ELSEVIER

Plant Science 157 (2000) 97–104

PLANT SCIENCE

www.elsevier.com/locate/plantsci

Distribution of microsatellites in relation to coding sequences within the *Arabidopsis thaliana* genome

Elena Casacuberta, Pere Puigdomènech *, Amparo Monfort

Departament de Genètica Molecular, Institut de Biologia Molecular de Barcelona, CID-CSIC, Jordi Girona, 18, 08034 Barcelona, Spain

Received 9 February 2000; received in revised form 4 April 2000; accepted 4 April 2000

Abstract

The distribution of repetitive sequences, or microsatellites, formed by either one or two base pairs and longer than eight units, has been studied in almost 1 Mb of the sequenced *Arabidopsis thaliana* genome. Except for those formed by only G and C residues, the repetitions are more abundant in the Arabidopsis genome than can be calculated from its nucleotide composition. They are distributed in proportions higher than expected in introns, and in the intergenic regions both proximal and distal to the coding sequences. In exons, only the TC/GA microsatellite seems to be particularly abundant. The AT/TA microsatellites produce more length variation between Arabidopsis ecotypes than the A/T repeated sequences. These two classes are more abundant per kilobase than coding sequences in the Arabidopsis genome. The results indicate not only that the presence of microsatellites is not an effect of random distribution of nucleotides, but that their resolution as molecular markers may be equivalent to the number of genes and also that they do not seem to be systematically linked to specific regulatory sequences proximal to genes.

Keywords: *Arabidopsis thaliana*; Microsatellites; Length polymorphism

1. Introduction

Genome projects are providing the structure of large portions of genome sequences in different species. Among eukaryotes, yeast [1] and *Caenorhabditis elegans* [2] genomes have already been completed by sequenced. The first plant genome to be fully sequenced will be the one from *Arabidopsis thaliana*. At the moment chromosome 2 and chromosome 4 have been sequenced [3,4]. The genome sequence provides information on the complete collection of genes, an important step to understand the genetic basis of physiology and development. However, the sequence also provides information on the structure of genomes themselves.

Plant genomes are an example of extreme variability in size among higher organisms. A factor of 1000 exists when comparing the DNA contents of different species [5] while there is no evidence of a comparable variability in the number of coding sequences. Therefore the difference between genome sizes lies in the length of the intergenic sequences. The nature of these sequences has been studied in maize where the analysis of a large genome section has shown that the intergenic regions in this species are essentially formed by repetitive DNA produced by the reproduction of mobile elements, mostly retrotransposons [6].

A. thaliana seems to have the most compact genome among plants. Extrapolation of the first sequencing data indicates that around 21 000 genes will be coded by approximately 100 Mbp of DNA coding regions [7]. This leaves little space for intergenic regions. However, careful analysis of genomic data has shown that Arabidopsis has all classes of repetitive sequences, including transpo-

* Corresponding author. Tel.: +34-93-4006100; fax: +34-93-2045904.

E-mail address: pprgmp@cid.csic.es (P. Puigdomènech).

0168-9452/00/$ - see front matter
PII: S0168-9452(00)00271-5

sons [8], retrotransposons [9–12] and MITEs [13]. It may be that the Arabidopsis genome has the same types of mobile elements as other species but they have been multiplied to a lower extent.

Microsatellites or simple sequence repeats (SSR) are the simplest type of repetitive sequence. They are stretches of a short nucleotide sequence that can be repeated many times in tandem. The interest in microsatellites lies in their variability. This is the reason why they are widely used as molecular markers for genome analysis [14–17] enabling the characterization of Arabidopsis ecotypes [18–20]. They are also one of the features of the intergenic regions in the genomes of higher organisms. The availability of the full sequence of large stretches of the Arabidopsis genome allows, for the first time, the use of these data to analyze the presence, nature and variability of microsatellites in this model species. This is the aim of the present article.

2. Materials and methods

2.1. Microsatellites searches

The values of the numbers of the different microsatellites were obtained using the 'findpatterns' program available in the UWGCG software package (Genetics Computer Group, Madison, WI USA). The search was made for all possible combinations of one and two nucleotides and for a minimum length of eight repetitions. The expected values were calculated using the composition of the corresponding sequence to obtain the frequency of each nucleotide repeat in each sequenced BAC.

2.2. Coding sequences predictions

The schemes of gene predictions were obtained running first the GENESCAN program [21,22], then fitting the predictions with the other gene prediction program X-grail [23,24].

2.3. Microsatellite amplification

Microsatellites were amplified by PCR in four different Arabidopsis ecotypes, Columbia, Landsberg erecta, RLD and WS. Seeds of each ecotype were grown at 22°C in a greenhouse with 8 h of light. Genomic DNAs were extracted using the protocol of Dellaporta [25]. PCR amplifications were made using external primers (see sequences below). The PCR amplifications were performed with: 20 ng of genomic DNA, 20 pmol of each primer, 2 mM $MgCl_2$, 0.2 mM of each dNTP, with a total volume of 50 μl. Annealing temperature and extension time were optimized in function of the primer pair and the ecotype, and ranged between 52 and 57°C. The PCR products were tested in 2% agarose gels and were checked for length polymorphism by 6% polyacrylamide gel electrophoresis.

2.4. Primer sequences

The external primers were chosen to be as close as possible to the repeat fragment in order to obtain a product with a length ranging between 62 and 250 bp.

Repeat fragment with 19 T/A:	5′-GGGTTGTTTCAGTCATTCTC-3′ / 5′-GGTGATATTACCTATTGATTTTG-3′.
Repeat fragment with 24 TA/AT:	5′-GTAGACGAAACATATAAGTAG-3′ / 5′-TGCGTTAGTCCATAGCC-3′.
Repeat fragment with 21 T/A:	5′-TGGTTATGGATGGGTTCTTTG-3′ / 5′-GATATCGATCACCTTTGC-3′.
Repeat fragment with 28 T/A:	5′-GAAGGATCATCTGCCTTATTTG-3′ / 5′-GTCTAATCAAGCATCGAGTCTC-3′.
Repeat fragment with 32 TA/AT:	5′-GAGACTCTGGCTGTTAGGTCTTG-3′ / 5′-GCCTGTTCATCTCAAACGC-3′.
Repeat fragment with 45 AT/TA:	5′-GACTCCATGCACATATGTGAAC-3′ / 5′-CCAAGTCTCTCATCCTCACAC-3′.
Repeat fragment with 23 A/T:	5′-CTTAGGATGAACTGGTGTAA-3′ / 5′-GAAGCTCTTAGTGTGATTATCC-3′.
Repeat fragment with 50 T/A:	5′-GAATGATGTCTCCTTTGTAC-3′ /5′-GGACATGGTTAAACTTTTGGTC – 3′.

3. Results

3.1. Relative frequency of different types of microsatellites

Microsatellite sequences have been used in plants as molecular markers [15–17,26,27] and in *A. thaliana* in order to characterize different ecotypes [14,18–20]. The scope of the present work is to use the availability of large stretches of genomic sequence of Arabidopsis in order to analyze the frequency of these repeated sequences and their location in relation to the predicted gene sequences within the genome. The sequences of eight BACs from four different chromosomes (I, II, IV, V) were used, which represent 0.87 Mbp of the Arabidopsis genome. At the beginning of this project no genomic sequences for chromosome III were available. The presence of microsatellites in these sequences was measured. For the purpose of this work a microsatellite was defined as a stretch of at least eight repetitions of a mononucleotide or dinucleotide sequence.

In Tables 1 and 2 the frequency of the different repetitions of one nucleotide (Table 1) and two nucleotides (Table 2) is shown compared with the values that can be expected from the nucleotide composition of the different BACs analyzed. It can be seen that, except for repeats formed only by C or G nucleotides, all the repetitions appear at frequencies higher than those expected in all BACs analyzed. And this is always the case if the whole set of sequences is considered. The total observed

Table 1
Analysis of mononucleotide microsatellites[a]

Name	A/T (observed)	A/T (expected)	C/G (observed)	C/G (expected)
Bac19p19	64	22	1	0
AC000107	71	24	1	0
AC000104	89	24	0	0
AF007271	70	25	0	0
AC00300	103	21	0	0
Z97337	164	46	0	0
Z97344	72	25	0	0
U78721	79	26	0	0
Total	712	213	2	0

[a] Observed values were searched using the findpatterns program with a minimum length of eight nucleotides. Expected values were calculated using the frequency of each nucleotide considering the composition of each BAC sequence.

Table 2
Analysis of dinucleotide microsatellite[a]

Name	TA (observed)	TA (expected)	TC/GA (observed)	TC/GA (expected)	TG /CA (observed)	TG/CA (expected)	CG (observed)	CG (expected)
Bac19p19	17	11	4	2	4	2	0	0
AC000107	29	12	13	1	9	2	0	0
AC000104	38	11	26	2	10	2	0	0
AF007271	20	12	18	2	5	2	0	0
AC00300	25	10	14	2	5	2	0	0
Z97337	53	23	44	4	9	4	0	0
Z97344	43	12	17	2	9	2	0	0
U78721	20	13	28	2	9	2	0	0
total	245	104	166	17	60	18	–	–

[a] Observed values were searched using the findpatterns program with a minimum length of eight nucleotides. Expected values were calculated using the frequency of each nucleotide considering the composition of each BAC sequence.

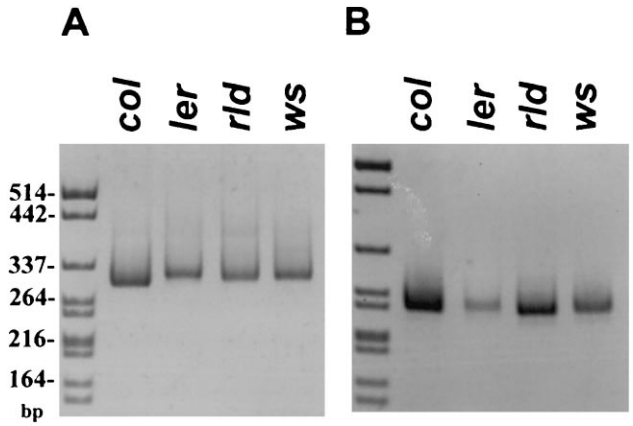

Fig. 1. Polymorphism analyses of mononucleotide microsatellites. Ecotypes used: *Columbia* (*col*), *Landberg erecta* (*ler*), *RLD* (*rld*), *WS* (*ws*). The size of genomic sequence is referred to *Arabidopsis thaliana Columbia* ecotype. Molecular weight marker λ digested with PstI. PCR products checked in a 6% polyacrilamide gel and stained with ethidium bromide solution. (A) 19 T/A nucleotides; (B) 23 T/A nucleotides.

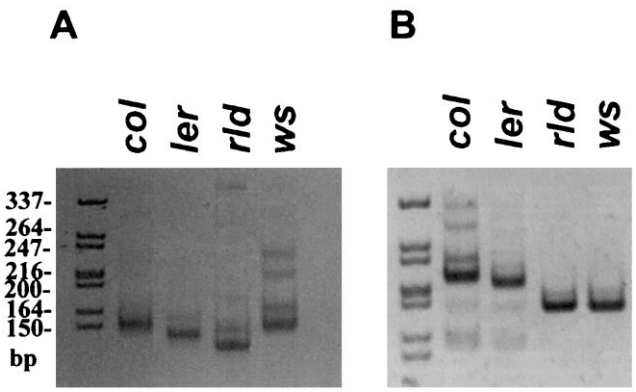

Fig. 2. Polymorphism analyses of dinucleotide microsatellites. Ecotypes used: *Columbia* (*col*), *Landberg erecta* (*ler*), *RLD* (*rld*), *WS* (*ws*). The size of genomic sequence is referred to *Arabidopsis thaliana Columbia* ecotype. Molecular weight marker λ digested with PstI. PCR products checked in a 6% polyacrilamide gel and stained with ethidium bromide solution. (A) 32 TA/AT nucleotides, (B) 50 TA/AT nucleotides.

values are two times higher than expected for A/T and AT/TA microsatellites, but more than two (three to nine) for TC/GA and TG/CA. The relative frequency for the different microsatellite sequences analyzed is very similar to those observed for Brassica species [28] and for soybean [15].

The variability of the different types of microsatellites was also examined using eight different specific SSR, having A/T or AT/TA sequences. Oligonucleotides adjacent to the SSR were designed and the regions were amplified from genomic DNA extracted from different Arabidopsis ecotypes. In Fig. 1 the results for two representative A/T SSR are presented and in Fig. 2 the same analysis for two representative AT/TA SSR. As can be seen from the figures, the A/T SSR shows a light variability within the Arabidopsis ecotypes while in all the examined cases the AT/TA shows a significant change in the length of the repeated sequence. The amplified fragments were sequenced and it was found that the variability did correspond to the microsatellite sequences and not to polymorphisms in the adjacent sequences (sequences not shown). The TC/GA and TG/CA SSR shows a similar variability that A/T SSR (data not shown). This result could be especially important in order to choose the more useful microsatellite AT/TA, when a high level of polymorphism is required.

3.2. Relation of microsatellites and coding sequences

In order to test the usefulness of microsatellites as molecular markers and whether any functional or structural relation could exist between microsatellites and coding sequences, it was of interest to compare their frequency with the frequency of genes that can be predicted in the same sequences. This is the result shown in Table 3 (no results are shown for repetitions of C/G and CG SSR due to their low number). The calculation has been done for each specific region analyzed and for the whole set of regions. The average frequency of predicted coding sequences in these regions is one gene for every 4.6 kb, a figure similar to that found in the 1.9 Mb fragment published [7]. For the four types of microsatellite sequences analyzed, TA/AT and TC/GA SSR have frequencies similar to this one, (one every 3.8 and 6.4 kb) while for T/A SSR the average distance between these repetitive sequences is lower than that of coding sequences (one every 1.2 kb), and the opposite holds for the TG/CA SSR (one every 15.5 kb). With small differences, these values are very similar within the eight regions of the Arabidopsis genome analyzed.

As described above, the number of microsatellites per kb is similar to that of coding sequences in the genomic sequences analyzed. Therefore it was of interest to study where the microsatellites were located in relation to the coding sequences. To do so, the distribution of SSR was calculated

in relation to their proximity or presence within coding sequences. Four regions were distinguished: exons, introns and two intergenic zones depending on whether the microsatellite is located within 0.6 kb of the gene or more than 0.6 kb. The value was chosen taking into account that the average intergenic region in Arabidopsis is 2.4 kb long.

As can be seen in Table 4, in the majority of cases the microsatellite sequences are present in intergenic regions (more than 60% in all cases) while they show the lowest proportions in exons. The only remarkable exception is the TC/GA SSR. In this case, more than 15% of the microsatellites are present in the exon sequences. The comparison of observed values in front of expected, in relation of coding sequences features show that the proportion of simple sequence repeats in exons is similar to the expected values, except for TC/GA (Table 5). The presence of microsatellites in introns seems to be higher than expected with the exception of T/A SSR. The proportion of SSR in intergenic regions (>0.6 kb, <0.6 kb) is significantly higher that expected values with only one exception, the TA/AT SSR has the proportion expected of region within 0.6 kb of the gene.

Table 3
Frequency of each type of microsatellite (one per X Kb)[a]

Name	Chromosome	Length	One predicted gene for X Kb	T/A	TA/AT	TC/GA	TG/CA
Bac19p19	IV	88 345	5.5	1.4	5.2	14.7	22
AC000107	I	95 108	5.2	1.3	3.2	7.3	10.5
AC000104	I	107 526	3.3	1.2	2.8	4.1	10.7
AF007271	V	90 000	4.5	1.3	4.5	5	18
AC00300	II	92 624	4.6	0.9	3.7	6.6	18.5
Z97377	IV	202 861	5	1.2	3.8	4.6	22.5
Z97344	IV	81 850	5.1	1.2	1.9	4.8	9
U78721	II	114 144	4	1.4	5.7	4.1	12.7
Total		872 425	4.6	1.2	3.8	6.4	15.5

[a] Total values observed were divided by the length of each BAC sequence to obtain the frequencies in Kb.

Table 4
Distribution rates of microsatellites in relation to coding sequences[a]

Type	Total	Exons (%)	Introns (%)	<0.6 Kb (%)	>0.6 Kb (%)
T/A	709	1.8	25.9	33.4	38.7
TA/AT	244	2.1	32.6	21.2	44.1
TC/GA	164	15.6	21.0	40.3	22.2
TG/CA	60	6.6	30.0	40.0	23.3

[a] Average of values observed within the regions defined using gene prediction programs.

Table 5
Analyses of microsatellite distribution, in relation to coding sequences[a]

Type	Exons observed	Exons expected	Introns observed	Introns expected	<0.6 observed	<0.6 expected	>0.6 observed	>0.6 expected
T/A	13	14	184	197	237	90	275	90
TA/AT	5	7	79	39	52	44	108	44
TC/GA	25	8	35	5.6	67	6.5	37	6.5
TG/CA	4	7	18	7	24	6.5	14	6.5

[a] The expected values in exons, introns and intergenic regions were calculated using the values of frequencies of each nucleotide suggested by Hebsgaard et al. [33]. The average of the observed values are shown in average in Table 4.

4. Discussion

With the availability of large fragments of genomic sequences it is possible to analyze the distribution of different types of sequences within a given genome. The *A. thaliana* genome, although it has an extremely compact genome, contains all the elements of the repetitive intergenic elements that characterize plant genomes, such as transposons [8] retrotransposons [9–12] and MITEs [13]. For this reason it is used as a starting model to analyze the mechanisms that produce and conserve the different classes of intergenic regions. Here we have analyzed the features of microsatellites or simple sequence repeats (SSR) in this plant species.

Except for the repeats containing only C or G residues, one and two nucleotide repetitions are present at frequencies that seem to be higher than the expected values, taking into account the base composition of each region. This result confirms that SSR are not merely the effect of random concentration of specific nucleotides but the consequence of specific processes producing them. The mechanisms involved in the production of SSR [29,30] are probably responsible for the observed values being higher than expected and it may be concluded that these mechanisms are active in Arabidopsis, as they are in other plant genomes.

The high variability found for TA/AT microsatellite is in agreement with the study about dinucleotide repeats made by Innan et al. [20]. In that study it has been demonstrated that microsatellite loci in *A. thaliana* are highly variable despite the selfing nature of this plant species. ATHCHIB and ATEAT1, which contain TA/AT microsatellite, are two loci with high variability in size (number of repeats). As it can be seen in Fig. 2 the AT/TA microsatellite presented in this study and sequenced shows variation only in the number of repeats.

The distribution of microsatellites in the eight regions examined, corresponding to four different chromosomes, seems relatively homogeneous. No specific concentration has been found in any given region. The same effect has previously been observed for coding regions [7]. Telomeric and centromeric regions of the chromosomes could be the only exception in this relatively homogeneous distribution. In the Arabidopsis genome, although there are regions lacking gene sequences, there do not seem to exist regions specifically rich in coding regions or, regions rich in small repetitive sequences, except in pericentromeric regions [3,4].

We have also found a relationship between the distribution of microsatellites and their location with regard to coding sequences. In exons, where their numbers seem to be similar to those expected, only the TC/GA microsatellite seems to be present at a high number. This may be due to the amino acid residues encoded by these sequences (Ser and Leu), that may be easily accepted in protein sequences, while those encoded, for instance, by T/A (Lys and Phe) may have a greater effect on the protein structure or function.

Finally, the distribution of SSR appears to be high in the three intergenic regions considered: introns, and those proximal and distal to coding regions. If no specific enrichment is found in any of these zones, this may suggest that these sequences do not take part in general mechanisms of control of gene expression, although it can not be excluded that in defined genes this may not be the case. For instance, it has been shown that in Drosophila a protein known as the GAGA-factor interacts with TC/GA sequences and seems to be important for heat shock gene regulation [31,32]. It is possible that the variability of microsatellites may either have no effect on the expression of genes or they may contribute to minor effects in their control. It is clear that in this case they may be a factor in the allelic variability within a given species.

The frequency of SSR is comparable to the frequency of genes. This is also true for the most variable SSR: TA/AT. When we consider the variability of microsatellite sequences found in different Arabidopsis ecotypes, and their frequency comparable to coding regions, it may be concluded that in theory it could be possible, on average, to have one microsatellite marker for every gene in Arabidopsis. This result confirms the high value of SSR as genetic markers and in particular for the AT/TA ones, at least in Arabidopsis.

Acknowledgements

The present work has been funded by grants from the European Union (Arabidopsis BIO4-CT96-0338 'ESSA2') and CICYT (grant BIO97-

1419-CE). The work has been carried out within the framework of Centre de Referència de Biotecnologia de la Generalitat de Catalunya.

References

[1] Goffeau et al., The yeast genome directory, Nature 387 (1997) 1–105.

[2] The *Caenorhabditis elegans* sequencing consortium: genome sequence of the nematode *C. elegans*: a platform for investigating biology, Science 282 (1998) 2012–2022.

[3] K. Mayer, C. Schuller, R. Wambutt, G. Murphy, G. Volckaert, T. Pohl, A. Dusterhoft, W. Stiekema, K.D. Entian, N. Terryn, B. Harris, W. Ansorge, P. Brandt, L. Grivell, M. Rieger, M. Weichselgartner, V. de Simone, B. Obermaier, R. Mache, M. Muller, M. Kreis, M. Delseny, P. Puigdomenech, M. Watson, W.R. McCombie, et al., Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*, Nature 402 (1999) 769–777.

[4] X. Lin, S. Kaul, S. Rounsley, T.P. Shea, M.I. Benito, C.D. Town, C.Y. Fujii, T. Mason, C.L. Bowman, M. Barnstead, T.V. Feldblyum, C.R. Buell, K.A. Ketchum, J. Lee, C.M. Ronning, H.L. Koo, K.S. Moffat, L.A. Cronin, M. Shen, G. Pai, S. Van Aken, L. Umayam, L.J. Tallon, J.E. Gill, J.C. Venter, et al., Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*, Nature 402 (1999) 761–768.

[5] C. Dean, R. Schimdt, Plant genomes: a current molecular description, Annu. Rev. Plant Physiol. Plant Mol. Biol. 46 (1995) 395–418.

[6] P. SanMiguel, A. Tikonov, Y.-k jin, et al., Nested retrotransposons in the intergenic regions of the maize genome, Science 274 (1996) 765–767.

[7] M. Bevan, I. Bancroft, E. Bent, et al., The *Arabidopsis* genome project analyses of 1.9 Mb of contiguous sequence from chromosome IV of *Arabidopsis thaliana*, Nature 391 (1998) 485–488.

[8] Y.-F. Tsay, M.J. Frank, T. Page, C. Dean, N.M. Crawford, Identification of a mobile endogenous transposon in *Arabidopsis thaliana*, Nature 260 (1993) 342–344.

[9] D.F. Voytas, F.M. Ausubel, A copia-like transposable element family in *Arabidopsis thaliana*, Nature 336 (1988) 242–244.

[10] A. Konieczny, D.F. Voytas, M.P. Cummings, F.M. Ausubel, A superfamily of *Arabidopsis thaliana* retrotransposons, Genetics 127 (1991) 801–809.

[11] T. Pélissier, S. Tutois, J.M. Deragon, S. Tourmente, S. Genestier, G. Picard, *Athila*, a new retroelement from *Arabidopsis thaliana*, Plant Mol. Biol. 29 (1995) 441–452.

[12] D.A. Wright, N. Ke, J. Smalle, B.M. Hauge, H.M. Goodman, D.F. Voytas, Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana*, Genetics 142 (1996) 569–578.

[13] E. Casacuberta, J.M. Casacuberta, P. Puigdomenech, A. Monfort, Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterization of the *Emigrant* family of elements, Plant J. 16 (1998) 79–85.

[14] C.J. Bell, J. Ecker, Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*, Genomics 19 (1994) 137–144.

[15] M. Morgante, A.M. Olivieri, PCR-amplified microsatellites as markers in plant genetics, Plant J. 3 (1993) 175–182.

[16] S.R. McCouch, X. Chen, O. Panaud, et al., Microsatellite marker development, mapping and applications in rice genetics and breeding, Plant Mol. Biol. 35 (1997) 89–99.

[17] D. Struss, J. Plieske, The use of microsatellite markers for detection of genetic diversity in barley populations, Theor. Appl. Genet. 97 (1998) 308–315.

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

[19] C.S. Hardtke, J. Müller, T. Berleth, Genetic similarity among *Arabidopsis thaliana* ecotypes by DNA sequence comparison, Plant Mol. Biol. 32 (1996) 915–922.

[20] H. Innan, R. Terauchi, N.T. Miyashit, microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*, Genetics 146 (1997) 1441–1452.

[21] C.B. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.

[22] C.B. Burge, S. Karlin, Finding the genes in genomic DNA, Curr. Opin. Struc. Biol. 8 (1998) 346–354.

[23] Y. Xu, R.I. Mural, M. Shah, E.C. Uberbacher, Recognizing exons in genomic sequences using GRAIL II, in: Genetic Engineering, Principles and methods, Plenum, London, 1994a, pp. 241–253.

[24] Y. Xu, R.I. Mural, E.C. Uberbacher, Constructing gene models from accurately predicted exons: an application of dynamic programming, in: Computer Applications in the Biosciences, 1994b, pp. 613–623.

[25] S.L. Dellaporta, J. Wood, J.B. Hicks, Molecular Biology of Plants: A Laboratory Course Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1984.

[26] M. Chen, J.L. Bennetzen, Sequence composition and organization in the *Sh2/A1* — homologous region of rice, Plant Mol. Biol. 32 (1996) 999–1001.

[27] G.G. Vendramin, B. Ziegenhagen, Characterization and inheritance of polymorphic plastid microsatellites in *Abies*, Genome 40 (1997) 857–864.

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

[29] G.P. Smith, Evolution of repeated DNA sequences by unequal crossover, Science 191 (1976) 528–535.

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

[31] H. Granok, B.A. Leibovitch, C.D. Shaffer, S.C.R. Elgin, Chromatin Ga-ga over GAGA factor, Curr. Biol. 5 (1995) 238–241.

[32] R.C. Wilkins, J.T. Lis, Dynamics of potentiation and activation: GAGA factor and its role in heat shock gene regulation, Nucl. Acids Res. 25 (1997) 3963–3968.
[33] S.M. Hebsgaard, P.G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze, S. Brunak, Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information, Nucl. Acids Res. 24 (1996) 3439–3452.