Cell. Mol. Life Sci. 65 (2008) 201 – 204
1420-682X/08/020201-4
DOI 10.1007/s00018-007-7369-3

Cellular and Molecular Life Sciences

Visions & Reflections (Minireview)

Conserved eukaryotic transposable elements and the evolution of gene regulation

J. Jurka

Genetic Information Research Institute, 1925 Landings Drive, Mountain View, California 94043 (USA),
e-mail: jurka@girinst.org

Received 13 August 2007; received after revision 8 October 2007; accepted 23 October 2007
Online First 20 November 2007

Abstract. Multiple remnants of transposable elements preserved in *cis*-regulatory modules may represent a record of mutations that were critical to the evolution of gene regulation and speciation.

Keywords. Transposable elements, gene regulation, conservation, speciation, evolution.

Transposable elements (TEs) represent a variety of relatively short DNA sequences (typically <25 000 bp) that use different strategies to replicate and insert at different genomic locations. This leads to the generation of multiple, mostly inactive copies of TEs in the host genome, which are called interspersed repetitive DNA. Some classes of TEs are also able to excise and reinsert at new genomic locations. From the moment of their discovery, TEs were known to affect gene expression, and Barbara McClintock literally referred to them as 'controlling elements' [1]. In 1969, Britten and Davidson [2] proposed an elaborate scheme of eukaryotic gene regulation based on the structural involvement of eukaryotic repetitive DNA [3]. These ideas remained dormant for decades, partially due to the fact that our understanding of repetitive DNA and its relationship to TEs was limited. Moreover, evidence linking TEs and repetitive DNA to any biological function in the genome was scarce prior to the era of DNA sequencing. This led to the dominance of the selfish DNA theory, which postulates that the only function of TEs is their own survival within genomes and that they act mostly as parasites [4, 5]. Indeed, the only function of any biological species may be survival within its ecosystem [6, 7]. However, Darwinian competition for survival led to the evolution of complex relationships within ecosystems, from parasitic to symbiotic, that are subjects of intellectually fertile and ultimately useful biological studies. Likewise, 'selfish' TEs have evolved a range of relationships, from parasitic [5] to symbiotic, such as *Drosophila* HeT-A and TART involved in maintaining telomeric regions [8] and domesticated single-copy elements which persist at orthologous loci in different organisms [9, 10]. Some complex evolutionary events, such as the origin of the V(D)J recombination in vertebrates [11], would probably have been unachievable without the involvement of TEs. It is tempting to speculate that only those eukaryotes that did not extinguish the persistent 'trial and error-like' insertions of TEs could undergo a spectacular evolutionary transformation in terms of morphological complexity and diversity [12]. In this context, the contribution of TEs to the evolution of eukaryotic gene regulation is of particular interest.

TEs are ideally suited to supply eukaryotic genomes with regulatory sequences inserted in different permutations and at different genomic locations. The original 'selfish DNA' theory did not preclude occasional recruitment or exaptation of such 'junk DNA', but minimized their overall contributions to genome evolution. However, anecdotal evidence supporting

such contributions continued to accumulate, particularly for mammalian species [13–19]. Subsequently, even older conserved repetitive families were identified in diverse vertebrates [20–22]. One element from the repetitive family shared between lungfish and land vertebrates (LF-SINE) was recruited as a distal enhancer and an ultraconserved exon [20]. Other recent evidence points to overrepresentation of transcription factor binding sites (TFBSs) in TEs [23]. Finally, two recent publications suggest a massive involvement of interspersed repetitive elements in gene regulation [24, 25]. The first paper [24] describes a comparative analysis of non-exonic sequences conserved in diverse placental mammals and reports that a large number of them originated from repetitive elements undergoing purifying selection in mammals. This extends and confirms similar conclusions derived from earlier studies of ancient MIR and *L2* families of mammalian repetitive elements [26]. Furthermore, the paper reports that many of the conserved repeats tend to be located relatively close (within 1 Mb) to genes involved in development and transcription regulation, suggesting massive involvement of repeats in the evolution of gene regulation [24]. The second paper reports studies of interspersed repetitive elements from *Monodelphis domestica*, including 76 new families of interspersed repeats preserved in mammals and chicken, and 7 families preserved in placental mammals and marsupials [25]. Most of these highly conserved repeats were shown to be overrepresented in predicted *cis*-regulatory modules (CRMs) [25]. The significance of these small conserved families for understanding a broader role for TEs in the evolution of gene regulation is the main focus of this essay.

The CRMs are typically <1000 bp long, and harbor one or more TFBSs for different transcription factors [27]. They are relatively well conserved and particularly resistant to indel-type mutations >20 bp long [28]. Predicted CRMs represent only ~2.9 % of the human genomic DNA [27], but contain 6.7 – 70 % of the total number of repeats from different families conserved between mammals and other vertebrates (Fig. 1, yellow circles). Among mammalian-only families previously identified as conserved [25, 29, 30], six are significantly overrepresented in CRMs (Fig. 1, triangles pointing upwards), and two (indicated by asterisks) are close to expected. The remaining 38 mammalian families listed in Figure 1 (triangles pointing downwards) are represented together by ~37340 copies in the genome, of which less than 500 copies (~1.3 %) are present in CRMs, as compared to 1083 (2.9 %) expected by chance. Also, less than 1 % of the copies from large mammalian families that are conserved in the proximity of genes, such as MIR, *L2* or *L3* [24], are present in the CRMs. Large ancient mammalian families could be underrepresented due to a potential 'saturation effect', i.e. even a small percentage of a large family represents a large number of repetitive elements in a limited number of CRMs. However, as indicated above, most small mammalian families are also underrepresented.

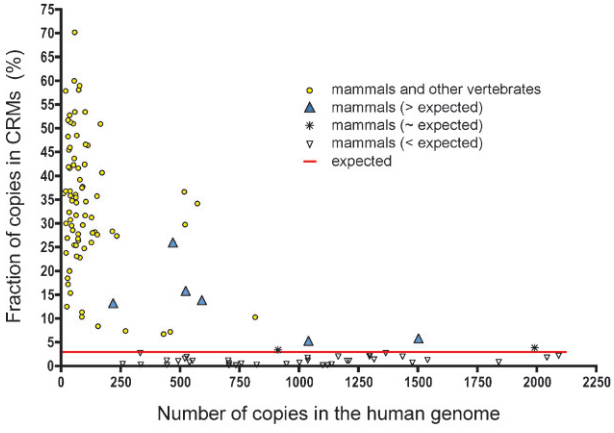

Figure 1. Proportions of repeats in CRMs relative to their total genomic copy numbers. Yellow circles indicate repetitive elements shared between mammals and other vertebrates [20–22, 25]. Triangles pointing upwards indicate the following mammalian families of repeats overrepresented in CRMs: X3_LINE, MER135, MER124, MARE3, MER121, MER128 [25, 29, 37]. Triangles pointing downwards indicate 38 mammalian families underrepresented in CRMs that include *Charlie6*, *Charlie26a*, LTR52, LTR65, LTR80–84(9), LTR86*(4), LTR88a-89, MER91B-C, MER92A-B, MER105–106B(3), MER110–110A, MERX, MamGypLTR1–3(6), MamRep488, MamRep1894, Tigger9b [37]. Overall, they represent 27 LTRs, 10 DNA transposons and one LINE-derived repeat (MERX). Asterisks indicate two putative SINE elements (MARE1 and MARE2), potentially conserved in mammals [25, 37].

The proportions of conserved repeats relative to their total genomic copy numbers are expected to grow over time due to elimination of the non-conserved copies and growing sequence divergence. Based on this premise, even repeats that are originally underrepresented in conserved CRMs may become overrepresented at some future time. However, this model does not easily account for the remarkable overrepresentation of the six mammalian families listed in Figure 1 (blue triangles), which is comparable to that of much older vertebrate repeats [20–22, 24, 25, 30] (Fig. 1, yellow circles). The six families (X3_LINE, MER135, MER124, MARE3, MER121 MER128) are present in ~4350 copies in the human genome, of which 458 copies (10.5 %) are located in CRMs. As discussed above, the majority of moderately abundant mammalian families are represented by ~1.3 % of their all genomic copies in CRMs (compared to 2.9 % expected by chance). Some of these families (e.g. *Gypsy* elements) are as old or older than the six

families. One possibility is that the six overrepresented families contributed more useful regulatory signals than the rest. However, carrying useful signals alone may be insufficient, because CRM modules appear to be resistant against any insertions >20 bp long [28]. Even if the repeat insertions did not target the pre-existing CRMs but introduced new ones, such fixations had to succeed multiple times during a limited lifespan of a few relatively small families of TEs.

To address the difficulties explaining the apparent bursts of TE fixations in CRMs, it has recently been proposed that the overrepresented families might have been abundantly expressed and exapted during episodes of speciation in small populations [30]. During such episodes the evolving species could be more susceptible to TE insertions that could help to create new CRMs and modify the pre-existing ones. Eventually, the regulatory systems would 'rigidify' in successful survivors of the speciation process and become less receptive to new insertions [28, 31]. If so, the overrepresented families may reflect at least a partial 'frozen record' of multiple regulatory changes that took place during speciation.

Regulatory changes have long been viewed as the primary cause of morphological changes occurring during speciation [32–34], and classical scenarios linking speciation to TEs [35] continue to be actively debated [30, 36]. The major difficulty in this debate is the scarcity of empirical evidence, because regulatory changes often lead to unstable organisms that are difficult to create experimentally and are unlikely to be preserved in fossil records. Therefore, if the 'frozen record' hypothesis is true, studies of TEs preserved in CRMs and other conserved regions could prove invaluable for understanding evolutionary changes underlying speciation events. In particular, the preserved remnants of TEs can be helpful to determine the approximate time of their insertion and the affected genes. The ultimate goal would be to correlate the affected genes with species-specific morphological changes.

In conclusion, 'selfish' TEs played multiple roles in eukaryotes, but probably their most important role was supplying evolutionary material and opportunistic bursts of mutations that were critical to the evolution of gene regulation during speciation.

Acknowledgements. I thank Jolanta Walichiewicz and Michael Jurka for help with editing the manuscript. This work was supported by the National Institutes of Health grant 5 P41 LM06252-09.

1 McClintock, B. (1956) Controlling elements and the gene. Cold Spring Harbor Symp. Quant. Biol. 21, 197–216.
2 Britten, R. J. and Davidson, E. H. (1969) Gene regulation for higher cells: a theory. Science 165, 349–357.
3 Britten, R. J. and Kohne, D. E. (1968) Repeated sequences in DNA: hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161, 529–540.
4 Doolittle, W. F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603.
5 Orgel, L. E. and Crick, F. H. C. (1980) Selfish DNA: the ultimate parasite. Nature 284, 604–607.
6 Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. Curr. Opin. Struct. Biol. 8, 333–337.
7 Jurka, J. and Kapitonov, V. V. (1999) Sectorial mutagenesis by transposable elements. Genetica 107, 239–248.
8 Pardue, M. L. and DeBaryshe, P. G. (1999) *Drosophila* telomeres: two transposable elements with important roles in chromosomes. Genetica 107, 189–96.
9 Miller, W. J., McDonald, J. F. and Pinsker, W. (1997) Molecular domestication of mobile elements. Genetica 100, 261–270.
10 Volff, J. N. (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28, 913–22.
11 Kapitonov, V. V. and Jurka, J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 3, e181.
12 Bowen, N. J. and Jordan, I. K. (2002) Transposable elements and the evolution of eukaryotic complexity. Curr. Issues Mol. Biol. 4, 65–76.
13 Brosius, J. (1991) Retroposons – seeds of evolution. Science 251, 753.
14 Britten, R. J. (1996) Cases of ancient mobile element DNA insertions that now affect gene regulation. Mol. Phylogenet. Evol. 5, 13–17.
15 Britten, R. J. (1996) DNA sequence insertion and evolutionary variation in gene regulation. Proc. Natl. Acad. Sci. USA 93, 9374–9377.
16 Brosius, J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238, 115–134.
17 Jordan, I. K., Rogozin, I. B., Glazko, G. V. and Koonin, E. V. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19, 68–72.
18 Medstrand, P., Lagemaat, L. N. v. d. and Mager, D. L. (2002) Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 12, 1483–1495.
19 van de Lagemaat, L. N., Landry, J. R., Mager, D. L. and Medstrand, P. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 19, 530–536.
20 Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J. and Haussler, D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441, 87–90.
21 Nishihara, H., Smit, A. F. and Okada, N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16, 864–874.
22 Xie, X., Kamal, M. and Lander, E. S. (2006) A family of conserved noncoding elements derived from an ancient transposable element. Proc. Natl. Acad. Sci. USA 103, 11659–11664.
23 Thornburg, B. G., Gotea, V. and Makalowski, W. (2006) Transposable elements as a significant source of transcription regulating signals. Gene 365, 104–110.
24 Lowe, C. B., Bejerano, G. and Haussler, D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc. Natl. Acad. Sci. USA 104, 8005–8010.
25 Gentles, A. J., Wakefield, M. J., Kohany, O., Gu, W., Batzer, M. A., Pollock, D. D. and Jurka, J. (2007) Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. Genome Res. 17, 992–1004.

26 Silva, J. C., Shabalina, S. A., Harris, D. G., Spouge, J. L. and Kondrashov, A. S. (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. Genet. Res. 82, 1–18.
27 Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., Coulombe, B. and Robert, F. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. 16, 656–668.
28 Cameron, R. A., Chow, S. H., Berney, K., Chiu, T. Y., Yuan, Q. A., Kramer, A., Helguero, A., Ransick, A., Yun, M. and Davidson, E. H. (2005) An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. Proc. Natl. Acad. Sci. USA 102, 11769–11774.
29 Kamal, M., Xie, X. and Lander, E. S. (2006) A large family of ancient repeat elements in the human genome is under strong selection. Proc. Natl. Acad. Sci. USA 103, 2740–2745.
30 Jurka, J., Kapitonov, V. V., Kohany, O. and Jurka, M. V. (2007) Repetitive sequences in complex genomes: Structure and Evolution. Annu. Rev. Genomics Hum. Genet. 8, 241–259.
31 Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. and Haussler, D. (2004) Ultraconserved elements in the human genome. Science 304, 1321–1325.
32 King, M. C. and Wilson, A. C. (1975) Evolution at two levels in humans and chimpanzees. Science 188, 107–116.
33 Carroll, S. B. (2001) Chance and necessity: the evolution of morphological complexity and diversity. Nature 409, 1102–1109.
34 Carroll, S. B. (2005) Evolution at two levels: on genes and form. PLoS Biol. 3, e245.
35 McClintock, B. (1984) The significance of responses of the genome to challenge. Science 226, 792–801.
36 Lonnig, W. E. and Saedler, H. (2002) Chromosome rearrangements and transposable elements. Annu. Rev. Genet. 36, 389–410.
37 Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.