SSR frequency and occurrence in plant genomes

L. Cardle, M. Macaulay, D.F. Marshall, D. Milbourne, L. Ramsay & R. Waugh

Until recently, bioinformatics could give only a vague picture of plant genomes due to the short fragmentary nature of the plant DNA sequences available from public databases. As large scale sequencing projects are revealing more and more long, contiguous DNA sequences, the overall genetic structure of plants will become increasingly clear, providing better genetic models for the development of strategies for experimental studies.

Here we demonstrate a bioinformatic analysis which takes advantage of the significant number of long, contiguous sequences recently deposited in the international sequence databases. We have re-evaluated the distribution of Simple Sequence Repeats (SSRs), a sequence feature of eukaryotic genomes which are an important source of polymorphisms for genetic studies, and shown that SSRs are far more common in plant genomes than previously estimated.

The ubiquity of SSRs 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. The usefulness of SSRs has also been demonstrated for a variety of plant species and this has prompted the initiation of SSR discovery programmes for the majority of agronomically important crops. 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.

Source	*Arabidopsis thaliana*	
Subgroup	Genomic (P1 & BAC)	ESTs
Number of sequences	306	36199
Number with ≥ 1 SSR	305	1040
Repeat type:		
Mononucleotide	1471 (33)	103 (10)
Dinucleotide	1333 (30)	254 (24)
Trinucleotide	1350 (30)	706 (66)
Tetranucleotide	236 (5)	7 (<1)
Pentanucleotide	83 (2)	0
Total SSR content	4473	1070
Total length (kb)	27011.3	14808.0
Average distance (kb)	6.04	13.83

Numbers in brackets show percentage of total SSR content

Table 1 SSR survey of *Arabidopsis* genomic and EST sequence data.

Previous analyses of plant DNA sequence database entries for all possible SSR motifs have revealed frequencies ranging from one every 29kb to 50kb, depending on species. Oligonucleotide hybridisation studies have suggested figures in the range of one SSR every 65kb to 80kb. These results contrast sharply with those for humans, with an estimate of one SSR every 6kb on average.

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, or developing selective SSR enrichment techniques, in order to generate sufficient numbers of SSRs. 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.

Frequency and distribution of SSRs in *Arabidopsis thaliana* Three hundred and six non-redundant, genomic DNA sequences longer than 10 kb, and over 36000 EST sequences were retrieved from the EMBL nucleotide database (on 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, and each clone possessed 10 SSRs on average (Fig. 1a, see Fig. 2 for an example of SSR distribution in a clone). In contrast, only 3% of ESTs contained an SSR, which is similar to the proportion previously found in rice ESTs[1]. Overall, the average distance between SSRs in *Arabidopsis* genomic DNA was approximately 6 kb compared to a figure of 14 kb for ESTs (Table 1).

By examining the detailed features tables available for 51 of the 306 *Arabidopsis* genomic sequences, a con-

siderable difference in the distribution of SSR motifs was found between introns, exons and intergenic regions (Fig. 1b). Almost two-thirds of SSRs were found in intergenic regions (and the majority of these were either mono- or di-nucleotides), 14% were found in exons, and 23% in introns.

Of the exonic SSRs, 91% were tri-nucleotides, reflecting repetitive amino-acid sequence motifs, although there was no simple pattern of motifs in relation to different protein classes. The remaining 9% was made up of 10 di-nucleotides, one mono- and one penta-nucleotide repeat. A more diverse range of motif lengths was found in introns, with similar proportions of repeat types being found in intergenic regions. The proportions found within the data examined indicated that over 40% of all trinucleotide repeats are exonic in *Arabidopsis.*

SSR distribution in other plants In similar analyses of the 52 genomic DNA sequences (>10 kb) 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, almost identical to that found in *Arabidopsis* alone.

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 was 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 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 towards 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 which make up 31% of the genomic sequence but contain only 14% of the

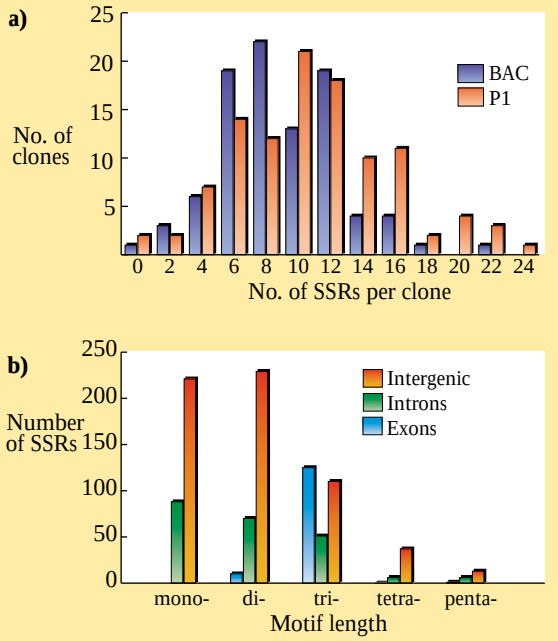

Figure 1 Frequency of SSRs in *Arabidopsis.* a) Actual number of SSRs in a collection of P1 and BAC clones, b) Number of SSRs of different repeat lengths in intron, exon and intergenic sequence.

SSRs, 91% of which are trinucleotide. This finding corresponds well with the lower frequency of SSRs and the preponderance of trinucleotide repeats found in EST sequences compared to genomic sequences[1].

An experimental approach We developed and tested 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. In demonstrating the approach in a barley BAC clone, only one SSR met our 'repeat length' definition and two were slightly short. Nevertheless, polymorphism at one SSR enabled the BAC clone to be mapped to a chromosomal position, which was confirmed by the use of another short SSR known to be upstream of the gene sequence. 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 estimated values above. However, the polymorphism found in these short SSRs and others, 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 which are intrinsically 'high value' for several reasons. BAC

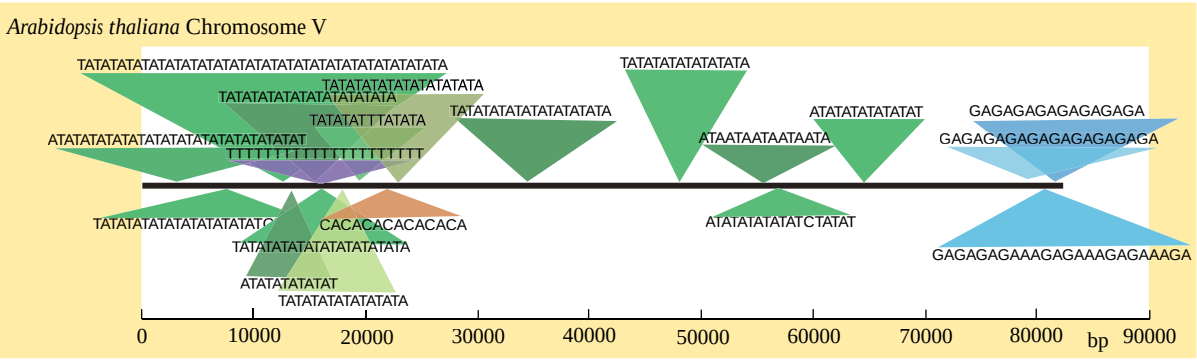

Figure 2 Distribution of SSR sequences in P1 clone MCL19 (84,510 bp).

clones linked to genes of interest can be selected by hybridisation 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. Even when the actual sequence of the target gene is known, the use of 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. They are also suitable for 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 (SNPs) in analyses.

Conclusion 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. However, differences in the 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. 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.

A complete description of the above study can be found in L. Cardle et al. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156(2), 847-854.

Reference

[1]Akagi H., Yokozeki Y., Inagaki A. & Fujimura T. (1996). *Theoretical and Applied Genetics* 93, 1071-1077.