13 Overy, S.A. *et al.* (2005) Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. *J. Exp. Bot.* 56, 287–296
14 Weckwerth, W. *et al.* (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7809–7814
15 Kemsley, E.K. (1998) *Discriminant Analysis and Class Modelling of Spectroscopic Data.* J. Wiley

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.08.002

Genome Analysis

In plants, highly expressed genes are the least compact

Xin-Ying Ren[1], Oscar Vorst[1], Mark W.E.J. Fiers[1], Willem J. Stiekema[2] and Jan-Peter Nap[1,2]

[1] Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands
[2] Centre for BioSystems Genomics, 6700 AA Wageningen, The Netherlands

In both the monocot rice and the dicot *Arabidopsis*, highly expressed genes have more and longer introns and a larger primary transcript than genes expressed at a low level: higher expressed genes tend to be less compact than lower expressed genes. In animal genomes, it is the other way round. Although the length differences in plant genes are much smaller than in animals, these findings indicate that plant genes are in this respect different from animal genes. Explanations for the relationship between gene configuration and gene expression in animals might be (or might have been) less important in plants. We speculate that selection, if any, on genome onfiguration has taken a different turn after the divergence of plants and animals.

Introduction

A major issue in relating genome structure to gene expression is the relationship between the relative activity of genes and their position and/or structure. In organisms as diverse as human [1–4] and *Caenorhabditis elegans* [1], highly expressed genes have fewer and shorter introns, shorter coding sequences and shorter intergenic regions [1–5]. This compact nature of highly expressed genes is explained by a selection for either transcriptional efficiency to reduce time and energy [1], a regional mutation bias that positions highly expressed genes in domains more prone to deletions [3] or by a genomic design into open chromatin [4]. We here present a whole genome analysis of the relationship between gene structure and gene expression for two widely diverged plant species, the monocotyledonous plant rice (*Oryza sativa*) and the dicotyledonous plant *Arabidopsis thaliana,* with data from two different expression platforms, massively parallel sequencing signature (MPSS) and microarrays. In both plant genomes, highly expressed genes have more and longer introns and a longer primary transcript. In short, they are less compact than the genes expressed at a low level. This contrasts with the relationship between gene expression and gene structure in human and *C. elegans*, although the absolute differences between plant genes are considerably smaller than for human genes. These findings could suggest that the outcome of selection has been different between animals and plants.

Analysis of plant gene expression in relationship to gene structure

The public domain MPSS expression data for *Arabidopsis* [6] (http://mpss.udel.edu/at/) and rice [7] (http://mpss.udel.edu/rice/) offer good genome-wide expression coverage in a range of different expression libraries and allow easy quantification. To correlate expression data with gene structure, we obtained *Arabidopsis* and rice genome sequences and annotations from The Institute of Genomic Research (TIGR). All genes annotated as either (retro)-transposons or pseudogenes were excluded from the analysis and, in cases of alternative splicing, the longest variant was used in the analyses. We mapped the MPSS expression data to their position in the *Arabidopsis* (TIGR5) and rice (TIGR version 3) genome and all 17 base MPSS tags with a unique position were taken into account. Genes without expression data were not included in the analysis.

To compare the levels of expression of genes in different expression libraries, we sorted the expression values in each library in an ascending order, then divided them into five groups, each containing 20% of the population, and assigned an expression rank from 1 (low expression) to 5 (high expression). Where the cutoff caused equal expression values to be in different rank groups (happening notably with zero expression), the expression values were placed in the lower rank group. For each gene, we averaged the expression ranks over all libraries. This averaged expression rank (rE) indicates the relative expression level of each gene under all conditions analyzed. Alternative methods of expression analysis (see the supplementary material online) give similar results as found for rE. As the rE can be influenced in part by the number of libraries

Corresponding author: Nap, J.-P. (janpeter.nap@wur.nl)
Available online 24 August 2006.

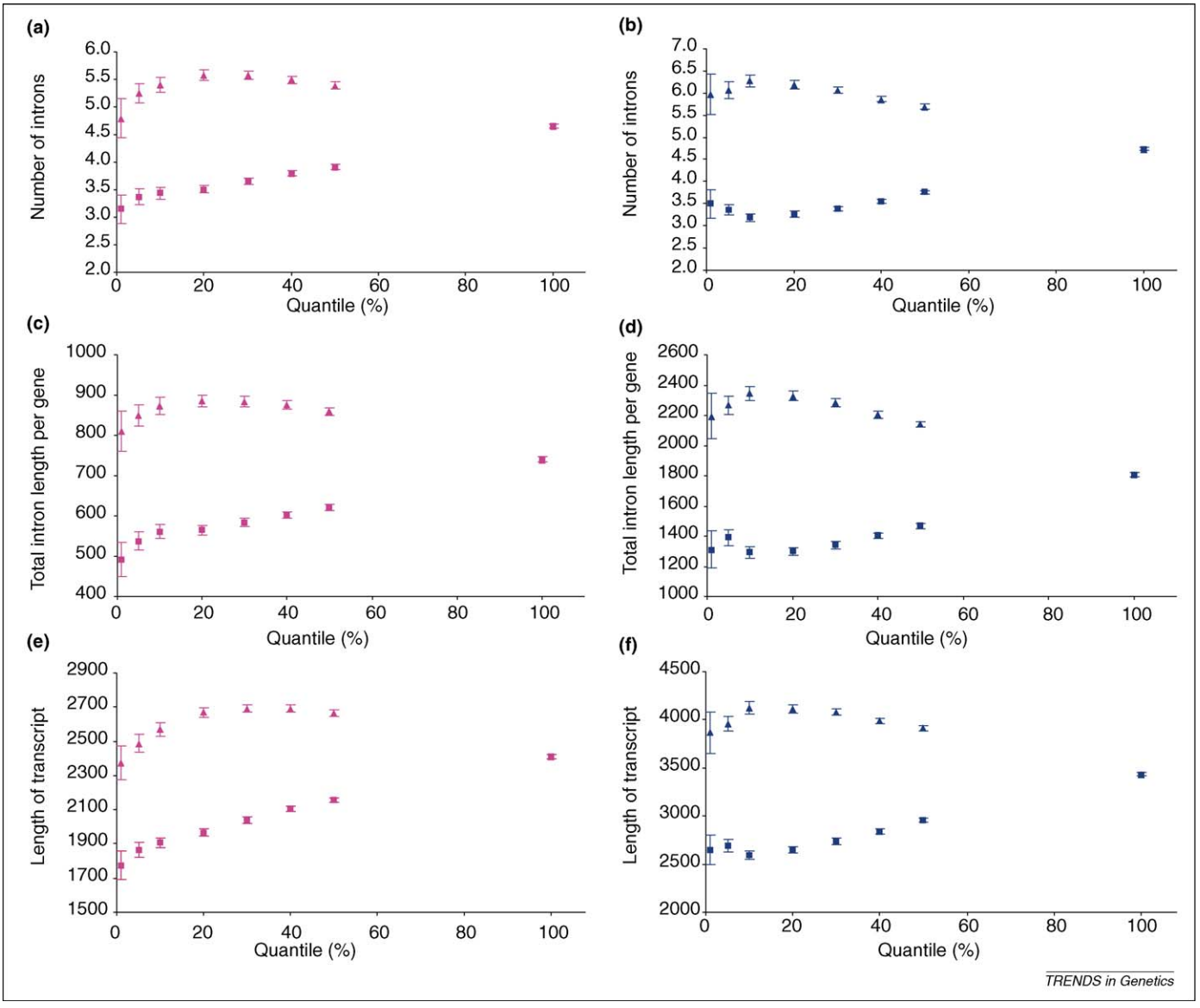

Figure 1. Relationships between the structural characteristics of plant genes and their expression. The panels show a structural parameter ± standard error of the mean versus the average expression rank (rE) in the 1%, 5%, 10%, 20%, 30%, 40% and 50% quantiles from both tails of the ranked population, as well as the value for the whole population (100%). Relationships are shown for three different parameters: the number of introns in **(a)** *Arabidopsis* and **(b)** rice, the total intron length per gene (**(c)** *Arabidopsis*; **(d)** rice) and the length of primary transcript (**(e)** *Arabidopsis*; **(f)** rice). The rice data are in blue and the *Arabidopsis* data in magenta. Higher expressed genes are plotted with a triangle and lower expressed genes with a square. In total, 18 394 *Arabidopsis* genes with expression in 14 different libraries (callus, inflorescence, leaf, root and silique under different experimental conditions and developmental stages) and 21 431 rice genes with expression in 18 different libraries (callus, leaf, root, seed, panicle, meristem, pollen, stem, seedlings, ovary and stigma under different experimental conditions and developmental stages) were included in these analyses. Additional data are available in the supplementary material online.

in which transcription is detected (the so-called breadth of expression [8]), the analyses were also performed with the highest rank of the gene over all libraries, the peak expression rank (pE) (see the supplementary material online).

We correlated the rE parameter with various structural characteristics of each gene, such as the number of introns per gene, the total length of the introns per gene, and others. The rE values were sorted in an ascending order and equal quantiles were taken from the two tails of the population. The top and bottom 1%, 5%, 10%, 20%, 30%, 40% and 50% quantiles were compared for the structural characteristic evaluated to avoid discussions over what level of expression should be considered 'high' and 'low'. For comparison, the data for the whole population (100%) is also given. The quantile comparisons for the parameters are shown in Figure 1 for *Arabidopsis* (Figure 1a,c,e) and for rice (Figure 1b,d,f). The corresponding quantitative data for the 40% quantile, representing 80% of all genes analyzed, are given in Table 1.

The differences between the means and medians (Table 1) indicate that the various parameters are not normally distributed, which is why we used the non-parametric Mann-Whitney test for comparisons. Analyses of the logarithmically transformed gene parameters confirmed the conclusions (see the supplementary material online). Both plant species have the same average number of introns per gene: 4.7 ± 0.04 s.e.m. (standard error of the mean). In both plants, and for each quantile analyzed, the higher expressed genes have significantly ($p < 10^{-4}$) more introns than the lower expressed genes (Figure 1a,b) and the total intron length per gene is significantly ($p < 10^{-4}$) longer (Figure 1c,d) in the higher

Table 1. Relationship between the structural characteristics of genes and their expression characteristics in *Arabidopsis* and rice (40% quantile)[a]

	Arabidopsis			Rice		
Expression level	**High**	**Low**	**All**	**High**	**Low**	**All**
Number of genes included	7358	7358	18 394	8572	8572	21 431
Number of introns	5.5 ± 0.07 (4)	3.8 ± 0.06 (2)	4.7 ± 0.04 (3)	5.9 ± 0.06 (4)	3.6 ± 0.05 (2)	4.7 ± 0.04 (3)
Average intron length per gene (bp)	164 ± 1.8 (133)	140 ± 2.1 (106)	152 ± 1.2 (120)	416 ± 4.4 (333)	359 ± 4.6 (250)	387 ± 2.9 (298)
Total intron length per gene (bp)	876 ± 11 (684)	603 ± 9.0 (367)	740 ± 6.3 (533)	2204 ± 23 (1818)	1405 ± 20 (816)	1805 ± 14 (1368)
Average exon length per gene (bp)	372 ± 5.0 (212)	479 ± 5.7 (293)	430 ± 3.5 (251)	329 ± 3.8 (203)	474 ± 5.4 (298)	405 ± 3.0 (244)
Total coding sequence length per gene (bp)	1396 ± 12 (1173)	1284 ± 9.9 (1113)	1350 ± 6.8 (1152)	1400 ± 11 (1164)	1251 ± 9.4 (1071)	1339 ± 6.6 (1128)
Length of primary transcript (bp)	2692 ± 20 (2313)	2105 ± 17 (1822)	2411 ± 12 (2082)	3988 ± 30 (3420)	2842 ± 25 (2277)	3435 ± 18 (2895)

[a]Values are given as mean ± standard error with median in parentheses. All genes that have a unique 17 b MPSS tag in at least one library and a protein translation in their annotation were taken into account. All parameters for higher expressed genes are significantly ($p < 10^{-4}$) different from lower expressed genes according to the *z* value approximation (http://www.texasoft.com/winkmann.html) of the non-parametric Mann-Whitney test for the comparison of two samples.

expressed genes. In plants, therefore, highly expressed genes have not only more introns, but also longer introns than genes expressed at a low level. The average intron length per gene is therefore also larger for highly expressed genes (Table 1).

Excluding the genes without introns, or removal of up to the first four introns, to correct for the tendency of introns to become smaller towards the 3′ end [9], all confirmed the same relationship between expression and gene characteristics, as did the analysis based on pE (see the supplementary material online). Therefore, the positive correlation between high expression and the number or length of introns is not due to the first introns only, nor can the correlation be an artifact of the averaged ranking (rE) over libraries.

To investigate the potential importance of transcription for expression, we analyzed the correlation between rE and the length of the primary transcript, including all introns and UTR sequences as annotated. For this structural parameter as for those analyzed previously, the highly expressed genes in both *Arabidopsis* (Figure 1e) and rice (Figure 1f) are significantly ($p < 10^{-4}$) longer than the genes expressed at a low level for all quantiles analyzed. None of the variations of the analyses described above affected the results (see the supplementary material online). In current genome annotations, UTRs can be missing from the gene model. Limiting the analyses to all genes with both 5′ and 3′ UTR sequences given in their gene model again confirmed the results (see the supplementary material online).

The length of the coding sequence per gene is larger in higher expressed genes than in lower expressed genes (Table 1), owing to the higher number of introns – and consequently also exons – in higher expressed genes, although the average exon length correlates negatively with expression level. Excluding all genes that have alternative splice forms in their annotation did not affect the results either, ruling out alternative splice variants as explanation (see the supplementary material online). No positive correlation was found between high expression and either short introns or short flanking intergenic regions (see the supplementary material online), whereas in human such a correlation motivated the regional mutation bias model [3] and the genomic design model [4]. Similar analyses on plant expression data from a microarray platform [10] used in another analysis [11] confirmed these results (see the supplementary material online). In both *Arabidopsis* and rice, higher expressed genes have larger primary transcripts with more and longer introns than lower expressed genes. In these plants, higher expressed genes are, in other words, less compact than lower expressed genes.

Are animal genes different from plant genes?

In animals, highly expressed genes have smaller primary transcripts with fewer and smaller introns [1–4]. The more compact nature of highly expressed animal genes is explained by transcriptional efficiency [1], regional mutation bias [3] or genomic design [4]. In plants, our data indicate that highly expressed genes tend to be significantly less compact than lowly expressed genes, although the absolute difference is much smaller than in animals. As highly expressed plant genes are not more compact than plant genes expressed at a low level, there is no need to hypothesize the existence of selection for such compactness in high expression. Neither transcriptional efficiency nor regional mutational bias or genomic design favoring open chromatin seems necessary, or appropriate, to explain the relationship between gene structure and gene expression in *Arabidopsis* and rice. In pollen-expressed genes of *Arabidopsis*, evidence for the efficiency hypothesis has been documented [9]; these results could indicate that expression in the male gametophyte of plants is more prone to selection on intron length than expression in the sporophyte (see the supplementary material online).

An important parameter to consider for the interpretation of these data is the relative length of introns per gene. The average intron length per gene in the human genome is ~5.5 kb [12], which is considerably larger than the average intron length per gene in the plant genomes analyzed here (*Arabidopsis*: 152 b; rice: 387 b; Table 1). Human genes have on the average also more introns (7.7 introns per gene [12]), so the total intron length per gene in human is ~42 kb, compared with 0.74 kb (1.8% of human) for *Arabidopsis* and 1.8 kb (4.3% of human) for rice. By contrast, the total exon length per gene in human (1.49 kb, with 8.7 exons per gene [12]) is of the same order of magnitude as the total exon length per gene in *Arabidopsis* (1.35 kb, with 5.7 exons per gene; Table 1) or rice (1.34 kb, also with 5.7 exons per gene; Table 1). Therefore, in plant genomes, not all gene parameters are smaller than in the human genome, but it is the intron size per gene (either

average or total) that is very different and makes the configuration of plant genes different from animal genes.

More genomes will have to be analyzed to show whether plant introns are under selection to stay relatively small or to become relatively small. The difference in total intron length between higher and lower expressed genes in the 40% quantile class is ~273 b for *Arabidopsis* and 799 b for rice (Table 1). This is between ~11% (*Arabidopsis*) and 23% (rice) of the average length of the primary transcript of the genes (averaged over both classes). When the 10% quantile is considered, these figures go up to 14% for *Arabidopsis* and 31% for rice (data not shown). These differences in total intron length per gene are significant ($p < 10^{-4}$), even when the first four introns are removed (see the supplementary material online).

The hypothesis of selection for efficiency in pollen using serial analysis of gene expression (SAGE) data was based on a (significant) difference of 16 b per intron and ~140 b in total [9]. Therefore, it seems reasonable to assume that the similar small differences here reported have also biological relevance. If so, they point to a different outcome of selection in plants and animals with respect to intron length and expression characteristics. It is feasible that the much larger differences in total intron length in the human genome cause the primary transcripts to be subject to other selective forces than the overall much smaller plant transcripts. Possibly, the difference in intron length between higher and lower expressed genes in plants is not relevant – or much less relevant – for a selection based on length. Introns are involved in a variety of regulatory phenomena, such as RNA stability [13–15], post-transcriptional gene regulation [15–17], nucleosome formation and chromatin organization [5,15,18,19], and/or separating functional domains of proteins [20,21]. Any or a combination of such phenomena could have shaped the structural configuration of higher expressed plant genes in comparison with lower expressed plant genes. Possibly, in plants longer introns with regulatory roles were necessary to achieve high(er) expression. Such a regulatory role of plant introns could have favored additional selective forces to keep plant introns relatively small to reduce the likelihood of interruption by transposons. There could be a preferred intron length for high expression, whereas selection, if any, for low expression would have been different between human and plant.

Highly expressed genes in various yeasts and other unicellular organisms also have longer introns [22]. Although these analyses were based on relatively low numbers of genes, they also suggest a functional role for intron length in gene expression [22]. A recent study on the evolution of intron number in a set of orthologous genes showed that *Arabidopsis* and human retained exceptionally more introns than other eukaryotes [23]. Unfortunately, rice genes were not covered and neither intron length nor expression characteristics was considered. Our results show that it might be worthwhile to include intron length and expression characteristics in further studies on the evolution of eukaryotic gene structure. Whatever selection, if any, has been responsible for more and longer introns in highly expressed plant genes, those selective forces must have taken a different turn after the split of plants and animals, some 1,600 million years ago [24].

Acknowledgements

We thank Cedric Simillion (VIB, Gent, Belgium), Nayelli Marsch-Martinez, Bas te Lintel Hekkert, Harrie Verhoeven and Roeland van Ham (Plant Research International, Wageningen, the Netherlands) for help, comments, suggestions and discussion, and Blake Meyers (University of Delaware, USA) for access to the MPSS data. This research was supported by a program subsidy from the Dutch Organization for Scientific Research (NWO) and by the Centre for Biosystems Genomics (CBSG), which is part of the Dutch Genomics Initiative.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2006.08.008.

References

1 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
2 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
3 Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264
4 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
5 Vinogradov, A.E. (2005) Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res.* 33, 559–563
6 Meyers, B.C. *et al.* (2004) *Arabidopsis* MPSS. An online resource for quantitative expression analysis. *Plant Physiol.* 135, 801–813
7 Nakano, M. *et al.* (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res.* 34, D731–D735
8 Urrutia, A.O. and Hurst, L.D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159, 1191–1199
9 Seoighe, C. *et al.* (2005) Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet* 1, e13
10 Birnbaum, K. *et al.* (2003) A gene expression map of the *Arabidopsis* root. *Science* 302, 1956–1960
11 Ren, X.-Y. *et al.* (2005) Local coexpression domains of two to four genes in the genome of *Arabidopsis*. *Plant Physiol.* 138, 923–934
12 Sakharkar, M.K. *et al.* (2004) Distributions of exons and introns in the human genome. *In Silico Biol.* 4, 387–393
13 Haddrill, P. *et al.* (2005) Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* 6, R67
14 Kirby, D.A. *et al.* (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc. Natl. Acad. Sci. U. S. A.* 92, 9047–9051
15 Shabalina, S. and Spiridonov, N. (2004) The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biol.* 5, 105
16 Carlini, D.B. *et al.* (2001) The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the Drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics* 159, 623–633
17 Liebhaber, S.A. *et al.* (1992) Translation inhibition by an mRNA coding region secondary structure is determined by its proximity to the AUG initiation codon. *J. Mol. Biol.* 226, 609–621
18 Mattick, J.S. and Gagen, M.J. (2001) The evolution of controlled multitasked gene network: the role of introns and other non-coding RNAs in the development of complex organisms. *Mol. Biol. Evol.* 18, 1611–1630
19 Zuckerkandl, E. (1997) Junk DNA and sectorial gene repression. *Gene* 205, 323–343

20 Choi, T. *et al.* (1991) A generic intron increases gene expression in transgenic mice. *Mol. Cell. Biol.* 11, 3070–3074
21 Duester, G. *et al.* (1986) Intron-dependent evolution of the nucleotide-binding domains within alcohol dehydrogenase and related enzymes. *Nucleic Acids Res.* 14, 1931–1941
22 Vinogradov, A.E. (2001) Intron length and codon usage. *J. Mol. Evol.* 52, 2–5
23 Roy, S.W. and Gilbert, W. (2005) Complex early genes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1986–1991
24 Sanderson, M.J. *et al.* (2004) Molecular evidence on plant divergence times. *Am. J. Bot.* 91, 1656–1665

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.08.008

Alu elements within human mRNAs are probable microRNA targets

Neil R. Smalheiser and Vetle I. Torvik

University of Illinois-Chicago, UIC Psychiatric Institute MC912, 1601 W. Taylor Street, Chicago, IL 60612, USA

Recently, we reported that four microRNAs show perfect complementarity with MIR/LINE-2 elements within human mRNAs. This finding raises the question of whether microRNAs might also target other genomic repeats and transposable elements. Here, we demonstrate that almost 30 human microRNAs exhibit typical short-seed complementarity with a specific site within Alu elements that is highly conserved within 3′ untranslated regions of human mRNAs. The results suggest that at least some Alu elements within human mRNAs serve as microRNA targets.

Introduction

The rules governing microRNA–target interactions are under study by many groups (for reviews see Refs [1–4]). It is established that many microRNAs have short, perfect 'seeds' of at least 6–8 bases (with no mismatches or G:U matches) near the 5′ end of the microRNA that are complementary to sequences within 3′ untranslated regions (UTRs) [1–3]. Although not all human microRNA-target interactions follow this consensus pattern [4,5], recent studies suggest that microRNA seeds are often at least 8 bases in length [6] and that bases 2–8 of the 5′ seeds are optimally placed to interact directly with targets [7,8].

Recently, we reported that four mammalian microRNAs show perfect complementarity with the MIR/LINE-2 class of repeat elements, which are present within a large number of human mRNAs and EST transcripts [9]. This finding raises the question of whether microRNAs might also target other genomic repeats and transposable elements in 3′ UTRs. Given that Alu is the most prominent repeat, expressed in >5% of all human 3′ UTRs [10], we asked whether a significant number of other human microRNAs show 5′ seed complementarity against Alu sequences.

We took all 313 human microRNAs listed in the Sanger miRNA Registry (http://microrna.sanger.ac.uk, Version 7.0, June 2005), obtained the 235 unique seed sequences beginning at position 2 and having length 8, and examined their complementarity against 3′ UTR regions of all human mRNAs listed in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/). Regions of exact complementarity (no G:U matches), called 'hits,' were scored according to whether they were within annotated Alu repeats or outside known repeats. Over three quarters of the microRNA seeds had ≤10 hits in Alu elements in the entire set of 3′ UTRs, and almost all showed a greater number of hits in 3′ UTR sequences outside known repeats. However, two seeds were significant outliers (>3 standard deviations above the mean of the distribution) and predominantly hit Alu sequences (Figure 1). These seeds (CAAAGUGC and AAGUGCUG) were highly overlapping and did not contain low-complexity sequence or unusual nucleotide composition.

As a different way of assessing whether hits in Alu sequences are likely to be due to some general property, such as distinctive nucleotide composition, all 235 microRNA seeds were tested for complementarity to the set of Alu sequences in 3′ UTRs subjected to scrambling 100 times (maintaining dinucleotide composition; see Supplementary Material, file 1). Most seeds showed no significant excess of hits in Alu compared with scrambled Alu. Nine seeds showed *z*-scores of 5–50; it is possible that these represent biologically relevant hits in Alu sequences, but they have not been studied further in the present study. Three seeds showed extremely high *z*-scores of 160–180: these represented the two outlier seeds identified in Figure 1 together with another highly overlapping seed sequence (CAAAGUGC, AAAGUGCU and AAGUGCUG). Thus, these microRNA seeds stand apart from all others, according to two independent lines of evidence.

These seeds shared a common 6-mer core sequence (AAGUGC) that was also shared in the 5′ seeds of a set of 27 different human microRNAs (Figure 2). Furthermore, additional sequences in the human microRNA set, extending on either side of the core sequence, also showed complementarity to the Alu consensus sequence (Table 1), and the 9-mer seeds in this set were even more significant outliers when plotted as in Figure 1 (not shown). The 5′ seed of another human microRNA, miR-150, did not share the 6-mer core but nevertheless overlapped with the microRNAs in this set and mapped to an immediately adjacent

Corresponding author: Smalheiser, N.R. (neils@uic.edu)
Available online 17 August 2006.