J. Plant Biotechnology (2005) Vol. 7 (1). pp. 17~25

Journal of Plant Biotechnology

Simple Sequence Repeat (SSR) and GC Distribution in the *Arabidopsis thaliana* Genome

Jennifer C Mortimer[1,2,3], Jacqueline Batley[1], Christopher G Love[1,2], Erica Logan[1,2], David Edwards[1,2*]

[1]Plant Biotechnology Centre, Primary Industries Research Victoria, La Trobe University, Bundoora 3086, Victoria, Australia; [2]Victorian Bioinformatics Consortium, Plant Biotechnology Centre, Primary Industries Research Victoria, La Trobe University, Bundoora 3086, Victoria, Australia; [3]Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

Abstract

We have mined each of the five *A. thaliana* chromosomes for the presence of simple sequence repeats (SSRs) and developed custom perl scripts to examine their distribution and abundance in relation to genomic position, local G/C content and location within and around transcribed sequences. The distribution of repeats and G/C content with respect to genomic regions (exons, UTRs, introns, intergenic regions and proximity to expressed genes) are shown. SSRs show a non-random distribution across the genome and a strong association within and around transcribed sequences, while G/C density is associated specifically with the coding portions of transcribed sequences. SSR motif repeat number shows a high degree of variation for each SSR type and a high degree of motif sequence bias reflecting local genome sequence composition. PCR primers suitable for the amplification of identified SSRs have been designed where possible, and are available for further studies.

Key words: Arabidopsis, GC distribution, Genome Structure, Genome topology, Simple Sequence Repeat (SSR)

Introduction

Since Walther Flemming published the first images of chromatin in 1882, there has been increasing research into this genetic material. The recent deduction of the complete coding sequence for the human, rice and *Arabidopsis thaliana* genomes represents the latest step in the quest to understand the basis of inheritance and biological function at the molecular level. The availability of the complete sequence of the *A. thaliana* genome enables detailed distribution analysis of features and motifs. An understanding of feature distribution has practical applications in genetic and physical mapping and diversity studies both in *A. thaliana* and related crop species. As well as coding sequences and the sequences which regulate their expression, other sequence features including transposons, GC rich isochores and SSRs (Simple Sequence Repeats) have been identified in the genomic landscape. SSRs, also known as microsatellites, are common, readily identified DNA features consisting of short (1-6 bp), tandemly repeated sequences, widely and ubiquitously distributed throughout eukaryotic genomes (Schlötterer and Pemberton 1994; Tóth et al. 2000) and have been found in all eukaryotic genomes so far analysed (Katti et al. 2001). They are highly polymorphic and informative markers, currently used for a wide range of applications including gene flow, genetic diversity, paternity exclusion and genetic mapping (Tautz 1989; Powell et al. 1996). SSRs were initially considered to be evolutionally neutral, (Awadalla and Ritland 1997), though recent evidence suggests an important role in genome evolution (Moxon and Wills 1999). SSRs are a source of abundant, non-deleterious mutations that provide variation in the face of stabilising selection, and their recognised role in the pro cess of evolutionary adaptation is predicted to increase as our knowledge of them expands (Kashi et al. 1997). SSR stability may be correlated with overall levels of genomic stability (Ross et al. 2003) as mutations which affect SSR stability, such as those involved in DNA mismatch repair, can also influence genomic stability.

* Corresponding author, E-mail: Dave.Edwards@dpi.vic.gov.au
Received Oct. 7, 2004; Accepted Mar. 15, 2005

The potential biological function and evolutionary relevance of SSRs is currently under scrutiny and leading to a greater understanding of genomes and genomics (Subramanian et al. 2003). Initial suggestions that the majority of DNA was either 'junk' or had no biological function are being challenged by the discovery of new functions for these sequences. Various functional roles have now been attributed to SSRs. For example, SSRs are believed to be involved in gene expression, regulation and function (Gupta et al. 1994; Kashi et al. 1997) and there are numerous lines of evidence suggesting that SSRs in noncoding regions may also be of functional significance (Kashi et al. 1997). Furthermore, SSRs provide hotspots of recombination, a variety of SSRs have been found to bind nuclear proteins and there is direct evidence that SSRs can function as transcriptional activating elements (Li et al. 2002).

SSRs are frequently identified from EST database searches or sequence data obtained from SSR-enriched genomic libraries, which provide little data about genomic distribution. Initial studies of SSR distribution in physical mapping, using *in situ* hybridisation, in fish and primates, showed clustering of SSRs on some chromosomes (Nanda et al. 1991). Further studies in *Drosophila* suggested the association of these sequences with euchromatin, with a reduced level of hybridisation observed around centromeric regions (Pardue et al. 1987; Lowenhaupt et al. 1989). In addition, this association was shown to be evolutionally stable, with similar results observed between two different *Drosophila* strains. Subsequent *in situ* hybridisation studies of genomic organisation of SSRs in plants have demonstrated a low density of SSRs around the centromeres (Schmidt and Heslop-Harrison 1996). However, the exclusion of certain SSR repeat types from rRNA gene clusters, centromeres and intercalary sites was observed and the level of hybridisation was reduced around the centromeres.

In contrast to an earlier study by Ramsay et al. (1999), Morgante et al. (2002) demonstrated the preferential association of SSRs with non-repetitive DNA sequences for a number of plant genomes. The availability of complete and annotated genome sequences of a number of organisms has provided an excellent opportunity to analyse SSRs in greater detail for their genomic locations, distributions and frequencies. Results from such analysis provide a useful basis for carrying out further investigations into the structure and function of SSRs (Sreenu et al. 2003). *A. thaliana* is a significant model species for investigating the genomic distribution of SSRs, since the entire sequence of the relatively small nuclear genome has been deduced and SSRs are known to be abundant. Therefore, to further understand the relationship between SSR distribution and other genomic features, we have undertaken a whole genome analysis of the five completely sequenced *A. thaliana* chromosomes, assessing the abundance and length distribution of di-, tri-, tetra- and penta-nucleotide repeats in relation to genomic position, association within and around transcribed sequences and local GC density.

Materials and Methods

Sequence datasets

Arabidopsis thaliana sequence datasets were obtained from TAIR (The Arabidopsis Genome Initiative 2000; http://www.arabidopsis.org). These data included five complete chromosome sequences, and multi fasta files representing exons, introns, coding sequence, intergenic regions, 3' UTRs (UnTranslated Regions), 5' UTRs and 1000 bp and 3000 bp regions flanking expressed genes. The original gene flanking sequence datasets represented mixtures of UTR and non-UTR containing sequences. These were used to identify a subset of 1000 bp genomic sequences flanking annotated UTRs. This dataset is available on request from the author.

Identification of SSRs and GC density

The program SPUTNIK (Abajian 1994, http://abajian.net/sputnik/) was applied, in combination with custom perl scripts, to identify SSRs within a series of overlapping windows of 100 Kb with a 1 Kb step, across the five *Arabidopsis* chromosomes and within each of the additional datasets. The criteria for SSR discovery was a minimum repeat length of 5 repeat units for dinucleotide repeats, 4 repeat units for trinucleotide repeats, 3 repeat units for tetranucleotide repeats and 2 repeat units for pentanucleotide repeats. We report motifs that represent the same SSR by one repeat motif, eg AC = CA, AC, GT, TG. The total length of each SSR was calculated and includes point mutations and indels. No distinction was made between perfect, imperfect and interrupted SSRs (Weber 1990). Custom perl scripts were also applied to calculate GC content using overlapping windows as above. SPUTNIK and custom perl scripts were processed using a SunOS 5.8 (2Gb RAM, 2x UltraSparc II, 400MHz). All custom scripts are available on request from the authors.

Identification of PCR Primer Pairs for SSRs

The program SSRPrimer (Robinson et al. 2004) was applied, in combination with custom perl scripts, to identify SSRs across the five *Arabidopsis* chromosomes. Primer specificity was set to design primers greater than 10bp either side of the identified SSR. The optimum size for the primers are 21 bases with a maximum of 23 bases. Optimum melting temperature is 55℃, minimum 50℃, maximum temperature is 70℃ and maximum GC content is 70%.

Results

Frequency and distribution of SSRs across the genome

This study identified a total of 44,316 SSRs over 116.89 Mb of the complete sequence of the five *A. thaliana* chromosomes, giving an overall density of 379.29 SSRs per Mb with SSRs making up a total of 0.5% of the *A. thaliana* genome.

SSRs were classified according to motif type and sequence. Overall, trinucleotide repeats were found to be the most abundant, followed by dinucleotide, pentanucleotide and tetranucleotide repeats, with a similar density of each of these motifs across each of the five chromosomes (Figure 1). A clear bias was observed in SSR motif sequence. For example 49% of all dinucleotide SSRs consisted of $[AT]_n$ repeats while $[GC]_n$ repeats constituted only 0.045% of the dinucleotide SSRs and 0.013% of the total SSRs. This pattern was observed for all other repeat lengths. For example, the most abundant trinucleotide repeat identified in this study was $[AAG]_n$ at 42% of the total, while $[CCG]_n$ was found to be the least abundant (8.5%) (supplementary data). This reflects the overall nucleotide and motif composition within each of the *A. thaliana* chromosomes. There was a reduced frequency of repeats containing either G or C nucleotides and a significant bias against CG or CNG sequence containing motifs (supplementary data). This again reflects an overall bias against these sequences in the *Arabidopsis* genome. This trend was further demonstrated in the analysis of the SSR repeat types which were not identified within the genome. In total, fourteen (9.4%) of the 149 possible repeat motifs were not identified within the *A. thaliana* genome. All of these unidentified motifs consist of a high G/C content (AGGC, GGCC, GGGC, ACAGG, ACCCT, ACGCT, ACGGC, ACGTC, AGCCT, AGCGC, AGGCC, AGGGC, CCCGC, CCGCG).

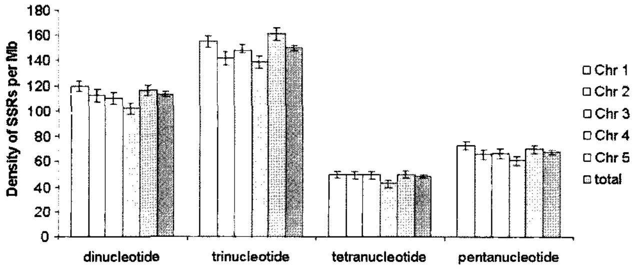

Figure 1. Distribution of SSR repeat types across *Arabidopsis thaliana* chromosomes.

Frequency and distribution of SSRs and GC distribution across individual chromosomes

The SSR frequency across each of the five chromosomes was assessed using 100 kbp overlapping windows with a 1 kbp step. SSR frequency showed a non-random distribution across each of the chromosomes with regions of high and low SSR density (Figure 2). To test for association of SSR density with other genome features, both gene density and GC content were measured, using the same overlapping windows (Figure 2a-e). GC content ranged from 31 to 45% and was observed to be higher in gene-rich regions. SSR density closely followed both GC and gene density across all five chromosomes, with the exception of around predicted centromeric regions and a region of heterochromatin on chromosome 4, in which significant reductions in gene and SSR density were accompanied by a rise in GC content (Figure 2).

Frequency and distribution of SSRs and GC distribution around coding regions

In order to examine, in detail, the association of SSRs with expressed genes, total genomic sequence was divided into expressed sequence and non-expressed sequence portions. Expressed sequence was again subdivided into coding sequence, introns, 5' UTRs and 3' UTRs, while non-transcribed sequence was subdivided according to distance either 5' or 3' from expressed genes.

Transcribed sequences exhibited a higher overall frequency of SSRs than non-transcribed sequences, with the highest SSR density observed in UTRs and introns and a significant reduction in SSRs within coding sequences (Figure 3). The high frequency of SSRs appears to extend into untranscribed sequence adjacent to transcribed regions, with SSR frequency decreasing by half within 100 bp from the annotated transcription start site (Figure 4). In addition, coding sequences maintained a high frequency of trinucleotide repeats with a reduced frequency of other repeat types compared with non-coding sequence (Supplementary data). The exons were found to be relatively GC rich (Figure 3), but introns, untranslated sequences and gene flanking regions exhibited a GC content similar to that observed within intergenic DNA sequences. The GC density dips sharply around 50 bp upstream from the 5' UTR (Figure 4). GC

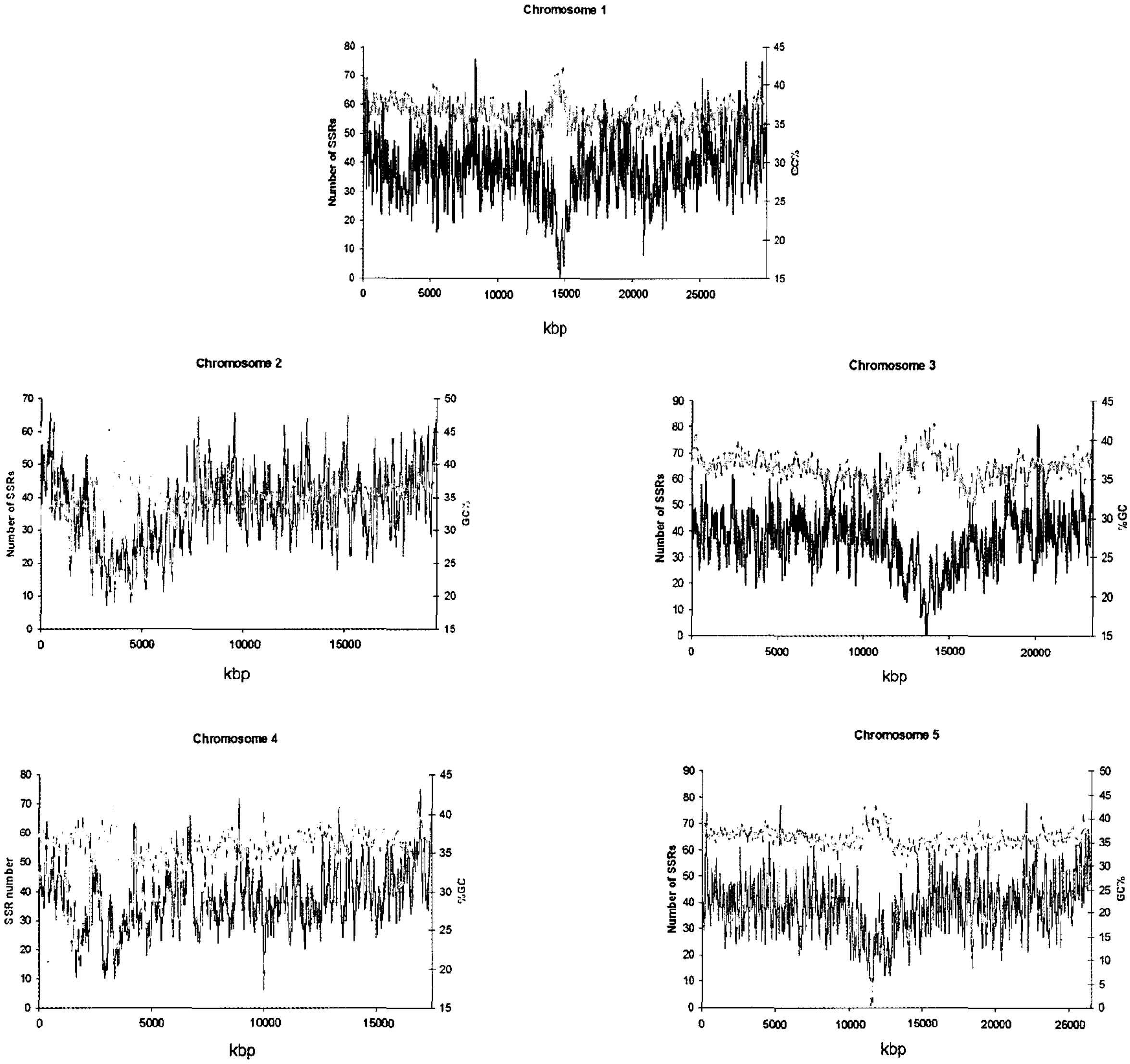

Figure 2 (a-e). SSR distribution and GC content across each *Arabidopsis* chromosome calculated using a 100 Kb sliding window (1 Kb step).

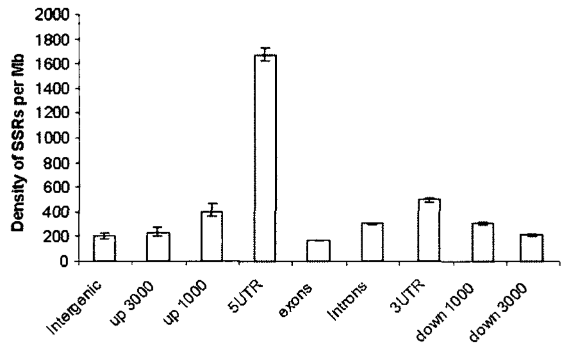

Figure 3a. Density of SSRs/Mbp for each of the individual genomic regions: intergenic, upstream 3000 bp of transcribed sequence, upstream 1000 bp of transcribed sequence, 5' UTRs, exons, introns, 3' UTRs, 1000 bp downstream of transcribed sequence and 3000 bp downstream of transcribed sequence.

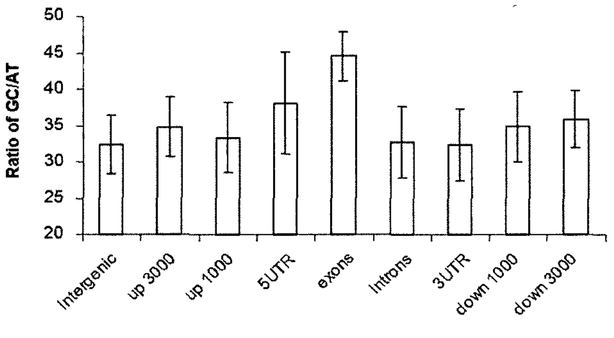

Figure 3b. GC content for each of the individual genomic regions: intergenic, upstream 3000 bp of transcribed sequence, upstream 1000 bp of transcribed sequence, 5' UTRs, exons, introns, 3' UTRs, 1000 bp downstream of transcribed sequence and 3000 bp downstream of transcribed sequence.

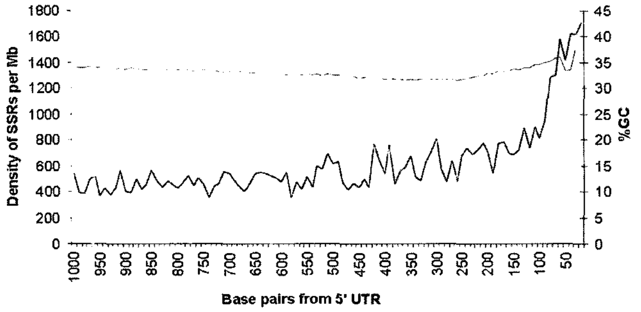

Figure 4. Density of SSRs/Mbp and GC content 1000 bp upstream of transcribed sequence

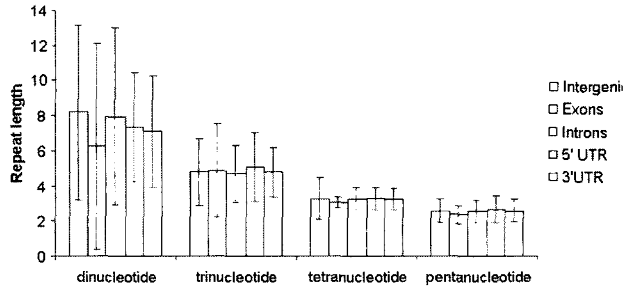

Figure 5. Average repeat length of di, tri, tetra and pentanucleotide SSRs (bp) in exons, introns, UTRs and intergenic regions of the genome.

density then trails to 31%, before increasing to 35% 1000 bp upstream of the 5' UTR.

Over the entire genome, SSR tandem arrays were found to have a mean length of 13.8 bp, with dinucleotide repeats longest at 15 bp, followed by trinucleotide repeats (14.5 bp), tetranucleotide repeats (12.9 bp) and pentanucleotide repeats (12.8 bp). Dinucleotide repeats were found to be generally shorter within coding sequence and longer within intergenic regions or introns (Figure 5). However, large standard deviations were recorded for all the mean SSR lengths.

Examination of the proportion of each motif sequence allowed the identification of SSR sequence bias between genomic regions. The proportions of dinucleotide repeat motif sequences were shown to be similar across all genomic regions, with the exception of the 5' UTR and exon subsets which contained 89.52% $[AG]_n$ dinucleotide SSRs, and 85.85% $[AT]_n$ dinucleotide SSRs respectively (supplementary data).

Within the trinucleotide repeat set, the 5' UTR demonstrated a higher frequency of the motif $[AAG]_n$ (73.13%) compared to a total genome frequency of 42.14%. The unusually high proportion of AG containing motifs in the 5' UTR was also observed for tetranucleotides ($[AAAG]_n$ - 41.77% compared with a genome average of 19.27%). However, this effect was also accompanied by a much reduced frequency of the $[AAAT]_n$ motif type (8.01% for 5' UTR compared with 32.71% for the total genome) (supplementary data).

SSR Primer Design

In order to provide a resource for further studies on polymorphism and mutation information of SSRs across the *A. thaliana* genome, the entire genomic sequence was processed using SSRPrimer (Robinson et al. 2004) to design PCR primers suitable for the amplification of the SSRs. PCR primers were designed to 33,274 (75%) of the total SSRs and the primer sequences are available on the web (http://hornbill.cspp.latrobe.edu.au) and as supplementary information. The proportion of SSRs with PCR primers successfully designed is similar between the five chromosomes, ranging from 73.8% (chromosome 2) to 75.7% (chromosome 5). However, it is noted that trinucleotide motif SSRs have a greater proportion of primers successfully designed to them (86%) than di-, tetra- and penta-nucleotides SSRs (65% to 69%).

Discussion

Large scale genomic sequencing provides the opportunity to evaluate the abundance and relative distribution of SSRs between transcribed and non-transcribed regions and the relationship of these to genomic features. In the present study we have analysed the occurrence and density of SSRs and GC distribution across the *A. thaliana* genome. We present data on an individual chromosome basis and each chromosomal dataset has been split into exonic, intronic and intergenic regions. The data have been analysed by identification of di, tri, tetra and pentanucleotide classes of repeats. The study of repeat density and its distribution pattern in the genome is expected to help with the understanding of the biological significance of SSRs.

The proportion of the *A. thaliana* genome found to be covered by the 44316 SSRs (0.5%), is less than that reported in the human genome, of which 3% is composed of SSRs (Subremanian et al. 2003). SSRs are thought to occur at a much higher frequency than would be predicted purely on grounds of base composition (Tautz and Renz 1984). We found the density of each repeat class to be comparable across various genomic regions, however, different repeat motifs showed substantial variation in density in different genomic regions. The observed abundance of the trinucleotide repeats followed by dinucleotide, pentanucleotide and tetranucleotide repeats in this study, is unsurprising and supports previous work by Katti et al. (2001) which examined repeats of 40 or less nucleotides in *A. thaliana* chromosomes 2 and 4.

The non-random pattern of SSR distribution observed in this analysis of *A. thaliana* has been noted in previous studies which had suggested non-random distribution of SSRs or bias towards certain motifs (Ramsay et al. 1999; Tóth et al. 2000; Kantety et al. 2002; Varshney et al. 2002). However, in these studies, the sequences were collated from a limited number of published genes (or ESTs) or from limited genomic sequence. These datasets are not representative of the genome as a whole and differences may reflect bias inherent in the dataset. The use of the entire genomic sequence removes such previous criticism. The strong bias observed here in SSR motif sequence, with 49% of all dinucleotide SSRs consisting of $[AT]_n$ repeats while $[GC]_n$ repeats made up only 0.045% of dinucleotide SSRs also validates previous studies which suggest the predominance of $[AT]_n$ repeats in both *Arabidopsis* and yeast (Katti et al. 2001).

High frequencies of $[GT]_n$ repeats have been identified in the human and *Drosophila*, genomes, while $[GA]_n$ repeats predominate in *Caenorhabditis elegans* (Schlötterer 2000; Katti et al. 2001) and vertebrates show a greater abundance of tetranucleotide repeats than many other groups of organisms (Toth et al. 2000). While this observation reflects overall nucleotide and motif composition within each of these genomes, the disparities in SSR representation cannot be wholly explained by these differences (Ross et al. 2003). The large variation observed between frequencies of different SSR motifs may, therefore, be partly explained by the location of the SSR within the genome. Morgante et al. (2002) found that dinucleotide repeat composition differed between genomic fractions: $[GA]_n$ abundance was higher in EST sequences (transcribed regions) and their frequency in 5' UTRs was an order of magnitude higher than in the genome as a whole. The $[GA]_n$ SSRs in UTRs are believed to be involved in antisense transcription. Furthermore these authors demonstrated that $[AT]_n$ repeats were typically located in non-transcribed regions. It is consequently possible that there are different mechanisms acting on SSR evolution in transcribed and non-transcribed regions.

The observation in this study of the reduced frequency of repeats containing either G or C nucleotides and the significant bias against CG or CNG sequence containing motifs reflects an overall bias against these sequences in the *Arabidopsis* genome. The reduced level of CG and CNG sequences has been previously observed in human, *Drosophila, C. elegans*, yeast genomes and a restricted *Arabidopsis* dataset (Katti et al. 2001; Morgante et al. 2002). SSR density was found to closely follow both GC and gene density across all five chromosomes, with the exception of the centromeric regions, where significant reductions in gene and SSR density were accompanied by a rise in GC levels. The relationship between SSR density, gene density and GC content observed in this analysis on a chromosomal basis was clarified when subsets of transcribed sequences were examined. While exons were shown to be relatively GC rich, reflecting their conserved codon triplet sequence, introns, untranslated sequences and gene flanking regions exhibited a lower GC content, similar to that observed within intergenic DNA sequence. This is in contrast to SSR density, which remains relatively high across transcribed sequences. This observation suggests that the pattern described in figure 1 is due not to a direct correlation between GC ratio and SSR frequency, but is an indirect relationship, in which SSRs are associated with transcribed regions, while the higher regions of GC content are associated with coding sequence. The direct correlation observed between GC density and gene density across all 5 chromosomes confirms previous studies which indicated that genic regions were rich in GC nucleotides (Barakat et al. 1999; Arhondakis et al. 2004). The observed dip in GC content, 50 bp upstream of the 5' UTRs (Figure 4) is likely to represent the conserved location of the AT rich TATA box motif. The observed reduction and subsequent increase in GC density moving away from the gene in the 1000 bp upstream of the 5' UTR (Figure 4) has not been noted previously. This may be associated with the location of AT rich transcription factor binding sites within this region or possibly the general requirement for proteins to access and modify DNA structure in this region.

The increased GC content in the exons and 5' UTR may also explain the bias in SSR primer design between motif types, with a greater number of primers designed to trinucleotide SSRs (85%) compared to di-, tetra- and pentanucleotide repeats (65% to 69%). The SSRPrimer design software requires a minimum GC content of 30% with an associated minimum primer annealing temperature of 50℃. Transcribed regions were shown to be rich in trinucleotide SSRs and also displayed the greatest GC density while in non-transcribed regions, which are relatively poor in trinucleotide SSRs, the GC density surrounding SSRs may frequently be too low for the design of suitable PCR primers. The availability of PCR primer sequences for SSR amplification is useful for selection of SSR marker loci around regions and genes of interest, as well as for studying the mutation and polymorphism of SSRs in relation to genomic location, repeat type and repeat length.

It was found that the transcribed sequences exhibited a higher overall frequency of SSRs than non-transcribed sequences for all the sections of the transcribed regions except for the exons. Although exons showed high densities of trinucleotide repeats, all other repeats occurred at very low frequencies. This may reflect evolutionary constraints inherent within sequences that encode functional proteins and can be explained by the complete impairment of protein function by the frameshift mutation which takes place when an exon receives insertion/deletion events whose length is not a multiple of codon length (Borštnik and Pumpernik 2002). Our findings are supported by the results of Morgante et al. (2002) who reported that trinucleotide SSRs doubled in frequency in coding regions of plants compared to non-coding regions, as a result of mutation pressure and possible positive selection for specific single amino acid stretches.

The high frequency of gene associated SSRs appears to extend into the untranscribed sequences upstream of expressed genes, with SSR frequency decreasing by half within 100 bp from the annotated transcriptional start site. SSRs have previously been identified within gene promoter regions (Holland et al. 2001). In addition, it has been shown that deletion/insertion of repeats in promoter regions may modulate the expression of genes (Khashnobish et al. 1999). The transcribed regions of the genome are those regions most subjected to the pressures of natural selection, and so the increased levels of SSRs observed here suggests that they may have some biological function. In particular the observed increase in SSR density 5' upstream from transcribed genes suggests a role for SSRs in the regulation of gene expression in *Arabidopsis*.

Centromeric heterochromatin regions showed a reduction in SSR density for each of the five chromosomes. Previous studies of the centromeric regions of animals and plants have demonstrated the presence of large blocks of tandemly repeated satellite sequences and retroelement-like components that are embedded in recombination-deficient heterochromatic regions (Li et al. 2002). The sequence of the *Arabidopsis* centromeric region shows that this consists of megabase extents of tandem repeat arrays with a 180 bp unit length that is not present elsewhere in the genome (The *Arabidopsis* Genome Initiative 2000). The presence of such repeats may result in the observed increased in GC density within the centromeric and heterochromatin regions and also the observed reduction in SSR density within the centromeric regions of the *Arabidopsis* genome.

There were no observed significant differences in the average SSR repeat length between the non-coding regions, introns, UTRs and exons. The dinucleotide SSRs had the longest average repeat length, which may be expected as it has been previously shown that dinucleotide repeats have the highest occurrence of rare long alleles (Ross et al., 2003). The length distributions of SSRs observed in A. thaliana indicates that the frequency of repeats decreases with repeat length. This may be because these SSRs have higher mutation rates and are less stable. This effect has been demonstrated by Xu et al. (2000) who report that compared to expansion mutation events, contraction mutation events occur more frequently with increases in allele size, and long alleles tend to mutate to shorter lengths to prevent their infinite growth. Of the SSRs longer than 40 bp, dinucleotide repeats were most common, perhaps because slippage frequency is greatest in dinucleotide repeats (Katti et al. 2001). DNA replication slippage rate is dependent on the number of repeats in the SSR as alleles of longer length (more repeats) are less stable than those with fewer repeats (Dieringer and Schlötterer 2003). Analysis of sequence data from human, mouse, C. elegans and yeast also show that dinucleotide repeats are typically longer in non-coding regions than coding regions (Li et al. 2002). It has been proposed that this is because there is a higher tolerance of non-coding DNA to mutations. The average repeat length of SSRs may also be associated with recombination (Brandes et al. 1997). SSRs, especially dinucleotide repeats, have been reported to provide hotspots for genetic recombination. However, the repeat number may also influence recombination and may also affect DNA replication. The number of repeat units within an SSR array is also reported to have an effect on gene expression. Some genes can only be expressed at a specific SSR repeat number or within a narrow range of SSR repeat numbers. Out of this range, gene expression is repressed. Other genes have a wide range of SSR repeat numbers with no impact on gene expression.

The analysis of SSRs in complete genome sequences provides information about the detailed distribution of features and motifs. We have used the complete *A. thaliana* genome sequence to further our understanding of SSR distribution across genome fractions. Details of all the *A. thaliana* SSRs are available on request from the corresponding author. This information may be useful for the selection of a wide range of SSRs for further studies. The availability of data on the SSR content of complete chromosome sequences should facilitate studies on the function of SSRs in genome organisation and gene expression and regulation. In consideration of the importance of SSR sequences, it seems inevitable that there will be a need to analyse and compare in detail the distribution of these repeats and genes associated with them in a range of organisms.

References

Abajian C (1994) SPUTNIK

Arhondakis S, Auletta F, Torelli G, D'Onofrio G (2004) Base composition and expression level of human genes. Gene 325: 165-169

Awadalla P, Ritland K (1997) Microsatellite variation and evolution in the Mimulus guttatus species complex with contracting mating systems. Mol Biol Evol 14: 1023-1034

Barakat A, Han DT, Benslimane AA, Rode A, Bernadi G (1999) The gene distribution in the genomes of pea, tomato and date palm. FEBS Lett 463: 139-142

Borštnik B, Pumpernik D (2002) Tandem repeats in protein coding regions of primate genes. Genome Res 12: 909-915

Brandes A, Thompson H, Dean C, Heslop-Harrison JS (1997) Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidopsis thaliana* L. Chromosome Res 5: 238-246

Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 1: 2242-2251

Gupta M, Chyi Y-S, Romero-Severson J, Owen JL (1994) Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theor Appl Genet 89: 998-1006

Holland JB, Hellend SJ, Sharopova N, Rhyne DC (2001) Polymorphism of PCR based markers targeting exons, introns, promoter regions, and SSRs in maize and introns and repeat sequences in oat. Genome 44: 1065-1076

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48: 501-510

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

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18: 1161-1167

Khashnobish A, Hamann A, Osiewacz HD (1999) Modulation of gene expression by (CA)(n) microsatellites in the filamentous ascomycete Podospora anserina. Applied Microbiol Biotech 52: 191-195

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11: 2453-2465

Lowenhaupt KY, Rich A, Pardue ML (1989) Nonrandom distribution of long mono-nucleotide and dinucleotide repeats in Drosophila chromosomes - correlations with dosage compensation, heterochromatin and recombination. Mol Cell Biol 9: 1173-1182

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30: 194-200

Moxon ER, Wills C (1999) DNA Microsatellites: Agents of Evolution. Sci Am 280: 94-99

Nanda I, Zischler H, Epplen C, Guttenbach M, Schmid M (1991) Chromosomal organisation of simple repeated DNA Sequences used for DNA fingerprinting. Electrophoresis 12: 193-203

Pardue ML, Lowenhaupt K, Rich A, Nordheim A (1987) (DC-DA)N.(DG-DT)N sequences have evolutionarily conserved chromosomal locations in Drosophila with implications for roles in chromosome structure and function. EMBO J 6: 1781-1789

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

Ramsay L, Macaulay M, Cardle L, Morgante M, degli Ivanissevich S, Maestri E, Powell W, Waugh R (1999) Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J 17: 415-425

Robinson AJ, Love CG, Batley J, Barker G, Edwards D (2004) Simple sequence repeat marker loci discovery using SSR-Primer. Bioinformatics (In Press)

Ross CL, Dyer KA, Erez T, Miller SJ, Jaenike J, Markow TA (2003) Rapid divergence of microsatellite abundance among species of Drosophila. Mol Biol Evol 20: 1143-1157

Schmidt T, Heslop-Harrison JS (1996) The physical and genomic organisation of microsatellites in sugar beet. Proc Natl Acad Sci USA 93: 8761-8765

Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Nucleic Acids Res 20: 211-215

Schlötterer C, Pemberton J (1994) The use of microsatellites for genetic analysis of natural populations. In: Scheirwater B, Streit B, Wagner GP, DeSalle R, (eds), Molecular Ecology and Evolution: Approaches and Applications. Birkhauser Verlag Basel, Switzerland, pp 71-86

Sreenu VB, Alevoor V, Nagaraju J, Nagarajaram HA (2003) MICdb: database of prokaryotic microsatellites. Nucleic Acids Res 31: 106-108

Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol 4: R13

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17: 6463-6471

Tautz D, Renz M (1984) Simple sequences as ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12: 4127-4138

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815

Tóth G, Gáspári Z, Jurka J (2000) Microsatellites in different eukaryotic genomes:survey and analysis. Genome Res 10: 967-981

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 7: 537-546

Weber JL (1990) Informativeness of human (DC-DA)n. (DG-DT)n polymorphisms. Genomics 7: 524-530

Xu X, Peng M, Fang Z, Xu X (2000) The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24: 396-399