Copyright © 2000 by the Genetics Society of America

# Computational and Experimental Characterization of Physically Clustered Simple Sequence Repeats in Plants

**Linda Cardle, Luke Ramsay, Dan Milbourne, Malcolm Macaulay, David Marshall and Robbie Waugh**

*Scottish Crop Research Institute, Dundee DD2 5DA, Scotland, United Kingdom*

Manuscript received January 31, 2000
Accepted for publication June 26, 2000

ABSTRACT

The type and frequency of simple sequence repeats (SSRs) in plant genomes was investigated using the expanding quantity of DNA sequence data deposited in public databases. In Arabidopsis, 306 genomic DNA sequences longer than 10 kb and 36,199 EST sequences were searched for all possible mono- to pentanucleotide repeats. The average frequency of SSRs was one every 6.04 kb in genomic DNA, decreasing to one every 14 kb in ESTs. SSR frequency and type differed between coding, intronic, and intergenic DNA. Similar frequencies were found in other plant species. On the basis of these findings, an approach is proposed and demonstrated for the targeted isolation of single or multiple, physically clustered SSRs linked to any gene that has been mapped using low-copy DNA-based markers. The approach involves sample sequencing a small number of subclones of selected randomly sheared large insert DNA clones (*e.g.*, BACs). It is shown to be both feasible and practicable, given the probability of fortuitously sequencing through an SSR. The approach is demonstrated in barley where sample sequencing 34 subclones of a single BAC selected by hybridization to the *Big1* gene revealed three SSRs. These allowed *Big1* to be located at the top of barley linkage group 6HS.

THE ubiquity of simple sequence repeats (SSRs) in eukaryotic genomes and their usefulness as genetic markers has been well established over the last decade. In mammalian systems, in particular, SSRs have been the marker of choice for several years, and well-developed SSR-based linkage maps are available for a number of species (Dib *et al.* 1996; Dietrich *et al.* 1996; Sverdlov *et al.* 1998). A high level of SSR informativeness has also been demonstrated for a variety of plant species and this has prompted the initiation of SSR discovery programs for the majority of agronomically important crops (Weising *et al.* 1989; Condit and Hubbell 1991; Akkaya *et al.* 1992; Zhao and Kochert 1992; Morgante and Olivieri 1993; Wu and Tanksley 1993; Röder *et al.* 1995; Liu *et al.* 1996; Panaud *et al.* 1996; Provan *et al.* 1996; Senior *et al.* 1996; Milbourne *et al.* 1997, 1998). However, to date, a number of limitations have existed with SSR discovery in plants, including a lack of DNA sequence in databases, a perceived low abundance of SSRs (compared to mammals), and differences in the most common types of repeat found.

Previous analyses of plant DNA sequence database entries for all possible SSR motifs have revealed frequencies ranging from one every 29 kb to 50 kb, depending on species (Lagercrantz *et al.* 1993; Morgante and Oliveiri 1993). SSR frequency in plants has also been assessed by oligonucleotide hybridization, and such studies have suggested figures in the range of one SSR every 65 kb to 80 kb (Panaud *et al.* 1996; Echt and Maymarquardt 1997). These results contrast sharply with those for humans, with an estimate of one SSR every 6 kb on average (Beckmann and Weber 1992).

Despite this relative difference in abundance, the perceived advantages of SSRs as markers are such that plant geneticists have resorted to screening large numbers of clones (Röder *et al.* 1995; Liu *et al.* 1996; Bryan *et al.* 1997) or developing selective SSR enrichment techniques (Edwards *et al.* 1996; Milbourne *et al.* 1998) to generate sufficient numbers of SSRs for implementation in genetic research (*e.g.*, Paglia *et al.* 1998; Röder *et al.* 1998). Given the interest of the plant genetics community in SSRs as genetic markers, we have been particularly concerned to establish methods of rapidly identifying robust and informative SSRs linked to genes of agronomic significance. Compared with genome-wide isolation approaches, gene-targeted strategies are more likely to yield SSRs that are relevant to the goals of marker-assisted selection and germplasm assessment. In the former, linkage disequilibrium between an SSR and a gene is fortuitous and frequently insufficient for transfer to other germplasm of interest.

Given these objectives, we have reassessed the frequency of SSRs in plant genomes, taking advantage of the significant number of long contiguous sequences (>10 kb) recently deposited in international sequence databases. The use of such data includes the intergenic

*Corresponding author:* Robbie Waugh, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, United Kingdom.
E-mail: r.waugh@scri.sari.ac.uk

regions that, even in Arabidopsis, make up about half the genome and would reduce the bias toward coding regions inherent in previous studies.

## MATERIALS AND METHODS

**Sequence data sources and analysis of SSR content:** Sequences were acquired through a Sequence Retrieval System search of EMBL and EMBL updates [on 10/08/98 for genomic (>10 kb long) and Arabidopsis expressed sequence tags (ESTs) and on 22/06/99 for other ESTs]. To locate SSRs in the sequence data, we used the C-program Sputnik (by Chris Abajian at Washington University, http://www.abajian.com/sputnik), which finds SSRs of between 2 and 5 bases long, incorporating a scoring system to detect imperfections in the pattern. In the subsequent calculations, only those SSRs with a perfect repeat pattern were included. The GCG (version 8.1-UNIX) program FINDPATTERNS was used to locate mononucleotide repeats in the Arabidopsis genomic and EST sequences.

For the searches, we define SSRs as being mononucleotide repeats >15 bp, dinucleotide repeats >14 bp, trinucleotide repeats >15 bp, tetranucleotide repeats >16 bp, and pentanucleotide repeats >20 bp. For comparison with other computerized searches for SSRs in plant genomic sequence, we also used the criteria of SSR motifs of 20 bp and over (Lagercrantz *et al.* 1993; Wang *et al.* 1994). A series of Perl5 scripts was written to analyze and summarize the SSR data obtained from Sputnik and FINDPATTERNS. We calculated the frequency of occurrence of SSR motifs from mono- to pentanucleotide in barley, rice, tomato, potato, and Arabidopsis genomic and Arabidopsis EST sequences and di- to pentanucleotides in rice, maize, poplar, tomato, and soybean EST sequences. We also classified all repeat types from mononucleotide to tetranucleotide, taking sequence complementarity into account, and examined their occurrence in Arabidopsis genomic DNA and Arabidopsis ESTs (Table 1 and Figure 1). We found 51 *Arabidopsis thaliana* EMBL entries that had features tables that were sufficiently detailed to estimate the occurrence of SSR motifs within exons and introns (Figure 1). This subset of the genomic sequences totaled 5,597,709 bp and included 7141 exons totaling 1,759,177 bp (31% of total length) and 5844 introns totaling 1,286,007 bp (23%).

**Bacterial artificial chromosome (BAC) isolation, subcloning, and sequencing:** The first 73,728 clones of a 7× BAC library of barley cv. Morex (http//www.genome.clemson.edu—R. Waugh, R. Wing and J. Tomkins, personal communication) were screened using a 363-bp fragment of the *Big1* gene (courtesy of J. Hargreaves, IACR). *Big1* shows high homology to leucine-rich repeat (LRR) motifs found in many resistance gene analogues (Jones and Jones 1997). A single BAC clone (BAC84c21), the most strongly hybridizing from a total of 24, was purified using a QIAGEN (Chatsworth, CA) plasmid midi kit according to a protocol modified by the manufacturers for the isolation of BAC DNA. A total of 20 μg purified BAC DNA in a volume of 200 μl d$H_2O$ was fragmented by sonication in an MSE (Sussex, UK) 150-W ultrasonic disintegrator with a nominal frequency of 20 kc/sec, using five 1-sec, 15-μm amplitude pulses. Fragments of 1–3 kb were size selected on a 1% agarose minigel, electroeluted, and resuspended at a concentration of 50 ng/μl in d$H_2O$. The ends of 1.0 μg sonicated DNA were "polished" using the Klenow fragment of DNA polymerase I by incubation at room temperature for 10 min followed by 75° for 10 min (in 50 mm Tris-HCl pH 7.2, 10 mm $MgSO_4$, 0.1 mm dithiothreitol, 20 μg/ml BSA, 40 μm dNTPs, 5 units Klenow) and ligated into *Sma*I cut pUC18 [5 ng pUC18, 500 ng sonicated DNA, 0.25 units Ligase, 1× buffer (Boehringer Mannheim, Indianapolis) in a total volume of 25 μl, incubated at 14° overnight]. One microliter of each ligation was transformed into DH5α competent cells by electroporation. Recombinant clones were picked off LB ampicillin (50 μg/ml), isopropylthio-β-d-galactoside, X-gal plates, colony purified, plasmid prepped by standard approaches, and sequenced in one direction (using PE "BigDye" sequencing reagents and the M13 forward primer).

**Sequence homology searches, SSR primer design, and analysis:** The sequences obtained were checked against public databases using BLASTN and BLASTX search algorithms (Altschul *et al.* 1997). Following the SSR analysis procedures given above, primers were designed to sequences flanking SSRs using the computer program PRIMER (v 3.0; E. Lander, Cambridge, MA). For the experimental verification, three primer pairs were designed to SSRs found in subclone sequences, BAC84c21_s02; forward (5′-TACAGGTAAAGGTTACTTGACG-3′) and reverse (5′-ATGAAA ACAAACGA GAAAAGA-3′), BAC84c21_s34; forward (5′-CATG GGCTAATCGTGCTT-3′) and reverse (5′-TGATTTTAACCTAT TGAGCTTTT-3′) and BAC84c21_s33; forward (5′-AATCCCTCT TCATAAATTAGTG-3′) and reverse (5′-TTCAAATCACTGAA CAAAAAC-3′). In addition, two primer pairs were designed to SSRs found in the sequence of *Big1* (AF166121); AF166121A; forward (5′-TTATGCTTCACACGGTGTAC-3′) and reverse (5′-GGCACAAAAAGACTGAAATAG-3′) and AF166121B; forward (5′-ATAAAGAAAGCTGGAGTACCC-3′) and reverse (5′-AACTT GTTGGTTGTACTCTGG-3′).

PCR reactions, using the direct incorporation of fluorescent oligonucleotides, were performed in a total volume of 15 μl following manufacturer's instructions (Perkin-Elmer, Norwalk, CT) and were run on a PE 9600. A PE 377 DNA sequencer with GeneScan software was used to visualize the results. Some reactions giving ambiguous results were confirmed by repeating the PCR reactions on a PE 9600 in a total volume of 10 μl and consisted of 20 ng genomic DNA, 1× PCR buffer, 0.3 units *Taq* polymerase (both from Boehringer Mannheim), 200 μm dNTPs, and 0.3 μm of forward and reverse primers, the reverse primer being end-labeled with [γ-$^{33}$P]ATP. An equal volume of 95% formamide electrophoresis loading buffer was added to the samples, which were then denatured, snap cooled on ice, and electrophoresed in 6% Easigel (Scotlab) according to standard procedures. An M13 sequencing marker was run to estimate product sizes and visualization of results was achieved by exposure of fixed, dried gels to X-ray film.

Linkage analysis of the Lina × *Hordeum spontaneum* Canada Park population was carried out by combining the segregation data with existing mapping data (Ramsay *et al.* 2000) and performed using Joinmap (version 2.0; P. Stam and J. W. Van Ooijen, Wageningen: CPRO-DLO).

## RESULTS

**Frequency, type, and distribution of SSRs in *A. thaliana*:** A total of 306 nonredundant genomic DNA sequences longer than 10 kb and 36,199 EST sequences were retrieved from the EMBL nucleotide database (release 24/06/98) and searched for the presence of SSR motifs. All but one of the long genomic DNA sequences contained at least one SSR. Figure 1a shows the distribution of number of SSRs per sequence from 198 P1 and BAC clones. The SSR content shows a normal distribution around a mean of 10 SSRs per clone. In contrast, only 3% of ESTs contained an SSR, which is similar to the proportion previously found in rice (Akagi *et al.* 1996). Table 1 shows the frequency of occurrence of SSRs according to repeat motif length.

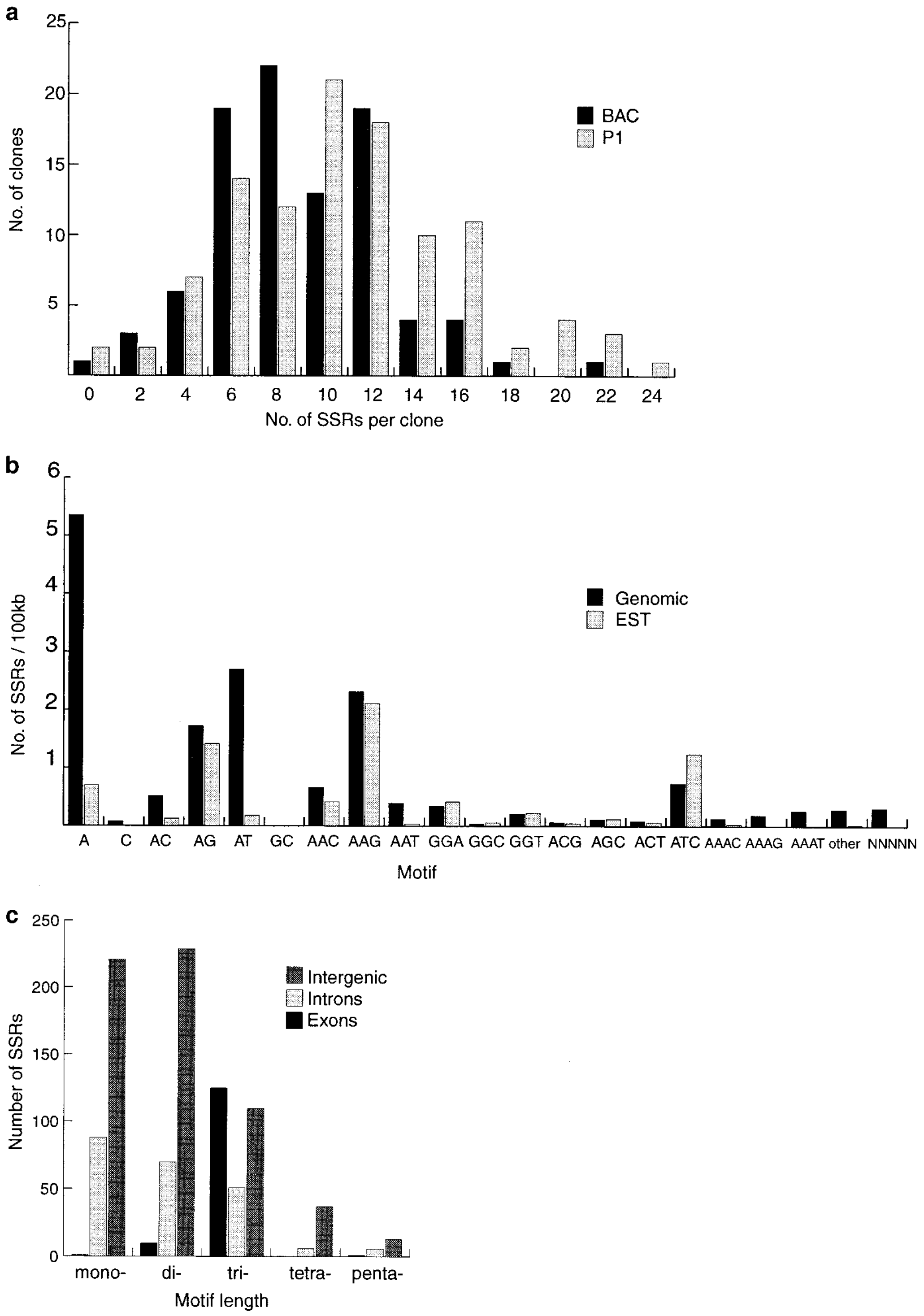

Figure 1.—Frequency of SSRs in Arabidopsis. (a) Actual number of SSRs in a collection of P1 and BAC clones. (b) Number and type of SSRs/100 kb of genomic and EST sequence. (c) Number of SSRs of different repeat lengths in intron, exon, and intergenic sequence.

For Arabidopsis genomic DNA, the average distance between SSRs was ~6.04 kb compared to 14 kb for ESTs. Compound repeats, which, if frequent, would affect the overall average, were surprisingly rare. Twenty-seven cases were found where two repeat regions were immediately adjacent and 47 where two repeats were within 5 bp of each other. These constitute only 3.3% of the SSRs found, making little difference to the average distribution.

The most common motif found in the Arabidopsis genomic DNA was the mononucleotide A/T, which comprises 32% of all SSRs found (Figure 1b). AT/TA repeats comprised 16% of the total SSR content, AAG/TTC, 14%, and AG/TC, 10%. Dividing the SSRs into repeating unit size classes, the SSR content was almost equally divided between mono- (33%), di- (30%), and trinucleotides (30%). The proportion of SSR repeat unit sizes in ESTs was different. The most common

**TABLE 1**

**SSR survey of Arabidopsis genomic and EST sequence data**

| Source | *A. thaliana* | |
|---|---|---|
| Subgroup | Genomic<br>(P1 and BAC) | ESTs |
| No. sequences | 306 | 36,199 |
| No. with ≥1 SSR | 305 | 1,040 |
| Repeat type: | | |
| Mononucleotide | 1,471 (33) | 103 (10) |
| Dinucleotide | 1,333 (30) | 254 (24) |
| Trinucleotide | 1,350 (30) | 706 (66) |
| Tetranucleotide | 236 (5) | 7 (<1) |
| Pentanucleotide | 83 (2) | 0 |
| Total SSR content | 4,473 | 1,070 |
| Total length (kb) | 27,011.3 | 14,808.0 |
| Average distance (kb) | 6.04 | 13.83 |

Numbers in parentheses show percentage of total SSR content.

repeats were trinucleotides (AAG/TTC, 29% and ATC/TAG, 17%) followed by the dinucleotides (AG/TC, 20%), and the mononucleotide A/T(10%). In both genomic DNA and ESTs, AAG/TTC repeats comprised 45% of the total trinucleotide content, followed by ATC/TAG (26% in genomic, 15% in EST). In genomic DNA the most common dinucleotide was AT/TA (in stark contrast to the total absence of any CG/GC repeats). In ESTs, AG/TC repeats were almost eight times as frequent as the AT/TA repeats. In genomic DNA, the AT-rich tetranucleotide repeats were more common than CG-rich repeats, with 13 tetranucleotide motifs never occurring: (CCCG), (CCGG), (AACG), (AAGT), (ACCG), (ACCC), (ACGG), (AGCC), (AGGC), (ATCC), (ACGC), and (AGCG) (*i.e.*, 9 out of 10 possible CG-rich motifs). The ESTs contained virtually no tetranucleotide repeats.

By examining the detailed features tables available for 51 of the 306 Arabidopsis genomic sequences, a considerable difference in the distribution of SSR motifs was found between introns, exons, and intergenic regions (Figure 1c). Almost two-thirds of SSRs were found in intergenic regions (608 out of total of 961), and the majority of these were either mono- or dinucleotides. A total of 14% (132/961) of the SSRs were found in exons and 23% (221/961) in introns.

Of the exonic SSRs, 91% were trinucleotides reflecting repetitive amino-acid sequence motifs, although there was no simple pattern of motifs in relation to different protein classes. The remaining 9% were made up of 10 dinucleotides, one mono-, and one pentanucleotide repeat. A more diverse range was found in introns (40% mono-, 32% di-, 23% tri-, 3% tetra-, and 3% pentanucleotides) with similar proportions of repeat types being found in intergenic regions, (36% mono-, 38% di-, 18% tri-, 6% tetra-, and 2% pentanucleotides). The proportions found within the data examined indicated that over 40% of all trinucleotide repeats are exonic in Arabidopsis.

The results of the SSR searches were used to extract flanking sequences to allow the design of primer pairs to over 4000 Arabidopsis SSRs for a range of annealing temperatures and product sizes. These primer sequences and associated information will be lodged at Arabidopsis Genome Resource at (http://synteny.nott.ac.uk/agr/agr.html).

**SSR distribution in other plants:** Of 52 genomic DNA sequences over 10 kb in length from species other than Arabidopsis, 38 were found to have at least one SSR motif. The overall average distance between SSRs for these species was 6.8 kb (38 SSRs in a total of 1075 kb), almost identical to that found in Arabidopsis alone. As with Arabidopsis, the most common motifs were A/T and AT/TA for mono- and dinucleotides; however, AAT was the most common trinucleotide motif. In the 52 sequences, only 7 out of the 33 possible tetranucleotide repeat motifs were found, most of which were AT-rich: (AAAG), (AAAT), (ACCC), (AATT), (ACAT), (AGAT), and (ATGC).

A number of contiguous sequences of over 30 kb were available for inclusion in this study (from barley, tomato, rice, and potato). Using all available data from these species, the estimated SSR frequency is one every 7.4 kb in barley, 7.1 kb in tomato, 7.4 kb in rice, and 6.4 kb in potato genomic DNA. Despite the relatively small number of sequences available, the similarity in SSR frequency with Arabidopsis suggests that one every 6–7 kb may be a good general estimate for SSR frequency in the type of plant DNA sequence studied here (*i.e.*, large insert DNA clone sequences containing a gene of interest).

The highest frequency of the EST-derived SSRs (excluding mononucleotide motifs) was found in rice at 3.4 kb between SSRs (which agrees closely with the estimate of Akagi *et al.* 1996) followed by soybean (7.4 kb), maize (8.1 kb), tomato (11.1 kb), poplar (14.0 kb), and cotton (20.0 kb). The overall average for these species was one SSR every 5.4 kb (7193 SSRs found in 38,502 kb of sequence).

**Targeted SSR isolation: the *Big1* gene of barley:** Given a known average inter-SSR distance, it is possible to construct a simple model to calculate the probability of finding an SSR in a contiguous DNA sequence as a simple function of the average DNA sequence read length and thereby calculate the probability of uncovering an SSR as a function of read length and the number of sequencing runs performed.

To test this hypothesis, a single barley BAC clone, chosen on the basis of hybridization screening with an LRR-containing gene fragment of the *Big1* gene, was chosen for sample sequencing. A total of 36 random subclones were sequenced from one end with each reaction yielding ∼400 bp of high-quality sequence. Nine sequences showed homology to known cereal retroelements, mostly LTR copia elements, *e.g.*, BARE-1 (Manninen and Schulman 1993) at the nucleotide level and three additional hits to plant retrotransposon polyprotein/integrase domains using translations of the sequences. Nine other sequences showed homology to regions upstream of known cereal genes. However, on closer inspection six of these homologies were due to the presence of miniature inverted repeat transposable elements, mostly of the Stowaway type (Bureau and Wessler 1994). Two sequences showed homology to regions upstream of the *g1* and *Ost1* genes present in the 60-kb contig around the *mlo* gene (Panstruga *et al.* 1998). The other subclone (BAC84c21_s24) showed identity to a region upstream of the barley *Big1* gene (AF166121) that was the source of the PCR fragment used for the hybridization with the BAC library (Y. Tokunaga, J. P. R. Keon and J. A. Hargreaves, unpublished results).

The sequences of the subclones were analyzed for the presence of SSRs as described above. Three subclones of the 36 sequenced showed the presence of an SSR, with BAC84c21_s02 containing $(CTT)_6$, BAC84c21_s34, $(AT)_6$, and BAC84c21_s33, $(T)_{10}$. In addition, two SSRs were discovered within the sequence (AF166121) upstream of the *Big1* gene: $(A)_{10}$ at position 2073–2082 (named AF166121A) and $(CTT)_5$ at position 2845–2859 (named AF166121B). Primers designed to the flanking regions were tested on the parents of a mapping population (Lina × *H. spontaneum* Canada Park) with both mononucleotide SSRs, BAC84c21_s33 and AF166121A, showing polymorphism. These two primer pairs were tested on the $F_1$ doubled haploid population derived from these parents with both SSRs mapping to the same position on the short arm of 6H, 11 cM from the distal end.

## DISCUSSION

The results presented here show that SSRs in plant genomic DNA are much more common than previous estimates suggest, indicating a frequency of 1 SSR every 6–7 kb, which is equivalent to that described in mammals (Beckmann and Weber 1992). The revision of this estimate is due to the recent submission of a large volume of contiguous DNA sequence emerging from the Arabidopsis genome sequencing project, allowing our search to be carried out on over 27,000 kb of genomic sequence. Previously Wang *et al.* (1994) proposed an average frequency of 1 SSR every 42.4 kb in Arabidopsis on the basis of the presence of 10 SSRs in the Arabidopsis sequence available at the time. The actual frequency is close to an order of magnitude higher, with over 5000 Arabidopsis SSRs identified by our search criteria. These markers provide a rich resource of informative sequence-tagged sites for future genetical studies in this species.

At present, ∼100 SSRs in Arabidopsis have been described previously [*A. thaliana* Genome Center (http://cbil.humgen.upenn.edu/∼atgc/SSLP_info/), Depeiges *et al.* 1995; Loridon *et al.* 1998], making the 5000 SSRs found during this study a very considerable increase in the number of available genetic markers. Even if a stricter definition of length of ≥20 bp is used (Bell and Ecker 1994), these new SSRs still represent an increase of greater than an order of magnitude in the number of SSRs available. It is noteworthy, however, that 5 of the 12 SSRs found to be polymorphic over four accessions by Depeiges *et al.* (1995) were compound/imperfect SSRs of <20 bp in length.

The analysis of large genomic sequences from other plant species demonstrated that the frequency of SSRs in Arabidopsis (every 6–7 kb) holds for other plants as well. Our findings agree with previous studies that the most common SSR motifs in plants are A/T rich. The most common dinucleotide repeat (AT/TA followed by AG/CT; Lagercrantz *et al.* 1993; Morgante and Olivieri 1993; Wang *et al.* 1994; Smulders *et al.* 1997) contrasts with the most common (AC/TG) found in mammals (Hamada and Kakunaga 1982; Stallings *et al.* 1991). Direct comparisons between estimates of SSR frequency in this and other studies are rendered difficult by differing minimal motif length criteria. Our slightly less stringent criteria are based on our experience that primers designed to the motif lengths we outline are generally successful in revealing polymorphism. Although these different criteria affect the absolute and relative frequency of the motifs found, it is clear that past estimates have also been affected by the bias in sequences lodged in databases.

The study reported here has made use of the recent submission of a large volume of contiguous DNA sequence emerging from the Arabidopsis genome sequencing project to allow an estimate to be made on sequence that is not skewed toward coding regions. This, together with the detailed annotation on a large proportion of the data, has shown that not only are SSRs at a higher frequency than previously estimated but also that the frequency of the SSRs varies within the genome, with exonic and intronic sequences making up roughly 55% of the genomic sequence but containing only 37% of the SSRs. This is particularly evident in exons that make up 31% of the genomic sequence but contain only 14% of the SSRs, 91% of which are trinucleotide. This corresponds well with the finding of this study and that of Akagi *et al.* (1996) of a lower frequency of SSRs in EST sequences with a preponderance of trinucleotide repeats. It also explains some of the differences between the absolute and relative frequencies of various repeat motifs in Arabidopsis found in this study and earlier estimates based mainly on coding regions (Depeiges *et al.* 1995).

Our comparison of SSR frequencies in ESTs from a range of plant species showed a considerable difference

TABLE 2

**SSR survey of rice, maize, poplar, tomato, cotton, and soybean EST sequences**

| Source | Rice | Maize | Soybean | Tomato | Cotton | Poplar |
|---|---|---|---|---|---|---|
| Di | 657 (13) | 140 (18) | 147 (30) | 84 (21) | 53 (22) | 38 (28) |
| Tri | 3,747 (73) | 478 (61) | 311 (63) | 289 (72) | 157 (66) | 83 (61) |
| Tetra | 498 (10) | 126 (16) | 30 (6) | 24 (6) | 21 (9) | 14 (10) |
| Penta | 230 (4) | 46 (6) | 9 (2) | 2 (1) | 8 (3) | 1 (1) |
| Total SSR | 5,132 | 790 | 497 | 399 | 239 | 136 |
| No. sequence | 45,033 | 14,950 | 9,611 | 9,100 | 8,083 | 4,809 |
| Average length | 380 | 430 | 380 | 490 | 590 | 390 |
| Total length (kb) | 17,304 | 6,411 | 3,675 | 4,444 | 4,788 | 1,880 |
| Average distance (kb) | 3.4 | 8.1 | 7.4 | 11.1 | 20.0 | 14.0 |

Numbers in parentheses show percentage of total SSR content.

in both the absolute and relative frequencies (Table 2). Unfortunately, insufficient large contiguous genomic sequences are available in the same species to determine whether these differences relate to artifacts of library construction (*e.g.*, tissue specificity, library redundancy, *etc.*) or to differences between species in SSR motif frequency generally. The estimate of total SSR frequency in plant species other than Arabidopsis presented here indicates similarity across plant species and therefore implies other factors (such as codon usage or nucleotide ratios) could be important.

It is possible that, as earlier estimates of SSR frequency were skewed by the preponderance of coding sequence in the databases, the estimates presented here are biased by the prevalence of low-copy sequences surrounding coding regions. This is particularly relevant when considering large genome species that contain a high proportion of repetitive DNA. Arabidopsis is atypical, given its general lack of high-copy number genomic DNA. In contrast, 70–80% of the genome content of large genome cereals is composed mainly of multiple copies of retrotranspons (San Miguel *et al.* 1996). Although SSRs are present in this highly repeated portion, it is possible that their frequency differs from that in low-copy regions. Indeed, Ramsay *et al.* (1999) found an intimate association between the SSRs and certain regions of retrotransposons and repetitive elements, implying that the distribution and frequency of SSRs in these high-copy regions of the genome might be constrained by these regions' retrotranspositional structure and evolution. It is, therefore, possible that the frequency of SSRs might be higher in the older gene-containing low-copy portion of such genomes, explaining the difference in the estimates of SSR frequency presented here and those based on random whole genome sequencing in maize (Morgante *et al.* 1999) and wheat (W. Powell, personal communication).

The high frequency found in these gene-rich areas allowed us to develop and test the hypothesis that sequencing random subclones (from, *e.g.*, a BAC clone) provides an effective strategy for identifying single or clustered SSRs in targeted genomic DNA. For a sample of $N$ sequencing runs, where the probability of finding an SSR per run is $p$, the probability of finding no SSRs is given by $(1 - p)^N$ and the probability of finding at least one SSR is $1 - (1 - p)^N$. Thus, from sequencing 20 different subclones, a simple calculation based on the SSR frequency indicated that a high probability of revealing an SSR ($P = 0.75$–$0.973$) can be achieved from average reads of 400–1000 bp, assuming a random occurrence of one SSR every 6.04 kb (Figure 2). This probability compares favorably to the attrition rate of current SSR isolation procedures (Bryan *et al.* 1997; Röder *et al.* 1998), particularly given the opportunity to identify multiple, clustered SSRs and the ability to choose which BAC or P1 to sequence by prior selection with molecular markers to identify a known chromosomal position.

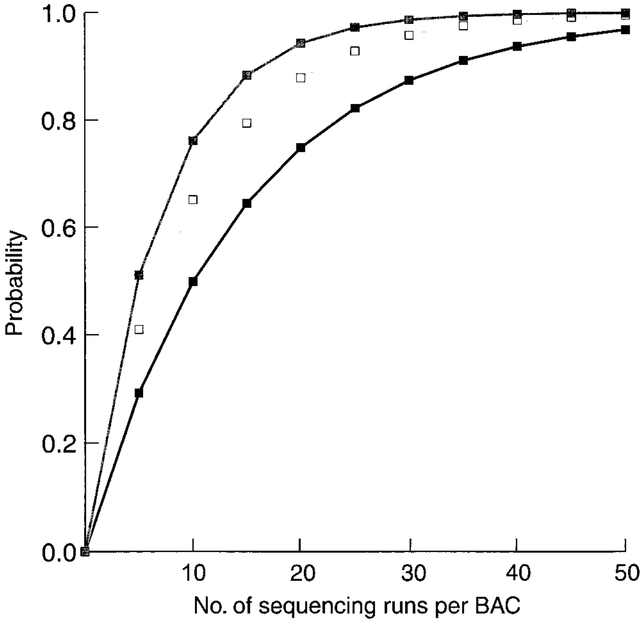

Figure 2.—Probability of sequencing through a simple sequence repeat in a random collection of subclones from a selected BAC or P1 clone with 400-, 600-, and 800-bp runs (shown by the bottom, middle, and top squares, respectively).

In demonstrating the approach in barley, only one SSR $(CTT)_6$ met our "repeat length" definition and two, $(AT)_6$ and $(T)_{10}$, were slightly short. Nevertheless, polymorphism of the $(T)_{10}$ SSR enabled the BAC clone to be mapped to chromosome 6HS and this position was confirmed by the use of another short SSR, $(A)_{10}$, known to be upstream of the gene sequence (AF166121). The discovery of one SSR that meets the criteria above in 36 runs of 400 bp represents a frequency of one SSR every 14.4 kb, somewhat lower than the value used in the calculation above. This may reflect the simplicity of the assumption of a random distribution of SSRs because the database survey indicated constraints on the distribution of SSRs in coding regions in comparison to that of intergenic regions. Moreover, roughly a third of the subclones showed homology to retrotransposons, which implies that a substantial proportion of the BAC used is composed of high-copy sequence. Importantly, repeats of lengths shorter than the criteria used for the sequence searches still prove useful in the discovery and confirmation of the map position of the PCR fragment of *Big1*, which was used to screen the BAC library. The polymorphism found with these short SSRs and others (White and Powell 1997; Milbourne *et al.* 1998) implies that the minimum repeat length used for the search of public databases was possibly too conservative.

In species where BAC or P1 libraries are already available, they represent a ready source of SSRs that are intrinsically "high value" for several reasons: BACs linked to genes of interest can be selected by hybridization directly or to any closely linked low-copy, DNA-based marker, and locus-specific SSRs can be developed quickly and efficiently by the sample sequencing approach described. Currently, the common approach to generating PCR-based markers for widespread application is to convert restriction fragment length polymorphism or amplified fragment length polymorphism markers into cleaved amplified polymorphic sequences—but these are generally of limited informativeness, frequently losing their diagnostic potential when transferred to germplasm other than that in which they were developed (Niewohner *et al.* 1995). Even when the actual sequence of the target gene is known, using this approach may produce SSR markers more easily deployable than markers based on the actual gene sequence.

In addition, the presence of multiple linked SSRs on a single BAC or P1 clone facilitates the detection of multiple SSR "haplotypes." Haplotypes are considerably more descriptive than single markers and are particularly suited for applications such as marker-assisted selection and many other areas of biology such as biodiversity assessment and population genetics with multiple multi-allelic SSRs giving comparable discrimination to that of single nucleotide polymorphisms in analyses (Chakraborty *et al.* 1999).

Another advantage of the use of large insert libraries is that all SSR motifs present can be sampled, unlike most SSR discovery programs that focus on AG/TC and AC/TG, which, in this study, represent <10% of those available. Here, all possible motifs are accessible, including the previously elusive AT/TA class, which has generally been suggested to be the most polymorphic. The sequencing necessary could be reduced by prescreening selected BACs through hybridization with suitable oligonucleotides (*e.g.*, Cregan *et al.* 1999). The use of hybridization might preclude the discovery of $(AT)_n$ SSRs (Depeiges *et al.* 1995) but could preferentially bias the finding of longer repeats (Cregan *et al.* 1999).

The findings presented here demonstrate that SSR frequency in plants is considerably higher than previous estimates with a frequency of one SSR every 6–7 kb, which is equivalent to that described in mammals (Beckmann and Weber 1992). However, differences in the absolute and relative frequency have been found between different regions of Arabidopsis, relating presumably to differing evolutionary constraints. Although the frequency of SSRs in gene-rich regions of other plant species appears to be similar to that of Arabidopsis, it is anticipated that genomic regions containing high-copy-number DNA will have a different profile. In addition, differences between SSR frequency and motifs in available EST data suggest that there are particular genome-specific constraints in coding regions. Simple sequence repeats are a manifestation of the deeper nonrandom nature of genomic sequences, with both direct tandem and symmetrical repeats being over-represented in eukaryotic genomes (Cox and Mirkin 1997) and with the constraints on the type and frequency of such repeats being, to some extent, genome specific (Antezana and Kreitman 1999). Outside these structural or evolutionary considerations, the reported findings have immediate practical value, with the development of a strategy for the targeted isolation of single or multiple, physically clustered SSRs near any mapped gene of interest. This will impact on genetic studies of the increasing number of plant species for which large insert libraries are available.

We thank Dr. John Hargreaves from IPSR for the use of the probe DNA and prepublication information, and Drs. W. Stiekema, H. Sandbrink, and E. Van der Vossen from CPRO, Wageningen, for searching an unpublished potato BAC sequence for SSR repeats. In addition, we thank Jane Davidson and Peter Davie for excellent technical support. The work reported here has been supported by grants from the Biotechnology & Biological Sciences Research Council (BBSRC), PAG 04430, GAT 09066, and EU CT95-0003, and through a grant-in-aid from the Scottish Executive Rural Affairs Department.

## LITERATURE CITED

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

Akkaya, M. S., A. A. Bhagwat and P. B. Cregan, 1992 Length polymorphisms of simple sequence repeat DNA in soybean. Genetics **132:** 1131–1139.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Antezana, M. S., and M. Kreitman, 1999 The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. **49:** 36–43.

Beckmann, J. S., and J. L. Weber, 1992 Survey of human and rat microsatellites. Genomics **12:** 627–631.

Bell, C. J., and J. R. Ecker, 1994 Assignment of 30 SSR loci to the linkage map of *Arabidopsis*. Genomics **19:** 137–144.

Bryan, G. J., A. J. Collins, P. Stephenson, A. Orry, J. B. Smith *et al.*, 1997 Isolation and characterization of microsatellites from hexaploid bread wheat. Theor. Appl. Genet. **94:** 557–563.

Bureau, T. E., and S. R. Wessler, 1994 *Stowaway*: a new family of inverted repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. Plant Cell **6:** 907–916.

Chakraborty, R., D. N. Stivers, B. Su, Y. X. Zhong and B. Budowle, 1999 The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. Electrophoresis **20:** 1682–1696.

Condit, R., and S. P. Hubbell, 1991 Abundance and DNA-sequence of 2-base repeat regions in tropical tree genomes. Genome **34:** 66–71.

Cox, R., and S. M. Mirkin, 1997 Characteristic enrichment of DNA repeats in different genomes. Proc. Natl. Acad. Sci. USA **94:** 5237–5242.

Cregan, P. B., J. Mudge, E. W. Fickus, L. F. Marek, D. Danesh *et al.*, 1999 Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor. Appl. Genet. **98:** 919–928.

Depeiges, A., C. Goubely, A. Lenoir, S. Cocherel, G. Picard *et al.*, 1995 Identification of the most represented motifs in *Arabidopsis thaliana* microsatellite loci. Theor. Appl. Genet. **91:** 160–168.

Dib, C., S. Fauré, C. Fizames, D. Samson and N. Drouot, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Dietrich, W. F., J. Miller, R. Steen, M. A. Merchant and D. Damron-Boles *et al.*, 1996 A comprehensive genetic map of the mouse genome. Nature **380:** 149–152.

Echt, C. S., and P. Maymarquardt, 1997 Survey of microsatellite DNA in pine. Genome **40:** 9–17.

Edwards, K. J., J. H. A. Barker, A. Daly, C. Jones and A. Karp, 1996 Microsatellite libraries enriched for several microsatellite sequences in plants. Biotechniques **20:** 758–760.

Hamada, H., and T. Kakunaga, 1982 Potential Z-DNA forming sequences are highly dispersed in the human genome. Nature **298:** 396–398.

Jones, D. A., and J. D. G. Jones, 1997 The role of leucine-rich repeat proteins in plant defences. Adv. Bot. Res. **24:** 89–167.

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

Liu, Z.-W., R. M. Biyashev and M. A. Saghai Maroof, 1996 Development of simple sequence repeat markers and their integration into a barley linkage map. Theor. Appl. Genet. **93:** 869–876.

Loridon, K., B. Cournoyer, C. Goubely, A. Depeiges and G. Picard, 1998 Length polymorphism and allele structure of trinucleotide microsatellites in natural accessions of *Arabidopsis thaliana*. Theor. Appl. Genet. **97:** 591–604.

Manninen, I., and H. A. Schulman, 1993 BARE-1, a copia-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol. Biol. **22:** 829–846.

Milbourne, D., R. Meyer, J. E. Bradshaw, E. Baird, N. Bonar *et al.*, 1997 Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. Mol. Breeding **3:** 127–136.

Milbourne, D., R. C. Meyer, A. J. Collins, L. D. Ramsay, C. Gebhardt *et al.*, 1998 Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol. Gen. Genet. **259:** 233–245.

Morgante, M., and A. M. Olivieri, 1993 PCR-amplified SSRs as markers in plant genetics. Plant J. **3:** 175–182.

Morgante, M., B. C. Meyers, D. R. Argentar, M. Ramaker, M. Dolan *et al.*, 1999 Analysis of 2Mb of random DNA sequence repeats structural and organisational features of the corn genome. Poster 506, Plant and Animal Genome, San Diego, p. 208, Abstracts.

Niewohner, J., F. Salamini and C. Gebhardt, 1995 Development of PCR assays diagnostic for RFLP marker alleles closely linked to alleles gro1 and h1, conferring resistance to the root cystnematode *Globodera rostochiensis* in potato. Mol. Breeding **1:** 65–78.

Paglia, G. P., A. M. Olivieri and M. Morgante, 1998 Towards second-generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.). Mol. Gen. Genet. **258:** 466–478.

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

Panstruga, R., R. Büschges, P. Piffanelli and P. Schulze-Lefert, 1998 A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res. **26:** 1056–1062.

Provan, J., W. Powell and R. Waugh, 1996 Microsatellite analysis of relationships within cultivated potato (*Solanum tuberosum*). Theor. Appl. Genet. **92:** 1078–1084.

Ramsay, L., M. Macaulay, L. Cardle, M. Morgante, S. Degli Ivanissevich *et al.*, 1999 Intimate association of microsatellite repeats with retrotransposons and other repetitive elements in barley. Plant J. **17:** 415–425.

Ramsay, L., M. Macaulay, S. Degli Ivanissevich, K. MacLean, L. Cardle *et al.*, 2000 A simple sequence repeat-based linkage map of barley. Genetics (in press).

Röder, M. S., J. Plaschke, S. U. Konig, A. Borner, M. E. Sorrells *et al.*, 1995 Abundance, variability and chromosomal location of microsatellites in wheat. Mol. Gen. Genet. **246:** 327–333.

Röder, M. S., V. Korzun, K. Wendehake, J. Plaschke, M. H. Tixier *et al.*, 1998 A microsatellite map of wheat. Genetics **149:** 2007–2023.

San Miguel, P., A. Tikonov, Y.-K. Jin, N. Motschoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

Senior, M. L., E. C. L. Chin, M. Lee, J. S. C. Smith and C. W. Stuber, 1996 Simple sequence repeat markers developed from maize sequences found in the GENEBANK database: map construction. Crop Sci. **36:** 1676–1683.

Smulders, M. J. M., G. Bredemeijer, W. Rus-Kortekaas, P. Arens and B. Vosman, 1997 Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivar and accessions of other *Lycopersicon* species. Theor. Appl. Genet. **97:** 264–272.

Stallings, R. L., A. F. Ford, D. Nelson, D. C. Torney, C. E. Hildebrand *et al.*, 1991 Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. Genomics **10:** 807–815.

Sverdlov, V. E., O. I. Dukhanina, B. Hoebee and J. P. Rapp, 1998 Linkage mapping of fifty-eight new rat microsatellite markers. Mamm. Genome **9:** 816–821.

Wang, Z., J. L. Weber, G. Zhong and S. D. Tanksley, 1994 Survey of plant short tandem repeats. Theor. Appl. Genet. **88:** 1–6.

Weising, K., F. Weigand, A. J. Driesel, G. Kahl, H. Zischer *et al.*, 1989 Polymorphic simple GATA/GACA repeats in plant genomes. Nucleic Acids Res. **17:** 10128.

White, G., and W. Powell, 1997 Isolation and characterization of microsatellite loci in *Swietenia humilis* (Meliaceae): an endangered tropical hardwood species. Mol. Ecol. **6:** 851–860.

Wu, K.-S., and S. D. Tanksley, 1993 Abundance, polymorphisms and genetic mapping of microsatellites in rice. Mol. Gen. Genet. **241:** 225–235.

Zhao, X. P., and G. Kochert, 1992 Characterization and geneticmapping of a short, highly repeated, interspersed DNA-sequence from rice (*Oryza sativa* L.). Mol. Gen. Genet. **231:** 353–359.

Communicating editor: C. Haley