Theor Appl Genet (1995) 91:160–168

A. Depeiges · C. Goubely · A. Lenoir · S. Cocherel
G. Picard · M. Raynal · F. Grellet · M. Delseny

Identification of the most represented repeated motifs in *Arabidopsis thaliana* microsatellite loci

Received: 20 May 1994 / Accepted: 18 October 1994

Abstract The major simple sequence repeats present in the *Arabidopsis* genome were identified by Southern hybridizations with 49 oligonucleotide probes matching all the possible combinations of motifs up to 4 nucleotides long. The method used allowed us to perform all the hybridizations under the same temperature conditions. A good correlation was observed with the data obtained from database analysis, indicating that the method can be useful for identifying the major classes of microsatellite loci in species for which few or no sequence data are available. AG/CT, AAG/CTT, ATG/CAT and GTG/CAC are the major motifs present in the *Arabidopsis* genome that can be used as convenient probes to isolate microsatellite loci by screening libraries. AAG/CTT is the more frequent of these motifs, and its relative frequency in *Arabidopsis* is much higher than averagely found in the plant kingdom. About 8% of the cDNA clones from an immature silique library contains AG/CT, AAG/CTT or ATG/CAT microsatellite loci. Several microsatellite loci were isolated by screening genomic and cDNA libraries. Twenty-six tri-nucleotide loci were PCR amplified from four different ecotypes, and polymorphism was observed for 12 of them; 10 loci showing two alleles and 2 loci showing three alleles.

Key words *Arabidopsis thaliana* · Microsatellite · Polymorphism · PCR amplification

Introduction

Microsatellites (Litt and Luty 1989) are a subclass of repeated sequences that consists of stretches of short tandemly repeated nucleotide motifs (1–5 bp) scattered throughout the genome of eukaryotes. These sequences have also been designated sequence-tagged microsatellite sites (STMSs) by Beckmann and Soller (1990), Williamson et al. (1992) and Serikawa et al. (1992), simple sequence repeats (SSRs) by Jacob et al. (1991) and short tandem repeats (STRs) by Edwards et al. (1991). Variations in number of repeated motifs generate extensive polymorphism, which makes microsatellite loci very useful markers for genetic mapping, DNA typing or population genetics studies (Beckmann and Soller 1990; Epplen et al. 1992).

Most of the knowledge concerning microsatellites has been obtained from the animal kingdom, mainly from mammals. In humans, it has been shown that the $(AC/GT)_n$ repeat, present at about $5 \cdot 10^4$ loci per haploid genome, is one of the most common motifs (Hamada et al. 1984; Moore et al. 1991; Beckmann and Weber 1992), while all trimeric and tetrameric repeats were estimated to account for a total of about $4 \cdot 10^5$ loci (Edwards et al. 1991). Amplification of trinucleotide repeats in several loci have been shown to be responsible for such human genetic diseases as X fragile syndrome (Fu et al. 1991) or myotonic dystrophy (Brook et al. 1992; Fu et al. 1992).

In plants, the presence of $(AC/GT)_n$ and $(AG/CT)_n$ repeats was first reported by Condit and Hubbell (1991), although the presence of similar alternating polymers had been detected much earlier in wheat and radish nuclear DNA by hybridization techniques (Delseny et al. 1983). Searches for microsatellites in plant DNA sequence databases indicated that the $(AC/GT)_n$ repeat is generally less frequently represented than in mammals, the most common motifs being $(AT/AT)_n$ followed in decreasing order by $(A/T)_n$ and $(AG/CT)_n$. Trinucleotide and tetranucleotide repeats are also present in plant genomes, the most frequent motifs being $(AAG/CTT)_n$ and $(AAT/ATT)_n$ (Akkaya et al. 1992; Lagercrantz et al. 1993; Morgante and Olivieri 1993; Wang et al. 1994). A wide range of variation in the abundance of most of these repeats was observed in various species. For

Communicated by J. S. Beckmann

A. Depeiges (✉) · C. Goubely · A. Lenoir · S. Cocherel
G. Picard
GDR977 BIOMOVE CNRS, Université Blaise-Pascal, 24, Avenue des Landais, 63177 Aubiere Cedex, France

M. Raynal · F. Grellet · M. Delseny
URA 565 and GDR 1003 CNRS, Université de Perpignan, 66860 Perpignan Cedex, France

example, $(CCG/CGG)_n$ is abundant in rice (Zhao and Kochert, 1993) while generally poorly represented in other higher plants (Lagercrantz et al. 1993; Wang et al. 1994). Moderate to high levels of polymorphism were reported for several loci in various species: soybean (Akkaya et al. 1992; Morgante and Olivieri 1993), *Brassica* (Lagercrantz et al. 1993), rice (Zhao and Kochert 1993, Wu and Tanksley 1993), grapevine (Thomas and Scott 1993), maize (Senior and Heun 1993), sunflower (Brunel 1994) and *Arabidopsis* (Bell and Ecker 1994). Plant DNA fingerprinting with various oligonucleotide probes such as $(TCC)_n$, $(GTG)_n$, $(GATA)_n$ and $(GACA)_n$ suggests that these repeats are present in plants in highly variable abundances according to species (Weising et al. 1991; Beyermann et al. 1992; Poulsen and Weising 1993).

The work reported in the present paper was initiated in order to search for microsatellite repeats present in the genome of *Arabidopsis thaliana* but not in the yeast *Saccharomyces cerevisiae*. Such repeats would be useful for identifying overlapping clones in an *Arabidopsis* YAC library by fingerprinting the clones, as has been reported by D. Cohen's group for a human YAC library, using LINE 1 and Alu families as probes (Bellanné-Chantelot et al. 1992). Southern hybridizations with oligonucleotide probes allowed us to detect the main repeats present in *Arabidopsis* and in yeast and revealed that every repeat detected in *Arabidopsis* is also present in the yeast genome. In *Arabidopsis*, a good correlation was observed between the hybridization data and the main classes of repeats found in databases. AG/CT, AAG/CTT, ATG/CAT and GTG/CAC are the major motifs present in the *Arabidopsis* genome that can be used as convenient probes to isolate microsatellite loci by screening libraries. The tri-nucleotide AAG/CTT is more represented in *Arabidopsis* than in most other plants. Several *Arabidopsis* microsatellite loci containing tri-nucleotide repeats were polymerase chain reaction (PCR)-amplified, and an inter-ecotype polymorphism was detected for about half of them.

Materials and methods

Hybridization with oligonucleotide probes

Arabidopsis thaliana DNA (ecotype C24) was extracted from 4-week-old plants by the CTAB method (Towner 1991) and purified by CsCl density gradient centrifugation. Yeast DNA (*Saccharomyces cerevisiae*, YNN295) was purchased from Clontech. Digestions by *Eco*RI and *Hin*dIII were performed at 37 °C for 2 h in buffers supplied by the manufacturer. Samples of about 1 µg of DNA were digested by one or both enzymes, electrophoresed in 1% agarose gels for 16 h at 3 V/cm in TBE buffer and transferred onto Hybound N nylon membranes.

The 49 oligonucleotide probes that were used (Table 1) match all the 316 possible different combinations of repeated motifs of 1, 2, 3 or 4 nucleotides long. Oligonucleotides were 3'-end-digoxigenin-labeled according to the instructions of the Boehringer Mannheim's DIG oligonucleotide labeling kit. Hybridizations were performed in 15 ml of 5 × SSPE, 0.5% SDS, 0.1% lauroyl-sarcosinate, 2% blocking reagent (Boehringer), 50 mM Tris pH 8.0, for 1 h at 37 °C, using 100 pmoles of oligonucleotide-labeled probe. Three 10-min non-stringent washings were done at room temperature in 3*M* tetra-methyl-ammonium-chloride (TMA), 2 m*M* EDTA, 50 m*M* Tris pH 8.0, followed by a 5-min stringent washing in the same buffer at 4°–5 °C below melting temperature (Tm). The use of TMA strengthens the interaction of A-T base pairs, which then becomes similar to that of the G-C base pairs. Consequently, the Tm value does not depend on the G+C content but only on oligonucleotide size (Wood et al. 1985). This allows the use of a single temperature for all the hybridization experiments involving oligonucleotides of the same length. Since only 18 mers have been used in this study (Table 1), the Tm value is 57.5 °C according to Wood et al. (1985).

Detection of labeled probes was achieved using the digoxigenin chemiluminescent kit supplied by Boehringer Mannheim. For every probe, the films were exposed for 5 h. When only very weak or no signals were detected, a new reaction was performed followed by a 16-h exposure. All positive signals were confirmed by a second independent hybridization experiment.

Table 1 List of the 49 oligonucleotide probes

(A) 18	(AAT) 6	(AAAT) 4.5	(AACC) 4.5	(AGGC) 4.5
(C) 18	(AAG) 6	(AAAG) 4.5	(GATA) 4.5	(AGCT) 4.5
	(AAC) 6	(AAAC) 4.5	(ATAC) 4.5	(AGCG) 4.5
(AT) 9	(ATG) 6	(AATT) 4.5	(GGAT) 4.5	(AGCC) 4.5
(AG) 9	(AGT) 6	(AATG) 4.5	(ATGC) 4.5	(ACGT) 4.5
(AC) 9	(TCC) 6	(AATC) 4.5	(ATCG) 4.5	(ACGG) 4.5
(GC) 9	(AGC) 6	(AAGT) 4.5	(GACA) 4.5	(ACGC) 4.5
	(ACG) 6	(AAGG) 4.5	(AGTG) 4.5	(ACCG) 4.5
	(GTG) 6	(AAGC) 4.5	(AGTC) 4.5	(ACCC) 4.5
	(GCC) 6	(AACT) 4.5	(AGGT) 4.5	(GGGC) 4.5
		(AACG) 4.5	(AGGG) 4.5	(GGCC) 4.5

Search in the GenBank database

The BISANCE software package (Dessen et al. 1990) was used to search the 2335 *Arabidopsis* DNA sequences of the GenBank database (version October 1993) for the presence of mono-, di-, tri-, and tetranucleotide repeats. The STRING 2 program was run to identify sequences corresponding to or complementary to the motifs present in the 49 oligonucleotide probes and carrying a minimum number of n repeats, with $n = 15$ for mononucleotides, $n = 8$ for dinucleotides, $n = 5$ for trinucleotides and $n = 4$ for tetranucleotides. No mismatch was allowed. The results were manually checked in order to eliminate redundant sequences, mitochondrial and chloroplastic sequences, rDNA-repeated units and poly(A) tails of cDNA sequences.

Cloning of microsatellite loci

Genomic DNA was digested to completion by *Mbo*I and ligated with T4 DNA ligase to the *Bam*HI site of the Bluescript SK plasmid vector (Stratagene). After transformation into *E. Coli* DH5α cells, about 2000 recombinant colonies were screened using various oligonucleotide probes. After purification of some positive clones, the nucleotide sequences of the inserts were determined using the dideoxychain termination method (Sanger et al. 1980).

Search of microsatellites in a cDNA library

A cDNA library derived from immature silique mRNAs (Giraudat et al. 1992) was used for this screening. It was prepared in the Lambda ZAP II vector. An aliquot of this library was in vivo excised and plated. One thousand colonies were picked at random, ordered in microtiter plates and replicated on nylon Hybond N filters. These filters were hybridized under the same conditions with the oligonucleotide probes as mentioned above. Positive clones were sequenced either manually or using an ABI 373 machine. The nucleotide se-

quences were translated into amino acid sequences, and the Fasta program (Pearson and Lipman 1988) was used for homology search in databases in an attempt to identify the corresponding proteins.

PCR conditions

Primer sequences were designed manually with the aim of producing well-matched primers, 15–20 nucleotides long, thereby avoiding obvious repeated sequences, with a G or C in the 3'-most position and with a Tm of 58° ± 3 °C. Primers were synthesized by Eurogentec. The specificity of each primer pair was optimized by the use of different annealing temperatures and in some cases by the addition of 2.5% formamide in the reaction solution. PCR reactions were performed in a volume of 25 µl containing 50 ng of genomic DNA, 2 m*M* $MgSO_4$, 10 m*M* KCl, 10 m*M* $(NH4)_2SO_4$, 20 m*M* Tris HCl (pH 8.8), 0.1% Triton X100, 1 µ*M* of each primer, 0.2 m*M* of each dNTP including [^{35}S]dATP (400 m*M*) and 0.25 U Vent DNA polymerase (Biolabs). After an initial denaturation step of 2 min at 94 °C, 25 cycles of 30 s at 94 °C, 30 s at the optimized annealing temperature and 30 s at 72 °C were carried out followed by a final extension of 4 min at 72 °C in a Techne PHC-3 thermocycler. One volume of a stop solution (deionized 98% formamide, 10 m*M* EDTA, 0.1% Xylene cyanol, 0.1% Bromophenol blue) was then added to the amplification products, and 4 µl of each reaction was analyzed on a 6% polyacrylamide denaturing gel containing 8.3 *M* urea. For sizing, a DNA sequencing ladder prepared from a template of a known sequence was run adjacent to the PCR product.

Results

Hybridizations with oligonucleotide probes

Both *Arabidopsis* and yeast genomic DNAs were hybridized with oligonucleotide probes corresponding to all possible repeated motifs 1, 2, 3 and 4 bp long, using a standardized method. *Arabidopsis* DNA cut by *Eco*RI, *Hin*dIII and *Eco*RI + *Hin*dIII and yeast DNA cut by *Eco*RI and *Eco*RI + *Hin*dIII were electrophoresed and transfered onto nylon membranes. Twelve identical membranes were prepared, each being successively hybridized with 4–6 of the oligonucleotide probes listed in Table 1. Most of the probes did not reveal any signal or only a few faint bands (see Fig. 1 the $(GATA)_{4.5}$ probe as an example), suggesting that the complementary repeats are either absent or present only in a very low amount in the genome. Only 9 probes revealed clear fingerprint patterns with at least one of the species (Fig. 1); 6 probes gave multiple band fingerprints in both species, with variable signal intensities: $(A)_{18}$, $(ATG)_6$, $(AAG)_6$, $(GTG)_6$, $(GGAT)_{4.5}$ and $(AAAC)_{4.5}$. In the *Arabidopsis* genome, in addition to the fingerprint pattern, the $(A)_{18}$ probe revealed one band with a higher signal intensity at the top of the gel. Since its mobility was not affected by *Eco*RI or *Hin*dIII digestion, this band did not correspond to a microsatellite locus but more likely to large clusters of tandemly repeated sequences carrying poly-A stretches but lacking the enzyme recognition sites. The same explanation can account for the presence of a second major band in the *Hin*dIII-digested sample only, except that the repeated sequences have to carry a *Eco*RI recognition site. The $(AG)_9$ probe also gave

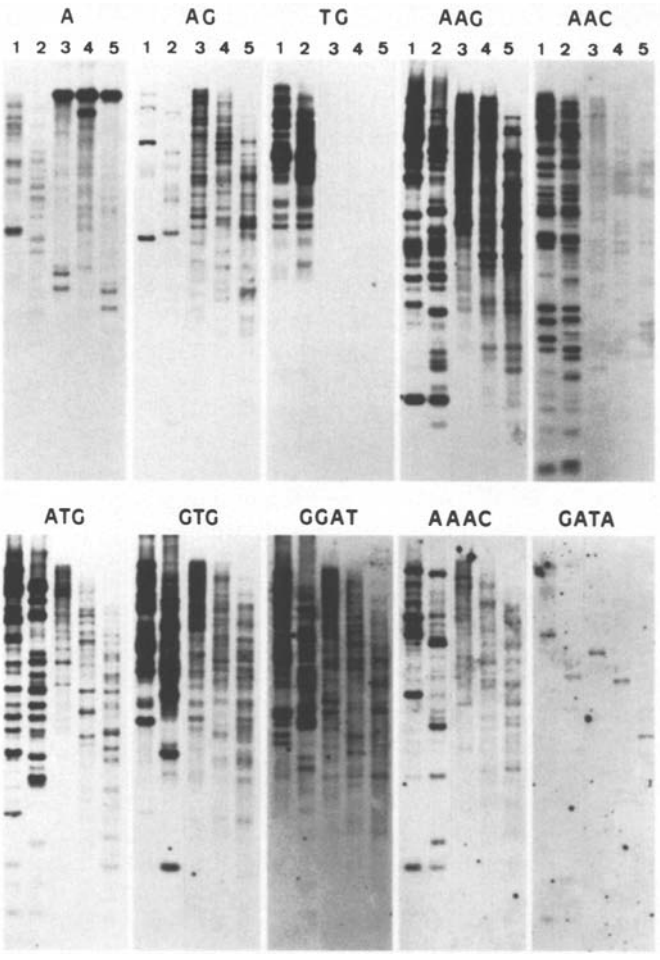

Fig. 1 Hybridization of *Arabidopsis* and yeast DNA with 18-mer oligonucleotide probes carrying the repeated motifs shown above each bloc. *Lanes 1 and 2* yeast DNA cut with *Eco*RI (*1*) or *Eco*RI + *Hin*dIII (*2*) *Lanes 3–5 Arabidopsis* DNA cut with *Eco*RI (*3*), *Hin*dIII (*4*) or both (*5*). All the pictures correspond to film exposures of 16 h, except for hybridizations with $(A)_{18}$ and $(AG)_9$ probes, which correspond to a 5-h exposure

signals in both species but only a few bands were revealed in yeast, suggesting that the repeat, largely present in the *Arabidopsis* genome, is less represented in yeast. $(TG)_9$ and $(AAC)_6$ probes revealed very clear fingerprints in yeast but only faint bands in *Arabidopsis* (visible on the film), indicating that the repeats are poorly represented in this species.

Search in the GenBank database

A search was conducted to identify microsatellite loci in the *Arabidopsis* genome. As shown in Table 2, 169 loci were identified that consist of perfect repeats of at least 15 mononucleotide, 8 dinucleotide, 5 trinucleotide and 4 tetranucleotide motifs. Only 7 motifs were found in more than 5% of the identified loci, and together these 7 motifs account for 85% of the loci. The most represented in the database sequences were by far AAG/CTT, found

Table 2 Number of microsatellite loci found in the Genbank database. (Only loci with a minimum number of repeated motifs were considered. The minimum number of repeats is 15 for mononucleotides, 8 for dinucleotides, 5 for trinucleotides and 4 for tetranucleotides. Numbers between parentheses are relative frequencies. Searches for all the tetranucleotide motifs not mentioned in the table did not reveal any loci)

A/T	23	(0.136)	AAT/ATT	0	
G/C	0		AAG/CTT	60	(0.355)
AT/AT	9	(0.053)	AAC/GTT	9	(0.053)
AG/CT	18	(0.106)	ATG/CAT	16	(0.095)
AC/GT	3	(0.018)	AGT/ACT	2	(0.012)
GC/GC	0		TCC/GGA	6	(0.035)
AAAG/CTTT	2	(0.012)	AGC/GCT	4	(0.024)
AAAC/GTTT	1	(0.006)	ACG/CGT	2	(0.012)
AAGG/CCTT	1	(0.006)	GTG/CAC	10	(0.059)
GATA/TATC	1	(0.006)	GGC/GCC	2	(0.012)

in 35.5% of the identified loci, followed by A/T (13.6%), AG/CT (10.6%), ATG/CAT (9.5%), GTG/CAC (5.9%), AT/AT (5.3%) and AAC/GTT (5.3%). The 42 remaining motifs either were found in only a few loci (9 motifs) or not found at all (33 motifs).

Figure 2 shows the size distribution of the identified microsatellite loci. Most of the 169 loci consist of less than 20 bp of perfectly repeated motifs. Among the 111 loci carrying trinucleotide motifs, only 14 display 7 repeats or more, the longest harboring a GTG/CAC cluster of 11 repeats. In the same way, the five loci carrying tetranucleotide motifs display 4 repeats (3 loci), 5 repeats (1 locus) and 6 repeats (1 locus). The distributions of the 30 loci carrying dinucleotide repeats are more heterogeneous. Among the 18 AG/CT loci, only 4 show more than 10 repeats while among the 9 AT/AT loci, 6 show more than 10 repeats, the longest harboring a 25-repeat cluster. The distribution of the 23 A/T loci is different from the others in that it is clearly bi-modal: 17 loci carry a poly A/T stretch shorter than 21 bp, while the 6 remaining loci have stretches of 35 bp or more, the longest reaching 79 bp.

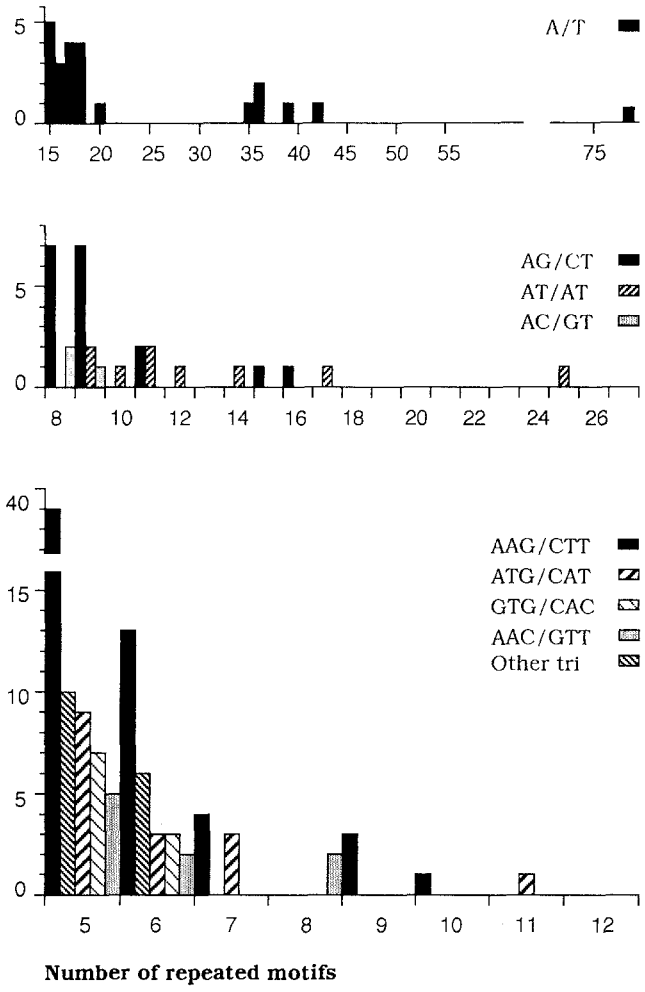

Fig. 2 Size distribution of *Arabidopsis* microsatellite loci found in the GenBank database. The *X axis* indicates the number of repeated motifs at each locus. The *Yaxis* indicates the number of loci. *Other tri* gives the pooled data obtained with the 6 remaining trinucleotide motifs. The 5 loci carrying tetranucleotide repeats are not shown

Sequences of microsatellite loci

In order to check that the signals detected by hybridization really reflect microsatellite loci, a *Mbo*I genomic library was constructed in the Bluescript plasmid vector and screened by hybridization with the $(AG)_9$, $(AAG)_6$ and $(ATG)_6$ oligonucleotide probes. Four positive clones were recovered with each of the 3 probes, but 1 was lost during the purification steps. The library was also screened with the $(GTG)_6$ probe, but positive clones cannot be detected because the probe hybridized to the plasmid vector. The nucleotide sequences of the 11 positive clones were determined, and their main features are presented in Table 3. As shown, every clone carries at least one cluster of repeats homologous to the probe. This confirms that the observed fingerprints in Fig. 1 reflect the presence of microsatellite loci. Two clones, pATag3 and pATaag4, harbor two clusters separated by 25 and 12 bp, respectively. pATag2 carries two clusters separated by more than 102 bp, one of them being linked to a $(TG)_6$ cluster.

Screening a cDNA library

Nine hundred and twenty-seven ordered cDNA clones from an immature silique library previously selected for partial random sequencing (Höfte et al. 1993) were screened for the presence of microsatellites using the $(ATG)_6$, $(AAG)_6$ or $(AG)_9$ probes. The results of this screening indicated that 49 colonies (5.3%) hybridized to $(AAG)_6$, 18 (1.9%) to $(ATG)_6$ and 10 (1.1%) to $(AG)_9$. These values roughly confirm the trend that was observed by screening the database (Table 2). Several positive clones were selected for further characterization by partial sequencing. Out of 38 that were analyzed in this way, we were able to confirm the presence of the expected repeats in 15 of them. The sequences of 8 of them

Table 3 Nucleotide sequence of microsatellite loci isolated from a *Mbo*-I genomic library. (Clones pATag, pATaag and pATatg were screened with the $(AG)_9$, $(AAG)_6$ and $(ATG)_6$ oligonucleotide probes, respectively. Coordinates originate from the 5'-end of the 11 sequences deposited in the EMBL database under the accessions X79419 to X79421 (PATaag2 to pATaag4); X79422 to X79425 (pATag1 to pATag4); X79426 to X79428 (pATatg1 to pATatg3). Only coordinates of nucleotides immediately preceding repeated motifs are indicated)

pATag1	1—3 $(GA)_7$,—301
pATag2	1—141 $(TG)_6$,—162 $(TC)_8$,—279 $(TC)_6TT(TC)_4TT(TC)_3$,—687
pATag3	1—137 $(AG)_5$,—173 $(GA)_{10}$,—400
pATag4	1—116 $(TC)_7$,—280
pATaag1	1—56 $(CTT)_7$,—349
pATaag2	1—74 $(AAG)_6$,—231
pATaag3	$GAAGAG(GAA)_4$,—202
pATaag4	1—577 $(TTC)_5$,—604 $(TTC)_5$,—821
pATatg1	1—95$GATGGATGAAT(GAT)_4$,—398
pATatg2	1—541 $(ATG)_3GTG(ATG)_3GTGATTCTG(ATG)_2$,—749
pATatg3	1—56 $(GAT)_7$,—236

Table 4 Nucleotide sequence of the microsatellite loci isolated from a cDNA library. (Sequence adjacent to the main cluster of repeats are given only when some similarity with the repeated motif can be seen. The location of the repeats with regard to the coding sequence is indicated as well as the putative function of the protein encoded by the cDNA when significant homologies were found between the predicted amino acid sequence and protein sequences in the database)

Clones	Pattern of repeats	Location	Putative function
ATTS0191	$(TGA)_7$	3'-untranslated	Calmodulin 1
ATTS0359	TGT $(TGA)_6$ TCT(TGA)GGA	3'-untranslated	Cyclophilin
ATTS2753	$(CTT)_5$	Coding	3-keto acyl CoA-thiolase
ATTS0184	$(TCT)_5$ CCTTCA	3'-untranslated	?
ATTS0185	CAT $(CTT)_5$ CGT	5'-untranslated	S 13 ribosomal
ATTS0200	AAA $(AAG)_5$	5'-untranslated	?
ATTS0262	$(GAA)_7$ CAA	3'-untranslated	Histone H2A
ATTS0316	$(TC)_2$ $CC(TC)_9$ $GTTTTCG(TC)_4$	3'-untranslated	?
ATTS2750	$(CAT)_6$ CAG	3'-untranslated	ATP-binding

with 5 or more repeats are shown in Table 4. A ninth cDNA (ATTS0262), previously selected by database screening and shown in Table 4, was confirmed in these experiments. The remaining 6 have only 4 repeats. Only 6 of the 38 clones analyzed were entirely sequenced. All of them carried the expected repeats. Thus, the most likely explanation for our failure to confirm the presence of repeats in 23 clones is that they escaped detection because they are located in the non-sequenced parts of the clones. For all the clones reported in Table 4 we could determine whether the repeats are located within the coding sequence or in the untranslated flanking regions. Only 1 clone had repeats in the coding region, and these determined a stretch of five serine residues. Two are located in 5'-non-coding regions, and the others are in the 3' non-coding sequences. An additional repeat, $(CAT)_6$, which was not selected by the hybridization procedure, was also observed by chance. We have also listed the putative function of the corresponding protein when it could be determined.

Detection of polymorphism

From the data extracted from the database or shown in Table 3 and 4, 36 microsatellite loci, which consist of tri-nucleotide repeats, were selected for the ability to design putatively good primers for PCR amplification. For each locus, PCR reactions were performed using genomic DNA from four different ecotypes (C24, Landsberg, Columbia and Bensheim) as template. After optimization of the PCR conditions, amplification products were obtained from the four ecotypes for 24 loci. One locus amplified only from three ecotypes and 1 only from two. The 10 remaining loci did not amplify or gave multiple non-specific products.

Polymorphisms obtained for 5 microsatellite loci are shown in Fig. 3. Each amplified locus displayed a specific pattern, including only one major band that corresponded to the expected size of the fragment. In addition, as usually observed in microsatellite PCR amplification products, a ladder of minor bands differing in length by one nucleotide was detected. These bands may result from both a template independent addition of one or several nucleotides to the 3'-end of the product strands and an incomplete extension by the polymerase (Weber 1990; Litt 1991). Slippage of the polymerase is another cause that can generate minor bands differing in length by 3 nucleotides (Weber and May 1989). The amplification products of loci ATPOSF21 (Fig 3 no. 2) and ATTS0262 (Fig. 3 no. 5), display a band with an intensity only slightly lower than that of the major one.

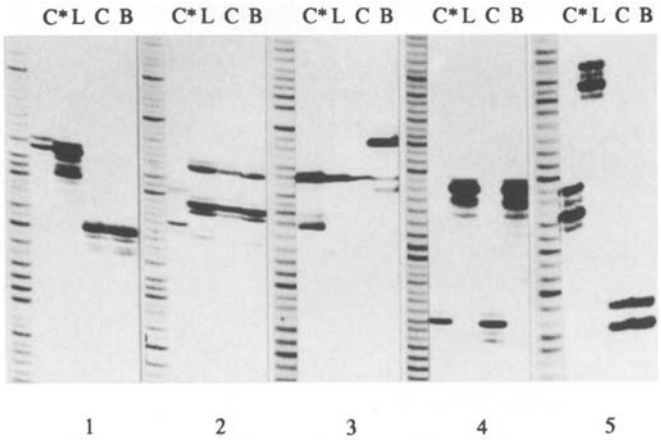

Fig. 3 Detection of polymorphism at microsatellite loci. Microsatellite loci were amplified by PCR from genomic DNA from four ecotypes (*C** C24, *L* Landsberg, *C* Columbia, *B* Bensheim), and the products separated on a 6% denaturing polyacrylamide gel. An illustration of the results obtained is given for 5 loci (*1* ATHGAPAB, *2* ATPOSF21, *3* PATaag1, *4* ATTS0392, *5* ATTS0262). For sizing, a DNA sequencing ladder was run adjacent to the PCR product

This additional band corresponds to a longer fragment than expected. In order to check whether these two bands reflect the presence of two different alleles, PCR amplification products of 15 individual plants from the same ecotype were obtained. The observed patterns (not shown) were the same for all plants and similar to those shown in Fig. 3, suggesting that the longer fragment results from a PCR artefact.

Table 5 is a summary of the data obtained for 26 amplified microsatellite loci that generate amplification products from at least two ecotypes. Twelve loci show polymorphisms after a screening of four *Arabidopsis* ecotypes. Three different alleles were detected at 2 loci (ATTS0191 and ATTS0262) and two for the remaining 10. Loci with variable levels of complexity in the repeated motifs were selected in this experiment. The simplest carry only one block of perfect repeats flanked by unrelated sequences (loci ATTIF4A2, PATaag2, ATTS0392 etc...), while the more complex carry several blocks of repeats separated and/or flanked by closely related tri-nucleotide motifs (loci ATHCOLR, ATPOSF21 etc...). The data of Table 5 show that both types detect polymorphism.

Table 5 Detection of polymorphism between four ecotypes

Microsatellite repeat type	Clones[a]	Pattern of repeats	Allelic distribution[b]			
			C*	L	C	B
	ATTIF4A2	$(GAA)_9$	N	N	N	N
	ATTS0184	$ATT(CTT)_4CTC\ (CTT)$	N	N	N	N
	ATHGAPAB	$CTC\ (CTT)_7CT\ (CTT)_2$	N	N	Z	Z
	pATaag1	$(CTT)_7\ (CCT)_2$	N	N	N	Z
	pATaag2	$(CTT)_6$	N	N	N	N
AAG/CTT	ATTS0392	$(GAA)_7$	N	Z	N	Z
	S45910	$CCT(TCC)_2(TTC)_7T(CTT)_2$	N	N	Z	Z
	ATHTUB9B	$(CTT)\ T\ (CTT)_5\ CTG\ (CTT)_2\ CCT\ (CTT)\ CTG$	N	N	N	N
	T04097	$(AAG)_4\ ATG\ (AAG)_3\ AGT$	N	...	Z	...
	ATTS966	$(CTT)_2\ (CAT)_3\ (CTT)_3$	N	N	N	N
	ATTS0262	$(GAA)_7CAA$	N	Z	T	T
	ATTS0185	$CAT\ (CTT)_5\ CGT$	N	N	N	N
	ATTS0191	$(TGA)_7$	N	Z	N	T
	ATTS0359	$TGT\ (TGA)_6\ TCT\ (TGA)\ GGA$	N	N	N	N
ATG/CAT	ATTOP1	$CTG\ (ATG)_7$	N	Z	Z	N
	ATCAN	$(CAT)_7\ GAC\ (CAT)\ C\ (CAT)_5\ (TAT)_2\ (CAT)$	N	N	N	N
	T04805	$CAC\ (CAT)_3\ (CCT)\ (CAT)_4\ (CCT)_2\ (CAT)_2$	N	N	N	N
	pATatg2	$(ATG)_3\ GTG\ (ATG)_3\ GTG\ (N)_5\ (GAT)_2$	N	N	N	N
GTG/CAC	ATHCOLR	$AAC\ (CAC)_2\ GTC\ (CAC)\ GAC\ (CAC)\ CAT\ (CAC)\ (CTC)_2\ (CAC)\ GAC\ (CAC)_3\ GACCTC\ (CAC)_4$	N	N	N	N
	ATPOSF21	$CAG\ (CAA)_8\ CAGCAACACCAA\ (CAG)_4\ (CAA)\ CAG\ (CAA)\ CAG$	N	Z	Z	Z
AAC/GTT	ATCSR12	$(CAA)_6\ CAT$	N	Z	Z	Z
	T04156	$(GTT)_2\ ATT\ (GTT)_4$	N	N	N	N
AGG/CCT	T04049	$(AGG)_6\ AGC$	N	N	...	N
	T04351	$(CTC)_6$	N	Z	Z	Z
AGC/GCT	ATTS0487	$(GAT)_3\ GAA\ (CAG)_7$	N	N	N	N
	ATHHANKA	$(AGC)\ AG\ (AGC)_5\ AGG$	N	Z	N	N

[a] Clones were named by their database mnemonics, except for clones originating from the screening of the genomic library (see Table 3)

[b] The allele found in the C24 genotype is arbitrarily indicated by N. The second allele found is designated by Z and the third by T.

Discussion

Because the original goal of this work was to search for simple sequence repeats present in the *Arabidopsis thaliana* genome but absent in yeast, we conducted a systematic search by the hybridization of oligonucleotide probes on Southern blots. Forty-nine different oligonucleotide probes were designed to potentially detect all the motifs corresponding to the 316 possible combinations of 1, 2, 3 and 4 nucleotides. Performing hybridizations with multiple oligonucleotide probes was made easier following the method originally described by Wood et al. (1985), which allows the standardization of experimental conditions by using the same Tm for all oligonucleotide probes of the same length. The only conspicuous difference between the two species was observed with $(TG)_9$ probe, which gives a fingerprint pattern in yeast but not in *Arabidopsis*. The lack of visible signal does not mean, however, that the AC/GT repeats are completely missing in the genome since a few loci were found in the GenBank database. Moreover, Bell and Ecker (1994) estimated the number of loci carrying AC/GT repeats in *Arabidopsis* to be about 230 (1 every 430 kb). Since the intensity of the hybridization signals are dependent on the length of the region of homology, the lack of detectable signal with the $(TG)_9$ probe very likely means that, under our experimental conditions, only microsatellite loci containing a large number of repeats can be detected on Southern blots.

In Arabidopsis, a good correlation was observed between the data from hybridization experiments and those from database analysis, for mono-, di- and trinucleotide repeats. Therefore, the hybridization method can be useful for identifying the major microsatellite repeats in species in which only a few or no nucleotide sequences are available. The five probes, $(A)_{18}$, $(AG)_9$, $(AAG)_6$ $(ATG)_6$ and $(GTG)_6$, which reveal clear fingerprint patterns in the hybridization experiment, correspond to the 5 more frequent repeats found in the GenBank database. The loci found in the database can be ordered according to their relative frequencies in the following way: AAG/CTT (35,5%), A/T (13,6%), AG/CT (10,6%), ATG/CAT (9,5%), GTG/CAC (5,9%), AAC/GTT (5,3%) and AT/AT 5,3%), the remaining repeats accounting for less than 4% of the total. The $(AAC)_6$ probe revealed multiple bands in the *Arabidopsis* genome (Fig. 1), but the signal intensities were very weak compared to those observed in the yeast genome. For the reason mentioned before, it can be suggested that in *Arabidopsis* no $(AAC/GTT)_n$ locus contains enough repeats to generate a hybridization signal of high intensity in our experimental conditions. The $(AT)_9$ probe did not reveal any signal. This can probably be explained by the self annealing nature of the probe, which reduces the efficiency of hybridization to a membrane-bound target. With respect to the tetranucleotide repeats, the number of loci found in the GenBank database is too low (5 loci consisting of 4 different repeated motifs) to determine whether a correlation with the hybridization data exists.

Three computer-assisted searches have recently been performed for microsatellite loci in EMBL and GenBank higher plants sequences databases (Lagercrantz et al. 1993; Morgante and Oliveri 1993; Wang et al. 1994). In two analysis, concerning loci carrying repeats with a total length of 20 bp or more, the different motifs were found at the following relative frequencies: AT/AT (43–38%), A/T (23–15%), AG/CT (15–8%), AAT/ATT (7–6%), AAC/GTT (7–5%), AAG/CTT (7–4%), AGC/GCT (4–5%), the frequencies of the remaining repeats being 3% or less of the total (data from Lagercrantz et al. 1993 and from Wang et al. 1994, respectively). Two major differences can be observed between these data and our observations on the *Arabidopsis* genome. The first concerns the AT/AT repeat, which is by far the most frequent in the bulk of higher plant sequences while it ranks only sixth in *Arabidopsis*, with a relative frequency about 8 times lower. However, a large part of this variation probably results from the choice of the parameter determining the minimum size of the loci, which was 20 bp for Lagercrantz et al. and Wang et al. in place of the 16 bp for our analysis. As shown in Fig. 2, 7 AT/AT loci out of the 9 identified (77%) have a total length of 20 bp or more, while the 4 major classes of repeats, AAG/CTT, A/T, AG/CT and ATG/CAT, give an average value of 19.5%. The percentages calculated for each of these 4 classes of repeats are not significantly different ($\chi^2 = 2{,}08$, $df = 3$, $0.5 < P < 0.7$). In contrast, the value of 77% observed for the AT/AT repeats differs significantly from 19.5% ($\chi^2 = 15.9$, using Yates' correction, $df = 1$, $P < 0{,}001$). Therefore, considering only loci with a minimum repeat length of 20 bp will greatly increase the relative frequency of those carrying the AT/AT repeat. The second major difference concerns the AAG/CTT repeat, which has a relative frequency 5–9 times higher in *Arabidopsis* than the average found in higher plants. It is by far the most frequent repeat present in the GenBank *Arabidopsis* sequences. However, this observation does not necessarily mean that the AAG/CTT motif is more frequent in the *Arabidopsis* genome than the mono- and di-nucleotide motifs A/T and AG/CT. It can be argued that differences in relative frequencies observed in databases between mono-, di- and tri-nucleotide repeats do not exactly reflect the situation in the genome because databases are enriched in coding sequences that are expected to better tolerate the presence of tri-nucleotide repeats than the presence of mono- or di- nucleotide repeats. In order to check this point, several assays were made to quantify the amount of AG/CT and AAG/CTT loci in the *Arabidopsis* genome by hybridization on a library constructed in lambda EMBL3 after *Mbo*I partial digestion of genomic DNA. Because a large variation in the signal intensity was observed, no accurate quantification could be done; however, in each assay, the number of estimated loci was higher for the AAG/CTT repeats than for the AG/CT

repeats. The estimations varied from 180 to 390 loci for AG/CT, which is roughly in agreement with the previously published estimation of about 400 loci (Bell and Ecker, 1994), and from 280 to 700 for AAG/CTT, confirming that AAG/CTT repeats are more frequent than AG/CT in the *Arabidopsis* genome. It is not possible to estimate the number of A/T loci by hybridization because, as shown in Fig. 1, the probe reveals large bands in Southern blots that do not correspond to microsatellites. Moreover, poly-A sequences are present in the intergenic spacer of the ribosomal DNA units (Gruendler et al. 1991).

Taken together, our results indicate that $(AAG)_n$, $(AG)_n$, $(ATG)_n$ and $(GTG)_n$ are the most convenient probes by which to isolate microsatellite loci in *Arabidopsis* by screening genomic libraries. $(AAC)_n$ may also be useful even if it is expected to mostly reveal loci with a small number of repeats. Probes corresponding to the other major repeats present in the genome are less suitable for this purpose: $(AT)_n$, because it is potentially self-annealing; $(A)_n$, because non-microsatellite poly-A tracks are present in the genome.

More than 8% of the cDNA clones gave hybridization signals with 1 of the oligonucleotide probes $(AAG)_6$, $(ATG)_6$ or $(AG)_9$. If, as suggested by our results, most of them carry a microsatellite locus, it will mean that among the 895 non-redundant cDNA clones partially sequenced in the Expressed Sequence Tag (EST) program (Höfte et al. 1993), about 70 carry a cluster of 1 of the 3 repeated motifs. Because the number of cDNA clones analyzed in the French and in the MSU (Michigan State University) EST programs is rapidly increasing, an important source of material will become available that will allow the identification of several hundred microsatellite loci. Some will be directly obtained from the partial nucleotide sequences obtained in the course of the EST programs, while others will require identification by hybridization with appropriate oligonucleotide probes and additional sequence analysis.

In Arabidopsis, Bell and Ecker (1994) reported that AG/CT loci are much more polymorphic than CA/TG loci and noted a correlation between the lack of polymorphism and the level of complexity of the loci. AG/CT loci, which mostly show a simple structure, are highly polymorphic, since 25 out of 30 revealed a polymorphism between the Columbia and Landsberg ecotypes. Our data do not reveal any obvious correlation between the degree of polymorphism of the tri-nucleotide loci and the repeat class or the locus complexity, but it is clear that the overall level of polymorphism is much lower. However, it cannot be concluded that $(AG)_n$ loci are much more polymorphic than tri-nucleotide loci. In humans, Weber (1990) reported that the degree of polymorphism of $(AC)_n$ and $(AG)_n$ loci tends to increase with the number of repeats. If this also applies to tri-nucleotide loci, the difference observed in the overall degree of polymorphism may just reflect the difference in the total length of repeated motifs. The $(AG)_n$ loci analyzed by Bell and Ecker have a total length of perfect repeats varying from 26 to 62 bp, while the tri-nucleotide analyzed in this report show a maximum value of 27 bp. Our data do not allow us to establish a clear correlation between the level of polymorphism and the total length of repeats. However, the proportion of polymorphic loci is higher among those carrying a block of 7 perfect repeats or more than among those carrying less than 7 repeats: 8/11 and 4/15, respectively ($\chi^2 = 5.41$, $df = 1$, $0.001 < P < 0.05$).

Acknowledgements We wish to thank the Nottingham Arabidopsis Stock Center for providing seeds of the ecotypes used in this work, Dr J. Giraudat for the gift of the cDNA library, S. Genestier for technical assistance, M Laudié and Dr. R Cooke for help in sequencing and analyzing the cDNA, and BIOCEM (Groupe LIMAGRAIN) for greenhouse facilities. This work was supported by CNRS, Université Blaise-Pascal, and GREG (Groupement de Recherches et d'Etudes sur les Génomes).

References

Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in Soybean. Genetics 132:1131–1139

Beckmann JS, Soller M (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence-tagged microsatellites sites. Biotechnology 8:930–932

Beckmann JS, Weber JL (1992) Survey of human and rat microsatellites. Genomics 12:627–631

Bell CJ, Ecker JR (1994) Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics 19:137–144

Bellané-Chantelot C, Lacroix B, Ougen P, Billault A, Beaufils S, Bertrand S, Georges I, Gilbert F, Gros I, Lucotte G, Susini L, Codani JJ, Gesnouin P, Pook S, Vaysseix G, Lu-Kuo J, Ried T, Ward D, Chumakov I, Le Paslier D, Barillot E, Cohen D (1992) Mapping the whole human genome by fingerprinting yeast artificial chromosomes. Cell 70:1059–1068

Beyermann B, Nürnberg P, Weihe A, Meixner M, Epplen JT, Börner T (1992) Fingerprinting plant genomes with oligonucleotide probes specific for simple repetitive DNA sequences. Theor Appl Genet 83:691–694

Brook JD, McCurrach ME, Harley HJ, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T, Sohn R, Zemelman B, Snell RG, Rundle SA, Crow S, Davies J, Shelbourne P, Buxton J, Jones C, Juvonen V, Johnson K, Harper PS, Shaw DJ, Housman DE (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell 68:799–808

Brunel D (1994) A microsatellite marker in *Helianthus annuus*. Plant Mol Biol 24:397–400

Condit R, Hubbell SP (1991) Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. Genome 34:66–71

Delseny M, Laroche M, Penon P (1983) Detection of sequences with Z-DNA forming potential in higher plants. Biochem Biophys Res Commun 116:113–120

Dessen P, Fondrat C, Valencien C, Mugnier C (1990) BISANCE: a french service for access to biomolecular sequence databases. Cabios 6:355–356

Edwards AL, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am J Hum Genet 49:746–756

Epplen JT, Melmer G, Schmidt P, Roewer L, Hundrieser J, Epplen C, Buitkamp J (1992) On the potential of simple repetitive DNA for fingerprinting in clinical, forensic, and evolutionary dynamic studies. Clin Invest 70:1043–1051

Fu YH, Kuhl DPA, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkerk AJMH, Holden JJA, Fenwick RG, Warren ST, Oostra BA, Nelson DL, Caskey CT (1991) Variation of the CGG repeat at the fragile X site results in genetic instability:resolution of the Sherman paradox. Cell 67:1047–1058

Fu YH, Pizzuti A, Fenwick RG, King J, Rajnarayan S, Dunne PW, Dubel J, Nasser GA, Ashizawa T, De Jong P, Wieringa B, Korneluk R, Perryman B, Epstein HF, Caskey CT (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science 255:1256–1258

Giraudat J, Hauge BM, Vallon C, Smalle J, Parcy F, Goodman HM (1992) Isolation of the *Arabidopsis* ABI 3 gene by positional cloning. Plant Cell 4:1251–1261

Gruendler P, Unfried I, Pascher K, Schweizer D (1991) rDNA intergenic region from *Arabidopsis thaliana*, structural analysis, intraspecific variation and functional implications. J Mol Biol 221:1209–1222

Hamada H, Petrino MG, Kakunaga T, Seidman M, Stollar D (1984) Characterization of genomic poly(dT-dG)/poly(dC-dA) sequences: structure, organization and conformation. Mol Cell Biol 4:2610–2621

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerrier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, De Marcillac G, Gigot C, Fleck J, Philipps G, Axelos M, Bardet C, Tremousaygue D, Lescure B (1993) An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA from *Arabidopsis thaliana*. Plant J 4:1051–1061

Jacob HJ, Lindpaintner K, Lincoln SE, Kusumi K, Bunker RK, Yi-Pei Mao, Ganten D, Dzau VJ, Lander ES (1991) Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell 67:213–224

Lagercrantz U, Ellegren H, Andersson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res 21:1111–1115

Litt M (1991) PCR of TG microsatellites. In: McPherson P et al. (eds) PCR, a practical approach. IRL press, pp 85–99

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle action gene. Am J Hum Genet 44:391–401

Moore SS, Sargeant LL, King TJ, Mattick JS, Georges M, Hetzel DJS (1991) The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in close related species. Genomics 10: 654–660

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 85:2444–2448

Poulsen GB, Kahl G, Weising K (1993) Abundance and polymorphism of simple repetitive DNA sequences in *Brassica napus* L. Theor Appl Genet 85:994–1000

Sanger F, Coulson AR, Smith AJM, Roc BA (1980) Cloning a single-stranded bacteriophage as an aid to rapid DNA sequencing. J Mol Biol 143:161–178

Senior ML, Heun M (1993) Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. Genome 36:884–889

Serikawa TS, Kuramoto T, Hilbert P, Mori M, Yamada J, Dubay CJ, Lindpainter K, Ganten D, Guénet JL, Lathrop GM, Beckmann JS (1992). Rat gene mapping using PCR-analyzed microsatellites. Genetics 131:701–721

Thomas MR, Scott NS (1993) Microsatellite repeats in grapevine reveal DNA polymorphisms when analysed as sequence-tagged sites (STSs). Theor Appl Genet 86:985–990

Towner P (1991) Purification of DNA. In: Brown TA (ed) Essential molecular biology, a practical approach, vol 1. IRL press, pp 54–55

Wang Z, Weber JL, Zhong G, Tanklsey SD (1994) Survey of plant short tandem DNA repeats. Theor Appl Genet 88:1–6

Weber JL (1990) Human DNA polymorphism based on length variations in simple sequence tandem repeats. In: Tilgman S, Davies K (eds) Genome analysis vol. 1: genetic and physical mapping. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 159–181

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weising K, Beyermann B, Ramser J, Kahl G (1991) Plant DNA fingerprinting with radioactive and digoxigenated oligonucleotide probes complementary to simple repetitive DNA sequences. Electrophoresis 12:159–169

Williamson RW, Bowcock A, Kidd K, Pearson P, Schmidtke J, Ceverha P, Chipperfield M, Cooper DN, Coutelle C, Hewitt J, Klinger K, Langley K, Beckmann J, Tolley M, Maidak B (1992) Report of the DNA committee and catalogues of cloned and mapped genes, markers formatted for PCR and DNA polymorphisms. Cytogenet Cell Genet 58:1190–1833

Wood WI, Gitschier J, Lasky LA, Lawn RM (1985) Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of higly complex gene libraries. Proc Natl Acad Sci USA 82:1585–1588

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

Zhao X, Kochert G (1993) Phylogenic distribution and genetic mapping of a (GGC)n microsatellite from rice (*Oriza sativa* L.). Plant Mol Biol 21:607–614