BMC Genomics

BioMed Central

Research article

Open Access

Exonization of active mouse L1s: a driver of transcriptome evolution?

Tomasz Zemojtel*†[1], Tobias Penzkofer†[2], Jörg Schultz[2], Thomas Dandekar[2], Richard Badge[3] and Martin Vingron[1]

Address: [1]Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany, [2]Department of Bioinformatics, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany and [3]Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Email: Tomasz Zemojtel* - zemojtel@molgen.mpg.de; Tobias Penzkofer - tobias.penzkofer@biozentrum.uni-wuerzburg.de; Jörg Schultz - Joerg.Schultz@biozentrum.uni-wuerzburg.de; Thomas Dandekar - Thomas.Dandekar@biozentrum.uni-wuerzburg.de; Richard Badge - rmb19@leicester.ac.uk; Martin Vingron - ingron@molgen.mpg.de

* Corresponding author †Equal contributors

Published: 26 October 2007

BMC Genomics 2007, **8**:392 doi:10.1186/1471-2164-8-392

Received: 18 June 2007
Accepted: 26 October 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/392

Abstract

Background: Long interspersed nuclear elements (LINE-1s, L1s) have been recently implicated in the regulation of mammalian transcriptomes.

Results: Here, we show that members of the three active mouse L1 subfamilies (A, G_F and T_F) contain, in addition to those on their sense strands, conserved functional splice sites on their antisense strands, which trigger multiple exonization events. The latter is particularly intriguing in the light of the strong antisense orientation bias of intronic L1s, implying that the toleration of antisense insertions results in an increased potential for exonization.

Conclusion: In a genome-wide analysis, we have uncovered evidence suggesting that the mobility of the large number of retrotransposition-competent mouse L1s (~2400 potentially active L1s in NCBIm35) has significant potential to shape the mouse transcriptome by continuously generating insertions into transcriptional units.

Background

LINE-1 elements (L1s) are by far the most abundant class of active autonomous transposons in mammalian genomes [1]. It has been established that active, i.e. retrotransposition-competent, L1 elements in the mouse genome [2,3] outnumber by many fold those found in the human genome [3,4]. This is reflected in more than an order of magnitude difference in the percentage of spontaneous mutations due to L1 activity in mice (~2.5%) compared to that in humans (~0.07%) [5]. As a result, based on recent experimental and bioinformatic data, one might speculate that the high insertional activity of mouse L1s could play a significant part in shaping the structure and expression of the mouse transcriptome [6,7]. Importantly, intronic L1 insertions have been shown to influence the expression of their host genes in a wide variety of ways including retardation of transcriptional elongation [6], transcriptional control [8-10], premature polyadenylation [11], and exon skipping [12].

The process by which L1 sequences inserted within introns are recruited into a mRNA, termed exonization, has been primarily studied by analysis of the human transcriptome [13-16] but to date little evidence has been col-

lected for the mouse [15,17]. In this study, we have assessed the exonization potential of currently active mouse L1 elements. Through detailed analysis of members of active L1 families and cDNA-supported L1 exonization events, we show this potential to be much more significant than previously appreciated. This finding, coupled with the much greater activity of L1s in mouse, suggests that not only have these elements dynamically modified the mouse transcriptome in the past, but continue to do so.

Results and discussion

Potentially active LINE-1s in the mouse genome

Using L1Xplorer [3], a suite of automated L1 annotation tools, we identified in the mouse genome sequence (NCBIm35) 2382 potentially active L1 elements (i.e. elements that are full-length and possess intact open reading frames (ORFs) (Fig. 1A, [18]). This contrasts strongly with the 1501 potentially active L1s obtained when analyzing the May 2004 genome release (mm5, build 33) [3] and reflects ongoing finishing of the mouse genome sequence.

Transcription of these open reading frames is driven by the mouse L1 internal promoter, which is built of a variable number of ~200 nt long repeats called monomers. The sequences of active mouse L1s contain related G_F and T_F monomers (F-type) as well as unrelated A-type monomers. Consequently, mouse L1s are classified into subfamilies based on the type of the monomer they harbor [2]. Notably we have established, by bioinformatic data mining, that whereas the number of potentially active L1s belonging to the G_F subfamily agrees with earlier estimates [2], the number of potentially active members of the T_F and A subfamilies is ~1.6-fold higher than previously (Fig. 1A) estimated. As a result, based on our analysis of genomic sequence data, we conclude that perhaps as many as ~4800 (2*2382) potentially active L1s reside in the diploid mouse genome.

Since L1 activity is expected to correlate with L1 expression level and the latter has been shown to correlate with the length of the internal promoter [19], we aimed to characterize the promoters of the 2382 potentially active L1 elements which we had discovered (Fig. 1B). We found that the G_F subfamily has the longest average promoter size (~5.5 monomers), followed by the A (~4.4 monomers) and T_F (~4.1 monomers) families. The annotation of promoter regions is available at [18].

Splice sites in L1s

Clearly any dispersed repeat commonly found in intragenic regions has the potential to be exonized due to the fortuitous occurrence of splicing signals in its sequence [20]. Because of their high number (2382 in NCBIm35) and the concomitantly increased potential for insertional activity, it is of particular interest to establish whether the sequences belonging to the currently active mouse L1 subfamilies (T_F, G_F and A) contain functional splice sites. We mined cDNA databases (RIKEN and NCBI) to discover putative mouse exonization events involving these

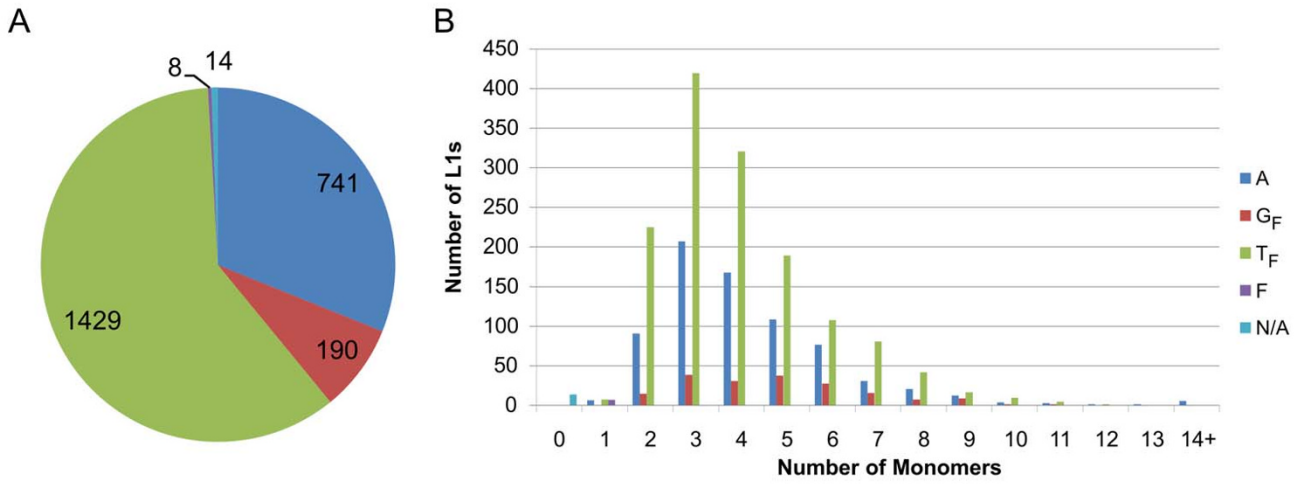

Figure 1

Classification of 2382 potentially active L1 elements residing in the mouse genome sequence (NCBI m35). **A**. Distribution of L1s among subfamilies. A, T_F and G_F correspond to active mouse L1 subfamilies. The small number of L1s that appear related to the inactivated F subfamily are marked with F and those lacking monomers are marked with N/A. **B**. Distributions of the lengths of the internal promoter regions among the three active families. The longest promoter discovered is composed of 28 monomers (here included in the 14+ class) and is a feature of a potentially active L1 element belonging to the A subfamily located on chromosome 2 (80663469-80649072).

sequences (see Materials and Methods). An overview of the different L1 exonization scenarios identified in this study is presented in Fig. 2.

Fig. 3A contains a summary of antisense (upper) and sense (lower) splice sites in L1 sequences, as identified by analysis of cDNA sequences. All of these splice donor sites (SD) and splice acceptor sites (SA) are consistent with the classical GT/AG splice junction motifs [21]. For details of 52 fully annotated examples see: [22].

Of the 52 discovered events 43 involved the exonization of antisense L1 sequences. Similarly, the authors of a very recent study investigated exonization of transposable elements and reported the greater number of the antisense orientation L1 exonization events [23]. The observed greater number of L1 exonizations in the antisense may result from the antisense orientation bias of intronic L1s (see **Antisense splice sites vs. antisense L1 insertional bias** and **Conclusion**).

The most frequently used acceptor and donor splice sites we have identified are SA-154 (located in the antisense strand of 5' UTR) and SD+52 (in the antisense strand of ORF1), supported by 13 and 16 different cDNA transcripts, respectively (Fig. 3A).

A classic example of sense orientation L1 exonization was previously reported when the insertion of a ~1100 bp 3' fragment of a L1 T_F element within an intron of the *beige* gene caused a disease-specific mutation in mouse [17]. Usage of the two SD sites, BG_SD+4694 and BG_SD+4903, identified in the latter study, was also evident in 6 and 3 different cDNAs, respectively, in our data set (see Fig. 3A and online annotation at [22]).

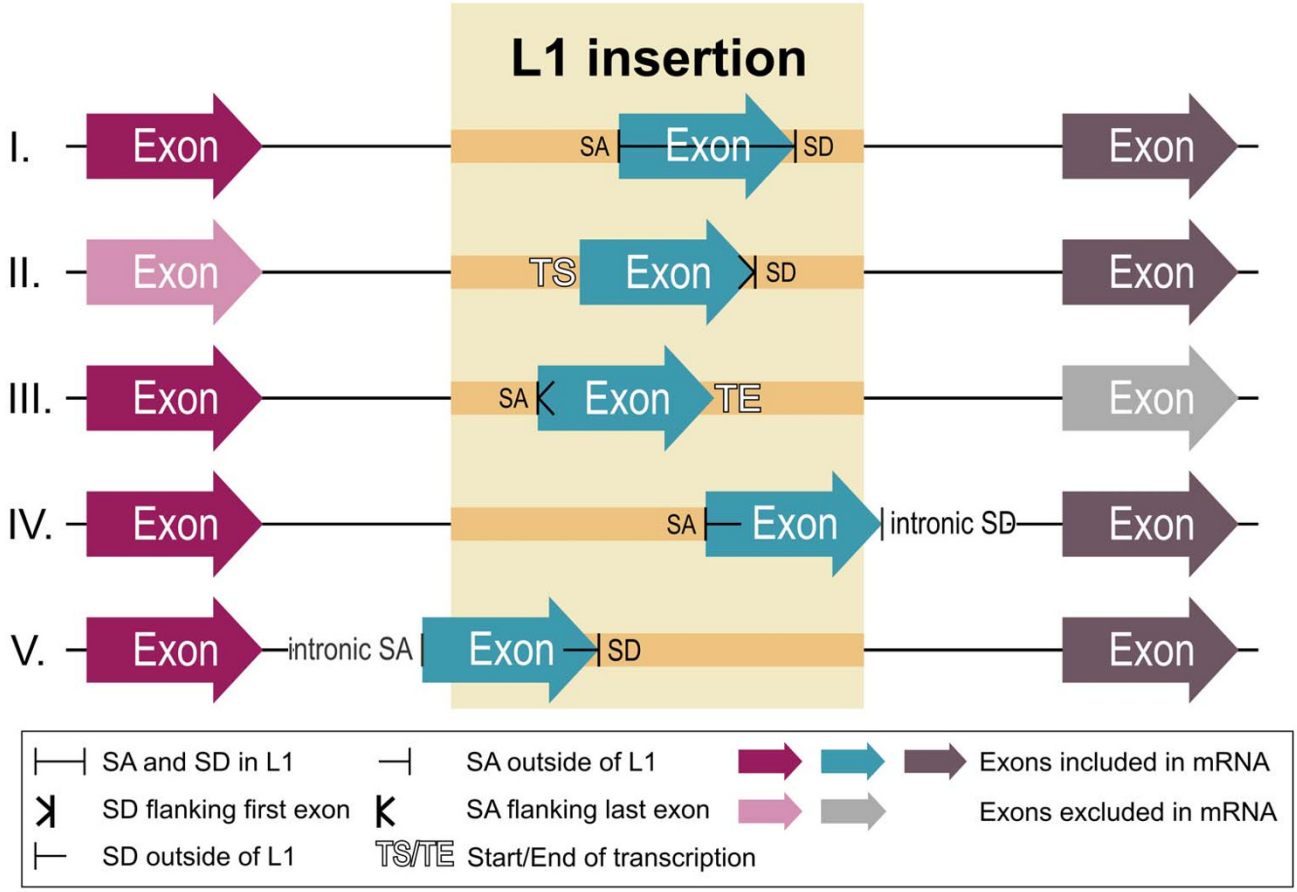

Figure 2
L1 exonization scenarios (I-V) involving sequences belonging to active L1 subfamilies A, T_F, G_F and related inactivated F subfamily, as identified in this study. The scenarios I-V are supported by 16, 26, 14, 6, and 2 exonization events, respectively (see Fig. 3 for details of cDNA sequences). SA: splice acceptor, SD: splice donor. In blue: L1-derived exons; in purple and gray: exons of transcriptional units; in light purple and light gray: exons which are not included in transcript due to L1 insertion.

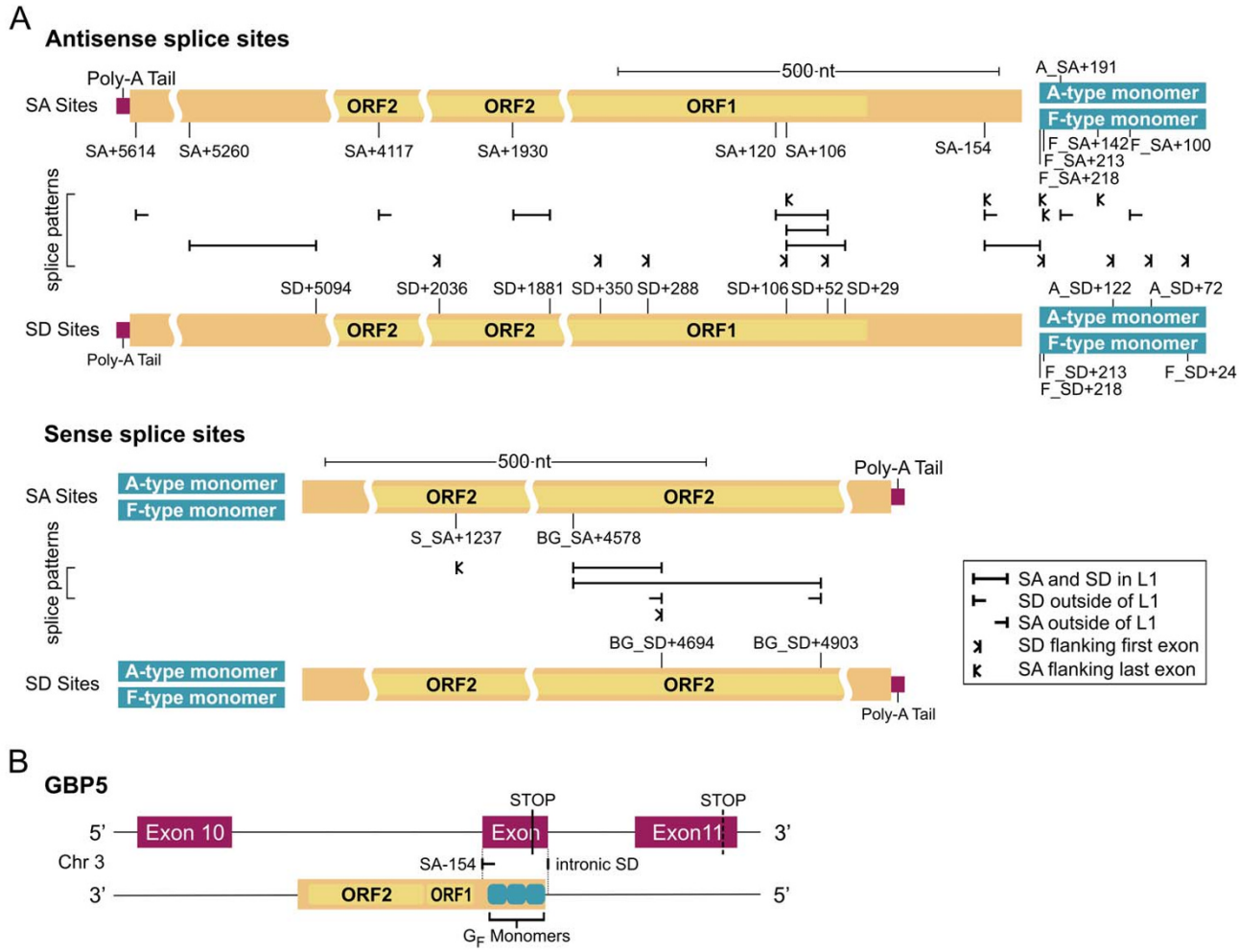

Figure 3

Multiple splice sites are present in antisense and sense L1 sequences (for annotated cDNA examples see [22], for exemplary cDNAs see below). L1Mda2 sequence M13002 was used as a coordinate reference. **A**. Diverse exonization patterns as supported by cDNA evidence. The names of the splice sites incorporate the following information: prefixes of "A_" and "F_" designate sites within A- and F -type (F, T_F, G_F) monomers, respectively; SD: splice donor, SA: splice acceptor; the numbering indicates the position of the base after which the cleavage occurs, relative to the start of L1 ORF1, or relative to the start of alignments for monomers (for the alignments see [2, 40]); prefix of "BG_" designates sites found in L1 inserted within an intron of the *beige* gene, prefix of "S" stands for sense splice sites. The blue boxes mark the monomers making up the internal promoter region. Exemplary cDNAs corresponding to the identified splice sites: F_SA+100: *AK017011*, *BC025138*; F_SA+142: *AI194597*, *AK079058*; F_SA+213: *AK081008*; F_SA+218: *AK015559*; A_SA+191: *AK028243*; SA-154: *BC056642*, *AF487898*, *BQ442932*, *AK039191*, *AK043154*, *BG144807*, *AK044020*, *AK145348*, *BB614554*, *BY733866*, *AK076999*, *AK015267*, *AK035725*; SA+106: *AK080034*, *AY167972*, *BG144807*, *BY733866*, *AK015267*, *AK007310*; SA+120: *AK006905*, SA+1930: *NM_177142*; SA+4117: *AK034994*; SA+5260: *AF529222*; SA+5614: *AK032656*; F_SD+24: *AK035725*; F_SD+213: *AK081008*; F_SD+218: *AK035725*; A_SD+72: *AK077067*, *AK015711*; A_SD+122: *AK032374*, *BC017615*, *AK015277*, *AK006354*; SD+29: *BY733866*; SD+52: *AK080034*, *AY167972*, *BG144807*, *AK006905*, *AK007235*, *AK161293*, *AK132928*, *AK135585*, *BB614554*, *AK016072*, *AK015559*, *AK076999*, *AK015267*, *AK015548*, *AK015778*, *AK015845*; SD+106: *AK015524*; SD+288: *AK076828*, *AK006905*, *AK015267*; SD+350: *AK017011*; SD+1881: *NM_177142*; SD+2036: *NM_177142*; SD+5094: *AF529222*; BG_SA+4578: insertion in *beige* gene (for sequence see the online annotation); S_SA+1237: *AK040102*; BG_SD+4903: *AK031201*, *AK032656*; BG_SD+4694: *AK134759*, *AK038418*, *DV059289*, *AK015958*, *AK034994*, insertion in beige gene. **B**. Insertion of L1 G_F element in the intron of *GBP-5* gene introduced a SA site (SA-154) and resulted in creation of a novel exon coding for the C-terminal and bearing a new stop codon (solid vertical line) (cDNA transcripts GBP-5a, b: gi: *24266664*, *26326418*).

Diverse lengths of exonized L1s

Clearly, truncated and rearranged L1s provide different splice sites. The shortest exonized L1 insertion we have annotated is only 164 bp long (see online annotation of AK032656) and provides the antisense SA+5614 site located at the polyadenylation signal (notably, the polypyrimidine tract here is derived from the polyadenylation signal of L1). As shown in this study (see [22]) the diverse range of lengths of exonized L1 insertions renders numerous possibilities for the combinatorial usage of the splice sites. Conversely, as evidenced by cDNA AK034994, two separate antisense (237 bp) and sense (1117 bp) L1 insertions cooperatively provide SD and SA sites to create a L1-derived exon.

Evidence for an antisense promoter in mouse L1s

As shown in Fig. 3A (scenario II in Fig. 2), we observed an exonization scenario in which first exons are created from intronic antisense L1 sequences (see [22] for examples). In these cases, first exons are spliced to the downstream exons via SD sites present within antisense L1 sequences. This might support the existence of regulatory motifs within mouse antisense L1 sequences that can initiate transcription, as has been reported for the antisense promoter in human L1s [8-10,24]. We identified a region of transcriptional start site within the antisense sequence of ORF1, which is supported by 17 different cDNAs (Fig. 4). This region could be the site of a novel mouse L1 antisense promoter, but experimental studies would be required to confirm this. Transcription of L1-derived first exons, that subsequently capture downstream exons which encode intact protein domains, may constitute a mechanism for the generation of novel functional coding transcripts through "gene breaking", as has been described in humans [24]. For example, in one of our annotated cDNA examples, AK006905, the SD+288 site present in ~6 kbp L1 T_F insertion is used to link L1-derived sequence with downstream exons encoding the C-terminus of DNA directed polymerase iota (gi: 6755273). This mRNA transcript, would encode a 254 amino acid (a.a.) protein containing two ubiquitin-binding motifs (UBM) [25]. Also in this example, two more splice sites originating from the same ~6 kbp L1 T_F. insertion, SA+120 and SD+52 generate an L1-derived exon. A similar exonization pattern, leading to first exon generation, is also evident in the transcript AK015267 where another ~6 kbp L1G_F insertion provides SD+288, SA+106 and SD+52 sites, as well as SA-154. In the latter case SA+120 is substituted for the SA+106 site located nearby (see online annotation).

Coding potential of antisense L1-derived exons

Provocatively it has been proposed that, in general, repeat exonization via alternative splicing may constitute a vehicle for the exaption of repeat sequences into novel functions [20,26,27]. In line with this, a recent experimental study demonstrated that arbitrary sequences can evolve towards functionality when fused with other pre-existing protein modules [28].

Our analyses of cDNAs have revealed that exonized antisense L1 sequences have the potential to code for parts of ORFs. For example, the alternative transcripts of the GBP-5 gene (gi: 24266664, 26326418) contain an L1-derived exon which contains sequences from three G_F monomers. The antisense sequences of each of the three monomers can be translated into peptides which are ~60 a.a. in length yielding a novel 174 a.a. long C-terminus of *GBP-5* protein (see Fig. 3B and also [22]). Although, clearly, the resulting protein variant is mouse-specific, it was noted that the alternative C-terminus variant of GBP-5 exists in humans (AF328727) and that both mouse and human

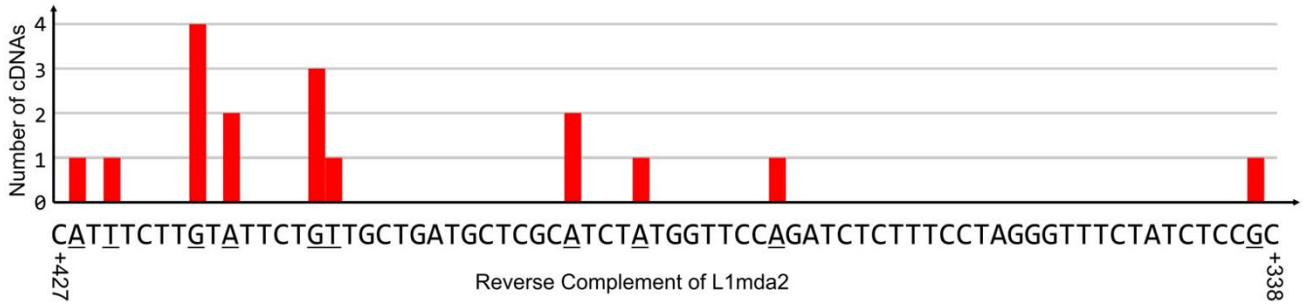

Figure 4

Distribution of transcriptional start sites within a region of antisense sequence of the ORF1. The coordinates are with respect to the start of the ORF1 in L1mda2 sequence (M13002). The Y axis represents the number of cDNAs supporting each transcriptional start site location. In total, 17 cDNAs support the TSs in this region: AK017011, AK076828, AK006905, AK007235, AK015524, AK161293, AK132928, AK135585, AK077067, AK016072, AK015559, AK076999, AK015267, AK015548, AK015778, AK015845, AK015266.

variants lack the C-terminal CaaX isoprenylation motif [29,30], which might be of physiological importance.

Exonization potential of active L1s

To examine the exonization potential of active L1 elements, we analyzed the residues at the positions corresponding to the cDNA-identified splice sites in a set of 2382 sequences of potentially active L1s. For this, we utilized the L1Xplorer annotation pipeline and created a set of customized modules (see Materials and Methods for details). Since all of the identified splice sites were classical GT-AG sites, we defined the presence of these nucleotides as a criterion for a functional splice site. In a total of 2382 full-length intact LINE-1 elements 45471 donor and 47848 acceptor splice sites were annotated (see Fig. 5B for summary on conservation of antisense splice sites, Additional File 1, for details on conservation of sense splice sites and for splice site annotation in L1 sequences see [18]).

We rendered sequence logos of the sequences corresponding to the identified acceptor and donor motifs in L1s (Additional File 2). In analyzing logos of the acceptor sites we observed polypyrimidine tracts which are typical of consensus splice sites in mouse genes [31] (see Fig. 5A). The presence of these motifs supports the putative functionality of annotated acceptor sites in active L1s.

As illustrated for the case of SA-154 site in Fig. 5A, family-specific patterns of splice motif conservations are observed in potentially active L1 sequences. More cases,

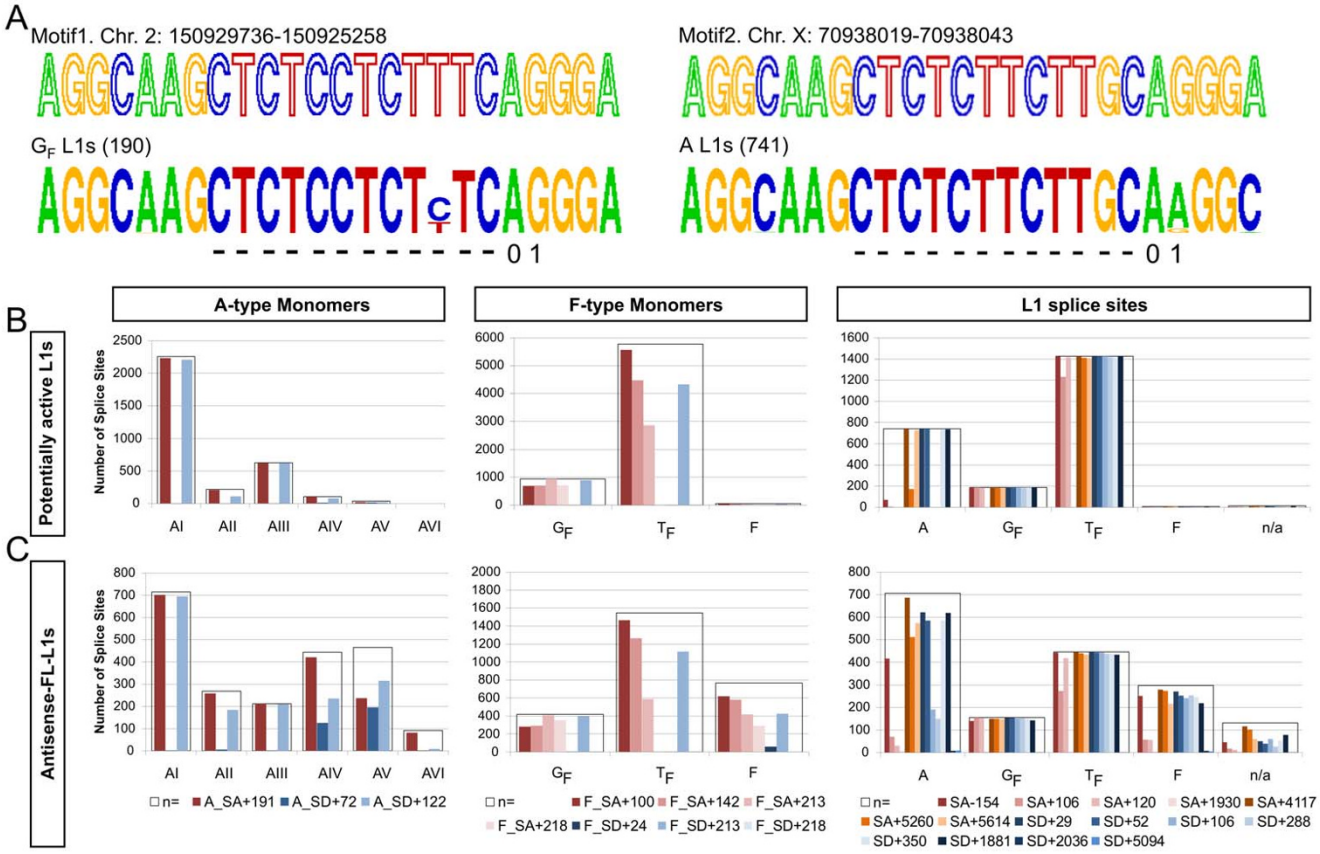

Figure 5

Annotation of antisense splice sites in different subfamilies of putatively active L1s and antisense intronic FL L1 insertions. **A**. For illustration the polypyrimidine tracts for the site SA-154 are shown. Here, the splice donor AG motif is present only in a small fraction of full-length intact elements belonging to the A subfamily (741), whereas it is intact in G_F subfamily (190). "01" marks the location of AG splice acceptor motif; "-" designates the position of the polypyrimidine tract. Exemplary Motif1 and Motif2 sequences, containing the functional SA-154 splice site, are evidenced by mapping of cDNAs (AK145348, BG144807, respectively) to the corresponding genomic locations containing L1s (NCBIm35). **B**. Conservation of antisense GT/AG splice motifs in potentially active L1s. **C**. Conservation of antisense GT/AG splice motifs in antisense intronic FL L1 insertions. "n = " indicates the number of annotated L1s/monomers. Legend: cDNA-identified splice sites.

such as that of SA+106, SD+288 sites that are AG/GT intact only in F-type mouse L1s, are highlighted in Fig. 5B.

We also found three splice sites SD+5094, SD+2036 and SA+1930 that are not intact in any subfamily of the potentially active L1 sequences and have been generated by single nucleotide mutations leading to functional "GT" and "AG" motifs (GG->GT, CT->GT and AT->AG respectively, see Additional File 2). The SA+1930 acceptor splice site contains the 11 bp-long polypyrimidine tract which is present in all sequences of potentially active L1s (Additional File 2). Thus the single nucleotide T->G mutation could activate this cryptic acceptor splice site in potentially active L1 sequences.

Splice sites in intronic full-length L1 sequences

The amount of L1 sequence residing in mouse introns is ~25% of the total genomic L1 content and as much as 8% of all intronic sequences are L1-derived nucleotides (based on our cumulative analysis of annotated transcriptional units found in Refseq, Known Genes and Ensembl from the UCSC mm7 Dataset). This sequence is comprised mostly of truncated L1 sequences: according to RepeatMasker annotation of NCBIm35 (UCSC mm7) more than 92% of intronic L1 sequences are less than 1000 nt long. As shown above, even these short sequences can be subject to exonization. It is expected, however, that intronic full-length (FL) L1 insertions have a much higher exonization potential since they contain multiple splice sites and, for example, only FL L1s will include 5' promoter regions containing splice sites.

We identified 1739 antisense and 1014 sense intronic FL (greater than 5 kbp in length) L1 insertions. As revealed by our L1Xplorer annotation, these belong in large part (75% in sense and 78% in antisense) to the active A, G_F and T_F families but some (17% in sense and 16% in antisense) belong to the inactivated F subfamily (Additional File 3). Similar to potentially active L1s, splice sites are largely intact in intronic FL L1 insertions (Fig. 5C, Additional File 1, Additional File 2). Multiple cDNAs confirm exonization of antisense sequences of FL elements (see [22]). In particular, antisense sequences of two intronic potentially active L1 T_F elements are exonized via the SD+52 site to create first exons in cDNAs AK132928 and AK007235.

By analysing three groups of gene annotations (cDNA and corresponding DNA), we identified as many as 1259 Ensembl Genes, 1436 UCSC Known Genes and 858 RefSeq Genes with at least one FL antisense intronic L1 insertion and 718 Ensembl Genes, 797 UCSC Known Genes and 464 RefSeq Genes that contained at least one intronic FL sense L1 insertion. Hence, the prediction of potential splice sites within FL intronic L1s may be an important consideration for researchers studying transcripts of particular genes. We have added the annotation of intronic FL L1 elements to L1Base, which is available at [18].

Antisense splice sites vs. antisense L1 insertional bias

The existence of antisense splice sites is highly intriguing, particularly in the light of ~2 fold antisense orientation bias of L1s located in introns of transcriptional units [32]. This orientation bias is especially evident when comparing regions immediately flanking transcriptional start sites (TSS) and transcriptional end sites (TES) (Additional File 4).

However, this global picture, which is based on all L1 insertions, does not provide information on whether the bias results from processes acting on a long time scales or rather is already reflected in young L1 insertions. To gain insight into this issue we specifically looked at young intronic FL L1 insertions.

The ratio of antisense to sense FL intronic insertions is ~1.7 and chi-square testing established that this ratio is significantly different from random insertion orientation model (chi^2: $p < 0.0001$), where either orientation is equally likely.

We set out to investigate, if this insertional bias is still evident among younger intronic insertions. For this analysis we utilized the set of the potentially active L1s (i.e. full-length elements with intact ORFs) that were inserted within introns. This is more stringent than analysing FL insertions with disrupted ORFs, since elements with intact ORFs are likely to be younger due to the L1 proteins' *cis*-preference towards their encoding RNA [33]. ~28% (657) of putatively active sequences reside in introns (393 in antisense, 270 in sense, 6 both in sense and antisense, ratio ~1.46). The chi-square test indicated that this is again highly significantly different from a random insertion model (chi^2: $p < 0.0001$). This result suggests that if insertion orientation is random for *de novo* insertions, and the observed bias towards antisense insertion occurs due to selection against sense insertions, this selection process is rapid.

The family distribution for the antisense intronic L1 insertions (Additional File 3A) is very similar to the FL L1s in intergenic regions (Additional File 3B). This is what one would expect to see under the assumption that the sequences of all L1 families equally impact the genes they insert into (i.e. there is no negative selection against any particular subfamily). However, we did observe a difference in the distribution of T_F and A subfamilies between intronic sense FL insertions (Additional File 3B) and intragenic FL insertions (chi^2: $p < 0.0001$). One might argue here that negative selection appears to have acted specifi-

cally on the intronic sense L1 insertions belonging to the A subfamily, but it is not clear why this might be.

Conclusion

We have shown that active mouse L1 elements contain functional splice sites within their antisense sequences using evidence from exonization events in mouse cDNA libraries. This is especially interesting in the light of the antisense insertional bias of *de novo* L1 insertions. A recent experimental study addressed the molecular nature of this phenomenon by showing that sense insertions of mouse L1 T_F element reduce transcript levels and impair their structure. By contrast antisense intronic insertions have little or no effect on transcript elongation and abundance [32]. These data suggest that the apparently benign nature of antisense intronic insertions and the presence of functional antisense splice sites can lead to the frequent exonization of L1 sequences, as we have observed in our dataset. Further, the conservation of these splices sites in L1 families known to be currently active in the mouse genome strongly implies that generation of L1-exonized transcripts is ongoing, and thus represents a driver of transcriptome evolution. However, it has to be said that this picture is complicated by many factors, particularly since the combination of intronic environment, the size and exact structure of the L1 insertion can impact upon its exonization potential, resulting in gene and insertion specific patterns of exonization. With this caveat, the evidence for inclusion of L1 sequences in transcripts and the high activity of the many LINE-1s in the mouse genome, suggests that their integration into introns has significant ongoing potential to shape the structure of the transcriptome, and ultimately, the proteome [26], in the course of evolution.

Methods

Identification of splice sites

We screened mouse cDNA databases (FANTOM3 [34] and NCBI) with RepeatMasker to identify cDNAs containing L1 sequences. The splice sites within L1 sequences were identified using a combination of the following tools. We used BLAT [35] to identify the genomic localization of the cDNAs on the mouse genome (NCBIm35). The genomic regions were extracted either from ENSEMBL [36] or the NCBI Nucleotide Database [37]. SPLIGN [38] was used to split the cDNAs into exons. RepeatMasker was used to identify the repeats in genomic regions corresponding to the mapped cDNAs. Family classifications of L1s were carried out with RepeatMasker and a customized version of the monomer search module of L1Xplorer, which uses Matcher from the EMBOSS package [39] and template sequences for A- and F-type monomers [2,40]. This allowed us to specifically identify the candidate splice sites that occur in sequences belonging to active L1 families. Because of their sequence similarity to the active G_F and T_F subfamilies, we also included in our analysis cases of exonization of sequences from the inactivated F subfamily. To expedite the splice site annotation process we developed, using PHP [41] and MySQL [42], a web interface which allowed for manual data curation. Furthermore a set of perl scripts has been developed to interact with the data and compute statistics. The online database containing annotations of cDNA transcripts with exonized L1 sequences [22] is a read-only version of the annotation system.

Annotation of Splice Sites in potentially active L1s and intronic FL L1s

We used a set of potentially active L1s, as identified in NCBIm35 [18], to examine the potential location of splice sites. In order to compile a set of FL intronic L1 insertions we extracted L1s residing in introns and spanning more than 5000 nt using the RepeatMasker annotation present in Ensembl (Mus musculus v38.35 [36]). The data was then split into sense and antisense L1s. To automatically annotate the presence or absence of the splice sites two L1Xplorer modules were developed: the first module utilizes the alignment-based search as determined by Matcher [39] for splice sites within monomers and the second utilizes a HMMer-based search [43] for splice sites within remaining parts of L1s.

Authors' contributions

TZ and TP designed the study and performed the analysis, drafted and wrote the manuscript. JS and TD participated in writing the manuscript. RB participated in the design of the study and wrote the manuscript. MV participated in the design of the study and wrote the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Annotation of sense splice sites in different subfamilies of potentially active L1s and sense intronic FL insertions. Conservation of AT/GT splice motifs. "n = " indicates the number of annotated L1s.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-392-S1.pdf]

Additional File 2

Logos of annotated splice sites motifs in sequences of potentially active L1s and FL intronic L1 insertions. Motif: sequences of functional splice sites identified via mapping of cDNAs to the mouse genome (see Fig. 3 and [22] for cDNA sequences).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-392-S2.pdf]

Additional File 3

Distribution of antisense (1739) and sense (1014) intronic full length L1s (A) and full length intergenic L1s (10671) (B) among subfamilies. See online annotation at [18].

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-8-392-S3.pdf]

Additional File 4

*Distribution of Line-1 elements around the transcriptional start sites (TSS) and transcriptional end sites (TES) of ~80000 transcriptional units (combined Ensembl, Refseq, UCSC mm7 annotations). **A**. L1 insertions found at +/-50 kbp from TSS and TES of transcriptional units. **B**. Intronic-only L1 insertions. X-axis of the left-hand chart in A and B: distance from TSS, X-axis of the right-hand chart in A and B: the distance from TES. The primary (leftmost) Y-axis shows the number of nucleotides (base pairs) of L1 sequence in sense (red) and antisense (yellow) orientation. The secondary (rightmost) Y-axis: shows the ratio of antisense to sense insertions (black line). Data are plotted in bins of 100 bp. Web data: Annotation of 52 exemplary cDNAs is available at [22]. The database containing the sequences and annotations of potentially active L1s and full length (FL) intronic L1s is available at [18].*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-8-392-S4.pdf]

Acknowledgements

Funding. TZ received founding from the European Commission within its FP6 Programme "Biosapiens", under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LHSG-CT-2003-503265.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Goodier JL, Ostertag EM, Du K, Kazazian HH Jr.: **A novel active L1 retrotransposon subfamily in the mouse.** *Genome Res* 2001, **11:**1677-1685.
3. Penzkofer T, Dandekar T, Zemojtel T: **L1Base: from functional annotation to prediction of active LINE-1 elements.** *Nucleic Acids Res* 2005, **33:**D498-500.
4. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr.: **Hot L1s account for the bulk of retrotransposition in the human population.** *Proc Natl Acad Sci U S A* 2003, **100:**5280-5285.
5. DeBerardinis RJ, Goodier JL, Ostertag EM, Kazazian HH Jr.: **Rapid amplification of a retrotransposon subfamily is evolving the mouse genome.** *Nat Genet* 1998, **20:**288-290.
6. Han JS, Szak ST, Boeke JD: **Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.** *Nature* 2004, **429:**268-274.
7. Han JS, Boeke JD: **LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression?** *Bioessays* 2005, **27:**775-784.
8. Matlik K, Redik K, Speek M: **L1 antisense promoter drives tissue-specific transcription of human genes.** *J Biomed Biotechnol* 2006, **2006:**71753.
9. Speek M: **Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes.** *Mol Cell Biol* 2001, **21:**1973-1985.
10. Nigumann P, Redik K, Matlik K, Speek M: **Many human genes are transcribed from the antisense promoter of L1 retrotransposon.** *Genomics* 2002, **79:**628-634.
11. Perepelitsa-Belancio V, Deininger P: **RNA truncation by premature polyadenylation attenuates human mobile element activity.** *Nat Genet* 2003, **35:**363-366.
12. Narita N, Nishio H, Kitoh Y, Ishikawa Y, Minami R, Nakamura H, Matsuo M: **Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy.** *J Clin Invest* 1993, **91:**1862-1867.
13. Nekrutenko A, Li WH: **Transposable elements are found in a large number of human protein-coding genes.** *Trends Genet* 2001, **17:**619-621.
14. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome.** *Nature* 2001, **409:**847-849.
15. Belancio VP, Hedges DJ, Deininger P: **LINE-1 RNA splicing and influences on mammalian gene expression.** *Nucleic Acids Res* 2006, **34:**1512-1521.
16. Kim DS, Kim TH, Huh JW, Kim IC, Kim SW, Park HS, Kim HS: **LINE FUSION GENES: a database of LINE expression in human genes.** *BMC Genomics* 2006, **7:**139.
17. Perou CM, Pryor RJ, Naas TP, Kaplan J: **The bg allele mutation is due to a LINE1 element retrotransposition.** *Genomics* 1997, **42:**366-368.
18. **L1Base** [http://l1base.molgen.mpg.de]
19. DeBerardinis RJ, Kazazian HH Jr.: **Analysis of the promoter from an expanding mouse retrotransposon subfamily.** *Genomics* 1999, **56:**317-323.
20. Brosius J: **RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements.** *Gene* 1999, **238:**115-134.
21. Breathnach R, Chambon P: **Organization and expression of eucaryotic split genes coding for proteins.** *Annu Rev Biochem* 1981, **50:**349-383.
22. **L1Exon** [http://l1base.molgen.mpg.de/l1exon]
23. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G: **Comparative analysis of transposed element insertion**

within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol* 2007, **8:**R127.
24. 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.
25. Bienko M, Green CM, Crosetto N, Rudolf F, Zapart G, Coull B, Kannouche P, Wider G, Peter M, Lehmann AR, Hofmann K, Dikic I: **Ubiquitin-binding domains in Y-family polymerases regulate translesion synthesis.** *Science* 2005, **310:**1821-1824.
26. Krull M, Brosius J, Schmitz J: **Alu-SINE exonization: en route to protein-coding function.** *Mol Biol Evol* 2005, **22:**1702-1711.
27. Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J: **Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs).** *Genome Res* 2007.
28. Hayashi Y, Sakata H, Makino Y, Urabe I, Yomo T: **Can an arbitrary sequence evolve towards acquiring a biological function?** *J Mol Evol* 2003, **56:**162-168.
29. Fellenberg F, Hartmann TB, Dummer R, Usener D, Schadendorf D, Eichmuller S: **GBP-5 splicing variants: New guanylate-binding proteins with tumor-associated expression and antigenicity.** *J Invest Dermatol* 2004, **122:**1510-1517.
30. Nguyen TT, Hu Y, Widney DP, Mar RA, Smith JB: **Murine GBP-5, a new member of the murine guanylate-binding protein family, is coordinately regulated with other GBPs in vivo and in vitro.** *J Interferon Cytokine Res* 2002, **22:**899-909.
31. Yeo G, Hoon S, Venkatesh B, Burge CB: **Variation in sequence and organization of splicing regulatory elements in vertebrate genes.** *Proc Natl Acad Sci U S A* 2004, **101:**15700-15705.
32. Chen J, Rattner A, Nathans J: **Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements.** *Hum Mol Genet* 2006, **15:**2146-2156.
33. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV: **Human L1 retrotransposition: cis preference versus trans complementation.** *Mol Cell Biol* 2001, **21:**1429-1439.
34. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW, Bult CJ, Fletcher CF, Forrest AR, Furuno M, Hill D, Itoh M, Kanamori-Katayama M, Katayama S, Katoh M, Kawashima T, Quackenbush J, Ravasi T, Ring BZ, Shibata K, Sugiura K, Takenaka Y, Teasdale RD, Wells CA, Zhu Y, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y: **Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs.** *PLoS Genet* 2006, **2:**e62.
35. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
36. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-61.
37. **NCBI Nucleotide Database** [http://www.ncbi.nlm.nih.gov]
38. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35:**D5-12.
39. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
40. Schichman SA, Adey NB, Edgell MH, Hutchison CA 3rd: **L1 A-monomer tandem arrays have expanded during the course of mouse L1 evolution.** *Mol Biol Evol* 1993, **10:**552-570.
41. **PHP** [http://www.php.net]
42. **MySQL** [http://www.mysql.com]
43. Eddy SR: **Multiple alignment using hidden Markov models.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:**114-120.