DNA Research 9, 199–207 (2002)

Development and Mapping of 2240 New SSR Markers for Rice (*Oryza sativa* L.)

Susan R. McCouch,[1,*] Leonid Teytelman,[2] Yunbi Xu,[3] Katarzyna B. Lobos,[3] Karen Clare,[3] Mark Walton,[3] Binying Fu,[4] Reycel Maghirang,[4] Zhikang Li,[4] Yongzhong Xing,[5] Qifa Zhang,[5] Izumi Kono,[6] Masahiro Yano,[7] Robert Fjellstrom,[8] Genevieve DeClerck,[9] David Schneider,[9] Samuel Cartinhour,[9] Doreen Ware,[2] and Lincoln Stein[2]

Plant Breeding Dept, Cornell University, Ithaca, NY 14853-1901, USA,[1] *Cold Spring Harbor Lab, P.O. Box 100, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA,*[2] *RiceTec, Inc., P.O. Box 1305, Alvin, Texas 77512, USA,*[3] *International Rice Research Institute, P.O. Box 933, Manila 1099, Philippines,*[4] *National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China,*[5] *Institute of the Society for Tecno-innovation of Agriculture, Forestry and Fisheries (STAFF), 446-1 Ippaizuka, Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan,*[6] *Applied Genomics Laboratory, Department of Molecular Genetics, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan,*[7] *USDA-ARS, Rice Research Unit, 1509 Aggie Dr., Beaumont, TX 77713, USA,*[8] *and USDA Center for Agricultural Bioinformatics, Theory Center, Cornell University, Ithaca, NY 14850-1901, USA*[9]

(Received 22 October 2002; revised 26 November 2002)

Abstract

A total of 2414 new di-, tri- and tetra-nucleotide non-redundant SSR primer pairs, representing 2240 unique marker loci, have been developed and experimentally validated for rice (*Oryza sativa* L.). Duplicate primer pairs are reported for 7% (174) of the loci. The majority (92%) of primer pairs were developed in regions flanking perfect repeats ≥ 24 bp in length. Using electronic PCR (e-PCR) to align primer pairs against 3284 publicly sequenced rice BAC and PAC clones (representing about 83% of the total rice genome), 65% of the SSR markers hit a BAC or PAC clone containing at least one genetically mapped marker and could be mapped by proxy. Additional information based on genetic mapping and “nearest marker” information provided the basis for locating a total of 1825 (81%) of the newly designed markers along rice chromosomes. Fifty-six SSR markers (2.8%) hit BAC clones on two or more different chromosomes and appeared to be multiple copy. The largest proportion of SSRs in this data set correspond to poly(GA) motifs (36%), followed by poly(AT) (15%) and poly(CCG) (8%) motifs. AT-rich microsatellites had the longest average repeat tracts, while GC-rich motifs were the shortest. In combination with the pool of 500 previously mapped SSR markers, this release makes available a total of 2740 experimentally confirmed SSR markers for rice, or approximately one SSR every 157 kb.

Key words: simple sequence repeats (SSR); rice (*Oryza sativa* L.); electronic PCR (e-PCR)

1. Introduction

The emergence of genomic sequences in rice[1,2] (http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl; http://www.usricegenome.org/; http://www.gramene.org/resources/) offers new opportunities to increase the density of locus-specific, polymorphic markers for high-resolution genetic analysis. Previous studies have contributed to the development of over 6000 DNA markers[3–5] (http://rgp.dna.affrc.go.jp/; http://www.gramene.org) that provide genome coverage of approximately one marker every 0.25 cM, or every 75–100 kb. Most are expressed sequence tags (ESTs) that are widely used as the basis for physical map construction, sequence assembly and comparative genome analysis, but polymorphism detection within the cultivated rice gene pool is inefficient. In addition to ESTs, approximately 500 simple sequence repeat (SSR) markers have previously been genetically mapped in rice.[6–9]

The technical efficiency and multiplex potential of SSRs makes them preferable for many forms of high throughput mapping, genetic analysis and marker-

Communicated by Satoshi Tabata

[*] To whom correspondence should be addressed. 240 Emerson Hall, Ithaca, NY 14853-1901. Tel. +1-607-255-0420, Fax. +1-607-255-6683, E-mail: SRM4@cornell.edu

assisted plant improvement strategies.[10–14] The fact that SSR markers are co-dominant, multi-allelic and can be reliably used to analyze both *indica* and *japonica* germplasm, as well as groups of AA genome *Oryza* species[15–19] makes them attractive as genetic markers and facilitates the integration of results from independent studies. In addition, the highly polymorphic nature of many microsatellites is of particular value when analyzing closely related genotypes, as is often the case in breeding programs working within narrowly adapted gene pools. Thus, the availability of a high-density SSR map is valuable as a public resource for studies aiming to interpret the functional significance of the rapidly emerging rice genome sequence information.

The International Rice Microsatellite Initiative (IRMI) was formed to increase the density and utility of the SSR map in rice. IRMI is comprised of an international group of researchers from both public and private sector institutions that worked collaboratively to augment the number of experimentally validated SSR markers. This initiative was motivated largely by the release of 6655 SSR-containing sequences by Monsanto Corp. (www.rice-research.com). By coordinating the development of this resource, IRMI encourages the use of a common reservoir of SSR markers and a common nomenclature system so that the results of genetic analysis can be readily interpreted and easily integrated into genome databases. By integrating all the information into Gramene, a comparative grass genome database, positional information about functional mutations and QTLs in rice that are mapped using SSRs can be readily referenced to the corresponding homologous locations in other cereals.

The results of this effort are summarized below and map positions of all reported SSRs in rice that align to fully sequenced BAC or PAC clones can be viewed in the Gramene database "Genome Browser" display[20] (www.gramene.org).

2. Materials and Methods

2.1. Sources of SSR-containing sequences

Two sources of SSR-containing sequences were used in this study. The first consisted of a set of 6655 sequences released by Monsanto in 2001 (www.rice-research.org) and available in GenBank (GenBank GI # 12700719–12707547) that consisted of perfect repeat motifs ($\geq$ 24 bp in length) flanked by 100 bp of unique sequence on either side of the SSR. The second source of SSR-containing sequences was genomic DNA sequence released by the International Rice Genome Sequencing Project (IRGSP) (http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl; http://www.usricegenome.org/).

2.2. Redundancy search

To minimize duplication, the set of sequences from Monsanto was searched for redundancy using minor variations of standard sequence analysis techniques. First, the microsatellite motifs were masked in each sequence by substituting the ambiguous nucleotide code 'N' for each position within the SSR sequence. Second, the masked sequences, each consisting of 200 informative bp of flanking sequence plus the masked residues, were used as a query against: a) all other entries in the Monsanto data set, and b) all previously reported SSR markers as summarized by Temnykh[9] and available at www.gramene.org. This analysis was conducted using BLASTN with an E-value cutoff of 10^{-30}, filtering the subject sequences for low-complexity regions and allowing gaps. The resulting BLASTN output was combined and parsed into subject identifier-query identifier pairs defining similarity relationships. These similarity relationships were used to cluster the sequences using a simple topological sorting algorithm. All clusters that contained sequences corresponding to known markers were identified and cross-referenced.

2.3. Primer design

Primer pairs were automatically designed using PRIMER 0.5 (http://www-genome.wi.mit.edu/ftp/pub/software/primer.0.5), with parameters essentially as described by Chen.[5] Primers were 18–34 nucleotides long, devoid of secondary structure or consecutive tracts of a single nucleotide, a GC content around 50% (Tm approximately 60°C) and preferably G- or C-rich at the $3'$ end. Sequences that provided no useful primer pairs were eliminated from further consideration.

2.4. Nomenclature

SSR-containing sequences that were used as the basis for primer design were given a Sequence ID (see Supplemental Table 1, http://www.dna-res.kazusa.or.jp/9/6/05/spl_table1/table1.pdf). For sequences released by Monsanto at www.rice-research.org, the Sequence ID is coded as MRG followed by a 4-digit code (MRG0001–MRG6655). In cases where publicly available BAC or PAC genomic sequence served as the template for primer design, the Sequence ID is coded as SSR followed by a 3-digit code (SSR002–SSR278). All sequences used as template for primer design also have a GenBank GI and/or a GenBank Accession number (see Supplemental Table 1).

Each primer pair that was used to amplify an SSR marker was considered a *marker reagent*. Marker reagents were assigned a compound name that included a sequence identifier followed by a suffix, with a dot (".") separating them. The suffix consisted of a two- to four-letter abbreviation that identified the research lab reporting the experimental information and a unique identifier for each specific pair of primers (see Supplemental Table 1). For example, MRG0330.IRRI0330 and MRG0330.RGP0330 are two different marker reagents (primer pairs) that are both derived from sequence MRG0330. In the first case, the marker amplifies optimally at an annealing temperature of 61°C, as reported by the group at IRRI, while the second primer pair has an optimal annealing temperature of 55°C, as reported by the RGP. This naming system allows for clear identification of different *marker reagents* designed from a single SSR-containing sequence and highlights the importance of reporting which specific *marker reagents* (primer pairs) were used in any particular experiment.

The SSR *locus* on the genetic or physical map of rice associated with each *marker reagent* was coded with the abbreviation *RM* (*R*ice *M*icrosatellite) followed by a unique identifier. Thus, a single pair of primers that amplified more than one locus retained the same *marker reagent name* but was assigned an independent *locus name*. This nomenclature system makes it easy to understand the relationship among and between sequence templates, reagents and mapped loci.

2.5. *PCR amplification*

Non-redundant primer pairs were used in PCR experiments to confirm amplification potential. All *marker reagents* (primer pairs) were originally tested at an annealing temperature of either 50°C or 55°C, and those that did not amplify well were then evaluated at 61°C and/or 67°C. With minor modifications, the PCR conditions used for detection of amplicons on silver-stained gels were performed in 20-μl reactions containing 25–50 ng of template DNA, 0.2 μM of each primer, 200 μM of each dNTP, 10 mM Tris-Cl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, and 1 unit of *Taq* polymerase. An MJ Research or PE9700 single or dual 96-well thermal cycler was used along with the following PCR profile: 94°C for 5 min (denaturation), followed by 35 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 2 min, and a final extension at 72°C for 5 min. PCR for detection of fluorescently labeled primers was routinely performed in a PE9700 dual 384 thermal cycler. PCR was performed in 10-μl reactions containing 15–25 ng of template, 1 pmol of each primer, 200 μM of each dNTP, 10 mM Tri-HCl (pH 9.0), 50 mM KCl, 2.5 mM $MgCl_2$, 0.1% Triton X-100 and 0.3 unit *Taq* DNA polymerase. The PCR profile was 94°C for 5 min, 35 cycles of 94°C for 15 sec, 50°C for 15 sec, 72°C for 30 sec, and a final extension at 72°C for 10 min. The temperature ramp speed was adjusted to 40% for all temperature changes. It should be noted that when using the 384-well thermal cycler, annealing temperatures are lowered by 5 degrees and cycle times are shortened compared to the 96-well format.

2.6. *Size evaluation of PCR product*

SSR markers whose *marker reagent* IDs indicate that they were evaluated by researchers at the International Rice Research Institute (IRRI) and Huazhong Agricultural University (HUA) were visualized on silver-stained polyacrylamide gels as described by Panaud,[18] and amplicon sizes were estimated in relation to a known size standard. Markers evaluated at RiceTec, Inc. (RT), the Rice Genome Program in Japan (RGP) and at the USDA Agricultural Research Service (ARS) lab in Texas were detected as fluorescent signals using an ABI3100 genetic analyzer with a 36-cm capillary array (4315931) using 10× buffer containing EDTA (PE # 402824), POP-4 polymer (# 4316355), Hi-Di formamide (4311320) and ROX 400 high-density size standard (402985). Running conditions included 60°C oven temp, 15 kV EP voltage, 130 μA EP current, 15 mW Laser Power and 5 A Laser Current.

2.7. *Estimates of polymorphism*

The level of polymorphism for 545 SSR markers was evaluated between cv Nipponbare and cv Kasalath using agarose gel electrophoresis. For PCR amplification, the following conditions were used: 20 μl reaction mixture containing 25–50 ng template DNA, 0.5 μM of each primer, 200 μM of each dNTP, 1.5 mM $MgCl_2$, 1 unit *Taq* polymerase and 2 μl of ×10 PCR reaction buffer (PE Applied Biosystems). Amplification was performed for 35 cycles (1 min at 94°C, 1 min at 55°C and 2 min at 72°C) followed by 5 min at 72°C. To detect polymorphism, the amplified product was electrophoresed on a 3% agarose gel (0.5 TBE).

2.8. *Mapping via e-PCR*

The Gramene database (www.gramene.org) presents an integrated view of the map positions of all publicly reported SSR markers in rice that align to fully sequenced BAC or PAC clones reported in GenBank. These positions were obtained by e-PCR[21,22] using the primer sequences specified for each marker against genomic sequence from 3284 publicly sequenced BAC/PAC clones, representing approximately 460 MB of sequence that was available as of November 12, 2002. We estimate that this sequence represents approximately 83% of the rice genome, based on a 23% sequence overlap predicted from aligning all available cv Nipponbare and cv 93–11 sequences from GenBank. The e-PCR was run with default parameters of a 50-bp margin on the product size and 0 mismatches allowed in the primers.

Identification of map position was accomplished by identifying BAC or PAC clones that simultaneously contained a hit from the SSR primer sequences generated in this study and a hit to at least one previously mapped genetic marker ("nearest marker"). Nearest markers consisted mostly of RFLP markers from the

JRGP map described in Harushima[4] and are available at http://rgp.dna.affrc.go.jp/, which were aligned *in silico* to the BAC/PAC clones. To place RFLP markers onto sequenced BAC/PAC clones, a minimum of 88% identity and 80% coverage of the marker was required. Hits from genomic markers with gaps greater than 20 bp and cDNA-based markers with gaps greater than 300 bp were discarded. In addition, RFLP markers were assigned to the rice clones only if the BAC/PAC chromosome assignment matched that of the marker. Nearest RFLP markers were designated for SSRs based on the single copy marker closest to the SSR on any given BAC/PAC.

2.9. Downloadable information resource

A downloadable file that presents information for all SSR markers that were amplified in wet lab experiments is available in Supplemental Table 1 (http://www.dnares.kazusa.or.jp/9/6/05/spl_table1/table1.pdf). For convenience, we also provide Summary Table 2 in the Supplement section as part of this manuscript that contains only a portion of the information available in the larger Supplemental Table 1. Supplemental Table 1 provides the following information for each marker: Locus Name, Locus Synonym, Sequence ID, Clone origin, GenBank GI, GenBank Accession, Marker Reagent ID, Chromosome (genetically mapped), Map position (genetically mapped), ePCR Chromosome (chromosomal location determined by ePCR), ePCR position (map position as determined by e-PCR), ePCR nearest marker(s), ePCR clone (BAC/PAC hit(s) by e-PCR), ePCR distance (position of the SSR on the BAC/PAC clone in nucleotides), Chromosome (information provided by Monsanto), Map position (Monsanto), Monsanto nearest marker, SSR motif, Number of repeat motifs, Motif length, Forward primer sequence, Forward primer Tm, Forward start (bp of primer start site), Forward end (bp of primer end site), Forward primer length (bp), Reverse primer sequence, Reverse primer Tm, Reverse start, Reverse end, Reverse primer length (bp), Annealing temperature, %GC content, Reference Variety (from which primer sequences were derived), PCR product size (bp in reference variety), Predicted Product Size (in reference variety), Lab name (lab that verified the PCR amplification), Alternative reagent (Binary code where "0" indicates no alternative *marker reagent* and "1" indicates that there is more than one *marker reagent* for that SSR locus), and Published reference.

3. Results and Discussion

Of the 6655 SSR-containing sequences released by the Monsanto Rice Genome (MRG) sequencing effort, primers were designed for 36% (2408) of the sequences and 85% of those primer pairs gave reliable amplification with PCR using genomic DNA from the *japonica* cultivar, Nipponbare, as template. An additional 199 SSRs were developed from sequences released by the International Rice Genome Sequencing Program (IRGSP). Sequences in the MRG data set contained at least one perfect di-, tri- or tetra-nucleotide repeat motif and had a total SSR length greater than or equal to 24 bp, while a range of perfect and imperfect motifs greater than or equal to 20 nt were represented in the SSRs designed from IRGSP sequences (Summary Table 2 in the Supplement section). Figure 1 summarizes the relative frequency of the 12 most abundant motifs among the 2240 newly developed markers. The largest proportion corresponds to poly(GA) motifs (36%), followed by poly(AT) (15%) and poly(CCG) (8%) motifs. All di-, tri- and tetra-nucleotide motifs are represented among the new markers, with the exception of three GC-rich classes, poly(GC) (for which only one marker was designed), poly(GGCC) or poly(GGGC) motifs.

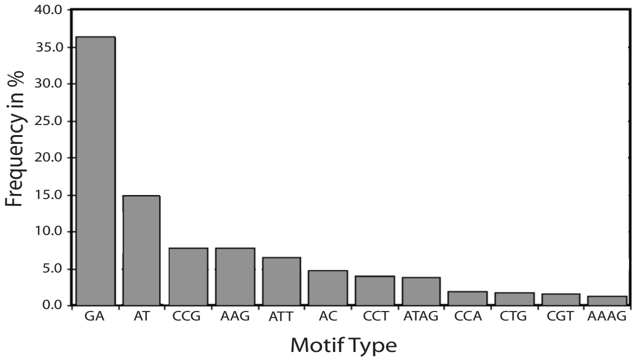

Figure 1. Frequency (in %) of the 12 most abundant SSR motifs in the data set of newly released SSR markers.

When microsatellites were divided into three classes, consisting of AT-rich (greater than 50% A or T in the motif), AT/GC-balanced (limited to di- and tetra-nucleotide motifs that fit this criterion), or GC-rich motifs, the longest repeat motifs corresponded to AT-rich microsatellites where the average number of repeat units was 19 and the average length of the repeated sequence was 48 nt. The AT/GC-balanced class averaged 17 repeats, while the GC-rich tri-nucleotides showed the least potential for expansion, with an average repeat tract length of 8.5 units, consisting of 26 nt (Fig. 2). The longest tract of perfect repeats corresponded to the di-nucleotide $(AT)_{128}$ (RM8135). These statistics are of interest because length of repeat has a demonstrated association with rates of polymorphism in rice.[9,23]

Annealing temperatures were standardized to support the potential for multiplex PCR, such that 95% percent of the new markers amplify well at annealing temperatures of either 50°C or 55°C (65% amplify best at 55°C and 30% at 50°C) (Summary Table 2 in the Supplement

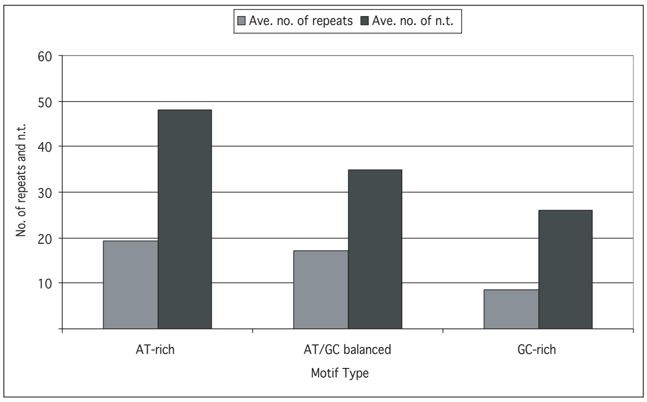

Figure 2. Relationship between motif type and length of repeat tract among the newly developed SSR markers. The AT-rich group (39% of markers) contains > 50% A or T in the motif; the AT/GC balanced group (43% of markers) consists only of di- and tetra-nucleotide motifs that fit this criterion; the GC-rich group (18% of markers) contains > 50% G or C in the motif.

section). Approximately 5% of the new markers were amplified at 61°C, and only 8 markers at 67°C. These annealing temperatures are consistent with optimum conditions for previously reported SSR markers in rice where a majority was reported to anneal well at 55°C.[7–9,18] In 174 cases, different pairs of primers were designed to amplify the same SSR and half of these differ in optimum annealing temperatures. The availability of alternative pairs of primers that amplify the same locus enhances opportunities to select the most favorable combination of allele mw and PCR conditions for efficient multiplexing of SSR markers in an automated system.

Using e-PCR,[21,22] 2000 (89%) SSR markers showed exact sequence matches to 51% (1674/3284) of the fully sequenced BAC or PAC clones, leaving 49% of sequenced clones with no SSR hit. These statistics suggest that the SSRs are not randomly distributed in the genome; if they were, we would expect them to follow a Poisson distribution, showing hits to 64% of sequenced BAC/PAC clones, and leaving only 36% with no hits. To further examine the distribution of SSRs along BAC/PAC clones, we compared observed with expected frequencies under a Poisson distribution for the number of SSR hits/clone. As summarized in Table 1, our data demonstrated that the observed distribution was significantly different from the predicted distribution, with fewer than expected BAC/PAC clones containing one or two SSR hits per clone, a roughly equal number with 3 hits per clone, and a greater than expected number of BACs/PACs containing more than 3 SSR hits per clone. These results suggest that there is significant local clustering of SSRs in the rice genome. This conclusion is consistent with reports by Temnykh[9] and Morgante[24] suggesting that specific motifs are found nonrandomly associated with GC-rich genic regions (which are, themselves, not randomly distributed), and often within specific components of genes, such as $5'$ and $3'$ UTRs, introns or exons, while other SSR motifs are strongly associated with AT-rich intergenic regions.

Table 1. Observed Distribution of SSRs per BAC/PAC clone compared with expected based on a Poisson distribution.

No. SSRs / clone	Number Observed	Number Expected	Percent Observed	Percent Expected	Chi-Square Test
0	1610	1167.7	0.49	0.36	167.51
1	817	1207.4	0.25	0.37	126.25
2	406	624.2	0.12	0.19	76.30
3	220	215.2	0.07	0.07	0.11
4	130	55.6	0.04	0.02	99.48
5	53	11.5	0.02	0.00	149.73
6	28	2.0	0.01	0.00	341.52
7	12	0.3	0.00	0.00	468.11
8	6	0.0	0.00	0.00	939.34
9	2	0.0	0.00	0.00	916.02

Of the 2000 SSRs that hit a sequenced BAC/PAC clone, 1449 (72%) hit clones that also contained the sequence of a previously mapped genetic marker, providing a unique map position for approximately 65% of the 2240 new SSR markers based on e-PCR. Positional information based on predictions of the “nearest marker” was also available from Monsanto for 1249 (61%) of the 2044 MRG-derived SSR markers. In approximately 31% (381) of cases, e-PCR provided positional information where none was available from Monsanto and in 28% (350) of cases, the reverse was true. When map position from the two sources were compared, they were found to be consistent 90% of the time, so a combination of e-PCR and Monsanto predictions were used to determine putative map positions for a total of 1825 (81%) of the SSR markers reported in this study. To confirm the validity of e-PCR predictions, 295 markers were genetically mapped and positions were compared to those predicted by e-PCR. The map positions were consistent 98% of the time.

A downloadable file illustrating the location of 1825 newly developed SSR markers that have been mapped to the 12 chromosomes of rice is available as Supplemental Figure 1 (http://www.dna-res.kazusa.or.jp/9/6/05/spl_figure1/fig1.pdf). The positions of the new SSRs are shown in relation to RFLP and STS markers on the genetic map of rice developed by the Rice Genome Program in Japan (http://rgp.dna.affrc.go.jp/). In addition, 217 previously reported SSR markers[9] have been positioned on the map using the same e-PCR strategy used for the IRMI markers. To illustrate one chromosome from Supplemental Figure 1, Figure 3 summarizes the location of 98 SSR markers that have been mapped to chromosome 10. The positions of the SSRs are shown in relation to 163 STS markers on the genetic map of rice developed by the Rice Genome Program in Japan. The ability to instantly integrate the new markers into

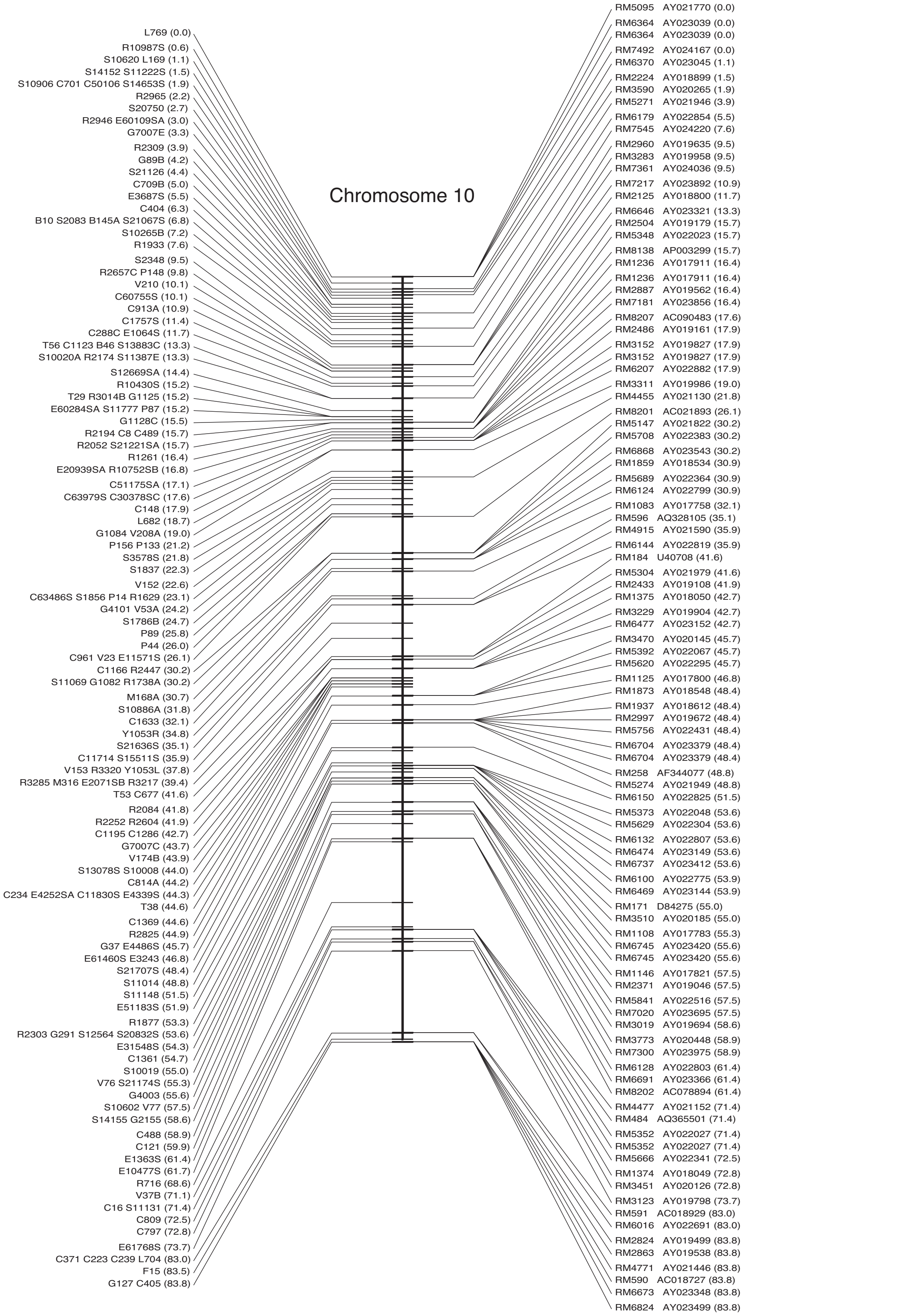

Figure 3. Integrated map of rice chromosome 10 showing relative positions of newly developed IRMI SSR markers (to the right of the chromosome bar) in relation to genetically-mapped RFLP/STS markers (left column). RFLP/STS marker names in left column are followed by genetic map position in parentheses. Rice Microsatellite (RM) locus_ID's to the right are followed by their corresponding GenBank Accession numbers, with cM positions in parentheses.

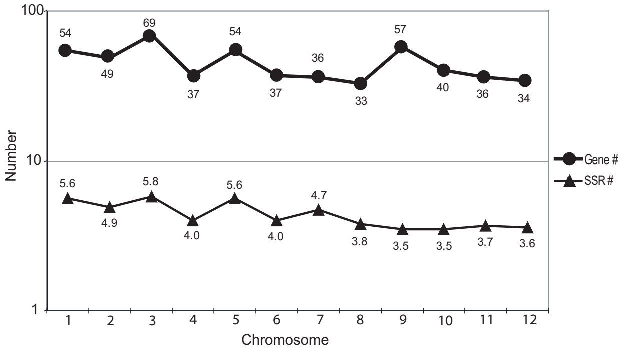

Figure 4. Distribution of SSRs (lower panel with triangles) and EST clusters (upper panel with circles) per Mb on each of the 12 chromosomes of rice. Numbers along the graphs represent the total numbers of SSRs or genes mapped to each chromosome by ePCR or based on TIGR gene indices, respectively, divided by the total length of genomic sequence available for each chromosome (as of November 2002).

existing genetic and physical maps is an obvious advantage of mapping via e-PCR. The number of SSRs that can be positioned along the rice map can be expected to increase as additional genomic sequence information becomes publicly available.

To investigate the distribution of SSRs per Mb on each of the 12 chromosomes, we divided the total number of SSRs mapped to each chromosome (as described above) by the total length of genomic sequence available for each chromosome at the time the e-PCR analysis was run (November 2002). As illustrated in Fig. 4 (lower panel), the highest density of SSR markers was found on chromsomes 1, 3, and 5 (with 5.6, 5.8, and 5.6 SSRs per Mb, respectively) and the lowest density was observed on chromosomes 8, 9, 10, 11, and 12 (3.8, 3.5, 3.5, 3.7, and 3.6 SSRs per Mb, respectively). When these figures were compared to the number of EST clusters/Mb on each chromosome identified by TIGR's Oryza Gene Index using the same genomic sequence data (upper panel), the density of genes was approximately 10 times the density of newly developed SSRs, but there was a significant correlation ($R_2 = 0.45$; $p < 0.015$) between the number of genes/Mb and the number of newly developed SSRs/Mb at the level of the chromosome. This data is consistent with observations of non-random associations between SSRs and genes that have been previously reported by Temnykh et al.[9] and Morgante et al.[24]

Approximately 2.8% (56) of markers hit BAC clones that mapped to independent positions on different chromosomes and appeared to be multiple copy. We investigated 40 SSR markers that fell into this category to determine whether there was evidence to suggest that they mapped to internally repeated regions of the rice genome or whether the multiple chromosomal assignment was more likely to be the result of misassigned BACs, given the emerging status of the physical map of rice. More than half of the putative multiple-copy SSR loci were found to reside in duplicated regions of the genome, as evidenced by the fact that the "nearest markers" had multiple map positions, as determined by genetic mapping. For example, RM1869 was detected on both chromosome 1 (nearest marker R480B) and chromosome 2 (nearest marker R480A), RM3495 was detected on both chromosome 2 (R1738B) and 10 (R1738A), and RM5496 was detected on both chromosome 1 (G89A) and 10 (G89B). In several cases, groups of multiple-copy SSR markers were found on the same or contiguous BAC clones and shared common, multiple-copy "nearest markers." For example, primer sequences for RM1292, RM2614, RM3844, RM5286, RM6053, RM6736 all hit BACs that mapped to both chromosome 3 and chromosome 5, and the nearest marker in each case was R2752, which has locus R2752A on chromosome 3 and R2752B on chromosome 5.[4] A cluster of 9 SSRs mapped to both chromosome 6 (nearest marker R1394B) and chromosome 8 (nearest marker R1394A).

In some cases, SSRs mapped to multiple locations on the same chromosome. A cluster of 5 markers (RM3346, RM3564, RM6084, RM5612, and RM5959) hit multiple BACs on chromosome 3 and the nearest marker was C393, which has one locus at 152.8 cM (C393A) and another at 156.6 cM (C393B), confirmed by both genetic and physical mapping experiments. While not all SSRs that show multiple chromosomal positions with e-PCR could be confirmed to fall in repeated regions of the genome, it is likely that more will be documented as higher resolution maps and finished sequence become available. This data is consistent with reports of multiple, small duplications in rice.[1,25]

While the proportion of multiple-copy SSR markers identified in rice agrees with previous reports based on wet lab experiments,[7,18] it is likely that current estimates underrepresent the amount of internal duplication in the rice genome. The requirement for 0 mismatches in primer sequence alignment during e-PCR restricts the number of hits observed, and obscures the similarity between regions where small genetic changes have occurred in regions flanking SSRs.

An understanding of the relative levels of SSR heterozygosity among and within different sub-species of rice will help guide users of these markers who are interested in working within specific ecotypes or restricted gene pools. A preliminary survey of polymorphism was undertaken for 545 markers to provide an estimate of heterozygosity in the two major sub-species of *O. sativa.* Eighty-six percent of markers detected size differences on 3% agarose gels between the *japonica* cultivar, Nipponbare, and the *indica* cultivar, Kasalath. This estimate can be considered a low threshold for detectable polymorphism, given the fact that significantly higher resolution can be achieved using polyacrylamide gels where 2 bp differences can be identified with ease. These results suggest that the markers provided here will be immediately useful for applications in rice breeding and genetics, to demarcate locations of genes along the rice chromosomes, to link the rice genetic and physical maps, and to provide location-specific anchors for linking out to comparative species maps.[12] A more in-depth analysis of marker polymorphism will be presented in a future study based on a standard panel of diverse *O. sativa* cultivars. SSRs can also be multiplexed on semi-automated systems and used in combination with bi-allelic markers detected as single nucleotide polymorphisms (SNP) or via transposon display (TD)[26–28] in association mapping and linkage disequilibrium studies. The unique power of SSRs lies in the fact that they are multi-allelic and largely co-dominant, providing a powerful assay for studies of genetic diversity, gene discovery and marker-based breeding.

While SSRs are readily detected computationally in genomic sequences, the value of this report is to provide experimentally confirmed marker reagents, most of which bracket perfect repeats, providing ready access to a highly polymorphic fraction of the rice genome. Continuing rounds of e-PCR will be performed as rice physical mapping and genome sequencing progresses, and this will contribute to a more robust positioning of the markers. Updates of the analysis will be released via the Gramene database (www.gramene.org). Use of a publicly available set of common, sequence-based reagents for genetic studies enhances the value of independent experiments, provides links to existing genotypic and phenotypic data and deepens our understanding of plant genome function and evolution.

Acknowledgements: We gratefully acknowledge N. Tao, W. B. Barbazuk, J. Liu, K. Wu, and G. F. Barry at Monsanto for the release of 6655 SSR-containing sequences in October 2000 which provided template for the majority of SSRs released here. We are grateful to Eric Christensen at the USDA-ARS Rice Research Unit in Beaumont, Texas for technical assistance testing primer pairs, Myriam Cristina Duque, Michael Morgan, Immanuel Yap, Hong Lu, and Greg Wilson for help with data management, to Mike Thomson for help with Fig. 3 and Supplemental Figure 1, to Svetlana Temnykh for critical review of this manuscript prior to submission and to Lois Swales for a devoted job of formatting.

References

1. Goff, S., Ricke, D., Lan, T-H. et al. 2002, A draft sequence of the rice genome (*Oryza sativa* L. spp. *Japonica*), *Science*, **296**, 92–100.
2. Yu, J., Hu, S., Wang, J. et al. 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indica*), *Science*, **296**, 79–92.
3. Causse, M., Fulton, T. M., Cho, Y. G. et al. 1994, Saturated molecular map of the rice genome based on an interspecific backcross population, *Genetics*, **138**, 1251–1274.
4. Harushima, Y., Jano, M., Shomura, A. et al. 1998, A high-density rice genetic linkage map with 2275 markers using a single F_2 population, *Genetics*, **148**, 479–494.
5. Wu, J., Maehara, T., Shimokawa, T. et al. 2002, A Comprehensive Rice Transcript Map Containing 6591 Expressed Sequence Tag Sites, *Plant Cell*, **14**, 525–535.
6. Akagi, H., Yokozeki, Y., Inagaki, A. et al. 1996, Microsatellite DNA markers for rice chromosomes, *Theor. Appl. Genet.*, **94**, 61–67.
7. Chen, X., Temnykh, S., Xu, Y. et al. 1997, Development of a microsatellite framework map providing genome-wide coverage in rice (*Oryza sativa* L.), *Theor. Appl. Genet.*, **95**, 553–567.
8. Temnykh, S., Park, W. D., Ayres, N. et al. 2000, Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.), *Theor. Appl. Genet.*, **100**, 697–712.
9. Temnykh, S., DeClerck, G., Lukashova, A. et al. 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.
10. Coburn, J., Temnykh, S., Paul, E., and McCouch, S. R. 2002, Design and Application of Microsatellite Marker Panels for Semi-automated Genotyping of Rice (*Oryza sativa* L.), *Crop Sci.*, **42**, 2092–2099.
11. Cregan, P. B., Jarvik, T., Bush, A. L. et al. 1999, An integrated genetic linkage map of the soybean genome, *Crop Sci.*, **39**, 1464–1490.
12. McCouch, S. R., Chen, X., and Panaud, O. 1997, Microsatellite mapping and applications of SSLP's in rice genetics and breeding, *Plant Mol. Biol.*, **35**, 89–99.
13. Ponce, M. R., Robles, P., and Micol, J. L. 1999, High-throughput genetic mapping in *Arabidopsis thaliana*, *Mol. Gen. Genet.*, **261**, 408–415.
14. Sharopova, N., McMullen, M. D., Schultz, L. et al. 2002, Development and mapping of SSR markers for maize, *Plant Mol. Biol.*, **48**, 463–481.
15. Chen, X., Cho, Y., and McCouch, S. R. 2002, Sequence divergence of rice microsatellites in *Oryza* and other plant species, *Mol. Gen. Genet.*, DOI.1007/s00438-002-0739-5.
16. Harrington, S. 2000, A survey of genetic diversity of eight AA genome species of *Oryza* using microsatellite markers, *MS Thesis*, Cornell University, Ithaca, New York.
17. Ni, J., Colowit, P. M., and Mackill, D. J. 2002, Evaluation of genetic diversity in rice subspecies using microsatellite markers, *Crop Sci.*, **42**, 601–607.
18. Panaud, O., Chen, X., and McCouch, S. R. 1996, Development of microsatellite markers and characterization of simple sequence length polymorphism (SSLP) in rice (*Oryza sativa* L.), *Mol. Gen. Genet.*, **252**, 597–607.
19. Yang, G. P., Saghai Maroof, M. A., Xu, C. G. et al. 1994, Comparative analysis of microsatellite DNA polymorphism in landraces and cultivars of rice, *Mol. Gen. Genet.*, **245**, 187–194.
20. Ware, D., Jaiswal, P., Ni, J. et al. 2002, Gramene: A Resource for Comparative Grass Genomics, *Nucl. Acids. Res.*, **30**, 103–105.

21. Schuler, G. D. 1997, Sequence Mapping by Electronic PCR, *Genome Res.*, **7**, 541–550.

22. Schuler, G. D. 1998, Electronic PCR: bridging the gap between genome mapping and genome sequencing [Focus], *Trends in Biotechnology*, **16**, 456–459.

23. Cho, Y. G., Ishii, T., Temnykh, S. et al. 2000, Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.), *Theor. Appl. Genet.*, **100**, 713–722.

24. Morgante, M., Hanafey, M., and Powell, W. 2002, Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes, *Nat. Genet.*, **30**, 174–200.

25. Salse, J., Piegu, B., Cooke, R. et al. 2002, Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project, *Nucl. Acids Res.*, **30**, 2318–2328.

26. Casa, A. M., Nrouwer, C., Nagel, A. et al. 2001, The MITE family heartbreaker (*Hbr*) molecular markers in maize, *Proc. Natl. Acad. Sci. USA*, **97**, 10083–10089.

27. Van den Broeck, D., Maes, T., Sauer, M. et al. 1998, Transposon display identifies individual transposable elements in high copy number lines, *Plant J.*, **13**, 121–129.

28. Waugh, R., McLean, K., Flavell, A. J. et al. 1997, Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphism (S-SAP), *Mol. Gen. Genet.*, **253**, 687–694.