Journal of Biology

BioMed Central

Minireview

Great exaptations

Kathleen H Burns* and Jef D Boeke†

Addresses: *Department of Pathology, The Johns Hopkins Hospital, 600 North Wolfe Street, Baltimore, MD 21287, USA.
†High Throughput Biology Center, Johns Hopkins University School of Medicine, 733 North Broadway, Baltimore, MD 21205, USA.

Correspondence: Jef D Boeke. Email: jboeke@jhmi.edu

Published: 15 February 2008

Journal of Biology 2008, **7**:5 (doi:10.1186/jbiol66)

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/7/2/5

Abstract

Long interspersed nuclear elements (LINEs) are among the most successful parasitic genetic sequences in higher organisms. Recent work has discovered many instances of LINE incorporation into exons, reminding us of the hazards they pose to genes in their vicinity as well as their potential to be co-opted for the host's purposes.

Picking up LINEs

The genomes of higher organisms are littered with repetitive sequence elements, and evidence for how this came to be is accumulating through a combination of bench work and bioinformatics. Long interspersed nuclear elements (LINEs or L1s) are retrotransposons that have generated integrated DNA copies of themselves through RNA intermediates. Intact L1 RNAs encode two proteins, ORF1p and ORF2p, which together associate with L1 RNA and direct its nuclear importation, reverse transcription to DNA, and integration at a random site within the host genome [1-3]. Sequence analyses show a succession of L1 families in mammalian genomes, each family expanding over time before being relegated to the status of fixed DNA by host-defense mechanisms and perhaps by direct competition from newer, more competent L1s. In humans, this history is traced in order through the inactive L1Pa5, L1Pa4, L1Pa3b and L1Pa2 families and the currently active L1Pa1 family [4,5]. In the mouse, the A, G_F and T_F families are currently active. Although typically regarded as genetic baggage, Martin Vingron and his colleagues (Zemojtel *et al.* [6]) estimate in a recent paper in *BMC Genomics*, that the mouse genome has approaching 5,000 intact active L1s in the diploid nucleus - nearly an order of magnitude more than in our own genome. This is a timely reminder of both the harmful and potentially beneficial effects of L1 elements to cause functional changes in the genome.

Zemojtel *et al.* [6] catalog examples of mouse transcripts that have incorporated portions of LINE sequences embedded in their genomic loci. Although such incorporation into exons - 'exonization' - and the related phenomenon of the co-option of transposon sequences for functional purposes - 'exaptation' - have been described previously (reviewed in [7]), the new work from Vingron's team provides an update on the prevalence of L1 exonization within the mouse transcriptome.

Making sense of strand bias

In most examples of exonization described by the authors, the L1 sequence provides a new start site of transcription (Figure 1) and provides exonic sequence followed by a GU splice donor site. In other examples, however, LINEs provide entire exons or AG splice acceptors with downstream transcript end points. Contributions of single L1-derived splice sites creating chimeric L1/non-L1 exons were found less frequently.

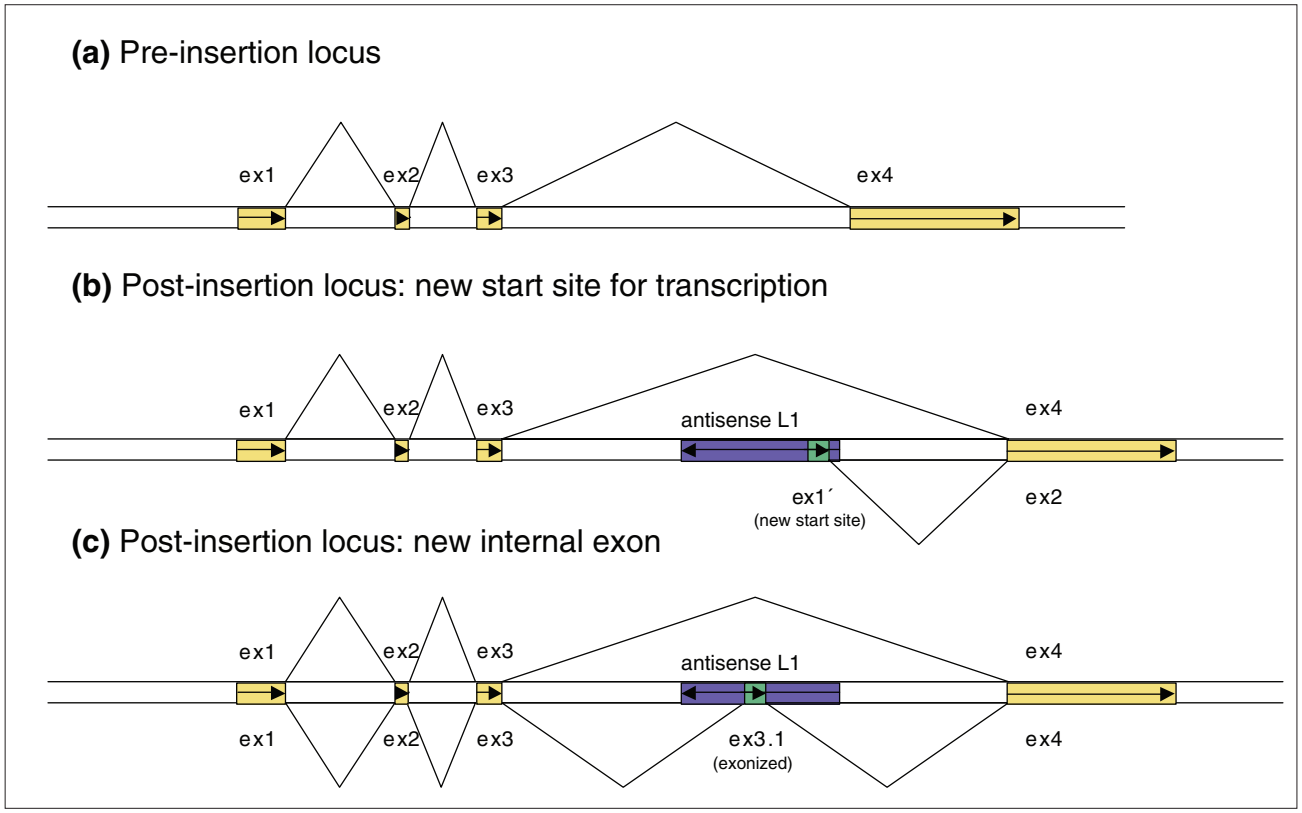

Figure 1
Exonization of mobile genetic elements. A hypothetical gene locus is diagrammed **(a)** before and **(b,c)** after insertion of an L1 in antisense orientation. The L1 insertion allows for two alternative transcripts. (b) In one of these, the L1 sequence provides a novel transcription start site, an alternative first exon (ex1') and a splice donor site. (c) In the second, an entire new exon (ex3.1) with splice sites on either side is derived from the inserted L1. ex, exon.

In 43 out of 52 L1 exonizations annotated by the authors, the L1 sequence is oriented in antisense with respect to the mRNA (Figure 1). This bias is attributable to several factors. First is the apparent potential of the antisense promoter in the ORF1p sequence to initiate transcription, as was observed for 17 cDNAs in this analysis. Second is the presence of a greater number of functional donor and acceptor splice sites on the antisense strand compared with the sense strand, totaling 18 (A-type L1s) and 22 (F-type L1s) functional antisense splice sites versus four functional splice sites in the sense orientation. Third is the repeated use of particular splice sites: SA -154 in the antisense 5′ untranslated region and SD +52 in antisense ORF1p were used in 13 and 16 cDNAs, respectively (SA (splice acceptor) and SD (splice donor) numbers refer to the distance in nucleotides from the beginning of the ORF1p coding sequence in sense-oriented L1Mda2) [8]. Finally, there is a relative prevalence of antisense intragenic L1 insertions, which are found 1.7 times more frequently than sense insertions.

It has been proposed that this last observation is the result of rapid selection against sense-oriented L1s that interrupt gene loci. Disruption of gene function by premature polyadenylation [9] may be one reason an intragenic L1 insertion is selected against, and AAUAAA sites in the sense direction of murine L1s outnumber those in the antisense orientation by a ratio of 9:2 using the M13002 L1Mda2 sequence as a reference [8]. Perhaps more importantly, sense-oriented ORF2 sequence interferes with transcriptional elongation, resulting in profoundly compromised transcript levels [10].

Neomorphic alleles

Disrupted transcription in the absence of true exonization can create hypomorphic and null alleles, but these are only two of many potential sequelae of an intragenic L1 insertion. Exonized L1s also have the important potential to generate neomorphic alleles - allelic changes resulting in

new functions by creating a novel transcription start or end site, or in the case of antisense L1s, both a premature polyadenylation site and a new transcript initiation site (a scenario described as 'gene breaking') [11]. Exonized repetitive sequences can also contribute to protein-coding sequence [12]. Zemojtel *et al.* [6] give the example of guanylate-binding protein 5a (encoded by *Gbp5*), which is predicted to have a 174-amino-acid carboxy-terminal extension derived from an exonized L1. Because of the activity of murine L1s, comparison across inbred mouse strains reveals many polymorphic L1 insertions (David Symer, personal communication). This observation, coupled with the genetic tractability of the mouse, will make it instrumental in our understanding of the functional aspects of endogenous L1s.

Interestingly, cell context is emerging as an important factor in the exonization of transposon sequences, with tissue- and tumor-specific transposon-derived cDNAs being identified [13]. As yet, the functional consequences of such exonizations are not well understood, but recognition that exonization might depend on cellular context lends special significance to the annotations of predicted intronic L1 splice sites by Zemojtel *et al.* [6], which can be seen in their online L1 annotation resource L1Base [14,15]. As similar projects are undertaken in the human genome, experimentalists with an interest in a particular locus in a given cell type or tumor context will have a new tool for predicting alternative L1-containing transcripts, even in the absence of expressed sequence tags indicating splice-site usage.

Neofunctionalized transposons

In addition to changes to an mRNA that result from transposon insertion, pressure created by acquisition of a novel function (neofunctionalization) may act on transposon sequences over evolutionary time. It has been proposed that transposon exonization is a prelude to sequence exaptation, and that alternative splicing of the exon initially allows either its loss or functional co-option [16-18]. Selective divergence of exonized sequence from the L1 consensus is not appreciable in the examples given by Zemojtel *et al.* [6]: that is, an exonized L1 sequence segment is equally similar to a consensus L1 as is the entirety of the L1 containing it (our unpublished analysis). Thus, the status of these relatively young murine L1 sequences may be similar to that of recently exonized Alu sequences in primates in that there is no appreciable sequence co-option [12,16].

Neofunctionalization of more ancient transposon sequences is, however, being increasingly recognized from examinations of mammalian gene structure. Examples include an alternatively spliced exon in poly(rC)-binding protein 2 (*Pcbp2*) that is borrowed from a Silurian period SINE (short interspersed nuclear element) transposon [19]; the recombination-activating gene 1 (*Rag1*), which may be derived from a *Transib* DNA transposon [20]; the primate SET domain and mariner transposase fusion gene (*Setmar*) related to the *mariner*-like *Hsmar1* transposon [21]; and the derivatives of Tf1/Sushi LTR retrotransposons, the *Mar* gene family, which includes an essential gene for mammalian development, paternally expressed gene 10 (*Peg10*) [22]. We expect that, along with identifying already neofunctionalized sequences, studies of recent exonization events will enhance our understanding of mobile element exaptation.

In summary, the recent paper by Vingron and colleagues [6] provides an updated view of the mouse genome and transcriptome with respect to already exonized L1 sequences and intronic L1s with the potential to become part of processed transcripts. This work furthers our understanding of the complex relationships between mammalian genes and the retroelements within them.

References

1. Kolosha VO, Martin SL: ***In vitro* properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition.** *Proc Natl Acad Sci USA* 1997, **94:**10155-10160.
2. Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A: **Reverse transcriptase encoded by a human transposable element.** *Science* 1991, **254:**1808-1810.
3. Feng Q, Moran JV, Kazazian HH Jr, Boeke JD: **Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.** *Cell* 1996, **87:**905-916.
4. Boissinot S, Furano AV: **Adaptive evolution in LINE-1 retrotransposons.** *Mol Biol Evol* 2001, **18:**2186-2194.
5. Boissinot S, Furano AV: **The recent evolution of human L1 retrotransposons.** *Cytogenet Genome Res* 2005, **110:**402-406.
6. Zemojtel T, Penzkofer T, Schultz J, Dandekar T, Badge R, Vingron M: **Exonization of active mouse L1s: a driver of transcriptome evolution?** *BMC Genomics* 2007, **8:**392.
7. Sorek R: **The birth of new exons: mechanisms and evolutionary consequences.** *RNA* 2007, **13:**1603-1608.
8. Loeb DD, Padgett RW, Hardies SC, Shehee WR, Comer MB, Edgell MH, Hutchison CA 3rd: **The sequence of a large L1Md element reveals a tandemly repeated 5' end and several features found in retrotransposons.** *Mol Cell Biol* 1986, **6:**168-182.
9. Perepelitsa-Belancio V, Deininger P: **RNA truncation by premature polyadenylation attenuates human mobile element activity.** *Nat Genet* 2003, **35:**363-366.
10. Han JS, Szak ST, Boeke JD: **Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.** *Nature* 2004, **429:**268-274.
11. Wheelan SJ, Aizawa Y, Han JS, Boeke JD: **Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution.** *Genome Res* 2005, **15:**1073-1078.
12. Gotea V, Makalowski W: **Do transposable elements really contribute to proteomes?** *Trends Genet* 2006, **22:**260-267.
13. Mersch B, Sela N, Ast G, Suhai S, Hotz-Wagenblatt A: **SERpredict: Detection of tissue- or tumor-specific isoforms generated through exonization of transposable elements.** *BMC Genet* 2007, **8:**78.
14. Penzkofer T, Dandekar T, Zemojtel T: **L1Base: from functional annotation to prediction of active LINE-1 elements.** *Nucleic Acids Res* 2005, **33:**D498-D500.
15. **L1Base** [http://l1base.molgen.mpg.de]

16. Krull M, Brosius J, Schmitz J: **Alu-SINE exonization: en route to protein-coding function.** *Mol Biol Evol* 2005, **22:**1702-1711.
17. Sorek R, Ast G, Graur D: **Alu-containing exons are alternatively spliced.** *Genome Res* 2002, **12:**1060-1067.
18. Wu M, Li L, Sun Z: **Transposable element fragments in protein-coding regions and their contributions to human functional proteins.** *Gene* 2007, **401:**165-171.
19. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D: **A distal enhancer and an ultra-conserved exon are derived from a novel retroposon.** *Nature* 2006, **441:**87-90.
20. Kapitonov VV, Jurka J: **RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons.** *PLoS Biol* 2005, **3:**e181.
21. Cordaux R, Udit S, Batzer MA, Feschotte C: **Birth of a chimeric primate gene by capture of the transposase gene from a mobile element.** *Proc Natl Acad Sci USA* 2006, **103:**8101-8106.
22. Brandt J, Veith AM, Volff JN: **A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes.** *Cytogenet Genome Res* 2005, **110:**307-317.