall frequency of the microsatellites excluding monomers in the unigenes of the five cereal species was 1/7.6 kb as against 1/6.4 kb in the EST sequences. The discrepancy in the frequency is due to the elimination of redundant sequences in the unigenes. The estimated redundancy was least in rice among the six species characterized. The unigene sequences, therefore, more accurately reflected the density of microsatellites in the expressed component of the genomes. It was maximum in rice (1/3.6 kb), followed by *Sorghum* (1/5.9 kb), *Arabidopsis* (1/7.5 kb), barley (1/8.9 kb), maize (1/9 kb) and wheat (1/10.6 kb). The percentage of unigenes that contained microsatellites excluding the monomers followed a similar trend as above (Table 1). When the monomeric microsatellites were included, the trend was altered with maize having the highest proportion (69.5%) of unigenes carrying microsatellites followed by wheat (65.5%), rice (58.8%), *Arabidopsis* (38.8%), barley (37%) and *Sorghum* (21.5%). The compound microsatellites were maximum in barley (8.6%), followed by rice (8.4%), wheat (5.9%), *Arabidopsis* (4.6%), *Sorghum* (1.2%) and maize (0.6%) (Table 1). These observations thus revealed considerable variations in the microsatellite sequence organization in the genic regions of the five cereal species and *Arabidopsis*.

Distribution of UGMS classes

The trinucleotide repeat motifs were the most prevalent class of microsatellites in the unigenes of the five cereal species as well as in *Arabidopsis*. The proportion of trinucleotide containing unigenes ranged from 60% in maize to 80% in rice (Supplementary Table 1). Trinucleotide motifs were followed by dinucleotide (17.6–35.2%), tetranucleotide (0.65–6.1%), pentanucleotide (0.15–1.2%) and hexanucleotide (0.27–0.6%) repeat motifs (Supplementary Table 1). This is similar to the earlier observations on the relative abundance of trinucleotide motifs in the EST sequences of cereals (Varshney et al. 2002; Kantety et al. 2002). Higher frequency of the trinucleotide repeat motifs than the other classes could be attributed to the selection against frameshift mutations that limits expansion of nontriplet microsatellites (Metzgar et al. 2000). Our analysis in rice revealed that 85% of the trinucleotide repeats were found within the ORFs, 12% in 5′ UTRs and 3% in 3′ UTRs and that 26–37% of the trinucleotide repeat containing genes across cereals were related to transcription and signal transduction. Any alteration in the sequence of these genes would have significant negative consequences (Young et al. 2000). In contrast, the dinucleotide repeats were more frequently observed in the 5′ (48%) and 3′ (30%) UTRs than the coding regions (22%) in rice. Microsatellite repeat motifs found in the 5′ UTRs could have the potential to function as factors in regulating gene expression (Bao et al. 2002; Fujimori et al. 2003).

Abundance of GA-rich dinucleotide and GC-rich trinucleotide microsatellites

Among the dinucleotide repeats, GA was most common that ranged from 15% in barley to 21.2% in rice while in *Arabidopsis* it was 32.4% (Supplementary Table 1). The least abundant dinucleotide motif was CG that ranged from 2.5% in barley to 3% in rice; while in *Arabidopsis* it was GT (1.7%). These results are in close agreement with earlier studies, where the GA motif was reported to be the dominant dinucleotide repeat in the EST-derived microsatellites in cereals (Varshney et al. 2002). These abundant dinucleotide repeat motifs are known to be the best source of useful microsatellite markers possibly because of their occurrence in regions with a balanced

Table 1 Comparative analysis on the distribution of microsatellites in ESTs and unigenes of five cereal species and *Arabidopsis*

Characters under study	Rice			Wheat			Maize			*Sorghum*			Barley			*Arabidopsis*
	ESTs	Unigenes	RF*	ESTs	Unigenes	RF*	ESTs	Unigenes	RF*	ESTs	Unigenes	RF*	ESTs	Unigenes	RF*	Unigenes
Total number of sequence examined	298,808	33,722	8.9	586,577	24,854	23.6	416,674	14,277	29.1	110,835	8,146	13.6	391,948	12,220	32.0	21,133
Total size (bp) of examined sequences	158,551, 602	48,840, 791	3.3	325,522, 991	17,703, 941	18.4	192,253, 889	13,283, 403	14.5	65,706, 682	5,829, 662	11.3	209,488, 807	9,491, 839	22.0	34,075, 702
Total number of identified microsatellites	172,075 (45.7%)	27,665 (58.8%)	6.2	155,675 (21.2%)	19,332 (65.5%)	8.0	115,349 (23.8%)	12,309 (69.5%)	9.4	63,825 (41.1%)	2,045 (21.5%)	31.2	79,371 (16.3%)	6,242 (37%)	12.7	11,129 (38.8%)
Number of SSR containing sequences	136,702	19,845	6.9	124,185	16,300	7.6	99,458	9,932	10.0	45,578	1,752	26.0	63,903	4,522	14.1	8,204
Number of sequences containing more than one SSR	21,547	5,445	4.0	18,645	2,253	8.3	13,538	1,918	7.0	18,247	262	69.6	8,786	984	8.9	2,174
Number of compound microsatellites	22,264 (7.4%)	2,862 (8.4%)	7.8	20,296 (3.4%)	1,469 (5.9%)	13.8	8,210 (2%)	886 (0.6%)	9.3	2,231 (2%)	104 (1.2%)	21.5	10,373 (2.6%)	1,057 (8.6%)	9.8	974 (4.6%)
Total number of monomers	137,130 (45.9%)	14,028 (41.6%)	9.8	112,940 (19.3%)	17,661 (71%)	6.4	90,329 (21.7%)	10,836 (75.8%)	8.3	6,101 (5.6%)	1,052 (13%)	5.8	51,279 (13%)	5,181 (42.4%)	9.9	
Total number of microsatellites excluding monomers	34,945 (20.3%)	13,637 (40.4%)	2.6	42,735 (27.4%)	1,671 (6.7%)	25.6	25,020 (21.7%)	1,473 (10.3%)	16.7	24,380 (22%)	993 (12.2%)	24.6	28,092 (7.1%)	1,061 (8.7%)	26.4	4,604 (21.7%)
Size (kb) of sequences containing one microsatellite	5	3.6	1.3	7.6	10.6	1.4	7.7	9	1.2	4.3	5.9		7.5	8.9	1.2	7.5
Number of primer pairs for unigene-microsatellites	–	13,230 (97%)	–	–	1,588 (95%)	–	–	1,401 (95.2%)	–	–	952 (96.8%)	–	–	1,009 (95%)	–	–

RF* (redundancy factor) = ESTs/unigenes

GC content (Temnykh et al. 2000; Cho et al. 2000) and absence of any association with the transposable elements (Temnykh et al. 2001). In the unigenes of rice, we also observed $(GA)_n$ polynucleotides usually in regions with balanced (40–50%) GC content, which favours robust PCR amplification (Temnykh et al. 2001) and therefore could be used for efficient genotyping applications.

Among the trinucleotide repeats, the motifs GCA/GCC/GCG/GCT coding for amino acid alanine were most abundant across all the cereal species (Supplementary Table 2) that ranged from 16.6% in barley to 25.7% in rice. The next abundant trinucleotide motifs were AGA/AGG/CGA/CGC/CGG/CGT coding for arginine that ranged from 14.6% in wheat to 20.4% in *Sorghum*. High frequency of (CCA/CCG/CCT) encoding proline was also found in rice (13.8%), maize (9.5%), *Sorghum* (11.8%) and barley (11.2%). The abundance of GC-rich repeats in the monocot genomes is reflected in their high GC content (Morgante et al. 2002) particularly in the coding sequences. In contrast, we observed AGC/AGT/TCA/TCC/TCG/TCT (16.6%) coding for serine as the most abundant motifs in *Arabidopsis*, followed by glutamic acid (GAA/GAG, 12.3%) and leucine (CTA/CTC/CTG/CTT/CTC/TTA/TTG, 10.9%) (Supplementary Table 2). The abundance of small/hydrophilic amino acid repeat motifs like that of alanine and serine in the unigenes of cereals and *Arabidopsis* was perhaps because these are tolerated in many proteins, while strong selection pressure possibly eliminates codon repeats encoding hydrophobic/other amino acids (Katti et al. 2001). This observation suggested that considerable sequence divergence, since their early separation about 200 million year ago (Wolfe et al. 1989), between monocot and dicot has led to differential amino acid repeat motifs in the proteins, and that the selection has played a significant role in greater retention of those which are tolerated more.

Class I microsatellites in the unigenes

The frequency of class I UGMS was highest in rice (1/17.1 kb), followed by *Sorghum* (1/21.4 kb), barley (1/31.6 kb), *Arabidopsis* (1/37.8 kb), wheat (1/42.2 kb) and maize (1/48.1 kb). The maximum number of class I UGMS was found in rice (2,851), followed by *Arabidopsis* (900), wheat (429), barley (300), maize (276) and *Sorghum* (272). This correlated with the size of the unigene database available in these species. In terms of the proportion of the total number of microsatellites found in the unigenes, it was highest in barley (30.2%), followed by *Sorghum* (27.4%), wheat (25%), rice (21%), *Arabidopsis* (19.5%) and maize (18.7%) (Table 2). The class I trinucleotide repeats expressed as the percentage of all the class I UGMS varied from 39.4% in wheat to 67.8% in rice. The proportion of class I dinucleotide repeats ranged from 19.2% in rice to 35.3% in wheat. It appeared that the potential of microsatellite expansion

Table 2 Proportionate distribution of class I and class II UGMS motifs in five cereal species and *Arabidopsis*

Repeat length	Rice		Wheat		Maize		*Sorghum*		Barley		*Arabidopsis*	
	Class I	Class II	Class I	Class II	Class I	Class II	Class I	Class II	Class I	Class II	Class I	Class II
Dinucleotides	546 (19.2%)	1,939 (17.6%)	148 (35.3%)	413 (30.8%)	90 (32.6%)	456 (36.1%)	54 (20%)	171 (21%)	93 (31%)	252 (30.1%)	313 (34.8%)	1,432 (37%)
Trinucleotides	1,933 (67.8%)	8,937 (81%)	169 (39.4%)	872 (65%)	114 (41.3%)	770 (61%)	156 (57.3%)	614 (75.2%)	129 (43%)	534 (63.8%)	544 (60.4%)	2,424 (62.5%)
Tetranucleotides	237 (8.3%)	169 (1.5%)	87 (20.3%)	56 (4.2%)	54 (19.6%)	37 (3%)	44 (16.1%)	31 (3.7%)	65 (21.7%)	50 (6%)	30 (3.3%)	23 (0.6%)
Pentanucleotides	68 (2.4%)	0	17 (4%)	0	12 (4.3%)	0	12 (4.4%)	0	9 (3%)	0	7 (0.8%)	0
Hexanucleotides	67 (2.3%)	0	8 (2%)	0	6 (2.1%)	0	6 (2.2%)	0	4 (1.3%)	0	6 (0.7%)	0
Total	2,851 (21%)	11,045 (81%)	429 (26%)	1,341 (80%)	276 (18.7%)	1,263 (85.7%)	272 (27.4%)	816 (82.2%)	300 (30.2%)	836 (78.8%)	900 (19.5%)	3,879 (84.2%)

in unigenes of five cereal species and *Arabidopsis* was not corresponding to their genome size/ploidy. For example, the diploid barley genome contained higher percentage of class I microsatellites as compared to the hexaploid wheat genome. The large diploid maize genome contained lesser percentage of class I UGMS than the rice genome, in spite of their closer evolutionary relationship.

Development of UGMS markers for the cereal genomes

To develop UGMS markers, primer pairs (forward and reverse) were designed from the flanking regions of the identified microsatellites in all the five cereal species. We designed 13,230 primer pairs in rice, 1,588 in wheat, 1,401 in maize, 952 in *Sorghum* and 1,009 in barley. The primer sequences with their T_m values and product sizes constitute a table that is too large to be provided as supplementary information, but are available with us that can be provided on request. The primer pairs could be designed for 95% in wheat and barley to 97% in rice of the total microsatellites identified. Our failure to design primers for the rest was due to either short sequence of the unigenes and/or lack of unique flanking sequences. Varshney et al. (2002) estimated that only 53–71% of the total microsatellites in the ESTs of cereal species had primer designing potential. The unigenes being longer in higher quality sequences thus offered advantages over the ESTs for the development of microsatellite markers. The designed primer pairs for the class I UGMS markers in rice (2,780), wheat (429), maize (273), *Sorghum* (266) and barley (297) have been given in the Supplementary Excel sheet 1. The primer pairs for the unigenes with known functions across all the five cereal species have been given in the Supplementary Excel sheet 2. On average these identified functional genes in cereal genomes contained more of trinucleotide microsatellite motifs and related mostly to transcription and translation factors, signal transduction genes and metabolic enzymes. In rice and wheat, the class I UGMS markers were present in genes predicted to have control over different structural, biochemical and morphological traits. Among these, the predominant classes were metabolic enzymes (36% in rice, 33% in wheat), structural proteins (34%, 30%), transcription and translation factors (22%, 13%), signal transduction genes (16%, 7%), cell growth and development factors (8%, 6%) and disease resistance genes (6%, 2%). These genic class I microsatellite markers would be useful for connecting genetic maps with physical maps, genomic sequences and ultimately with phenotypic variation.

Genome coverage and density of mapped class I UGMS markers in rice

Out of 2,851 class I UGMS markers in rice, 2,622 (92%) showed significant hit ($e = 0$) on the rice chromosomes. The matches were considered absent when the bit score was less than 100. Most of these markers (2,140) had unique positions in the rice genome. Multiple locations were observed for 174 markers of which six mapped to all the 12 rice chromosomes. The physical positions with details of BLAST results have been given in the Supplementary Excel sheet 3.

The overall map density of class I UGMS markers for the whole rice genome was estimated as 141 kb and thus considered as a high-density physical map. The size of the marker interval varied from 174 bp to 3.26 Mb (Supplementary Table 3). Chromosome 3 had maximum UGMS markers (364) and thus the highest average map density (99 kb) followed by chromosome 1 (355 markers, 121 kb). It could be due to their greater physical size (http://www.rgp.dna.affrc.go.jp/IRGSP/Build3), presence of least heterochromatinized sequences and least transposon association as compared to the other rice chromosomes. The least number of markers (128) was placed on chromosome 11 (Supplementary Table 3) giving the lowest map density (221 kb) that might be due to higher transposon association, even though it had larger physical size than chromosomes 9 and 10. Large physical gaps (≥1 Mb) present in the UGMS marker based rice physical map can be attributed to the absence of class I microsatellites in the genes present in these intervals and/or lower frequency of the available unigenes corresponding to these regions.

We constructed a framework physical map of rice genome (Supplementary Figure 1) consisting of 825 class I UGMS markers, with an average interval of 500 kb. The chromosome 10 framework map for instance, consists of 56 well distributed markers at an average interval of about 405 kb (Fig. 1). These markers would be of immediate use in various applications including tagging of useful genes, marker-assisted background as well as foreground selection, germplasm characterization and variety identification in rice.

Density and anomalies of mapped class I UGMS markers in wheat-bins

Of the 429 class I UGMS markers developed in wheat, 157 (37%) mapped to wheat chromosome bins (Supplementary Excel sheet 1), while the rest did not find any significant homology with the bin-mapped ESTs. Fifty-six (36%) wheat UGMS markers mapped to unique chromosome bins and the rest mapped to two or more bins revealing intra and inter chromosomal duplications (Supplementary Table 4). A maximum of 47 (30%) UGMS markers were located on the group 3 homoeologous chromosome bins giving 94 loci with a mean of 2.00 loci per bin (Supplementary Table 4). The number of mapped loci, however, varied from chromosome to chromosome with an average of 1.96 loci per bin across all the homoeologous chromosomes (Supplementary Figure 2). There were anomalies in terms of a lack of correspondence of the order of the

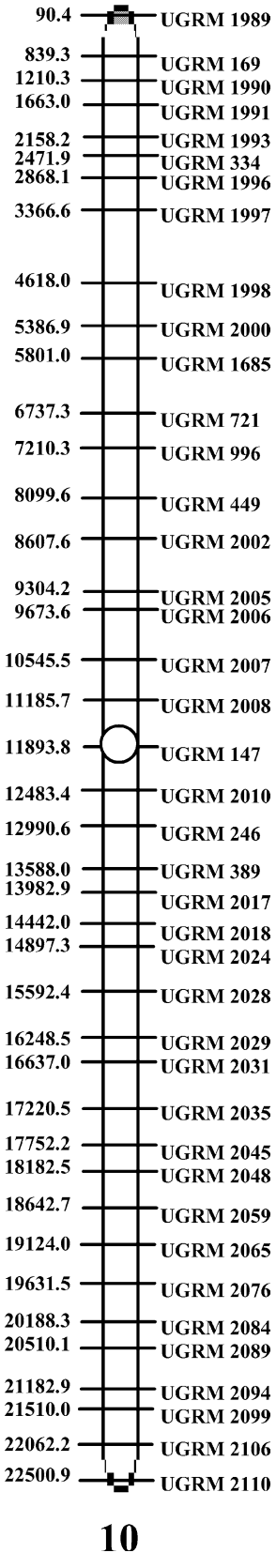

Fig. 1 A framework physical map of rice chromosome 10 based on 56 unigene derived class I microsatellite markers placed at an average interval of 405 kb. The positions of the markers in kb are indicated on the left side. The identities of the markers are given on the right side that corresponds to the GenBank Unigene ID of NCBI as given in the Supplementary Excel sheet 3. *UGRM* **U**ni**G**ene derived **R**ice **M**icrosatellites. The size of the chromosome

mapped class I UGMS markers in the homoeologous chromosome bins. A maximum of 17 such anomalies was observed in the group 3 homoeologous chromosome bins (Supplementary Figure 3), possibly due to the presence of a greater number of mapped loci. Of these, eight were between chromosome 3A and 3B, two were between 3A and 3D and seven were between 3B and 3D. These could be the result of actual biological events such as chromosomal rearrangements, transposition and gene duplications. There was also conservation of marker order across homoeologous groups as in the case of chromosomes 4, 5 and 7, which could be due to translocations involving 4AL, 5AL and 7BS chromosome arms (Nelson et al. 1995). Another type of anomaly was observed among 4AL, 4BS and 4 DS chromosome arms that might be due to pericentric inversion in the chromosome 4A reported earlier by Mickelson-Young et al. (1995). These bin-mapped wheat UGMS markers provide invaluable information for a targeted mapping of genes for useful traits, comparative genomics and sequencing of gene-rich regions of the wheat genome.

Comparative mapping using class I UGMS markers between rice and wheat

Comparative mapping was done by evaluating the significant homeology of mapped class I UGMS markers between rice and wheat. Sorrells et al. (2003) compared the bin-mapped ESTs of all homoeologous groups of wheat with the predicted gene sequences of rice and showed that 81% of the rice BAC/PAC clones were matched by wheat ESTs. In the present study, analysis of the BLASTN results for the wheat groups 1 and 3 chromosome bin maps against the rice genome (Supplementary Excel sheet 4) indicated that the mapped markers in group 1 chromosome bins shared the highest level of correspondence with rice chromosome 5 (65.4%), followed by chromosome 10 (32.4%), while the mapped markers in the group 3 chromosome bins revealed maximum correspondence with rice chromosome 1 (91.2%). Most of the UGMS markers in the proximal two bins (1AS3-0.85-1.00 and 1AS1-0.47-0.85) of the short arm and the distal two bins (1AL1-0.17-0.61 and 1AL3-0.61-1.00) of the long arm of chromosome 1A of wheat showed correspondence with the short arm of rice chromosome 5 in the same marker order, while the two bins (C-1AS1-0.47 and C-1AL1-0.17) in the centromeric region had correspondence with the long arm of rice chromosome 10 (in the same marker order). Markers in the 1B and 1D chromosome bins were

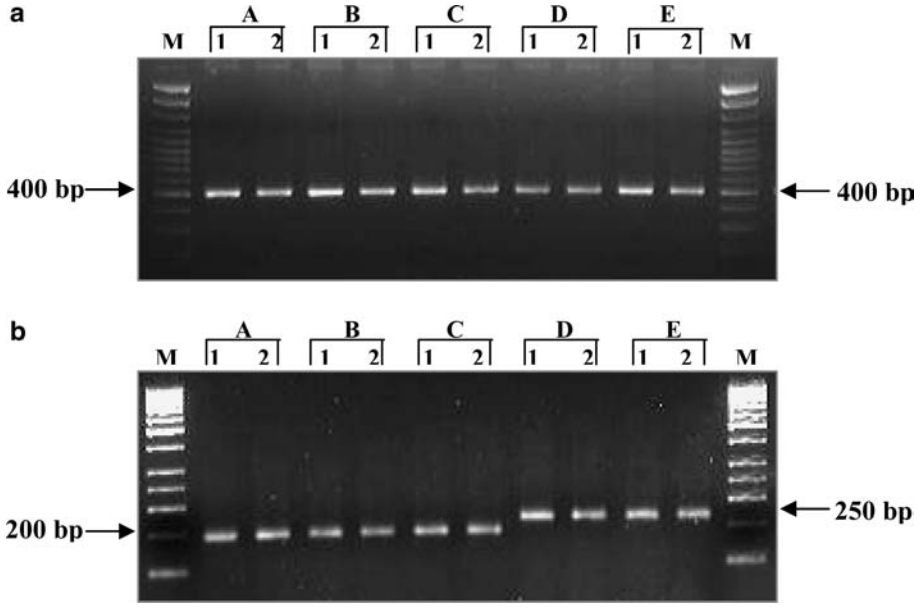

Fig. 2 Validation of conservation and cross-transferability of class I UGMS-COS markers designed from the unigenes for sucrose phosphate synthase **(a)** and β-tubulin **(b)** containing $(CGG)_7$ and $(GATC)_5$ microsatellite motifs, respectively. Two varieties were used in each of the five cereal species. *A* rice (1-IR64 and 2-IRBB60), *B* Maize (1-CA14509 and 2-CA14510), *C Sorghum* (1-Pusa Chari6 and 2-Pusa Chari9), *D* Wheat (1-Kalyansona and 2-Sonalika) and *E* Barley (1-K551 and 2-RD2505). *M* 100 bp DNA ladder plus. The marker based on sucrose phosphate synthase was monomorphic, while β-tubulin derived UGMS marker showed two alleles. A 200 bp fragment was amplified in rice, maize and *Sorghum*, whereas a 250 bp fragment was present in wheat and barley

mostly homoeologous to rice chromosomes 5 and 10 (Peng et al. 2004), but the correspondence did not follow a distinct pattern like chromosome 1A. The mapped UGMS markers in the entire 3A and 3B chromosome bins gave maximum synteny with rice chromosome 1. Six (75%) mapped markers in the distally located bin (3DL3-0.81-1.00) of the long arm of chromosome 3D were syntenic to rice chromosome 3, but the rest 17 (85%) of the bin-mapped markers were syntenic to chromosome 1 (Supplementary Figure 3).

Results from the comparison of rice physical maps against wheat revealed that for an individual rice chromosome, there is a clear correspondence with a specific homoeologous wheat chromosome (Supplementary Table 5). The mapped class I UGMS markers in rice chromosome 1 gave maximum synteny with wheat group 3 chromosomes, whereas rice chromosomes 2 and 3 were related to wheat groups 6 and 4, respectively (Supplementary Excel sheet 4). There was synteny of two rice chromosomes to a single wheat homoeologous chromosome. For instance, rice chromosomes 4 and 7 gave clear correspondence with wheat group 2, rice chromosomes 5 and 10 to wheat group 1 and rice chromosomes 6 and 8 to wheat group 7. The observed synteny between rice and wheat genomes is similar to the previous studies of Sorrells et al. (2003) and La Rota and Sorrells (2004) based on all rice genes. The identified rice UGMS markers syntenic to wheat chromosomes would be useful for transferring information about genes from the model species rice to wheat.

Conservation and cross-transferability of class I UGMS markers in the cereal genomes

The conservation and cross-transferability of identified class I UGMS markers across five cereal species as well as with *Arabidopsis* based on significant sequence homology is given in the Supplementary Table 6. At a bit score of 100 ($\leq$ 1e-15), the class I microsatellite containing unigenes of rice, wheat, maize, *Sorghum* and barley gave high degree of homology with maize (1,308, 47%), barley (221, 51.5%), *Sorghum* (244, 88.4%), maize (236, 87%) and wheat (195, 65%), respectively. However, under such conditions the conservation of the class I microsatellite repeat motifs and unigene sequences flanking these motifs in the syntenic cereal species was not found, as shown in the Supplementary Figure 4A. Our results are in contrast to the earlier studies on the identification and evaluation of the conserved orthologous set (COS) markers (Fulton et al. 2002) and interspecific cross-transferability (Varshney et al. 2005), where the existence of conservation was inferred based on very low expectation value ($\leq$ 1e-15) and bit score (100). When the bit score was increased to $\geq$500 ($\leq$ 1e-141), the conservation of the class I microsatellite motifs and their flanking unigene sequences improved, as shown in Supplementary Figure 4B. Based on this strategy, we identified the class I UGMS-COS markers. The COS markers of rice, wheat, maize, *Sorghum* and barley followed a similar homology trend as observed under lower stringency conditions (Supplementary Table 6). These relationships

are similar to the earlier phylogenic studies involving the cereal genomes, which described the orthologous and paralogous synteny among the Panicoideae (maize and *Sorghum*), Triticeae (wheat and barley) and the Oryzoideae (rice) groups (Gale and Devos 1998). The results further agreed well with contemporaneous divergence of the cereal genomes from a common ancestor (Paterson et al. 2004; Swigonova et al. 2004). A maximum of 9.5% of rice class I UGMS markers showed homology with maize genome followed by wheat (7.6%) indicating the existence of evolutionary closeness as suggested by Wolfe et al. (1989) that rice diverged from an ancestor of *Sorghum* and maize about 50 million years ago, whereas from an ancestor of wheat about 70 million years ago. In contrast, least homology was observed between the class I UGMS markers of cereals and *Arabidopsis*, as they diverged from each other 200 million years ago. However, among the cereal genomes the rice class I UGMS markers were more homologous to *Arabidopsis*. To validate our predicted conservation and cross-transferability, a set of 22 class I UGMS-COS markers of maize were used (Supplementary Table 7). All except four UGMS markers amplified the same size of DNA fragments (Fig. 2A) in all the five cereal species, whereas four primer pairs amplified different size of DNA fragments (Fig. 2B), thus experimentally validating our prediction. Our study thus identified class I UGMS-COS markers that are cross-transferable among the five cereal species and *Arabidopsis*. At high stringency conditions ($e=0$), we have identified a set of 18 rice class I UGMS-COS markers those were conserved across all the five cereal species and six markers conserved across all the five cereals and *Arabidopsis* (Supplementary Table 8). The UGMS-COS markers were identified in different genes associated with basic metabolic processes of energy generation, biosynthesis and degradation of cellular building blocks such as β-galactosidase, ATPase, ribosomal proteins and transcription factors. These markers would be useful for comparative mapping and phylogenetic analysis, and facilitate the development of syntenic networks across cereals necessary for understanding the evolution of genes, genomes and gene functions. A database of UGMS markers that would amplify orthologous loci across species would be very useful to breeders and geneticists, especially for minor or under-funded crop species belonging to the grass family.

In summary, our study revealed the structure and organization of microsatellites in the genic regions of five major cereal genomes, which could be compared with those of the dicot plant species *A. thaliana*. A large number of UGMS markers were developed for the five cereal genomes. The use of these markers would reduce the cost of future microsatellite marker development and facilitate the production of comparative genetic and physical maps, study of genetic diversity, gene mapping, marker-aided selection and eventually positional cloning of useful genes in cereals and other important members of the grass family.

Acknowledgements The work presented in the manuscript was partly carried out under the Sugarcane Genomics Project and National Bioscience Project awarded to the corresponding author (TM) by the Department of Biotechnology (DBT), Government of India. We are thankful to NCBI and RGP for making their databases available and Institute of Plant Genetics and Crop Research (IPK) for the availability of microsatellite search tool *MISA*.

References

Bao J, Corke H, Sun M (2002) Microsatellites in starch-synthesizing genes in relation to starch physicochemical properties in waxy rice (*Oryza sativa* L.). Theor Appl Genet 105:898–905

Chen X, Temnykh S, Xu Y, Cho YG, McCouch SR (1997) Development of a microsatellite framework map providing genome-wide coverage in rice (*Oryza sativa* L.). Theor Appl Genet 95:553–567

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, Park WD, Ayres N, Cartinhour S, McCouch SR (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:713–722

Fujimori S, Washio T, Higo K, Ohtomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S, Tomita M (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. FEBS Lett 554:17–22

Fulton TM, Hoeven RV, Eannetta NT, Tanksley SD (2002) Identification, analysis and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

Gale DM, Devos KM (1998) Comparative genetics in grasses. Proc Natl Acad Sci USA 95:1971–1974

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, *Sorghum* and wheat. Plant Mol Biol 48:501–510

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

La Rota M, Sorrells ME (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals complexity of genome relationships between rice and wheat. Funct Integr Genomics 4:34–46

McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing Y, Zhang Q, Kono I, Yano M, Fjellstrom R, DeClerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development of 2,240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:199–207

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72–80

Mickelson-Young L, Endo TR, Gill BS (1995) A cytogenetic ladder-map of the wheat homoeologous group 4 chromosomes. Theor Appl Genet 90:1007–1011

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Nelson JC, Van Deynze AE, Autrique E, Sorrells ME, Lu YH et al (1995) Molecular mapping of wheat homoeologous group 3. Genome 38:525–533

Panaud O, Chen X, McCouch SR (1996) Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (*Oryza sativa* L.). Mol Gen Genet 252:597–607

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

Peng JH, Zadeh H, Lazo GR, Gustafson JP, Chao S et al (2004) Chromosome bin map of expressed sequence tags in homoeologous group 1 of hexaploid wheat and homoeology with rice and *Arabidopsis*. Genetics 168:609–623

Powell W, Morgante M, Andre C, Hanfey M, Vogel J, Tingey S, Rafalsky A (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238

Roder MS, Korzun V, Wendehake K, Plaschke J, Tixier MH, Leory P, Ganal MW (1998) A microsatelllite map of wheat. Genetics 149:2007–223

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Davis G, Coe EH (2002) Development and mapping of SSR markers for maize. Plant Mol Biol 48:463–481

Singh NK, Raghuvanshi S, Srivastava SK, Gaur A, Pal AK, Dalal V, Singh A et al (2004) Sequence analysis of the long arm of rice chromosome 11 for rice-wheat synteny. Funct Integr Genomics 4:102–117

Sorrells ME, La Rota M, Bermudez-Kandianis CE, Greene RA, Kantety R et al (2003) Comparative DNA sequence analysis of wheat and rice genomes. Genome Res 13:1818–1827

Swigonova Z, Lai J, Ma J, Ramkrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of *Sorghum* and maize genome progenitors. Genome Res 14:1916–1923

Temnykh S, Park WD, Ayers N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:697–712

Temnykh S, Declerk G, Lukashover A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length-variation, transposon associations and genetic marker potential. Genome Res 11:1441–1452

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) *In-silico* analysis of some cereal species. Cell Mol Biol Lett 7:537–546

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23:48–55

Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc Natl Acad Sci USA 86:6201–6205

Young ET, Sloan JS, Van Riper K (2000) Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics 154:1053–1068