Y.G. Cho · T. Ishii · S. Temnykh · X. Chen
L. Lipovich · S.R. McCouch · W.D. Park · N. Ayres
S. Cartinhour

Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.)

Received: 5 May 1999 / Accepted: 16 August 1999

Abstract The growing number of rice microsatellite markers warrants a comprehensive comparison of allelic variability between the markers developed using different methods, with various sequence repeat motifs, and from coding and non-coding portions of the genome. We have performed such a comparison over a set of 323 microsatellite markers; 194 were derived from genomic library screening and 129 were derived from the analysis of rice-expressed sequence tags (ESTs) available in public DNA databases. We have evaluated the frequency of polymorphism between parental pairs of six inter-subspecific crosses and one inter-specific cross widely used for mapping in rice. Microsatellites derived from genomic libraries detected a higher level of polymorphism than those derived from ESTs contained in the GenBank database (83.8% versus 54.0%). Similarly, the other measures of genetic variability [the number of alleles per locus, polymorphism information content (PIC), and allele size ranges] were all higher in genomic library-derived microsatellites than in their EST-database counterparts. The highest overall degree of genetic diversity was seen in GA-containing microsatellites of genomic library origin, while the most conserved markers contained CCG- or CAG-trinucleotide motifs and were developed from GenBank sequences. Preferential location of specific motifs in coding versus non-coding regions of known genes was related to observed levels of microsatellite diversity. A strong positive correlation was observed between the maximum length of a microsatellite motif and the standard deviation of the molecular-weight of amplified fragments. The reliability of molecular weight standard deviation (SDmw) as an indicator of genetic variability of microsatellite loci is discussed.

Key words Allelic diversity · Simple sequence repeat (SSR) · Microsatellite marker · Rice (*Oryza sativa* L.)

Communicated by F. Salamini

The first three authors made equal contributions to this manuscript

Y.G. Cho · T. Ishii · S. Temnykh · X. Chen · L. Lipovich
S.R. McCouch (✉)
Department of Plant Breeding, 252 Emerson Hall,
Cornell University, Ithaca, NY 14853–1901, USA

W.D. Park · N. Ayres · S. Cartinhour
Department of Biochemistry and Biophysics,
Texas A&M University, College Station, TX 77845, USA

Present addresses:
Y.G. Cho, Department of Agronomy,
Chungbuk National University, Chongju 361–763, Korea

T. Ishii, Laboratory of Plant Breeding,
Faculty of Agriculture, Kobe University,
Nada-ku, Kobe 657–8501, Japan

L. Lipovich, Department of Molecular Biotechnology,
University of Washington, 1705 NE Pacific St., Box 357730,
Seattle, WA 98195–7730, USA

S. Cartinhour, USDA/ARS and Department of Plant Breeding,
G16 Bradfield Hall, Cornell University,
Ithaca, NY 14853–1901, USA

Introduction

Microsatellites, also known as simple sequence repeats (SSRs), are tandem arrays of short nucleotide repeats from 1 to 5 bases per unit. Simple sequence length polymorphisms (SSLPs) are based on the difference in the number of the DNA repeat units at a given locus and provide a valuable source of genetic markers. Microsatellites have been extensively exploited for genome mapping and for a wide range of population and evolutionary studies in human (Bowcock et al. 1994; Dib et al. 1996), mouse (Dietrich et al. 1996), *Drosophila* (Goldstein and Clark 1995; Schlötterer et al. 1997; Schug et al. 1998), *Arabidopsis* (Innan et al. 1997), rice (Yang et el. 1994), and other animal and plant species (Jarne and Lagoda 1996; Powell et al. 1996). In addition to its clear utility for practical applications such as genetic mapping and fingerprinting, information about the distribution and variability of microsatellite sequences in the genome of a given species can elucidate the genetic history of the species from the standpoint of evolution and artificial selection. Population and evolutionary studies in human and *Drosophila* have shown that highly polymorphic microsatellite markers can provide information on the differentiation of populations and can detect recent selective sweeps and bottle-neck events (Di Rienzo et al. 1994; Slatkin 1995; Sclötterer et al. 1997). On the other hand, more stable SSR markers with lower variability can be used to reconstruct more ancient evolutionary events (Meyer et al. 1995). Since mutation rates at microsatellite sequences vary drastically between species (reviewed by Schug et al. 1997) and between loci (Schlötterer et al. 1997; Brinkmann et al. 1998; Harr et al. 1998), it is important to investigate factors influencing microsatellite variability and to account for these factors when using SSLP markers.

The structure and length of simple sequence repeats are considered to be the major factors affecting microsatellite variability (McMurray 1995; Brinkmann et al. 1998). In general, SSLP loci with more repeats tend to be more polymorphic and have a larger amplitude of variation (Weber 1990; Goldstein and Clark 1995; Innan et al. 1997; Schug et al. 1998). Interestingly, in *Drosophila melanogaster* a stronger correlation was observed between the maximum repeat count and variance than between the mean repeat count and variance, suggesting that the mutation rate increases with repeat count (Goldstein and Clark 1995; Schug et al. 1998). The increased probability of microsatellite DNA length variation over longer tracts of SSR motifs is in good agreement with the proposed mechanism of DNA expansion via slipped-strand mispairing (Levinson and Gutman 1987). Direct estimations of mutation rates for SSLP loci in *Drosophila* demonstrated that dinucleotide repeats mutate more frequently than tri- and tetranucleotide repeats (Schug et al. 1998). A similar difference in mutation rate between di-, tri-, and tetranucleotide SSR motifs was observed in humans (Chakraborty et al. 1997).

In rice, the availability of a large number of microsatellite markers with different SSR motifs developed from genomic libraries and extracted from DNA databases (Wu and Tanksley 1993; Panaud et al. 1996; Akagi et al. 1996, Chen et al. 1997, Temnykh et al. 1999) made it possible to investigate the occurrence and variability of simple sequence repeats at the whole-genome level. In the study reported here, a total of 323 microsatellites, including 194 markers isolated from small-insert genomic libraries, hereafter referred to as “genomic library-derived microsatellites”, and 129 derived from the GenBank database, or “GenBank-derived microsatellites”, were characterized using 14 diverse rice accessions. Of these, 13 varieties represented the two major cultivated subspecies of rice (*O. sativa indica* and *O. sativa japonica*), and 1 accession represented a wild rice species from Africa, *Oryza longistaminata*. The number of alleles, polymorphism information content (PIC) and variation in allele size were estimated for each SSR locus and used to determine factors affecting microsatellite variability in rice. We aimed to evaluate the frequency of polymorphism between widely used pairs of mapping parents and to compare markers with different di- and trinucleotide SSR motifs and originating from the two sources (random genomic clones versus expressed DNA sequences) for diversity and variability.

Materials and methods

Plant materials

Seven widely used pairs of rice mapping parents from rice research programs in the US, China, Japan, Korea, and the Philippines were used for the evaluation of microsatellite allelic diversity (Table 1). Most of the mapping populations were derived from inter-subspecific crosses involving *indica* and *japonica* varieties of *O. sativa*, with the exception of the IR36/N22 combination, which was an *indica/indica* cross, and the BS125/WL02 combina-

Table 1 Levels of polymorphism between parental varieties of seven mapping populations detected by 300 SSR markers

Population ID[a]	Parental lines	Species/subspecies	Percentage polymorphism		
			Genomic	cDNA	Total
IN	IR36 N22	*O. sativa* (*indica*) *O. sativa* (*indica*)	58.7	34.1	48.2
DH1	IR64 Azucena	*O. sativa (indica*) *O. sativa* (tropical *japonica*)	87.8	51.2	72.1
DH2	Zhai-Ye-Qing 8 Jing-Xi 17	*O. sativa* (indica*)* *O. sativa* (*japonica*)	88.9	54.3	74.1
RIL1	Milyang 23 Gihobyeo	*O. sativa* (*indica/japonica*) *O. sativa* (*japonica*)	83.7	49.6	69.1
RIL2	Lemont Teqing	*O. sativa* (tropical *japonica*) *O. sativa* (*indica*)	83.7	53.5	70.8
JRGP	Nipponbare Kasalath	*O. sativa* (*japonica*) *O. sativa* (*indica*)	87.8	55.8	74.1
SL	BS125 WL02	*O. sativa* (*indica*) *O. longistaminata*	95.9	79.8	89

[a] Populations are from the following institutions: IN, Cornell University, USA; DH1, International Rice Research Institute, Philippines; DH2, Academia Sinica, China; RIL1, National Institute of Agricultural Science and Technology, Korea; RIL2, Texas A & M University, USA; JRGP, Rice Genome Program, Japan; SL, Cornell University, USA & ORSTOM, W. Africa

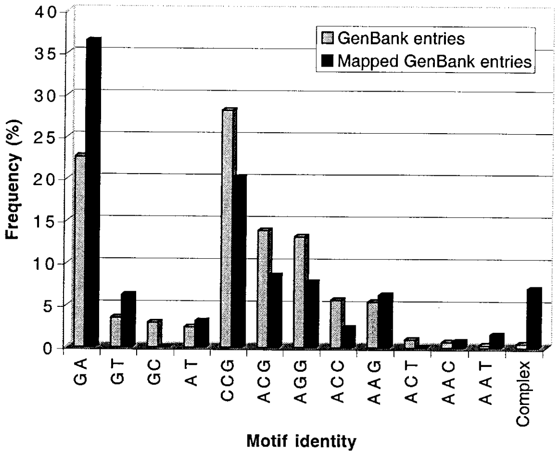


Fig. 1 Frequency of individual microsatellite motifs derived from the survey of rice GenBank sequences compared to those in the subset mapped in this study

tion, which was an interspecific cross between *O. sativa (indica)* and *O. longistaminata*, a wild rice species from Africa. The 13 *O. sativa* varieties were used to analyze genetic diversity of SSR loci in this paper, while the wild *O. longistaminata* accession served as an outgroup. Total DNA was extracted from fresh leaves of a single plant from each of the 14 rice varieties by the potassium acetate method (Dellaporta et al. 1983).

Microsatellite markers from genomic libraries and GenBank

Microsatellites from three genomic libraries and from the GenBank database were used in this study. Of the markers 11 had been previously described: *RM122, RM148, RM164, RM167,* and *RM168* from a 15 kb genomic library (Wu and Tanksley 1993); *RM1–RM80* from a genomic library of physically sheared subcloned fragments (Panaud et al. 1996; Chen et al. 1997); and *RM201–RM263* from a *Tsp509*-digested genomic library (Chen et al. 1997). Eighty-four new microsatellites with different SSR motifs (23 GT, 31 GA, 19 CTT, 9 CAT and 2 ATT) were isolated from genomic libraries and are reported in the companion paper by Temnykh et al. (1999). These 194 loci had the following frequencies: 70.1% poly(GA), 12.9% poly(GT), 9.8% poly(CTT), 4.6% poly(TCT) and 2.6% poly(ATT).

One hundred twenty-nine GenBank microsatellites were included in this survey, of which 59 markers contained dinucleotide motifs (GA, GT and AT), 61 contained six different trinucleotide motifs, and 9 markers had complex patterns of repeated units. Frequencies of markers developed from GenBank sequences corresponded to the relative frequencies of their occurrence in the database (Fig. 1). The majority of markers contained either poly(GA) or GC-rich trinucleotide repeats (CCG, ACG, AGG, ACC).

Polymerase chain reaction (PCR) amplification and silver staining

PCR amplification and microsatellite detection were performed as described in Temnykh et al. (1999). For some GenBank microsatellites which have specific annealing temperatures, touchdown PCR was used as follows: 5 min at 94°C; 10 cycles of 1 min at 94°C, 40 s at 65°C minus 1°C /cycle, 1.5 min at 72°C; 30 cycles of 1 min at 94°C, 40 s at 55°C, 1.5 min at 72°C; and 5 min at 72°C for final extension.

Allele scoring

After silver staining of polyacrylamide gels, a cluster of two to five discrete bands (stutter) was apparent for most of the markers. The size (in nucleotides) of the most intensely amplified band for each microsatellite marker was determined based on its migration relative to molecular weight (mw) size markers (10-bp and 25-bp DNA ladders from Gibco BRL, Gaithersburg, Md.). IR36 was useful as a mw reference because a sequence-based estimate of allele size in this variety was available, as described in Panaud et al. (1996) and Chen et al. (1997). For the GenBank microsatellites, cv. Nipponbare was used as a reference for allele calling because it was the variety most commonly used as the source of rice sequences in GenBank. For some markers, two or more bands amplified with equal intensity, and in those cases the molecular weight of the fragment nearest to the predicted size for IR36 (in the case of genomic library microsatellites) or Nipponbare (for GenBank microsatellites) was selected as the representative allele at that locus. Null alleles were assigned to varieties for which no amplification product was generated in controlled experiments.

Evaluation of polymorphism

The frequency of microsatellite polymorphism between pairs of mapping parents was calculated based on the presence (1) or absence (0) of common bands. Heterozygous banding patterns, in which two distinct MW bands were visible in the same lane, were scored by giving one half-value to each of the alleles compared to a value of 1 assigned to homozygous alleles. The number of alleles per locus was based on an evaluation of the 13 *O. sativa* cultivars and did not include the wild *O. longistaminata*. The polymorphism information content (PIC) value described by Botstein et al. (1980) and modified by Anderson et al. (1993) for self-pollinated species was calculated as follows:

$$\mathrm{PIC}_i = 1 - \sum_{j=1}^{n} P_{ij}^2$$

where p_{ij} is the frequency of the jth allele for marker i, and summation extends over n alleles.

Determination of “maximum repeat count”

To obtain the maximum repeat count (“max repeat count”) for each microsatellite locus, the following formula was used: max repeat count = [(max allele MW – reference allele MW) / x] + reference repeat count (x=2 or 3 for dinucleotide and trinucleotide repeats, respectively). The reference repeat count was taken directly from the known sequence of the reference variety (IR36 for genomic library-derived or Nipponbare for GenBank microsatellites). This formula assumes that the PCR fragment size variation is due solely to the repeat number variation within the SSR loci and not to insertions or deletions in the sequences flanking the SSR. The assumption appears reasonable in most cases for cultivated rice based on unpublished results of direct sequencing of different rice alleles (X. Chen, personal communication).

Results

Polymorphism detected by microsatellite markers in rice

In this study, 323 microsatellite markers were surveyed on the panel of rice varieties summarized in Table 1. The highest proportion of polymorphic loci (89%) was detected between the interspecific parents, BS125 and WL02, and the least amount of polymorphism (48%) was detected between the 2 *indica* varieties, IR36 and N22. The five other *indica/japonica* mapping populations were intermediate, with between 69–74% polymorphism. The Milyang23/Gihobyeo combination was slightly less polymorphic than the other four, probably

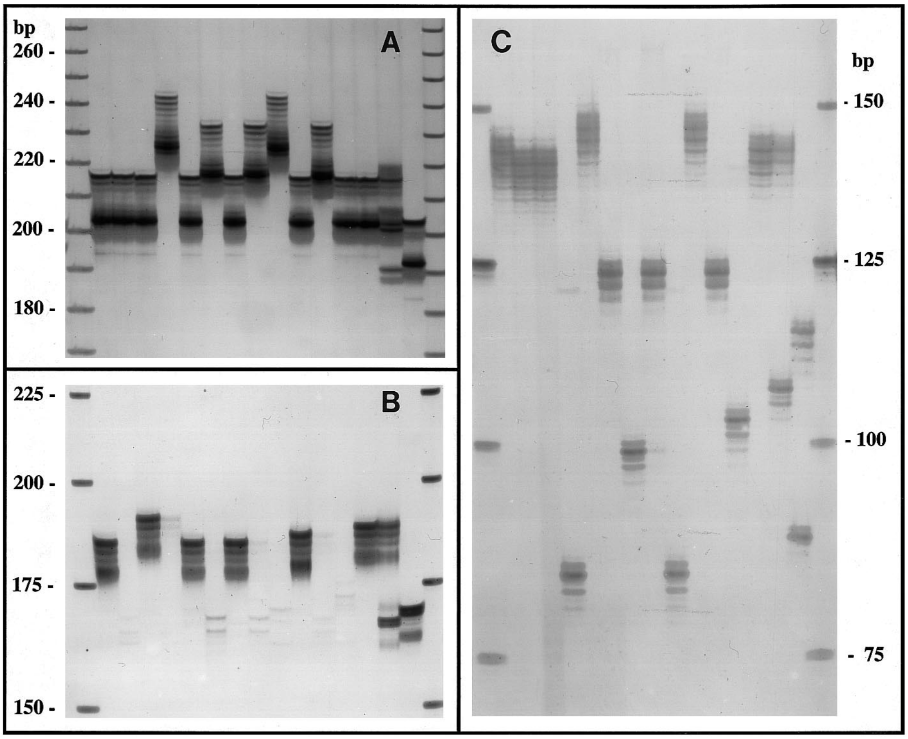


Fig. 2A–C Polyacrylamide gel electrophoresis patterns of microsatellites derived from genomic libraries and GenBank sequences on a panel of 15 rice varieties: lane 1, size marker; lane 2, IR36; lane 3, N22; lane 4, IR64; lane 5, Azucena; lane 6, Zhai-Ye-Qing 8; lane 7, Jing-Xi 17; lane 8, Milyang 23; lane 9, Gihobyeo; lane 10, Lemont; lane 11, Teqing; lane 12, Nipponbare; lane 13, Kasalath; lane 14, BS125; lane 15, WL02. **A** GenBank-derived microsatellite (*RM162*) distinguishes the *indica* and *japonica* subspecies; **B** Null alleles are observed in N22 and in japonica varieties at *RM269*, **C** large standard deviation in allele molecular weight is observed at *RM276*

due to the fact that Milyang 23 was, itself, derived from a *japonica/indica* cross.

There was an average of 71% polymorphism for the 323 microsatellites among the seven pairs of mapping parents. The average level of polymorphism was significantly higher for genomic library microsatellites (83.8%) than for GenBank microsatellites (54.0%) (T = 4.40, *P* <0.01). This tendency is also reflected in the number of alleles per locus and the PIC value of the microsatellites examined. Diversity data for markers whose identity and map position has been previously published, including loci *RM1–RM80* and *RM201–RM263*, can be found in the RiceGenes Database (http://genome.cornell.edu/rice/) Diversity data for new markers whose map position is unknown (due to monomorphism) is presented in Table 2 of this paper. Genetic diversity characteristics for newly mapped markers can be found in Table 2 of the companion paper by Temnykh et al. (1999). While the range was similar in the two groups (2–11 alleles per locus in genomic library-derived microsatellites, and 1–10 alleles per locus in GenBank microsatellites), the average number of alleles was 5.13 for genomic library microsatellites, and only 2.78 for GenBank microsatellites (*P* <0.01) among the 13 *O. sativa* cultivars examined. As illustrated in Fig. 2, the GenBank microsatellites often had only 2 or 3 alleles which served to distinguish the *indica* and *japonica* subspecies (Fig. 2A, *RM162*), while the genomic library-derived microsatellites tended to detect more alleles which resolved within-subspecies variation (Fig. 2C, *RM276*).

Interspecific variation at microsatellite loci was clearly observed upon comparison of the outcrossing wild species *O. longistaminata* (accession WL02) from Africa with the 13 varieties of cultivated Asian species, *O. sativa.* The WL02 accession showed a heterozygous allele pattern at 58 (33.9%) of the genomic library loci and 23 (17.8%) of the GenBank microsatellite loci, frequently having alleles with an unusually low or high molecular weight, that were uncommon among *O. sativa* varieties. Null alleles were also more frequently detected in *O. longistaminata*. About half of the 54 microsatellite loci with null alleles were identified as markers giving no amplification with the accession WL02 only, while the other 50% detected null alleles among *O. sativa* and *O. longistaminata* varieties. Overall, null alleles were found in 8% and 9% of the genomic library and GenBank microsatellites, respectively, with from 1 to 6 varieties harboring the null allele (for example, see Fig. 2B, *RM269*).

Microsatellite markers derived from genomic libraries detected a higher level of heterozygosity than did microsatellite markers derived from GenBank sequences, while null alleles were observed in roughly equal frequencies between the two. This suggests that mutations affecting the length of amplified fragments occur more frequently in microsatellite sequences isolated from random genomic clones, while mutations in unique flanking regions of microsatellite sequences that cause non-amplification occur with similar frequencies among SSLP loci of these two groups.

Table 2 EST-derived microsatellite markers with low informativeness

Marker name[a]	Accession	Motif	Number of alleles	PIC	Forward primer	Reverse primer	PCR fragment	Annealing temperature (°C)
m16	D15151	$(GCT)_5$	1[c]	0.00	gcggaagctgctgctgctgctc	gagcatgccccaaccggacctc	258	55
m36	D15172	$(CGG)_5$	1	0.00	atctccgagctccacctcggcg	ctcgatcgccgagaatgtgcgg	131	61
m13	D15349	$(AAG)_5$	1	0.00	gaggagaggagacgaggcgccg	cgaaaccaacacgacgcaaccc	93	61
m54	D15502	$(GA)_7$	2	0.14	ttgagaccgtagagagagagagag	ccacgccttccttgtgctcccc	160	55
m17	D15715	$(CAA)_5$	2	0.14	ggtcagcgtctctcaccatgtcg	agccctggctgttcccctgacg	161	61
m21	D15734	$(GT)_6$	1	0.00	cgctgccaacggtgcttgtctg	cgtactcttgtcgacacacacacacc	214	61
m33	D15853	$(GGC)_5(TGG)_5$	2	0.14	aggcggcggctcagatctggac	gaccaccaccaccaccaaccgc	146	61
m35	D15858	$(GGC)_5$	1	0.00	gagctcggtggcatggcgatgc	ttggtcttcttgtccccgcgcc	131	61
m47	D15897	$(CT)_8$	2	0.50	actccgcctcatcacccgaggc	cgcccgcgtcatcctcttccte	209	55
osm77	D22182	$(CCT)_8$	1[c]	0.00	tcttcgtcctctgcggcagggc	ttgccggccggctattccttcc	99	55
m10	D22576	$(CGG)_5$	2	0.47	aaatctcgcgaccaagcggcg	gtcggggcggaactcgaactcg	174	61
m43	D23753	$(AGC)_8$	1[c]	0.00	cgattcccactcctctcgcgg	gtaggccaccagggacgcctcg	171	55
m64	D23856	$(CT)_7$	1	0.00	atcctccgacgcctgtcctggc	cgccagatccgaatcctgcacg	281	55
osm75	D24140	$(CTG)_5$	1	0.00	ttggtgacgtgtgccctgctgc	ctgggggcacttgtcgcggtag	99	67
m63	D24432	$(TGC)_3(CT)_7(GAA)_3$	1[c]	0.00	agcaaagcaccactgctgctgc	gctgcctggagctatgccatgg	160	55
m22	D24445	$(CCG)_5$	1	0.00	gcgcacggcgggttagccatac	acagggcgatgtttctccggcg	139	61
m28	D24454	$(CCG)_6$	1[c]	0.00	aggcctgagcaccaatggcacg	caagcgcccggaggccaagaac	135	61
m59	D24468	$(CCG)_6$	2	0.26	ggtctcaccccaaactcgccgc	ggcttccaggaggggacgaggg	190	55
m34	D24787	$(CCG)_4$	1[c]	0.00	agccagatctcgctcgctcgcc	ggttgcccatccccaggagcac	103	61
m15	D24879	$(CGG)_5$	1[c]	0.00	ttgtggtggtgttcatggcggc	tggaagaggaggcggacgaggc	175	61
m19	D25367	$(GA)_9$	2	0.33	acggctggcaaatgacgttgcc	aagtacagaggcgttgccaggtag	147	61
m14	D25505	$(GA)_7$	1	0.00	aatgcagtgctgctcccgccag	gcacagcaagctctcgtctctctc	206	61
m11	D28316	$(TCG)_8$	1	0.00	ctcgtcgtcgtcgtcgtcgtcg	cccttgtcgaagaggtcggcgg	189	61
m50	D38829	$(ACG)_7$	1[c]	0.00	cggacgccatggatgcttctcg	aagctcgtccggctccggagtg	266	55
m60	D39427	(CCG)	1[c]	0.00	ttagccatggcggccttctccg	aaacgatacggggtgagcggcg	73	55
m29	D39752	$(CGG)_6$	1[c]	0.00	ggctgatgctgcggtgattggc	caccgcggagaccgtatcaccg	124	61
m52	D39993	$(GA)_7$	1	0.00	gttcttggtttcctggcattgtgc	ccctggtgtccccccttgccaag	121	55
m42	D40193	$(GA)_8$	1	0.00	ctcccaattgcccacctaccgg	ccgccgcccttactcctgcatc	306	55
m61	D40429	$(CGG)_7$	1	0.00	cgagccacctctcctcctcgcc	gcactccgcccattccacctcc	222	55
m27	D41022	$(CGG)_5$	1	0.00	ggtggctgctatggcgagcacg	gaacggtacggatcccgccgac	169	61
m31	D41023	$(GA)_6$—$(TGG)_3$	2	0.50	gattaagggagagagagagatcatg	cttcgcgaacgtcaccatcggc	229	61
m45	D4568	$(CCG)_8$	3	0.34	aaacctagctccgcctccgccg	cccgccactagctctagcgcgc	239	55
m62	D41695	$(CGT)_7$	1	0.00	cgccgcaaaggggtctcgtctc	gtctcccggatcgacctccccc	97	55
osm81	D46170	$(CT)_7$	3	0.52	tttcagctcgatcgatccacagc	acgacatgctgctcgcgaaggc	108	61
m55	D47131	$(AGC)_7$	1[c]	0.00	cgcacgtagcagcagcagcagc	cctcccgaaggtcggggaggac	128	55
m53	D47426	$(GA)_7$—$(CGT)_5$	1[c]	0.00	tcctagccgccaccgcttcacc	tcctgggcgaggagaaggcgag	182	55
m38	D48517	$(GA)_9$	2	0.14	cggcttcttggagcgagcgagc	gatttcttcaccgggagccgcg	297	61
m49	D48927	$(TG)_8$	2	0.15	gagggggttggcaaggcagcag	gaagaacacaagctgactcacactg	209	55
m4	U08404[b]	$(CCT)_7$	1[c]	0.00	gtgcttcgccactgtcaccccg	gcgagagtcggcgcacgagaac	316	55
m24	U37133[b]	$(CGG)_5$	2	0.14	gcgacgaactcgcgctgctctc	ggagcgcagcagcagcttcacc	173	61
m40	U49113[b]	$(CT)_3$	2	0.26	cacggtgttcagcgcccaaac	gcttgggtcgaggagggaaccc	273	61
m9	X53596[b]	$(GGA)_4$	1	0.00	ggtggaggtggagggggtggtg	ttccacttccattgccgccacc	151	55

[a] Temporary names of unmapped markers

[b] Accessions correspond to known or putative genes (see Table 4)

[c] Markers are monomorphic among 13 *Oryza sativa* varieties but polymorphic between BS125 and WL02 (*O. longistaminata*)

Table 3 Mean values[a] for measures of genetic diversity of microsatellite markers derived from genomic libraries and from the rice ESTs within our panel of 13 *O. sativa* cultivars

Class of microsatellite markers	*n*	Number of repeat units in SSR	Allele count per locus	PIC	Allele size range (bp)	SD of molecular weight (Bp)
Genomic poly(GA)n	136	17.6	5.5	0.70	25.0	8.44
cDNA-derived poly/GA)n	47	10.4	3.1	0.46	10.5	4.09
Genomic GT	25	13.1	4.3	0.62	24.5	8.7
Genomic poly(AAT)n and poly(CTT)n	24	14.5	5.0	0.67	33.8	12.9
Genomic poly(CAT)n	9	8.4	3.1	0.56	13.0	4.5
cDNA-derived GC-rich trinucleotides	48	6.7	2.0	0.28	6.3	2.4

[a] Both polymorphic and monomorphic microsatellites were included in the calculations

Factors affecting the variability of microsatellite sequences

SSR motif and origin of microsatellite sequences

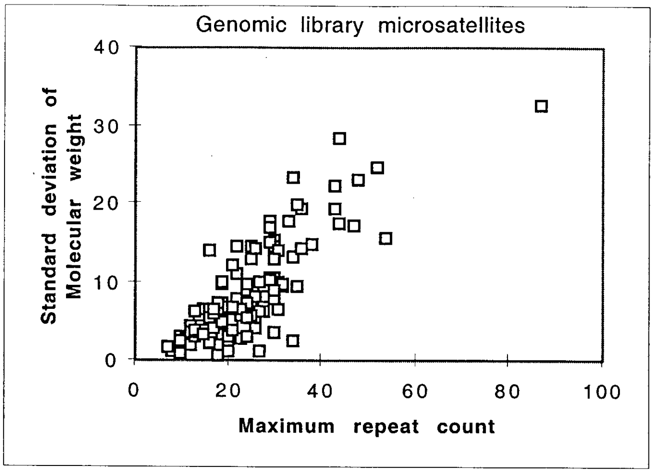
To account for the differences in variability of microsatellite markers, we compared the set of 136 poly(GA)n microsatellites from genomic libraries with the set of 47 poly(GA)n from GenBank to ensure that conclusions were drawn only from inherent differences between these two sources of microsatellite markers and not from differences due to motif. On the other hand, to investigate the effect of SSR motifs, we compared four groups of microsatellite markers isolated from genomic libraries. The first group consisted of 136 markers with the GA motif, the second of 25 markers with the GT motif, the third contained 9 markers with the CAT motif and the fourth, designated “non-CAT, AT-rich genomic”, was a combination of 24 markers with (ATT)n and (CTT)n poly-trinucleotides. These two classes were combined due to their rare occurrence and similar variability values. As a point of comparison, we also considered the set of 48 GC-rich trinucleotide motifs (those containing at least two-thirds G/C bases in each repeat unit and located in open reading frames of the genes) from the GenBank data. The mean values of statistical descriptors of variation for each class of the microsatellite markers are presented in Table 3.

The comparison of GA-containing microsatellites revealed that the SSR markers from genomic libraries were more variable in all respects (including allele count, PIC value, PCR size range) than their counterparts from the GenBank collection. Among poly dinucleotides of genomic origin, GA-containing SSLP markers were more variable than markers with the GT motif. All differences were statistically significant at P <0.05. When genomic library-derived microsatellite markers with trinucleotide motifs were compared, the CAT sequences showed lower variability than the other AT-rich trinucleotides. The non-CAT, AT-rich genomic group of markers was comparable with the highly variable GA polynucleotides of the same origin with respect to allele count per locus and for the PIC value. It had an even wider range of size variation (as would be expected due to the longer size of the repeat unit), which was 33.8 bp for the class of ATT and CTT motifs compared to 25.0 bp for the GA motif. Finally, the lowest values of all genetic variability parameters were detected for GC-rich polytrinucleotides derived from GenBank sequences, which on average detected only 2 alleles per locus, and the average difference between the highest and the lowest mw alleles was 6.3 bp.

Structure/length relation

The characteristics of genetic variability at each microsatellite were all related to the length of the microsatellite repeat, expressed in number of repeat units. There was a good correspondence between average repeat count and the mean values of the variability measures for a given class of markers, with longer microsatellites having higher values for genetic variability parameters (Table 3).

We also sought to develop a measure of microsatellite locus variability that could be used to make comparisons within a class of repeat types, i.e., dinucleotides or trinucleotides, and would incorporate, and give equal weight to, both the population components of diversity (such as PIC) and the length variation. Standard deviation of molecular weight (SDmw) is a suitable candidate for comparing the variability of different loci within a class because it accounts for both the allele size range and for the frequency of occurrence of various alleles of this locus. Mean values for SDmw for different classes of microsatellite markers are given in the last column of Table 3. On average, the SDmw expressed in base pairs was 8.44 for genomic library poly(GA)n and 4.09 for GenBank-derived poly(GA)n (P <0.01). For the GC-rich GenBank-derived trinucleotides, the standard deviation was only 2.37. The ATT- and CTT- containing microsatellites of genomic library origin had the highest SDmw (12.9 bp on average). While the average number of alleles and PIC values can be compared between polytrinucleotides and polydinucleotides, the SDmw is useful only within a class. This is because trinucleotide variation would naturally encompass a greater range of mw than dinucleotide variation for the same number of alleles and the same PIC value.



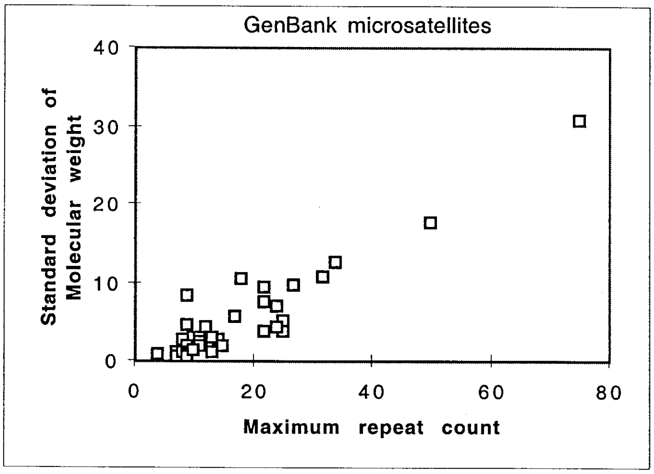


Fig. 3 Correlation between the standard deviations of molecular weight and maximum repeat count for genomic library and GenBank-derived poly(GA)n microsatellites

A strong positive correlation was observed between the SDmw and maximum repeat count for the 120 genomic library polydinucleotides (GA)n SSRs (r = +0.79) and for the 39 GenBank poly(GA)n SSRs (r = +0.93) for which both variables were available (Fig. 3). Similar correlations were obtained for the other groups of microsatellite markers. For SSLP loci with the GT motif and for AT-rich trinucleotides the coefficients of correlation were r = +0.73 and r = +0.70, respectively. We infer that SDmw is an excellent measure of SSRs' potential for array contraction or expansion and that the probability of expansion is therefore length-dependent.

Functional constraints on microsatellite variability

Table 4 shows the positions of microsatellite repeats in 27 cloned and completely sequenced genes in rice. Eleven microsatellites having di-nucleotide motifs and 2 with tri-nucleotide motifs are positioned in 5' or 3' untranslated regions (UTR), and 6 microsatellites are located in introns. Eight microsatellites (29.6%) were found in exons or open reading frames (ORFs), and these were all GC-rich, tri-nucleotide motifs. These 8 trinucleotide markers show very low levels of polymorphism, reflected in low PIC values and a small number of alleles, i.e., from 1 to 3 per locus. On the contrary, microsatellite sequences located in introns or in 5' and 3' untranslated regions (mostly poly(GA) or AT-rich, di- and trinucleotides) are much more polymorphic, with the number of alleles varying from 3 to 10.

Discussion

To characterize polymorphism at the 323 microsatellite loci developed for rice, we used a panel of 14 genotypes representing a diverse array of rice germplasm. This panel included 13 cultivars of *Oryza sativa* representing the *indica* and the *japonica* subspecies, as well as a wild accession of *Oryza longistaminata* to serve as an outgroup. The study focused on a comparative evaluation of marker polymorphism, emphasizing the differences in genetic variability of microsatellite sequences with different SSR motifs and originating from different sources (random genomic clones versus expressed sequence tags). It has been shown in early studies in human (Weber 1990), and confirmed in several different organisms, that the variability of microsatellite markers correlates well with the length of the tandem arrays (Goldstein and Clark 1995; Innan et al. 1997). In rice, it has been clearly demonstrated that SSLP markers with many repeat units also tend to embody large size differences among alleles.

For all variability characteristics, GenBank-derived microsatellites had lower values than genomic library microsatellites. In studies by Chin et al. (1996) and Becker and Heun (1995) on maize and barley microsatellite markers, respectively, relatively low levels of polymorphism were detected for markers developed from the GenBank sequence information. From 2 to 4 alleles per SSR locus were identified in maize (Chin et al. 1996) and from 2 to 6 with a mean of 2.7 alleles per marker in barley (Becker and Heun 1995). A large proportion of rice microsatellite markers developed based on the GenBank database contained polytrinucleotide motifs, preferentially with GC-rich SSR motifs such as GCC and GAC. These EST-derived markers were generally less polymorphic than dinucleotide-containing SSLPs. In this respect, the results of our study are in agreement with published data from other plant and animal species, which have reported low variability of most trinucleotide SSR loci (Chakraborty et al. 1997; Schug et al. 1998). CAT motifs, in particular, had low levels of polymorphism for reasons more fully discussed in the companion paper by Temnykh et al. (1999). In contrast to GC-rich trinucleotides, markers with ATT and CTT motifs derived from random genomic clones were much more variable. They were comparable with the class of the most polymorphic GA-containing microsatellites of "genomic" origin in number of alleles per locus and PIC value and had an even higher

Table 4 Position of microsatellite sequences in known genes rice

Rice map locus	GenBank accession number	Gene name	SSR motif	Position
RM101	D17586[a]	Carboxypeptidase I	(CT)37	3’ UTR
RM102	D17586[a]	Carboxypeptidase I	(CGG)8	exon
RM103	D16221	Endochitinase	(GAA)7	5’ UTR
RM120	M36469[b]	Alcohol dehydrogenase (adh-2)	(CT)9	3’ UTR
RM143	D78609	bZIP protein	(CGG)7	ORF
RM145	D16340	Aspartate aminotransferase	(GA)30	intron
RM146	X58877	Beta glucanase	(CT)11(TC)7	5’ UTR
RM144	Xg7711	Heat-shock protein 70	(ATT)11	intron
RM149	Z11920[a]	Heat-shock protein 82	(TA)10	5’ UTR
RM150	D14000	Lipoxigenase	(CGT)6(CGG)5	ORF
RM151	L37528[a]	MADS-box protein (MAD23)	(TA)23	3’ UTR
RM155	X07515[a]	Ribulose biphosphate carboxylase		
RM158	U12171	Anther-specific gene	(GGC)9	ORF
RM165	U33175	Sucrose-phosphate synthase	(CT)13	5’ UTR
RM173	D30794	Ferredoxin	(GA)9	5’ UTR
RM176	X64619	Alpha-amylase (Am-2)	(CCG)8	ORF
RM181	D78506[a]	w-3 fatty acid desaturase	(CT)13(AT)19	intron
RM182	L10346[a]	Beta-amylase	(AT)16	intron
RM180	D63901	13-kDa prolamin	(ATT)10	5’ UTR
RM184	U40708	Glycine-rich cell wall protein (Angrp-1)	(CA)7	5’ UTR
RM226	M29259[b]	Oryzacystatin	(AT)38	intron
RM190	X65183	Starch synthase; waxy gene	(CT)11	5’ UTR
	U08404[c]	Chloroplast carbonic anhydrase	(CCT)7	ORF
	U31771[c]	Orys 1	(AT)19	3’ UTR
	X53596[c]	Glycine-rich cell wall protein (GSA)	(GCA)4	ORF
	U37133[c]	Receptor kinase-linke protein (Xa21)	(CCG)5	exon
	U49113[c]	Protein phosphatase 2A	(CT)3	3’ UTR

[a] Microsatellite-containing sequences previously reported by Akagi et al. (1996)

[b] Microsatellite-containing sequences previously reported by Wu and Tanksley (1993)

[c] Markers were not mappable

range of variation in allele size. Therefore, our study breaks new ground in trinucleotide microsatellite analysis. We clearly demonstrate that these sequences comprise a heterogeneous group, where each class of sequences with a particular SSR motif has its own potential for variability. The prevalence of AT-rich repeats in non-coding regions of the genome coupled with the high level of variability suggest that these microsatellite sequences experience far fewer selective constraints than the GC-rich trinucleotides positioned in, or near, coding regions. This is consistent with reports from primates (Jurka and Pethiyagoda 1995), *Drosophila* (Goldstein and Clark 1995) and potato (Milbourne et al. 1998).

The location of specific SSRs in known gene sequences offers an opportunity to investigate the biological significance of microsatellite expansion and contraction on the functional aspects of the genes themselves. Economically significant phenotypic variation for grain quality associated with the expansion of a poly(CT) microsatellite in the 5’ UTR of the *waxy* gene has been reported in rice (Ayers et al. 1997). It remains to be seen whether any unusual phenotypic variation may be associated with the expansion of SSR in coding regions as has been reported with respect to several diseases in humans (McMurray 1995). While the same mechanism may play a role in generating phenotypic diversity in plants, variation associated with deleterious characters is less likely to be represented in the germplasm collections of agricultural species than among natural populations because undesirable mutations are commonly culled from agricultural populations.

The genotypes selected for this study represent a reference panel of mapping parents widely used in rice genome analysis. Microsatellite markers distinguish the two major sub-species (*indica* and *japonica*) and, in addition, detect higher levels of intra sub-specific variation (McCouch et al. 1997; Akagi et al. 1997). The inclusion of the wild species, *O. longistaminata* in this study confirms previous reports that reliable amplification can be achieved with these microsatellite primers on diverse members of the *Oryza* genus (Wu and Tanksley 1993; Panaud et al. 1996; Ishii, unpublished data, this lab). In this study, 90% of the primer pairs generated clean amplification products on this distantly related wild species. In addition, the inclusion of *O. longistaminata* demonstrated that even when a purely monomorphic pattern of SSR was observed between *indica* and j*aponica* varieties, polymorphism could be detected outside the cultivated gene pool. It remains to be confirmed how often the polymorphism detected in *O. longistaminata* was the result of expansion or contraction of the microsatellite motif itself and how often it was due to an accumulation of sequence differences coupled with insertion/deletion events in the flanking regions outside the SSR domains. The prevalence of null alleles in *O. longistaminata* as well as alleles with an unusually low or high molecular weight in comparison to common alleles for *O.sativa*, strongly suggests that sequence divergence in flanking regions may play a significant role in interspecific SSR variation.

This study introduces the use of the standard deviation of allele molecular weight (SDmw) to predict the variability of microsatellite markers. As it is a derivative of the number and frequency of alleles (the basis for PIC) on the one hand and the size range of the PCR fragments at each SSR locus on the other, it provides a comprehensive measure of microsatellite genetic diversity within a class. Standard deviation of allele molecular weight was most closely related to the maximum number of repeat units in rice microsatellites. This is similar to the observation made by Goldstein and Clark (1995) in *Drosophila*, except that the much larger number of data points in this study provides stronger validation of this trend. The use of standard deviation instead of variance reveals a linear, rather than exponential, positive correlation between the maximum length of an SSR allele and its potential for variation. It is important to remember that the number of units in a perfect array of tandem repeats, rather than the absolute length of the array in base pairs, is the principal determinant of a given microsatellite's propensity to mutate. The observation is consistent with microsatellite mutation models based on slipped-strand mispairing and polymerase slippage, because the likelihood of de novo mutations in these models is determined primarily by the number of units in the microsatellite array. It can be inferred that the length-dependent nature of DNA array expansion is likely to be a universal biological property of microsatellites from different organisms.

Note Primer sequences and images of silver-stained polyacrylamide gels containing amplification products in each of the 14 rice genotypes included in this study are available over the RiceGenes database (http://genome.cornell.edu/rice/). The images indicate clearly the intensity of amplification and the degree of stutter associated with each of the primer pairs. In association with the PIC value, number of alleles and the map location for each of the markers, this information should allow researchers to make rational decisions about the relative usefulness of specific markers for particular projects. The primers are also available as "Rice Map-Pairs" through Research Genetics (http://www.resgen.com).

Acknowledgements We gratefully acknowledge Olivier Panaud, Malcolm Schug, and Charles Aquadro for critical review of this paper, Takuji Sasaki for DNA from Nipponbare and Kasalath, Lihuang Zhu for DNA from Zhai-Ye-Qing 8 and Jing-Xi 17, Matthew Blair for DNA from N22, and Carole Morehouse for help formatting this manuscript. We are also grateful for financial support from the USDA (NRI Competitive Grant 93–37300–8703) and the Rockefeller Foundation (RF95001 #315 and RF95001 #342). This work was also funded by the USDA-ARS, the Texas Advanced Technology Development Experiment Station, and the Institute of Biosciences and Technology at Texas A&M University.

References

Akagi H, Yokozeki Y, Inagaki A, Fujimura T (1996) Microsatellite DNA markers for rice chromosomes. Theor Appl Genet 93: 1071–1077

Akagi H, Yokozeki Y, Inagaki A, Fujimura T (1997) Highly polymorphic microsatellites of rice consist of AT repeats, and a classification of closely related cultivars with these microsatellite loci. Theor Appl Genet 94: 61–67

Anderson JA, Churchill GA, Sutrique JE, Tanksley SD, Sorrels ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36: 181–186

Ayers NM, McClung AM, Larkin PD, Bligh HFJ, Jones CA, Park WD (1997) Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. Theor Appl Genet 94: 773–781

Becker J, Heun M (1995) Barley microsatellites: allele variation and mapping. Plant Mol Biol 27: 835–845

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314–331

Bowcock AM, Ruiz-Linares A, Tomfohrde J,Munch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymoyphic microsatellites. Nature 368: 455–457

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62: 1408–1415

Chakraborty R, Kimmel M, Strivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di- tri- and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94: 1041–1046

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

Chin ECL, Senior ML, Shu H, Smith JSC (1996) Maize simple repetitive DNA sequences: abundance and allele variation. Genome 39: 866–873

Dellaporta SL, Wood T, Hicks TB (1983) A plant DNA mini preparation: version II. Plant Mol Biol Rep 1: 19–21

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Mirissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380: 152–154

Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ, Evans CA, DeAngelis MM, Levinson DM, Kruglyak L, Goodman N, Copelang NG, Jenkins NA, Hawkins TL, Stein L, Page DC, Lander ES (1996) A comprehensive genetic map of the mouse genome. Nature 380: 149–152

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational process of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91: 3166–3170

Goldstein DB, Clark AG (1995) Microsatellite variation in North American populations of *Drosophila melanogaster*. Nucleic Acids Res 23: 3882–3886

Harr B, Zangerl B, Brem G, Schlötterer C (1998) Conservation of locus-specific microsatellite variability across species:a comparison of two *Drosophila* sibling species, *D. melanogaster* and *D. simulans*. Mol Biol Evol 15: 176–184

Innan H, Terauchi R, Miyashita T (1997) Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana.* Genetics 146: 1441–1452

Jarne P, Lagoda PJL (1996) Microsatellites from molecules to populations and back. Trends Ecol Evol 11: 424–429

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

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4: 203–221

McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Cho YG, Huang N, Ishii T, Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35: 89–99

McMurray CT (1995) Mechanisms of DNA expansion. Chromosoma 104: 2–13

Meyer E, Wiegand P, Rand SP, Kuhlmann D, Brack M, Brinkmann B (1995) Microsatellite polymorphisms reveal phylogenetic relationships in primates. J Mol Evol 41: 10–14

Milbourne D, Meyer RC, Collins AJ, Ramsay LD, Gebhardt C, Waugh R (1998) Isolation, characterization and mapping of simple sequence repeat loci in potato. Mol Gen Genet 259:233–245

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

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1: 215–222

Schlötterer C, Vogl C, Tautz D (1997) Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics 146: 309–320

Schug MD, Mackay TFC, Aquadro CF (1997) Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nat Genet 15: 99–102

Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TFC, Aquadro CF (1998) The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster. Mol Biol Evol 5: 1751–1760

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457–462

Temnykh S, Park W, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (1999) Mapping and genome organization of microsatellites in rice (*Oryza sativa* L.). Theor Appl Genet 100:698–712

Weber JL (1990) Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. Genomics 7: 524–530

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