A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription[1]

Shigeo Fujimori[a,b], Takanori Washio[a,*], Kenichi Higo[c], Yasuhiro Ohtomo[d], Kazuo Murakami[d], Kenichi Matsubara[d], Jun Kawai[e], Piero Carninci[e], Yoshihide Hayashizaki[e], Shoshi Kikuchi[c], Masaru Tomita[a,f]

[a] *Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan*
[b] *Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa 252-8520, Japan*
[c] *National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan*
[d] *Foundation for Advancement of International Science (FAIS), Tsukuba, Ibaraki 305-0062, Japan*
[e] *Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan*
[f] *Department of Environmental Information, Keio University, Fujisawa, Kanagawa 252-8520, Japan*

Received 2 July 2003; accepted 11 September 2003

First published online 6 October 2003

Edited by Robert B. Russell

Abstract A computer-based analysis was conducted to assess the characteristics of microsatellites in transcribed regions of rice and *Arabidopsis*. In addition, two mammals were simultaneously analyzed for a comparative analysis. Our analyses confirmed a novel plant-specific feature in which there is a gradient in microsatellite density along the direction of transcription. It was also confirmed that pyrimidine-rich microsatellites are found intensively near the transcription start sites, specifically in the two plants, but not in the mammals. Our results suggest that microsatellites located at high frequency in the 5′-flanking regions of plant genes can potentially act as factors in regulating gene expression.

Key words: Microsatellite; Tandem repeat; Full-length cDNA; Transcribed region; Transcriptional regulation

1. Introduction

Tandemly repeated sequences are distributed across the whole genome in many eukaryotes. Their unit length often classifies these sequences. Microsatellites or simple sequence repeats, one of the major classes of repeats, are each composed of several short (one to six nucleotides) units. Because of their ubiquitous distribution in the genome, microsatellites have been used as genetic markers [1,2]. Some recent studies have focused on their roles inside cells. Typical examples of these are the studies on triplet repeat disorders in human [3–6]. Triplet repeat disorders are caused by excessive expansions of triplet repeats located near or within coding regions. In particular, studies on the mechanisms of diseases caused by excessive expansions of triplet repeats in intronic or untranslated regions (UTRs) have given us new insight into the functional roles of microsatellites in transcription or translation. The excessive expansion of $(CGG/CCG)_n$ in the 5′-UTR of the *FMR 1* gene leads to hypermethylation in the microsatellites and the adjacent CpG-rich promoter, resulting in the repression of gene expression [7]. Another recent study has also reported the effects of repeats on the transcriptional activity of the gene [8]. In addition, repeated amino acid motifs are also observed in protein structures [9,10]. Therefore, in some cases microsatellites in coding sequences can be translated as amino acid motifs.

Because microsatellites have generally been considered 'junk DNA', useless DNA sequences in intergenic regions of the genome, there have been few reports of the functional features of microsatellites in the transcribed regions of plants. Statistical analysis is one approach to revealing the characteristics of microsatellites. Although many statistical surveys of microsatellites in genome sequences have been conducted in various eukaryotes [9–11], no clear overview of microsatellites in plants has been available. As a result of the development of sequencing technology, plant sequence databases have been rapidly expanded. The growth of these databases gives us many opportunities to analyze in detail the microsatellites of plants [12]. Rice is a model species for monocotyledonous plants, and *Arabidopsis* is a model for dicotyledonous plants. The availability of sufficient amounts of both genomic and cDNA data allows us to make a comparative study between rice and *Arabidopsis*.

In this report, we describe the characteristics of microsatellites within transcribed regions in rice and *Arabidopsis*, and simultaneously in human and mouse for a comparative analysis.

2. Materials and methods

2.1. Sequence data sources

We used 28 469 full-length cDNA sequences [13] for rice (*Oryza sativa* ssp. japonica cv. Nipponbare) from the National Institute of Agrobiological Sciences (NIAS), 13 095 full-length cDNA sequences for *Arabidopsis thaliana* from ftp://ftp.tigr.org/pub/data/a_thaliana/ceres/CeresTigr (as of 2 March 2001) and ftp://pfgweb.gsc.riken.go.jp/rafl/sequence/cDNA_020509_full_length.txt.gz (as of 9 May 2002), 21 245 full-length cDNA sequences for human (*Homo sapiens*) from http://www.ddbj.nig.ac.jp/, and 60 770 full-length cDNA sequences for mouse (*Mus musculus*) from ftp://fantom2.gsc.riken.go.jp/fantom/2.1/fantom2.00.seq.gz (FANTOM DB 2.1) [14]. We retrieved

*Corresponding author. Fax: (81)-466-47 5099.
E-mail address: washy@sfc.keio.ac.jp (T. Washio).

[1] Supplementary data associated with this article can be found at doi:10.1016/S0014-5793(03)01041-X

0014-5793/03/$22.00
doi:10.1016/S0014-5793(03)01041-X

genomic sequences for rice from http://rgp.dna.affrc.go.jp/ (as of 16 October 2002) and http://www.tmri.org/. We downloaded genomic sequences for *Arabidopsis* as of 31 January 2003 from ftp://ftp.ncbi.nih.gov/genomes/A_thaliana/, for mouse from ftp://wolfram.wi.mit.edu/pub/mouse_contigs/MGSC_V3/ (Mouse Genome Sequencing Consortium (MGSC) whole genome mouse assembly, version 3) and for human from ftp://ftp.ensembl.org/pub/human-9.30/data/golden_path/. ORF information corresponding to each cDNA was obtained from http://www.ncbi.nlm.nih.gov/entrez/. Redundancies of cDNA sequences in rice were reduced by Blastn homology searches (ftp://ftp.ncbi.nih.gov/blast/). PolyA tracts that present at the 3′-ends of cDNA sequences were removed to preclude their effects on the results.

2.2. Mapping of cDNA to the genomes

cDNA sequences were mapped to genomic sequences using Blastn homology searches for each species. Exon–intron junctions were detected by SIM4 [15] (http://globin.cse.psu.edu/ftp/dist/sim4/). In this process, cDNA sequences with less than 95% homology with genomic sequences in 95% of the length were eliminated.

2.3. Repeat detection

The repeat finding program was written in C. We set the criteria for detection of tandem repeats in each sequence as follows. The minimum and maximum lengths of the repeat motif were set to 1 bp and 6 bp, respectively. The total size of each repeat was more than 11 bp, and the minimum number of units was three. Only perfect repeats, those without insertion, deletion, or substitution, were counted in this work. Repeat motifs consisting of different frames (e.g. AGC, GCA, CAG) were regarded as the same type of repeat.

3. Results

3.1. Density and composition of microsatellites

Differences in the overall tendencies of microsatellites were surveyed among the four species. The number of various types of microsatellites and their densities in cDNA and intronic regions are given in Tables S1–S4, which can be found only online. The microsatellite densities in cDNA sequences of plants were higher than those of mammals, and those of rice were especially significant. In contrast, it was observed that the densities of microsatellites in intronic regions of plants were lower than those of mammals (Table S1). The compositions of repeat unit lengths were surveyed in both cDNA sequences and intronic regions for the four species (Fig. 1a,b). Differences in the composition of repeat unit length between cDNA sequences and intronic regions were observed within taxa. The most characteristic composition in these differences was the trinucleotide repeat (TNR). In the cDNA sequences of the plants, TNR frequencies were significantly higher than those of other microsatellites, but this was not the case in mammals (Fig. 1a). In the two mammals, frequencies of the TNR in intronic regions were more suppressed than those in cDNA sequences. This tendency corresponds to that observed in a previous study of mammals [16].

We counted microsatellites that were frequently found in cDNA sequences and intronic regions (Table S2). Microsatellites, including polyA/T tracts, $(AAAN/NTTT)_n$, and $(AAAAN/NTTTT)_n$, appeared in the cDNA sequences in all of the species. GC-rich microsatellites and $(AG/CT)_n$ were abundant in cDNA sequences of rice. Some of these microsatellites frequently exist in the genomic region [17]. The most common microsatellite in the cDNA sequences of *Arabidopsis* was $(AAG/CTT)_n$ followed, in order, by $(A/T)_n$, $(AG/CT)_n$, and $(ATC/ATG)_n$. In cDNA sequences of both human and mouse, $(A/T)_n$ and $(AC/GT)_n$ were abundant. The most abundant microsatellites differed in cDNA sequences of the two

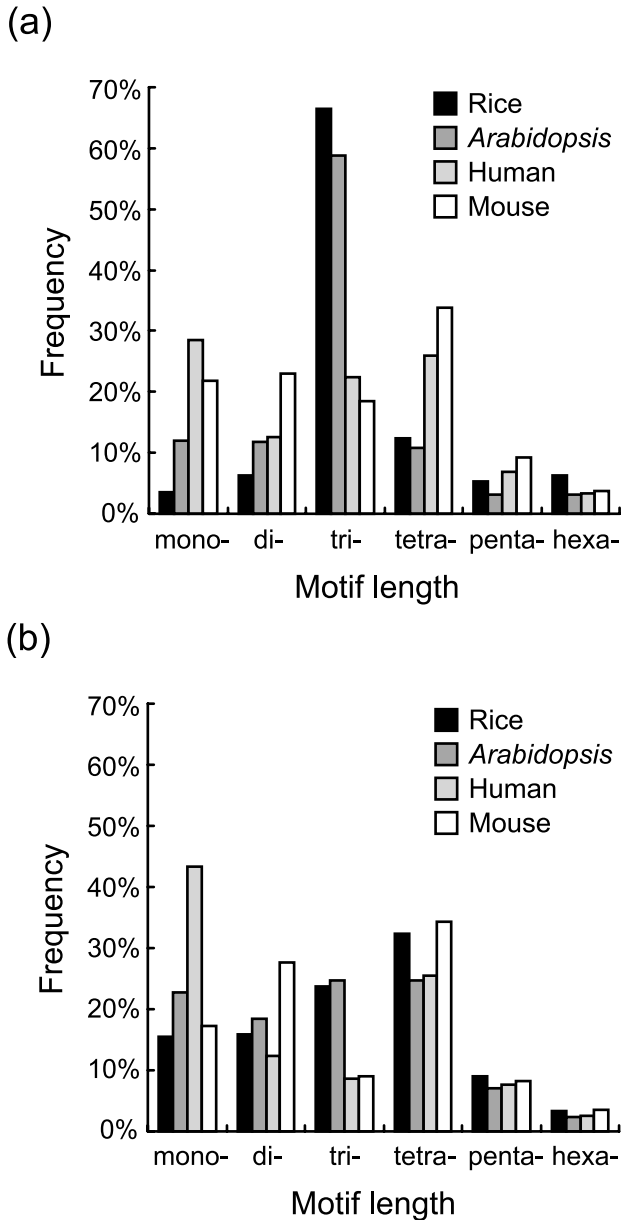

Fig. 1. Frequencies of microsatellites by unit length in rice, *Arabidopsis*, human and mouse. a: Frequency of repeats by unit length in cDNA sequences. b: Frequency of microsatellites by unit length in intronic regions.

plants, and the differences between them were greater than was the case in mammals. Among all four species, $(A/T)_n$ was the most or second most abundant microsatellite in intronic regions. Intriguingly, CG in the intronic regions of all four species was suppressed compared to other dinucleotide repeats. Similar tendencies have been reported in the case of ESTs and GenBank data [16,18,19].

3.2. Strand specificity of microsatellites in cDNA sequences

We compared the numbers of two complementary microsatellites in cDNA sequences. The two-by-one contingency χ^2 test was adapted to detect the presence of any biases. In plants, the frequencies of microsatellites consisting of pyrimidine-rich motifs, which contain CpT dinucleotides, were higher than those of complementary microsatellites with a few

exceptions. For example, the $(CCTCT)_n$ frequency was 135, whereas that of $(AGAGG)_n$ was 58, $(CCT)_n$ 1670 vs. $(AGG)_n$ 1372 and $(CT)_n$ 674 vs. $(AG)_n$ 533 in rice (see Table S3 for details). In *Arabidopsis*, the proportions of $(CT)_n$:$(AG)_n$ and $(CTTT)_n$:$(AAAG)_n$ were 390:226 and 125:77, respectively. We also confirmed a mammal-specific feature: microsatellites containing polyC or polyT (e.g. $(GTTTT)_n$) were positively selected more frequently than were their complements. It is worth noting that the frequencies of $(ATG)_n$, which has the potential to be a translation initiation and termination codon, were remarkably higher than those of their complementary microsatellites in all species.

3.3. Positions of microsatellites in genes and adjacent regions

In order to reveal the positional specificity of microsatellite densities in the transcribed region and the adjacent region (intergenic region), the relative position of each microsatellite in various areas was calculated. We divided the transcribed region, which consist of exons and introns, into the first exon/intron, the second exon/intron, the middle exon/intron, the second last exon/intron, and the last exon/intron. Fig. 2 shows the distributions of microsatellites in the transcribed regions and their up-/downstream regions. The most notable and highest density region was the plants' first exon (Fig. 2a,b). A positional bias toward the transcription start site (TSS) was observed in the plants' first exon, and the magnitude of this bias was predominant in rice. This tendency was also observed in the first/second intron and the second exon. The plants' microsatellite density began increasing gradually from the upstream region, and dramatically rose immediately upstream of the TSS and reached a peak (about 2500–4700 counts/Mb) downstream of the TSS, then subsequently decreased gradu-

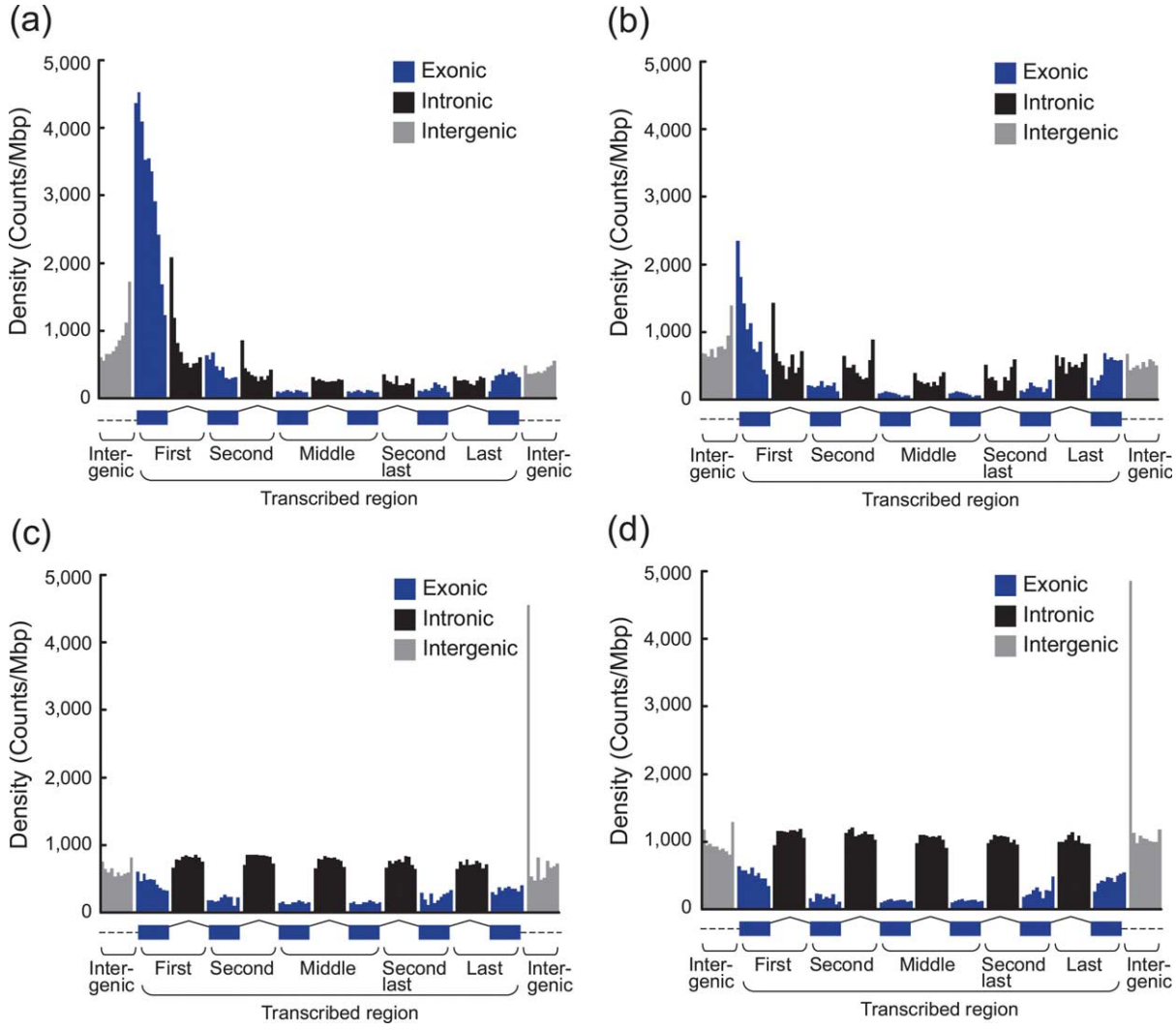

Fig. 2. The microsatellite density in transcribed regions and their up-/downstream regions of (a) rice, (b) *Arabidopsis*, (c) human, and (d) mouse. Figures illustrate the microsatellite density in specific regions, which were defined as follows: transcribed regions, which were interrupted intronic regions, and adjacent regions were determined by mapping full-length cDNAs to the genomic sequences. Exons and introns were classified into five categories: the first exon/intron, the second exon/intron, the middle exon/intron, the second last exon/intron, and the last exon/intron. Only transcribed regions consisting of more than six exons and five introns were considered to avoid the overlap in measurements. In addition, 0.5 kb upstream of TSSs and 0.5 kb downstream of the 3′-end of the transcribed regions were defined as intergenic regions. The numbers of analyzed loci in rice, *Arabidopsis*, human, and mouse were 5125, 2989, 4837, and 6235, respectively. All analyzed sequences possessed a more than 0.5 kb upstream region from the TSS and a more than 0.5 kb downstream region from the 3′-end of the transcribed region. The relative position of the microsatellite in each exon or intron was calculated as the sequence length upstream of the microsatellite divided by the length of each exon or intron excluding the microsatellite. The microsatellite density in the graphs was defined as a count per Mb.

ally along the direction of transcription. In contrast, in transcribed regions mammals were symmetrically distributed (Fig. 2c,d). The microsatellite density in mammalian exons gradually decreased from the upstream region to the first exon and increased from the last exon to the downstream region. However, no significant biases were seen in the introns of mammals. Additionally, no significant bias toward the TSS was detected in *D. melanogaster*; this tendency was similar to that of mammals (data not shown). The occupancies of the protein-coding region in the first and second exons were not particularly significant among four species (44.7%, 56.9%, 38.6% and 41.4% in the first exon, and 80.2%, 91.0%, 74.4% and 81.6% in the second exon for rice, *Arabidopsis*, human and mouse, respectively), indicating that the proportions of the protein-coding region do not reflect the differences of the microsatellite density among these four species. Another conspicuously high-density region is a downstream region flanking the 3′-end of the transcribed region in mammals. Most of the microsatellites detected in this region were polyA tracts. We have no clear explanation for this high-density region, but considering that the human/mouse genome draft sequences include ESTs and mRNAs, it may be an artifact of genome assembly.

As previously stated, we have observed positive selections for specific microsatellites in whole regions of the cDNA sequences of plants. In addition, the strong bias toward the TSS was also confirmed. Based on these observations, microsatellites that presented preferentially near TSSs in the plants were counted (see Table S4 for detail). Microsatellites including CpT dinucleotides such as $(CT)_n$, $(CCT)_n$, $(CTT)_n$, $(CCTT)_n$, $(CGCT)_n$, $(CCTCG)_n$, $(CCTCT)_n$, $(CGTCT)_n$, and $(CTCTT)_n$ were found in both *Arabidopsis* and rice. Some of their complementary microsatellites, such as $(AG)_n$, $(AGCG)_n$, and $(AGAGG)_n$, are also listed in *Arabidopsis* and rice. But GC-rich microsatellites containing CpG dinucleotides ($(CG)_n$, $(AGCG)_n$, $(CCGT)_n$, $(CGCT)_n$, $(AGGCG)_n$, $(CGTCT)_n$, $(CCGCG)_n$, $(CCTCG)_n$, and $(CCGCCT)_n$) were seen only in rice.

4. Discussion

A recent study revealed that microsatellites in the transcribed regions of *Arabidopsis* present more frequently in 5′-UTRs than in coding regions or 3′-UTRs [20]. In the present research, similar results were obtained not only in *Arabidopsis* but also in rice. Furthermore, we confirmed a novel plant-specific feature in which a gradient in the microsatellite density exists along the direction of transcription. This plant-specific feature was observed not only at the level of cDNA; exon, but also at the level of the genomic DNA; exon and intron. But the question arises as to why there exist only microsatellites in the 5′-flanking regions, especially in plants (rice and *Arabidopsis*), under such strong positive pressure. It might be hypothesized that increasing the number of micro-

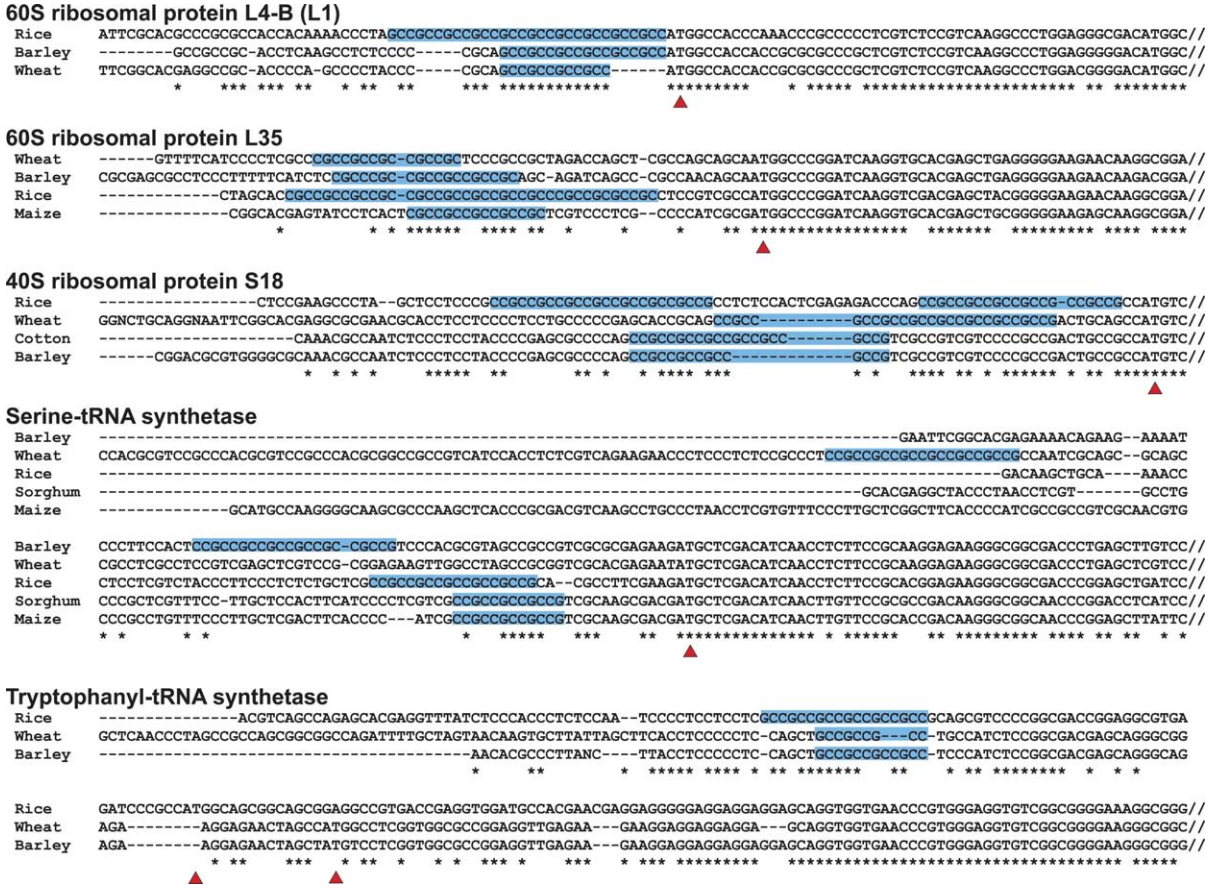

Fig. 3. Conserved microsatellites in orthologous genes. Predicted start codons are indicated by triangles. Sequences except for rice were retrieved from TIGR gene indices (http://www.tigr.org/tdb/tgi/). Orthologous genes were aligned using the CLUSTALW program. DDBJ IDs of rice cDNAs are AK060711 for 60S ribosomal protein L4-B (L1), AK059262 for 60S ribosomal protein L35, AK058881 for 40S ribosomal protein S18, AK069764 for serine-tRNA synthetase, and AK060440 for tryptophanyl-tRNA synthetase.

satellites in regions downstream of the TSS provide some benefits for the biological system, although most such fragile regions are neutral or disadvantageous for these systems.

In some cases, microsatellites in regions downstream of the TSS affect transcriptional and/or translational control. The pyrimidine-rich microsatellites, which contained CpT dinucleotides, were the most notable type of microsatellite in the two plants studied. It is known that increasing the length of the $(CT)_n$ in the promoter region leads to activation of the promoter, while substitution of $(CT)_n$ with purines decreases the activity [21]. It is known that higher eukaryotic genes have regulatory elements such as promoters, enhancers and silencers located in their 5′-UTRs [22,23]. $(AG)_n$ is also detected with significant bias in TSS-flanking regions, although their frequency is less than that of CTs. Both polypyrimidine and polypurine tracts such as $(CT/AG)_n$ and $(CTT/AAG)_n$ can form non-B-DNA, and these unusual DNA structures potentially play some important roles in transcription.

The frequent existence of GC-rich microsatellites such as $(CCG)_n$ only in rice is due to the high G+C content, which was higher than in the other three species analyzed (Table S1). The higher G+C content at the 5′-ends of rice genes was reported from the analyses of genome sequences [24], this tendency corresponds to GC-rich microsatellites being frequently detected in regions downstream of TSSs. $(CCG)_n$ were observed in 5′-UTRs of many translation-related genes in rice, and these microsatellites exist in orthologous genes (Fig. 3). Evolutionarily highly conserved $(CCG)_n$ is evidence of its essential function in these plants. It was found that 46.5% of translation-related genes that were annotated as ribosomal proteins, translation initiation factors, translation elongation factors, or aminoacyl tRNA synthetase, which are indispensable housekeeping genes, contain some microsatellites in their predicted 5′-UTRs. In addition, 54.8% of these genes, having some microsatellites in the 5′-UTR, contain $(CCG)_n$. Some ribosomal protein genes of maize (ZmrpS21A, ZmrpL39, ZmrpP0) possessing $(CCG)_n$ in their 5′-UTRs are believed to be involved in the regulation of fertilization [25].

A more comprehensive and reasonable explanation for the biased distribution of microsatellites is the cytosine methylation in plants, which has been reported frequently in recent years. In plant genomes, cytosines in CpG and CpXpG sites are methylated and maintained by DNA methyltransferase, MET1 and CMT, respectively. Furthermore, non-CpG/CpXpG asymmetric cytosines are also reported to be methylated in plants by DRM [26]. This fact indicates that microsatellites, which were frequently detected in plant genes, come to contain a large number of cytosine methylation target sites such as CCG, CCT, CTT, and other pyrimidine-rich motifs. The interesting point is that transcription elongation is inhibited by cytosine methylation in transcribed regions in fungi and plants [27,28], and it has been reported that cytosine methylation in the 5′-end of coding regions inhibits elongation more effectively in plants [28]. These facts concerning cytosine methylation in transcribed regions may be related to the distribution gradient of microsatellites along the direction of transcription in plants, suggesting that the microsatellite can potentially act as a factor for repressing gene expression in response to the cytosine methylation.

These microsatellites that were found to be positively selected in regions downstream of the TSS, are highly conserved among plants, and commonly exist in housekeeping genes, perhaps have a significant functional meaning for plants. It is expected that a thorough investigation into length polymorphisms and the variations in microsatellite methylation near the TSS within individual species in plants will provide new evolutional and/or functional information with regard to these microsatellites.

Acknowledgements: We would like to thank Keitaro Senda, Koji Numata, Atsushi Sakurai, Hiromi Kochiwa, Haruo Suzuki, Hitomi Ito and members of the Institute for Advanced Biosciences for their helpful discussion during the course of this work. We also thank TMRI for offering genomic sequence data. This work was supported by the Ministry of Agriculture, Forestry, and Fisheries of Japan (Rice Genome Project).

References

[1] Beckmann, J.S. and Soller, M. (1990) Biotechnology (NY) 8, 930–932.
[2] Morgante, M. and Olivieri, A.M. (1993) Plant J. 3, 175–182.
[3] The Huntington's Disease Collaborative Research Group (1993) Cell 72, 971–983.
[4] Ashley Jr., C.T. and Warren, S.T. (1995) Annu. Rev. Genet. 29, 703–728.
[5] Nelson, D.L. (1995) Semin. Cell Biol. 6, 5–11.
[6] Timchenko, L., Monckton, D.G. and Caskey, C.T. (1995) Semin. Cell Biol. 6, 13–19.
[7] Grabczyk, E., Kumari, D. and Usdin, K. (2001) Brain Res. Bull. 56, 367–373.
[8] Iwashita, S., Koyama, K. and Nakamura, Y. (2001) J. Hum. Genet. 46, 717–721.
[9] Groves, M.R. and Barford, D. (1999) Curr. Opin. Struct. Biol. 9, 383–389.
[10] Katti, M.V., Ranjekar, P.K. and Gupta, V.S. (2001) Mol. Biol. Evol. 18, 1161–1167.
[11] Tautz, D. (1994) Curr. Opin. Genet. Dev. 4, 832–837.
[12] Yuan, Q., Quackenbush, J., Sultana, R., Pertea, M., Salzberg, S.L. and Buell, C.R. (2001) Plant Physiol. 125, 1166–1174.
[13] Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, C.J., Ohtsuki, K., Shishiki, T., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuki, Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizuno, K., Yokomizo, S., Niikura, J., Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashidume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konno, H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M. and Hayashizaki, Y. (2003) Science 301, 376–379.
[14] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D.P., Bult, C., Hume, D.A., Quackenbush, J., Schriml, L.M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K.W., Blake, J.A., Bradt, D., Brusic, V., Chothia, C., Corbani, L.E., Cousins, S., Dalla, E., Dragani, T.A., Fletcher, C.F., Forrest, A., Frazer, K.S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I.J., Jarvis, E.D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R.M., King, B.L., Konagaya, A., Kurochkin, I.V., Lee, Y., Lenhard, B., Lyons, P.A., Maglott, D.R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W.J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J.U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J.C., Reed, D.J., Reid, J., Ring, B.Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C.A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M.S., Teasdale, R.D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang,

Y., Watanabe, Y., Wells, C., Wilming, L.G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E.S., Rogers, J., Birney, E. and Hayashizaki, Y. (2002) Nature 420, 563–573.

[15] Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) Genome Res. 8, 967–974.

[16] Toth, G., Gaspari, Z. and Jurka, J. (2000) Genome Res. 10, 967–981.

[17] Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) Science 296, 92–100.

[18] Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. and Waugh, R. (2000) Genetics 156, 847–854.

[19] Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and McCouch, S. (2001) Genome Res. 11, 1441–1452.

[20] Morgante, M., Hanafey, M. and Powell, W. (2002) Nat. Genet. 30, 194–200.

[21] Xu, G. and Goodridge, A.G. (1998) Arch. Biochem. Biophys. 358, 83–91.

[22] Curie, C., Axelos, M., Bardet, C., Atanassova, R., Chaubet, N. and Lescure, B. (1993) Mol. Gen. Genet. 238, 428–436.

[23] Lin, Y., Ince, T.A. and Scotto, K.W. (2001) Biochemistry 40, 12959–12966.

[24] Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L. and Yang, H. (2002) Science 296, 79–92.

[25] Dresselhaus, T., Cordts, S., Heuer, S., Sauter, M., Lorz, H. and Kranz, E. (1999) Mol. Gen. Genet. 261, 416–427.

[26] Wada, Y., Ohya, H., Yamaguchi, Y., Koizumi, N. and Sano, H. (2003) J. Biol. Chem. (in press).

[27] Colot, V. and Rossignol, J.L. (1999) BioEssays 21, 402–411.

[28] Hohn, T., Corsten, S., Rieke, S., Muller, M. and Rothnie, H. (1996) Proc. Natl. Acad. Sci. USA 93, 8334–8339.