Theor Appl Genet (2004) 108:392–400
DOI 10.1007/s00122-003-1457-z

ORIGINAL PAPER

Can Li · Yu Zhang · Kai Ying · Xiaolei Liang · Bin Han

Sequence variations of simple sequence repeats on chromosome-4 in two subspecies of the Asian cultivated rice

Received: 9 April 2003 / Accepted: 13 August 2003 / Published online: 17 October 2003

Abstract Computational screening of the chromosome-4 sequence of the rice cultivar Nipponbare (*Oryza sativa* L. *japonica*) revealed 1,844 tandem simple sequence repeats (SSRs) or microsatellites with SSR motifs $\geq$20 bp and repeated unit length of 1–6 base pairs. Thus SSRs occur once in every 18.8 kb, on the average, on the chromosome with one SSR per 23.8 kb and 16 kb on the short and long arms, respectively. No SSR was detected in the core region of the centromere. Poly(AT)$_n$ repeats represented the most abundant and length polymorphic class of SSRs on the chromosome, but it did not occur in the exons. GC-rich trinucleotide repeats were most abundant in the coding regions, representing 71.69% of the SSRs identified in the exons. Two hundred and twenty four SSRs were associated with the repetitive DNA sequences, most of them were poly(AT)$_n$ tracts. Sequence variations of SSRs between two cultivars, representing the two subspecies of the Asian cultivated rice *indica* and *japonica*, were identified, revealing that divergence and convergence of the two subspecies could be traced by the analysis of SSRs. These results provide a great opportunity for SSR-based marker development and comparative genome analysis of the two subspecies of the Asian cultivated rice.

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-003-1457-z.

Communicated by Q. Zhang

C. Li · Y. Zhang · K. Ying · X. Liang · B. Han (✉)
National Center for Gene Research,
Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences,
500 Caobao Road, Shanghai 200233, China
e-mail: bhan@ncgr.ac.cn
Fax: +86-21-64825775

Introduction

Simple sequence repeats (SSRs) or microsatellites are tandemly repeated DNA sequences of short DNA motifs of 1–6 base pairs (bp) in length. SSRs have become a popular type of co-dominant molecular marker in genetic analyses and breeding applications (Yang et al. 1994; Olufowote etal. 1997; Cho et al. 2000). Generation and instability of SSRs are primarily because of DNA polymeraseslippage (Richards and Sutherland 1992) and unequal recombination (Wells 1996; Sia et al. 1997; Wells et al. 1998; Jakupciak and Wells 1999). It was proposed that SSRs have played an important role in genome evolution (Tautz et al. 1986; Kashi et al. 1997; Toth et al. 2000).

SSRs have also been useful for integrating the genetic, physical and sequence-based maps of rice, and provide breeders and geneticists with an efficient tool to link phenotypic and genotypic variations (Wu and Tanksley 1993; Akagi et al. 1996; Panaud et al. 1996; Temnykh et al. 2001). A total of 2,740 microsatellite markers based on SSRs have been developed and integrated into the existing genetic maps for rice (Chen et al. 1997; McCouch et al. 2002). Approximately one SSR marker in every 157 kilo-base (kb) pairs was experimentally confirmed in the rice genome (McCouch et al. 2002). Most of these studies were based on the analysis of either random rice DNA sequence databases including rice ESTs, and genomic sequences or rice SSRs-containing sequences isolated from several small-insert genomic libraries (Temnykh et al. 2000).

The draft sequences have been available for both *japonica* and *indica* subspecies of *Oryza. sativa* L. (Barry 2001; Goff et al 2002; Yu et al. 2002). Moreover, a comprehensive rice transcript map including a large number of mapped expressed sequence tags (ESTs), and an integrated rice physical and genetic map, were constructed, respectively (Chen et al. 2002; Wu et al. 2002). These advances greatly facilitated the construction of a minimal tilling path of the large-inserted bacterial artificial chromosome (BAC) or phage (P1)-derived

artificial chromosome (PAC) clones to be used for complete genome sequencing. Recently, sequences of two chromosomes were completely determined (Feng et al. 2002; Sasaki et al. 2002). These progresses provide a great opportunity for identifying SSRs in the rice genome, and also potential uses of SSRs in marker development.

The rice chromosome-4 sequence of Nipponbare (*O. sativa* L. ssp. *japonica*) has been completely determined with an accuracy rate of 99.99%, and the eight sequence contigs covered 97.4% of the chromosome (Feng et al. 2002). Clearly, these sequences with known chromosomal map position provide a great opportunity for high-throughput *in silico* identification of SSR sites on the chromosome, and the highly efficient genetic marker development based on SSRs.

In this study, we screened the rice chromosome-4 sequence for the presence, chromosomal positions, and composition of SSR tracts. The objective was to produce a detailed map displaying all possible SSR motifs on rice chromosome-4 and use them for comparative genome analysis of the *indica* and *japonica* subspecies.

Materials and methods

DNA sequences

We constructed a comprehensive clone-based physical map of chromosome-4 for the cultivar Nipponbare through an integrated approach (Zhao et al. 2002), and identified a tiling path that consisted of 287 BAC clones and two PAC clones. Each clone was sequenced by a random shotgun approach on both strands of small insert subclones to achieve a ten-fold coverage. Sequence assemblies were verified by comparing *Not*I and *Hin*dIII restriction profiles predicted from the sequence against experimentally determined restriction profiles. Sequence overlaps of adjacent clones were used to assess potential sequence differences between them, which resolved the sequence with less than 1 mismatch per 14,000 base pairs (bp); these mismatches were corrected further by manual checking or by additional sequencing reactions. We estimated that the overall accuracy of our finished sequence exceeded 99.99% (Feng et al. 2002). The sequences of the collinear regions from *indica* variety Guangluai 4, which were completely sequenced, were used for comparative analysis (Feng et al. 2002).

Since a working draft of an *indica* variety 93-11 genome with 4.3 × coverage redundancy was obtained (Yu et al. 2002; http://btn.genomics.org.cn/rice/), we assessed the degree of homology between the two *indica* varieties. We searched the draft sequence database with 44 contigs with a total length of 430-kb contiguous sequences from chromosome 4 of Guangluai 4 with a cut off value of $e>10^{-3}$.

The rice EST sequences were obtained from GRAMENE (ftp://ftp.gramene.org/pub/gramene/blast_databases/). The rice repetitive sequences were obtained from TIGR (ftp://ftp.tigr.org/pub/data/o_sativa/osa1/). The contigs of 93-11 were obtained from the rice genome database (http://btn.genomics.org.cn/rice/).

SSR analysis

All sequences were scanned for various SSRs using computer programs written in Perl. Briefly, a DNA sequence was considered as a string, $N_1N_2N_3N_4N_5$–N_k–$N_{i-1}N_i$. To detect a tandem repeat of size n (1–6) at position k, the sequence N_k–N_{k+n-1} was compared with subsequent sequences starting at positions N_{i+n}, N_{i+2n}, N_{i+3n}, N_{i+4n}, ... A repeat was detected and extended further when a certain minimum number of units [$(N)_{20}$, $(NN)_{10}$,$(NNN)_7$, $(NNNN)_5$, $(NNNNN)_4$ and $(NNNNNN)_4$ for mono-, di-, tri-, tetra-, penta- and hexa-nucleotide motifs, respectively] were repeated tandemly. While scanning for di-, tri-, tetra-, penta- and hexa-nucleotide repeats, motifs involving runs of the same single nucleotide were not considered. Similarly, for tetranucleotide repeats, combinations representing perfect dinucleotide repeats were ignored. For hexanucleotide repeats, combinations representing perfect trinucleotide repeats were ignored. The results were validated by manual check of a randomly selected sequence of BAC clones.

A polyA repeat is the same as a polyT repeat on a complementary strand. Similarly, $(AC)_n$ is equivalent to $(CA)_n$, $(TG)_n$ and $(GT)_n$, while $(AGC)_n$ is equivalent to $(GCA)_n$, $(CAG)_n$, $(CTG)_n$, $(TGC)_n$ and $(GCT)_n$ in different reading frames or on a complementary strand (Jurka and Pethiyagoda 1995).

Results

SSRs (≥20 bp) on chromosome 4

To detect SSRs in the rice chromosome-4 sequences, we selected minimum repeating units such that a repeat spanned a minimum of 20 nucleotides (nt). Although previous studies used threshold repeat length ≥12 nt, longer repeats (≥20 nt), though unstable, have more implications in genome organization, genetic variation and protein evolution on a relatively shorter evolutionary time scale.

We scanned for various SSRs through a total length of 34,689,786 bp in eight contigs, and a total of 1,844 SSRs with length ≥20 nt were identified (Table 1 and Fig. 1 in the electronic supplementary data that can also be found at http://www.ncgr.ac.cn/ssr/ssr.htm), revealing a frequency of one SSR in every 18.8 kb on average (Table 1). Dinucleotide repeats (DNRs) are the most frequent, while mononucleotide SSRs are the least frequent. The most frequent mono-, di-, tri-, tetra-, penta- and hexa-nucleotideSSRs are C/G, AT/TA, CCG/CGG, ATAG, ATATA

Table 1 Numbers and frequencies of SSRs on chromosome-4

SSR	Occurrence	Percentage	Frequency (kb/SSR)	No. of repeats	Most frequent type
$(N)_n$	61	3.30%	568.2	20 to 47	C/G
$(NN)_n$	838	45.40%	41.4	10 to 90	AT/TA
$(NNN)_n$	404	21.90%	85.8	7 to 114	CCG/CGG
$(NNNN)_n$	178	9.70%	194.7	5 to 29	ATAG
$(NNNNN)_n$	246	13.30%	140.9	4 to 6	ATATA
$(NNNNNN)_n$	117	6.30%	296.3	4 to 13	ATACAT
Total	1,844	100%	18.8		

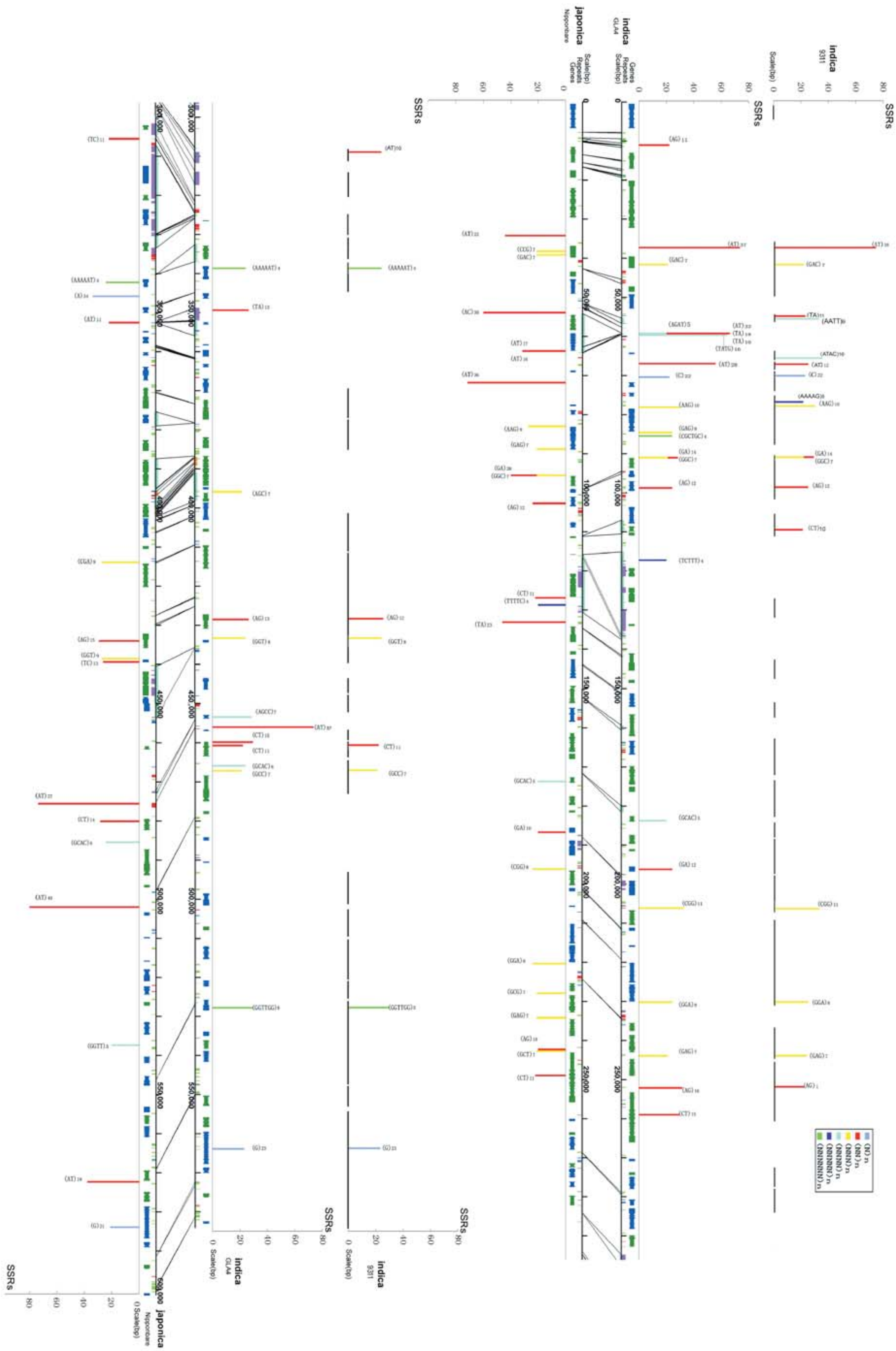
indica
9311
indica
GLA4
japonica
Nipponbare
SSRs
Scale(bp)
Genes
Repeats
(N) n
(NN) n
(NNN) n
(NNNN) n
(NNNNN) n
(NNNNNN) n

Table 2 Distributions of SSRs on the short and long arms of chromosome-4

Location	Length (kb)	No. of SSRs	$(N)_n$	$(NN)_n$	$(NNN)_n$	$(NNNN)_n$	$(NNNNN)_n$	$(NNNNNN)_n$
Short arm	9,327	392	11	204	72	36	53	16
Percentage		100.00%	2.80%	52.00%	18.40%	9.20%	13.50%	4.10%
Frequency (kb/SSR)		23.8	848	45.7	129.5	259.1	176	583
Centromere	131	0	0	0	0	0	0	0
Long arm	25,203	1,452	50	634	332	142	193	101
Percentage		100.00%	3.40%	43.70%	22.90%	9.80%	13.30%	7.00%
Frequency(kb/SSR)		16	464.1	36.6	69.9	163.4	120.2	229.7

and ATACAT, respectively. The longest SSR, a trinucleotide repeat (TNR) AAT motif with 114 repeated units, was 342 bp in length. Poly$(GC/CG)_n$ blocks were the least frequent in all regions. Only three of the $(GC/CG)_n$ repeats were detected on the chromosome. There were more $(AG/TC)_n$ repeats than $(AC/TG)_n$ tracts. The frequency of SSRs in the long arm of the rice chromosome-4 was one SSR per 16.0 kb on average, but one SSR per 23.8 kb in the short arm. However, no SSR was detected in the core region (130 kb in length) of the centromere (Table 2). Although the SSRs were quite evenly distributed in the genome, the density of SSRs seems to be biased towards the long arm of rice chromosome-4.

Since the sequence has been integrated into the existing genetic map, the identified SSR sites were therefore mapped along with the available markers on chromosome-4. This map will be useful for developing more polymorphic markers based on SSRs (Fig. 1 in the electronic supplementary data).

The repetitive DNA sequences have been identified by using the Repeat Mask Program based on the Rice Repeat Database of TIGR (ftp://ftp.tigr.org/pub/data/o_sativa/osa1). Eventually, 22,472 individual repetitive sequences with a total length of 8.5 Mb, representing 24.6% of the chromosome-4 sequence, were identified. Two hundred and twenty four SSRs were identified to be associated with 27 of the 76 repetitive DNA sequence subfamilies on chromosome-4 (Table 3). Particularly, 30, 71 and 40 SSRs were detected in 24 of the 4,756 Ty3-gypsy retrotransposons, 53 of the 1,250 CACTA transposons and 26 of the 137 Micro-Mite subfamilies, respectively (Table 3). Thus, the SSR motifs may be used in classifing the repetitive DNA sequence families. The total length of the 224 SSRs was 7,030 bp, representing 0.08% of the repetitive sequences of chromosome-4. Because of association with the repetitive DNA sequences, these 224 SSRs may not be very useful for microsatellite marker development.

Fig. 1 Comparative analysis of sequences of Guangluai 4 (583 kb), Nipponbare (602 kb) and 93-11 (430 kb). The length and classes of SSRs are indicated in the vertical lines with *different colors and motif names*. Insertions and deletions are indicated in *light blue*. Genes on the two strands are indicated in *green and blue*, respectively. Polymorphic regions are indicated in *orange*, and other types of repeats are indicated in *short cross-lines* with different colors

Relationship between SSRs and the predicted genes

To characterize the possible relationship between SSRs and genes predicted by using FGENESH, we investigated the distribution of SSRs in the rice chromosome-4 sequences relative to the occurrence of: (a) exons, (b) untranslated regions (UTRs), (c) introns, and (d) intergenic regions.

Of the 1,844 SSRs identified, 219 and 315 SSRs were found in exons and introns respectively, 479 SSRs were in UTRs, and the remaining 825 SSRs were found in intergenic regions (Table 4).

In exons, GC-rich trinucleotide SSRs (TNRs) were the most abundant SSR class, representing 71.69% (157/219) of the SSRs identified; hexanucleotide repeats constituted the second-most frequent type of SSRs, whereas, mononucleotide, dinucleotide, tetranucleotide and pentanucleotide SSRs are very rare. GC-rich TNRs were also the most abundant in 5′UTRs. Dinucleotide SSRs were detected mostly in introns, UTRs and intergenic regions. Only one $(AG)_n$ SSR was detected in exon (Table 4). Poly$(AT/TA)_n$ blocks were detected in introns, 3′UTRs and intergenic regions, but not in exons. DNRs were the most abundant and length-variable in the rice chromosome-4 sequence. Comparing with other motifs, $(AT/TA)_n$ repeats are most abundant and have hyper-variable length (Table 1 in the electronic supplementary data).

We also examined SSRs in the public databases of rice EST sequences. A total of 2,391 SSRs with a length of 20 nucleotides or longer were identified in the rice ESTs database of Gramene (ftp://ftp.gramene.org/pub/gramene/blast_databases/). TNRs are the most abundant SSRs (Table 5). However, hexanucleotide SSRs were not the second-most frequent type, presumably because the rice ESTs contained UTRs. The $(N)_n$ tracts, detected in ESTs, made up 21.8% of all identified SSRs in ESTs. This was due to the poly$(A)_n$ of the mRNA that was included in the ESTs. Similarly, GC-rich TNRs were found to be the most abundant in the ESTs.

Table 3 Association of SSRs with various classes of repetitive sequences. The number of SSRs observed in each class of repetitive sequences is listed in parenthesis

Sub-family	Name	Number	Length (bp)	SSRs observed		
				Number	Total length (bp)	Most frequent type
Retrotransponsons	Ty1-copia	502	525,835	4	85	CCGGC
	Ty3-gypsy	4,756(24)	4,244,802	30	913	AT/TA
Transposons	Ac/Ds	136	22,128	1	24	TCGGAA
	CACTA, En/Spm	1,250(53)	896,922	71	2,395	AT/TA
	Mutator (MULE)	113	20,652	1	30	CTCTCC
	Other transposons	1,223	259,800	7	234	GA,CGC
(MITEs)	Tourist	1,043	250,177	7	190	AT/TA
	Castaway	323	72,395	3	96	
	Ditto	581	91,555	4	150	
	Wanderer	888	12,9541	4	125	
	Snabo	127	22,054	2	42	
	MITE-adh, type Aa	576	10,0521	1	24	CGCCGA
	MITE-adh, type B	1,512	266,168	11	417	
	MITE-adh, type D	666	10,897	3	63	
	MITE-adh, type H	54	13,714	1	21	G
	MITE-adh, type I	581	93,954	7	170	AT/TA
	MITE-adh, type J	624	51,517	5	128	CT,GA
	MITE-adh, type K	366	52,148	1	30	GGCGGA
	MITE-adh, type M	199	27,474	1	20	GAGGC
	MITE-adh-5	333	68,911	1	20	CCAT
	MITE-adh-11	129	45,671	5	192	TTA,TAT
	MITE-adh-12	150	14,172	1	28	AG
	ID-4	33	5,174	1	21	GGT
	Stola	20	5,907	1	24	TA
	Youren	64	13,300	2	53	CGG,TC
	Micron	137(26)	37,362	40	1,296	AT/TA
	Others	2,668	361,124	9	239	
Total		19,054	7,703,875	224	7,030	AT/TA

Table 4 Distribution of SSRs (≥20 bp) in coding and non-coding regions of rice chromosome-4

SSR class	Exons	Percent-age	Introns	Percent-age	5′UTRs	Percent-age	3′UTRs	Percent-age	Inter-genic	Percent-age
A/T	0	0.00%	7	2.22%	4	1.27%	3	1.84%	6	0.73%
C/G	2	0.91%	7	2.22%	9	2.85%	1	0.61%	21	2.55%
AC/TG	0	0.00%	12	3.81%	7	2.22%	6	3.68%	23	2.79%
AG/CT	1	0.46%	68	21.59%	67	21.20%	17	10.43%	134	16.24%
AT/TA	0	0.00%	87	27.62%	37	11.71%	69	42.33%	303	36.73%
CG/GC	0	0.00%	0	0.00%	2	0.63%	1	0.61%	0	0.00%
GC-rich TNRs	157	71.69%	31	9.84%	77	24.37%	7	4.29%	25	3.03%
AT-rich TNRs	14	6.39%	21	6.67%	15	4.75%	10	6.13%	47	5.70%
$(NNNN)_n$	2	0.91%	30	9.52%	24	7.59%	14	8.59%	107	12.97%
$(NNNNN)_n$	6	2.74%	38	12.06%	58	18.35%	29	17.79%	115	13.94%
$(NNNNNN)_n$	37	16.89%	14	4.44%	16	5.06%	6	3.68%	44	5.33%
Total	219	100%	315	100%	316	100%	163	100%	825	100%

Table 5 SSRs in rice ESTs as found in the GRAMENE database (Dec 2002)

SSR	Occurrence	Percentage	Frequency (kb/SSR)	Length of SSR motifs (bp)	Most frequent type
$(N)_n$	639	21.80%	60.4	20 to 121	A/T
$(NN)_n$	404	13.80%	95.5	10 to 51	AG/CT
$(NNN)_n$	1,077	36.70%	35.8	7 to 29	CCG/GGC
$(NNNN)_n$	299	10.20%	129	5 to 18	TAAA
$(NNNNN)_n$	313	10.70%	123.2	4 to 6	TTTGG
$(NNNNNN)_n$	199	6.80%	193.8	4 to 6	TCGCCG
Total	2,931	100.00%	13.2		

Table 6 SSRs, SNPs and indels in the sequences of Guangluai 4 (*indica*) and Nipponbare *(japonica)* chromosome-4

Item	Guangluai 4	Nipponbare
Total length (bp)	583,234	601,764
Number of SNPs	2,472	2472
SNPs in exon	353	576
SNPs in intron	403	554
SNPs in intergenic region	1,716	1342
Number of indels	23	44
Indels >10 kb	1	3
Indels 1–10 kb	13	8
Indels <1 kb	9	33
Length of indels (kb)	64	86
Repeats length in indels (kb)	28	33
No. of SSRs (≥20 bp)	37	39
No. of conserved SSRs in the two subspecies	9	9
No. of SSRs with length variations at the same loci	15	15
No. of absent SSRs to another species	13	15

Variation of SSRs in *indica* and *japonica* varieties

While single nucleotide polymorphisms (SNPs) and insertions or deletions (indels) between the two cultivated subspecies *indica* (Guangluai 4) and *japonica* (Nipponbare) have been characterized in a previous study (Feng et al. 2002), we compared distributions, and length variations of SSRs, between the two subspecies in this study.

Variations of SSRs in the two varieties were analyzed through a comparative search of the 583,234-bp sequence of Guangluai 4 and its collinear region of the 601,764-bp sequence of Nipponbare. We aligned the two contiguous sequences with the annotated genes, repetitive sequences, indels and SSR motifs (Fig. 1, more details of Fig. 1 can also be viewed at http://www.ncgr.ac.cn/ssr/ssr.htm). The total numbers of SNPs, indels and SSRs from the collinear regions were summarized in Table 6. A total of 2,472 SNPs were identified from the collinear 583,234 bp of Guangluai 4 and 601,764-bp of Nipponbare sequences, indicating an average of 1 SNP per 235 bp in Guangluai 4 and one SNP per 243 bp in Nipponbare. Furthermore, the SNPs and indels detected were in 200 exons and 220 introns, predicted in the Nipponbare chromosome-4 sequence. The average densities of SNPs in exons, introns and intergenic regions are one SNP per 2.5, 2.0 and 3.0 kb, respectively. The frequencies of indels in exons, introns, and intergenic regions are one per 9.0-, 5.5- and 8.0-kb, respectively.

Moreover, 37 and 39 SSRs were identified in the collinear regions of Guangluai 4 and Nipponbare, respectively. Comparative analysis of the densities of SSRs in Guangluai 4 and Nipponbare revealed length and motif variations of SSRs on chromosome-4. No conspicuous differences of SSR densities have been discovered between the two cultivars in this region (Table 6). The

Table 7 Comparison of SSRs in the two indica varieties and one japonica variety at the same region

Locus	Nipponbare	Guangluai 4	93-11	Locus	Nipponbare	Guangluai 4	93-11
1		$(AG)_{11}$	No sequence	30	$(GCAC)_5$	$(GCAC)_5$	No sequence
2	$(AT)_{22}$	$(AT)_{37}$	$(AT)_{38}$	31	$(GA)_{10}$	$(GA)_{12}$	
3	$(CCG)_7$			32	$(CGG)_8$	$(CGG)_{11}$	$(CGG)_{11}$
4	$(GAC)_7$	$(GAC)_7$	$(GAC)_7$	33	$(GGA)_8$	$(GGA)_8$	$(GGA)_8$
5			$(TA)_{11}$	34	$(GCG)_7$		
6			$(AATT)_9$	35	$(GAG)_7$	$(GAG)_7$	$(GAG)_7$
7	$(AC)_{30}$			36	$(AG)_{10}$	$(AG)_{16}$	$(AG)_{11}$
8		$(AGAT)_5$	No sequence	37	$(GCT)_7$		
9		$(AT)_{33}$	No sequence	38			$(AT)_{10}$
10		$(TA)_{18}$	No sequence	39	$(CT)_{11}$	$(CT)_{15}$	
11		$(TA)_{10}$	No sequence	40	$(TC)_{11}$		
12		$(TATG)_{16}$	No sequence	41	$(AAAAAT)_4$	$(AAAAAT)_4$	$(AAAAAT)_4$
13	$(AT)_{17}$			42	$(A)_{34}$		
14	$(AT)_{16}$			43	$(AT)_{11}$	$(TA)_{13}$	No sequence
15			$(ATAC)_{10}$	44		$(AGC)_7$	No sequence
16	$(AT)_{36}$	$(AT)_{28}$	$(AT)_{12}$	45	$(CGA)_9$		
17		$(C)_{22}$	$(C)_{22}$	46	$(AG)_{15}$	$(AG)_{13}$	$(AG)_{12}$
18			$(AAAAG)_5$	47	$(GGT)_9$	$(GGT)_8$	$(GGT)_8$
19	$(AAG)_9$	$(AAG)_{10}$	$(AAG)_{10}$	48	$(TC)_{13}$		
20	$(GAG)_7$	$(GAG)_8$		49		$(AGCC)_7$	No sequence
21		$(CGCTGC)_4$		50	$(AT)_{37}$	$(AT)_{37}$	No sequence
22	$(GA)_{20}$	$(GA)_{14}$	$(GA)_{14}$	51		$(CT)_{15}$	
23	$(GGC)_7$	$(GGC)_7$	$(GGC)_7$	52	$(CT)_{14}$	$(CT)_{11}$	$(CT)_{11}$
24	$(AG)_{12}$	$(AG)_{12}$	$(AG)_{12}$	53	$(GCAC)_6$	$(GCAC)_6$	
25			$(CT)_{10}$	54		$(GCC)_7$	$(GCC)_7$
26		$(TCTTT)_4$	No sequence	55	$(AT)_{40}$		No sequence
27	$(CT)_{11}$			56	$(GGTT)_5$		
28	$(TTTTC)_4$			57		$(GGTTGG)_5$	$(GGTTGG)_5$
29	$(TA)_{23}$			58	$(AT)_{19}$		
				59	$(G)_{21}$	$(G)_{23}$	$(G)_{23}$

SSRs from the two cultivars can be classified into three groups as following: (1) SSRs were identical in length and motifs, at the same loci; (2) SSRs had the same motifs, but different lengths at the same loci; (3) SSRs were present in one cultivar but not in the other cultivar. Nine SSRs had the same motifs and length at the same loci in the two varieties. Fifteen SSRs were different due to changes in the number of repeats of the motifs at the same loci (Table 6). Among the SSRs that were present in one cultivar (subspecies), but absent in the other, six were located in five predicted genes, with one found in an exon; three were located in the insertion regions from both sequences. Several of the presence/absence types of SSRs were caused by base substitution or deletion of the repeat units. For example, the analysis of the sequences showed that the $(CCG)_7$ motif in the *japonica* sequence was changed into the $(CCG)_2CCT(CCG)_4$ motif in the *indica* sequence, because of a base substitution of CCG to a CCT. The shortened motif was therefore excluded in the identification of the SSRs with a cutoff of $\geq$20 bp in the computer search. The base substitution usually caused the change of one amino acid to another. In this case, the base substitution resulted in a change of an alanine residue to serine.

Since a sequence-draft of the 93-11 (*indica*) was available, we also aligned 44 contigs of 93-11 with a total length of 430 kb, with the two collinear contiguous sequences from Guangluai 4 and Nipponbare. Twenty five SSRs were found from the 93-11 sequences in these regions. It is interesting to note that 15 out of the 25 SSRs from the 93-11 were identical to those of Guangluai 4, representing 60% of identified SSRs in 93-11. Of these 15 SSRs, six were also observed in the Nipponbare sequence with the same length and motifs. Three sites, $(C)_{22}$, $(GCC)_7$ and $(GGTTGG)_5$ at loci of 17, 54 and 57, were found only in the two *indica* varieties (Table 7; Fig. 1). Length variations were detected between the two *indica* cultivars, and Nipponbare at the same loci, for six SSRs. In addition, four SSRs were found with the same motifs at the same loci but with different lengths among the three varieties (Table 7). Since the 93-11 sequence only covered about 74% of the Guangluai-4 sequence, the other three motifs that were conserved in Guangluai 4, and Nipponbare sequences were not found in the 93-11 sequence. It was obvious that SSRs of Guangluai 4 were more similar with 93-11 than with Nipponbare.

Discussion

In this study, we have identified 1,844 SSRs on rice chromosome-4, showing a frequency of one SSR per 18.8 kb. Compared to the current mapping result of one SSR marker per 157 kb in the rice genome, this finding provides a great opportunity for the SSR-based marker development. More importantly, these SSR tracts have known chromosomal positions. Since the SSR-based markers provide more polymorphisms, complementary to the classical RFLP markers, the association of the SSR sites with the predicted genes will help for designing more polymorphic markers to map the genes on the chromosome. The 224 SSRs, that have been clearly identified to be associated with the repetitive DNA sequences on the chromosome, could be removed from the list for the potential SSR marker development. Considering the difficulty of PCR amplification of $(AT)_n$ tracts, the frequency of amplifiable SSRs is probably one per 25–30 kb on this chromosome.

We observed that the SSRs were biased towards the long arm of the chromosome, where there were less repetitive DNA sequences and more genes. No SSRs were found in the centromeric region, indicating more unique sequences in this region. The distributions of the different SSR classes are clearly different. Interestingly, although poly(AT) SSRs are the most abundant on the chromosome, no such SSRs have been found in the coding regions. Only three $(CG/GC)_n$ SSRs were found, revealing that this motif was the rarest among the SSRs on this chromosome. In total, only one DNR was found to be located in an exon. GC-rich TNRs were the most abundant in the coding regions, representing about 72% of the identified SSRs in the exons. These preferences may have utility in the genome annotation. Almost half of the SSRs were located in intergenic regions. Most abundant SSRs were found to be located in UTRs (one SSR per 11.7 kb), indicating the potential roles of SSRs in the regulation of the translational process.

The Asian cultivated rice *O. sativa* is classified into two major subspecies *indica* and *japonica*, that represent most of the rice crop in the world. These two subspecies have partial sexual reproductive-isolation. Intraspecific phenotype variations are apparent between these two cultivated subspecies including growth, development and environmental adaptations. Comparative analysis between the two subspecies has been used to construct a fine physical map of chromosome-4 in parallel with chromosomal sequencing (Zhao et al. 2002). Comparative genome analysis of the two subspecies has revealed their extensive micro-collinearity in gene order and content. However, deviations from the collinearity are frequent with insertions or deletions (Feng et al. 2002; Han and Xue 2003). Intraspecific sequence polymorphisms frequently occur in both coding and non-coding regions. Subsequently, these variations often affect gene structures and may contribute to intraspecific phenotypic adaptations. These progresses represent major advances in understanding of the content and complexity of the rice genome, and have revealed several types of DNA sequence polymorphisms between the two subspecies and provide some important insights into the molecular basis of intraspecific phenotypic variations.

Comparative analysis of SSRs polymorphisms in landraces and cultivars of rice has been reported (Yang et al. 1994). Sequence divergence of rice microsatellites from six different species within the genus *Oryza* has been characterized based on PCR-electromorph size polymorphisms (Chen et al. 2002). It was found that allelic diversity within a species was predominately due

to changes in the number of repeats in the SSR region, but the frequency of indels and base substitutions increased as the genetic distance between samples increased (Chen et al. 2002).

In addition to the identification of SNPs and indels between the sequences, we evaluated the SSR variations at the DNA sequence level in the collinear regions of the *indica* and *japonica* varieties. Only 25% of the SSRs were conserved at the same loci in Guangluai 4 (*indica*) and Nipponbare (*japonica*). Most of the identified SSRs were found to vary in the number of repeats and have base substitutions at the same loci. Not surprisingly, each variety had a number of SSR motifs that were not found in the others. Interestingly, of the 25 SSRs identified in the 93-11 (*indica*) sequences, 15 SSRs (60%) were found to be conserved with the same length and motifs in the Guangluai 4 sequences, and only six SSRs of them were found to be conserved in the Nipponbare sequences. The 4 × draft sequence of 93-11 is likely to have some errors. It would be difficult to correctly compare the SNPs between the draft and the more accurate genome sequences. A high-quality sequence of 93-11 is needed for the identification of sequence differences between the two *indica* varieties, and between *indica* and *japonica* varieties. However, we demonstrated here that SSRs with length and motif variations would have some advantages to distinguish the differences of the closely related subspecies, even using the available draft sequences. Three unique SSR motifs were identified to be unique to, and identical in, Guangluai 4 and 93-11, while four SSRs were varied in length at the same loci in the three varieties. Divergence of the SSR motifs and the length between the *indica* and *japonica* sequences was observed. But more convergence of SSR motifs was clearly identified in the two *indica* varieties. Sequence analysis revealed that the length and motif variations of SSRs among the related subspecies were mostly due to base substitutions or deletions.

This comparative information about the chromosome organization in the two closely related subspecies has important implications in the development of new molecular markers for genetic mapping. We conclude that the sequence variations including the motif and length variation of SSRs can be used for the analysis of the genetic variation in both *indica* and *japonica*. Variation of SSRs can provide adequate information in studies of taxonomically related species. Molecular dissection of these variations will increase our understanding of phenotypic adaptations that can be helpful for the genetic improvement of cultivated rice and other crops.

Acknowledgements The authors acknowledge the support for rice functional genome study (Grant No. 2002AA2Z1003) from the Ministry of Science and Technology of China.

References

Akagi H, Yokozeki Y, Inagaki A, Fujimura T (1996) Microsatellite DNA markers for rice chromosomes. Theor Appl Genet 94:61–67

Barry G (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125:1164–1165

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu J, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo M,Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

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

Chen X, Cho YG, McCouch SR (2002) Sequence divergence of rice microsatellites in *Oryza* and other plant species. Mol Genet Genomics 268:331–343

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

Feng Q, Zhang YJ, Hao P, Wang SY, Fu G, Huang YC, Li Y, Zhu JJ, Liu YL, Hu X, Jia P, Zhang Y, Zhao Q, Ying K, Yu S, Tang Y, Weng Q, Zhang L, Lu Y, Mu J, Lu Y, Zhang LS, Yu Z, Fan D, Liu X, Lu T, Li C, Wu Y, Sun T, Lei H, Li T, Hu H, Guan J, Wu M, Zhang R, Zhou B, Chen Z, Chen L, Jin Z, Wang R, Yin H, Cai Z, Ren S, Lv G, Gu W, Zhu G, Tu Y, Jia J, Zhang Y, Chen J, Kang H, Chen X, Shao C, Sun Y, Hu Q, Zhang X, Zhang W, Wang L, Ding C, Sheng H, Gu J, Chen S, Ni L, Zhu F, Chen W, Lan L, Lai Y, Cheng Z, Gu M, Jiang J, Li J, Hong G, Xue Y, Han B (2002) Sequence and analysis of rice chromosome-4. Nature 420:316–320

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, T Migue, Paszkowski U, Zhang S, Colbert M, Sun W, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 296:92–100

Han B, Xue YB (2003) Genome-wide intraspecific DNA-sequence variations in rice. Curr Opin Plant Biol 6:134–138

Jakupciak JP, Wells RD (1999) Genetic instabilities in (CTG·CAG) repeats occur by recombination. J Biol Chem 274:23468–23479

Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol 40:120–126

Kashi Y, King D, Soller M (1997) Simple sequence repeats as a source of quantitative genetic variation. Trends Genet 13:74–78

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

Olufowote JO, Xu Y, Chen X, Park WD, Beachell HM, Dilday RH, Goto M, McCouch SR (1997) Comparative evaluation of within-cultivar variation of rice (*Oryza sativa* L.) using microsatellite and RFLP markers. Genome 38:1170–1176

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

Richards RI, Sutherland GR (1992) Dynamic mutations: a new class of mutations causing human disease. Cell 70:709–712

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, Antonio BA, Kanamori H, Hosokawa S, Masukawa M, Arikawa K, Chiden Y, Hayashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, Harada C, Hijishita S, Honda M, Ichikawa Y, Idonuma A, Iijima M, Ikeda M, Ikeno M, Ito S, Ito T, Ito Y, Ito Y, Iwabuchi A, Kamiya K, Karasawa W, Katagiri S, Kikuta A, Kobayashi N, Kono I, Machita K, Maehara T, Mizuno H, Mizubayashi T, Mukai Y, Nagasaki H, Nakashima M, Nakama Y, Nakamichi Y, Nakamura M, Namiki N, Negishi M, Ohta I, Ono N, Saji S, Sakai K, Shibata M, Shimokawa T, Shomura A, Song J, Takazaki Y, Terasawa K, Tsuji K, Waki K, Yamagata H, Yamane H, Yoshiki S, Yoshihara R, Yukawa K, Zhong H, Iwama H, Endo T, Ito H, Hahn JH, Kim HI, Eun MY, Yano M, Jiang J, Gojobori T (2002) The genome sequence and structure of rice chromosome-1. Nature 420:312–316

Sia EA, Jinks-Robertson S, Petes TD (1997) Genetic control of microsatellite instability. Mutat Res 383:61–70

Temnykh S, Park WD, Ayres 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, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch SR (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

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967–981

Tautz D, Trick M, Dover G (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Wells RD (1996) Molecular basis of genetic instability of triplet repeats. J Biol Chem 271:2875–2878

Wells RD, Bacolla A, Bowater RP (1998) Instabilities of triplet repeats: factors and mechanisms. Results Problem Cell Diff 21:133–165

Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J, Fujii F, Shomura A, Ando T, Kono I, Waki K, Yamamoto K, Yano M, Matsumoto T, Sasaki T (2002) A comprehensive rice transcript map containing 6,591 expressed sequence tag sites. Plant Cell 14:525–535

Wu KS, Tanksley SD (1993) Abundance polymorphism and genetic mapping of microsatellites in rice. Mol Gen Genet 241:225–235

Yang GP, Saghai Maroof MA, Xu CG, Zhang Q, Biyashev RM (1994) Comparative analysis of microsatellite DNA polymorphism in landraces and cultivars of rice. Mol Gen Genet 245:187–194

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica).* Science 296:79–92

Zhao Q, Zhang Y, Cheng Z, Chen M, Wang S, Feng Q, Huang Y, Li Y, Tang Y, Zhou B, Chen Z, Yu S, Zhu J, Hu X, Mu J, Ying K, Hao P, Zhang L, Lu Y, Zhang LS, Liu Y, Yu Z, Fan D, Weng Q, Chen L, Lu T, Liu X, Jia P, Sun T, Wu Y, Zhang Y, Lu Y, Li C, Wang R, Lei H, Li T, Hu H, Wu M, Zhang R, Guan J, Zhu J, Fu G, Gu M, Hong G, Xue Y, Wing R, Jiang J, Han B (2002) A fine physical map of the rice chromosome-4. Genome Res 12:817–823