frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 09 June 2015
doi: 10.3389/fbioe.2015.00071

Transposable element insertions in long intergenic non-coding RNA genes

Sivakumar Kannan[1†], Diana Chernikova[2†], Igor B. Rogozin[1†], Eugenia Poliakov[3], David Managadze[1], Eugene V. Koonin[1] and Luciano Milanesi[4]*

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA, [2] Department of Genetics, Institute for Quantitative Biomedical Sciences, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA, [3] Laboratory of Retinal Cell and Molecular Biology, National Eye Institute, National Institutes of Health, Bethesda, MD, USA, [4] Institute for Biomedical Technologies, National Research Council, Segrate, Italy

Transposable elements (TEs) are abundant in mammalian genomes and appear to have contributed to the evolution of their hosts by providing novel regulatory or coding sequences. We analyzed different regions of long intergenic non-coding RNA (lincRNA) genes in human and mouse genomes to systematically assess the potential contribution of TEs to the evolution of the structure and regulation of expression of lincRNA genes. Introns of lincRNA genes contain the highest percentage of TE-derived sequences (TES), followed by exons and then promoter regions although the density of TEs is not significantly different between exons and promoters. Higher frequencies of ancient TEs in promoters and exons compared to introns implies that many lincRNA genes emerged before the split of primates and rodents. The content of TES in lincRNA genes is substantially higher than that in protein-coding genes, especially in exons and promoter regions. A significant positive correlation was detected between the content of TEs and evolutionary rate of lincRNAs indicating that inserted TEs are preferentially fixed in fast-evolving lincRNA genes. These results are consistent with the repeat insertion domains of LncRNAs hypothesis under which TEs have substantially contributed to the origin, evolution, and, in particular, fast functional diversification, of lincRNA genes.

Keywords: mobile elements, molecular domestication, exaptation, junk DNA, long non-coding RNA, repetitive elements

OPEN ACCESS

Edited by:
Marco Pellegrini,
Consiglio Nazionale delle Ricerche,
Italy

Reviewed by:
Roderic Guigo,
Center for Genomic Regulation,
Spain
Justin Blumenstiel,
University of Kansas, USA

***Correspondence:**
Luciano Milanesi,
Institute for Biomedical Technologies,
National Research Council, Via Fratelli
Cervi 93, 20090 Segrate, Italy
luciano.milanesi@itb.cnr.it

†Sivakumar Kannan, Diana Chernikova and Igor B. Rogozin have contributed equally to this work.

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *10 December 2014*
Accepted: *06 May 2015*
Published: *09 June 2015*

Citation:
Kannan S, Chernikova D, Rogozin IB, Poliakov E, Managadze D, Koonin EV and Milanesi L (2015) Transposable element insertions in long intergenic non-coding RNA genes.
Front. Bioeng. Biotechnol. 3:71.
doi: 10.3389/fbioe.2015.00071

Introduction

Traditionally, genomes have been perceived mostly as repositories of protein-coding genes. Although this might be largely true in the case of viruses, prokaryotes, and unicellular eukaryotes, numerous recent studies on the genomes of multicellular eukaryotes, particularly animals, have revealed a vast non-coding RNome, i.e., numerous genes encoding various classes of non-coding RNAs (ncRNAs) (Carninci et al., 2005; Mattick and Makunin, 2006; Ponting et al., 2009; Derrien et al., 2012; Amaral et al., 2013). Strikingly, the total number of genes for ncRNAs that are expressed from a mammalian genome seems to exceed the number of protein-coding genes several fold (Mattick and Makunin, 2006; Amaral et al., 2013). The classification of ncRNAs and validation of their functionality remain matters of intensive investigation and debate (Van Bakel and Hughes, 2009; Ponting and Belgard, 2010; Graur et al., 2013). Among many distinct classes of ncRNAs, the long non-coding RNA (lncRNA) is probably the most enigmatic group. The definition of a lncRNA is based solely on the transcript size: lncRNAs are defined as ncRNAs longer than 200 nt (Mattick and Makunin, 2006; Ponting et al., 2009).

Many lncRNAs are spliced, 5′capped, and polyadenylated (Okazaki et al., 2002; Carninci et al., 2005; Kapranov et al., 2007; Ponjavic et al., 2007). Based on the localization in the genome, lncRNAs can be divided into two distinct classes: (i) transcripts that overlap protein-coding genes, many of which are likely to be involved in sense–antisense regulation (Chen et al., 2005; Ponting and Belgard, 2010; Rinn and Chang, 2012) and (ii) long intergenic non-coding (linc)RNAs that are transcribed from genome regions separating protein-coding genes (Ponjavic et al., 2007; Mercer et al., 2008; Ponting et al., 2009).

The current knowledge on the functions of long intergenic non-coding RNAs (lincRNAs) is scarce because very few of the lincRNAs have been experimentally characterized. Nevertheless, the functional range of this class of ncRNA is believed to be broad on the basis of indirect evidence (Bertone et al., 2004; Ponjavic et al., 2007; Mercer et al., 2008; Ponting and Belgard, 2010; Ulitsky et al., 2011; Glazko et al., 2012; Ng et al., 2013). It has been proposed that lincRNAs could be involved in the regulation of many cellular processes (Mattick and Makunin, 2006; Loewer et al., 2010; Wang et al., 2011; Rinn and Chang, 2012). For example, they can affect transcription locally on the gene level (Martens et al., 2004; Martianov et al., 2007; Osato et al., 2007; Hirota et al., 2008) as well as target transcription regulators and thus affect transcription of many genes (Feng et al., 2006; Goodrich and Kugel, 2006). They can also target RNA polymerase II in human and mouse (Espinoza et al., 2007; Mariner et al., 2008) and thus affect the expression of an even broader range of genes. Furthermore, lincRNAs participate in the regulation of splicing (Munroe and Lazar, 1991; Beltran et al., 2008) and translation (Wang et al., 2005; Centonze et al., 2007). Well-characterized examples of lincRNAs involved in epigenetic processes are *Xist* (Brockdorff et al., 1992; Elisaphenko et al., 2008), *Kcnq1ot1* (Umlauf et al., 2004; Pandey et al., 2008), and *Air* (Nagano et al., 2008).

It is well established that, compared to protein-coding sequences and structural RNAs, lincRNAs are weakly conserved in evolution. Many early studies, therefore, branded the lincRNAs "transcriptional dark matter" and considered them to be generally non-functional (Van Bakel and Hughes, 2009; Robinson, 2010). However, low level or lack of detectable conservation does not necessarily imply that these molecules have no function (Pang et al., 2006). A case in point is the best-characterized, functionally important lincRNA gene, *Xist*, which is weakly conserved although it does contain evolutionary constrained regions (Elisaphenko et al., 2008). In general, lincRNAs show reduced substitution and insertion–deletion rates, which has been attributed to purifying selection (Ponjavic et al., 2007; Managadze et al., 2011). Taking into account that some lincRNA genes originated from protein-coding genes [for example, *Xist* (Duret et al., 2006; Elisaphenko et al., 2008)], it appears likely that many properties of lincRNAs would generally resemble those of protein-coding genes, despite the typically lower level of constraint. In particular, protein-coding genes that are highly expressed in many tissues typically evolve slower than genes with lower expression level and breadth (Duret and Mouchiroud, 2000; Krylov et al., 2003; Drummond and Wilke, 2008), and a similar dependence has been observed for lincRNA genes (Managadze et al., 2011). Taken together, these findings imply that an unknown but substantial fraction of lincRNAs are functional molecules rather than transcriptional noise and have evolutionary properties similar to those of protein-coding genes. However, the number of functionally characterized lincRNAs remains scarce (Amaral et al., 2013).

The origin of lincRNA genes generally remains enigmatic. However, analysis of the well-characterized *Xist* lincRNA has revealed fragmentary homology to a protein-coding gene *Lnx3* suggesting that the *Xist* genes emerged in early eutherians via integration of transposable elements (TEs) into the *Lnx3* gene, which gave rise to simple tandem repeats (Duret et al., 2006; Elisaphenko et al., 2008). The *Xist* gene promoter region and 4 of its 10 exons retain homology to exons of the *Lnx3* gene. The remaining six *Xist* exons including those containing simple tandem repeats show similarity to different TEs (Elisaphenko et al., 2008). Integration of TEs into the *Xist* gene apparently had been occurring throughout the course of evolution of this gene and most likely continues in contemporary eutherian species. Additionally, it has been shown that the combination of remnants of protein-coding sequences and TEs is not unique to the *Xist* gene but is also found in neighboring genes that encode non-coding nuclear RNAs (Elisaphenko et al., 2008; Kolesnikov and Elisafenko, 2010).

The discovery of the pivotal contribution of TEs to the evolution of the *Xist* gene prompts the question on a possible general role of TEs in the evolution of lincRNAs. Diverse TEs are widespread and abundant in the genomes of most eukaryotes (Smit, 1996; Brosius, 1999; Kidwell and Lisch, 2001; Deininger and Batzer, 2002). Different classes of TEs include mobile retrovirus-like elements, or retrotransposons, which transpose within the genome via RNA intermediates, and DNA transposons, which can relocate directly. Retrotransposons including long interspersed repetitive elements (LINEs), short interspersed repetitive elements (SINEs), and long terminal repeat (LTR) retrotransposons are widely represented in mammals (Smit, 1996; Deininger and Batzer, 2002). The LINEs are transcribed by RNA polymerase II and contain open reading frames (ORFs) (Temin, 1985). A complete and transpositionally active L1 element (the most common variety of LINEs) is ~7 kb long and contains a 5′-untranslated region (UTR) with an internal promoter, two ORFs (ORF1 and ORF2) and a 3′-UTR terminated by a polyadenylate-rich tail (Smit, 1996; Deininger and Batzer, 2002). The ORF1 encodes a putative RNA-binding protein ~40 kDa in size (Martin, 2006) whereas ORF2 encodes a protein with endonuclease and reverse transcriptase (RT) activities that generates cDNAs from RNA transcripts of the element (Loeb et al., 1986). The mobility of the LINE elements had been demonstrated in mouse and human genomes (Kazazian et al., 1988; Boccaccio et al., 1990). The SINEs are characterized by the presence of a split intragenic RNA polymerase III promoter and a 3′A-rich region often followed by an oligo(A) tail (Smit, 1996; Rogozin et al., 2000; Kapitonov and Jurka, 2003). The SINEs do not contain long ORFs and do not encode enzymes for transposition. Instead, transposition of SINEs apparently requires RT encoded by other TEs, in particular, LINEs (Smit, 1996; Deininger and Batzer, 2002). The LTR retrotransposons have LTRs that range from ~100 bp to over 5000 bp in size (Smit, 1996; Deininger and Batzer, 2002). The LTR retrotransposons are similar to retroviruses in organization, with transcriptional regulatory sequences located in the flanking LTRs, a RT priming site that is typically located immediately downstream of an first LTR, and several ORFs encoding proteins involved in retrotransposition, in particular, RT and integrase (Smit, 1996; Deininger and Batzer, 2002).

The TEs are the primary contributors to the bulk of the genomic DNA in many eukaryotes, in particular mammals, and have the potential to contribute to the evolution of the hosts by providing novel regulatory or coding sequences (Makalowski, 2000). Different classes of regulatory regions in the human genome have been surveyed for the presence of TE-derived sequences (TES) to systematically assess the potential contribution of TEs to the regulation of human genes, and almost 25% of the analyzed promoter regions have been found to contain TES (Jordan et al., 2003; Feschotte, 2008; Bourque, 2009). In addition, numerous examples where experimentally characterized *cis*-regulatory elements are derived from TE sequences have been identified (Jordan et al., 2003; Bourque et al., 2008; Faulkner et al., 2009). Thus, thousands of human (and other mammalian) genes appear to be regulated, at least in part, by sequences derived from TEs (Jordan et al., 2003; Feschotte, 2008; Bourque, 2009). The TES are likely to have substantially contributed to evolutionary change in both gene specific and global patterns of mammalian protein-coding gene regulation (Makalowski, 2000; Jordan et al., 2003).

In light of the regulatory and structural effects that some TEs exert on host protein-coding and lncRNA genes (Makalowski, 2000; Jordan et al., 2003; Elisaphenko et al., 2008; Mattick et al., 2010; Wang et al., 2011; Kapusta et al., 2013; Johnson and Guigo, 2014), we sought to examine the contribution and conservation of TES to regulatory regions, exons and introns of human and mouse lincRNA genes. We found that introns of lincRNA genes contain the highest fraction of TES, followed by exons. The promoters of the lincRNAs contain the lowest fraction of TES but the largest fraction of ancient TES that are conserved between primates and rodents. The content of TES in lincRNA genes is substantially greater than in protein-coding genes, particularly in exons and promoter regions. These results are compatible with the view that TEs are major contributors to the origin and evolution of lincRNAs. We further sought to assess the potential utility of TES as an "evolutionary variable" by analyzing the correlations between the TES content, lincRNA expression, and sequence conservation.

Materials and Methods

Human and mouse lincRNA genes, the corresponding genomic alignments and expression data were taken from our previous work (Managadze et al., 2011) where the procedures of data processing are described in full details. Briefly, complete mouse and human probe sets were downloaded from the NRED database (Dinger et al., 2009) in the tab delimited and browser extensible data (BED, containing genomic coordinates) formats. The probe sets from platform GNF Atlas 2 (Mouse and Human), with the target classification "Non-coding Only," were used for further analysis. This protocol yielded 917 human and 5444 mouse probe sets. Only the probe sets that mapped to intergenic regions of the human and mouse genomes (i.e., between two adjacent protein-coding genes) were used for further analysis. The resulting list of lincRNAs was further filtered: sequences shorter than 200 nt were removed. This procedure yielded the final set of NCBI GenBank Accession IDs of 2390 mouse and 589 human lincRNAs and their corresponding microarray expression probe sets. The genomic coordinates and sequences of exons and introns of lincRNA and protein-coding genes were downloaded from the UCSC Table Browser (Karolchik et al., 2004), specifically, from "all_mrna" tables of mouse mm8 and human hg18 assemblies. Multiple alignments of these regions were fetched from Galaxy (Goecks et al., 2010). For the detection of TES, lincRNA and protein-coding genes were analyzed using RepeatMasker version open-3.1.3[1] with the following parameters: -w -s -no_is -cutoff 255 -frag 20000 -gff -species mouse/human. A TE insert was considered ancient if the pairwise alignment between human and mouse orthologous TE sequences was longer than 100 bp and contained <5% insertions/deletions (stringent definition) or 25% insertions/deletions (relaxed definition). Microarray data for normal (non-cancerous) tissues (73 human and 61 mouse tissues) were used to analyze the lincRNA expression. Log2-normalized median values of expression for each probe set across the tissues were calculated (Managadze et al., 2011). As an alternative method of measuring expression levels, the mouse RNA-seq data for eight tissues (the ENCODE project; modENCODE Consortium) were downloaded from the UCSC genome browser Web site[2] and pooled together. The RPKM value was calculated for each mouse lincRNA (Managadze et al., 2011). Pairwise evolutionary distances for human–macaque and mouse–rat lincRNA alignments were calculated using the DNADIST program from the PHYLIP package (Felsenstein, 1996), with the Kimura nucleotide substitution model. The lists of lincRNA genes and expression data are available at ftp://ftp.ncbi.nlm.nih.gov/pub/managdav/paper_suppl/TEs_lincRNA/.

One of the problems in the analysis of lincRNAs is that there is little overlap between lincRNA sets produced in different studies (Ulitsky et al., 2011; Chew et al., 2013; Managadze et al., 2013; Schuler et al., 2014). We used human and mouse datasets because these curated lincRNA sets have known evolutionary and gene expression properties (Managadze et al., 2011, 2013). Another reason for this choice is that we sought to analyze lincRNA datasets as different as possible from those used in previous studies (Kapusta et al., 2013; Johnson and Guigo, 2014) and to check how small sample size of human lincRNA set influences results. As shown below, the sample size does not perceptibly affect the conclusions of this study.

Results

Transposable Elements in Human and Mouse lincRNA Genes

Transposable element-derived sequences (TES for short) comprise at least half of the mammalian genomes, and in particular, are found in most lincRNAs. We identified TES in 69% of the human lincRNAs and 51% of the mouse lincRNAs. These values are somewhat lower than the previously reported 83% of TES in human lincRNAs (Kelley and Rinn, 2012) but nevertheless clearly show the importance of TE for lincRNA evolution. The

[1] http://www.repeatmasker.org

[2] http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrRnaSeq/

distribution of TES in 5′ flanking regions (putative core promoter regions), lincRNA exons, and introns is shown in the **Figure 1**. The lowest fraction of TES was found in the predicted core promoter regions (100 bp upstream regions), and the highest fraction of TES was observed in introns, whereas exonic sequences showed intermediate densities of TES (**Figure 1A**). This distribution of TES is compatible with the previously described general tendency of TES avoidance in functionally important regions of protein-coding genes (Jordan et al., 2003). In particular, similar to the protein-coding genes, the TES density in extended promoter regions has been found to be significantly greater than that in core promoter regions (Jordan et al., 2003). Notably, the fractions of TES in introns of lincRNAs and protein-coding genes are nearly identical, suggestive of comparable (weak) functional constraints. By contrast, in the exons and the core promoter regions of lincRNA genes, the fractions of TES are substantially and statistically significantly ($P < 10^{-5}$ according to the Fisher exact test) higher than in the respective regions of protein-coding genes (compare **Figures 1A,B**). These findings are consistent with the results of a previous study that employed different datasets of lincRNA genes (Kapusta et al., 2013), indicating that the distribution of TES in lincRNA genes is a robust feature. A more detailed analysis of the distribution of TES across lincRNAs is shown in Figure S1 in Supplementary Material. The most prominent feature of this distribution is the high fraction of lincRNAs with a low TES content: in 66% of human lincRNAs and 78% of mouse lincRNAs, the fraction of TES is <20% (Figure S1 in Supplementary Material).

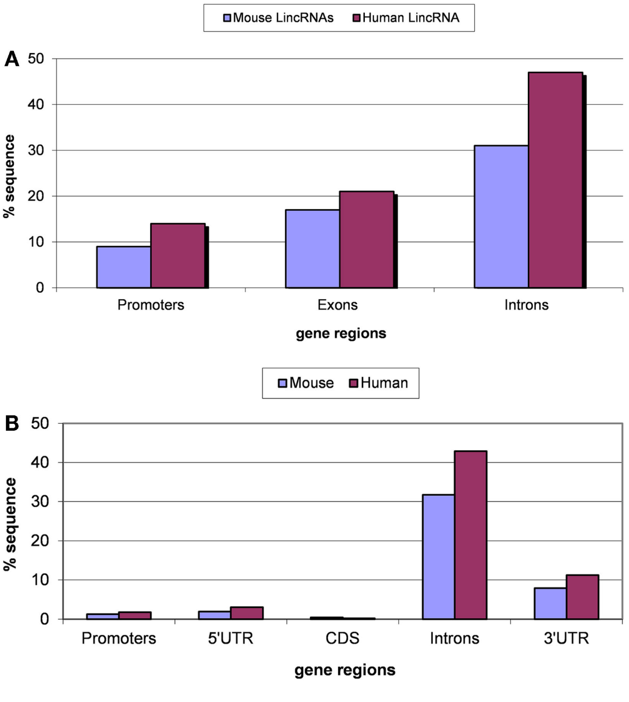

FIGURE 1 | Fractions (proportions) of lincRNA gene regions (concatenated promoters, exons, and introns) (A) and protein-coding gene regions (B) occupied by TE-derived sequences. The differences for pairwise comparisons "promoters vs. introns" and "exons vs. introns" are statistically significant for both classes of genes ($P < 10^{-5}$ according to the Fisher exact test; the raw counts of nucleotides in TES vs. the raw counts of nucleotides in TE-free regions was used as the input for 2 × 2 contingency tables).

The avoidance of TES in lincRNAs is consistent with purifying selection, which is an important feature of lincRNA evolution (Ponjavic et al., 2007; Managadze et al., 2011). The significant positive correlation between the evolutionary rate and the content of TES was observed for both human and mouse lincRNA sets (for human and mouse, respectively, the Pearson correlation coefficient are 0.183 and 0.337, $P < 10^{-5}$ for both comparisons) (**Figure 2**). We also tested the correlation between the expression level and the content of TES (Figure S2 in Supplementary Material). In many independent previous studies, it has been shown that protein-coding genes that are highly expressed in many tissues typically evolve slower than genes with lower expression level and breadth (Duret and Mouchiroud, 2000; Krylov et al., 2003; Drummond and Wilke, 2008), and a similar dependence has been observed for lincRNA genes (Managadze et al., 2011). Consistent with these observations, here we found a significant negative correlation between the content of TES and the expression level of lincRNAs (Figure S2 in Supplementary Material; for mouse RNA-seq data, Pearson correlation coefficient is −0.158, $P < 10^{-5}$; for mouse microarray data, Pearson correlation coefficient is −0.07, $P < 10^{-5}$; for human microarray data, Pearson correlation coefficient is −0.253, $P < 10^{-5}$).

Different Classes of Transposable Elements in lincRNA Genes

Analysis of different classes of TEs indicates that the fractions of each class are similar for introns of lincRNA and protein-coding

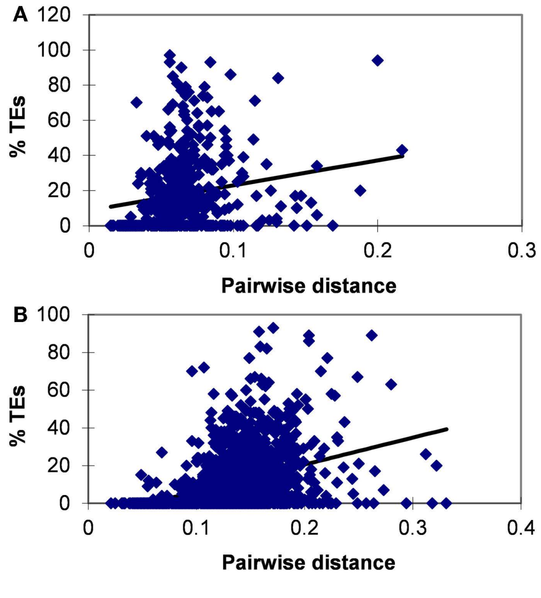

FIGURE 2 | Correlation between the fraction (proportion) of TEs in concatenated exons and evolutionary rate for human (A) and mouse (B) lincRNAs. Pearson correlation coefficient is 0.183 for human and 0.337 for mouse ($P < 10^{-5}$ for both comparisons).

genes and whole genomes (**Figures 3** and **4**). In each case, the fraction of LINEs is substantially greater than those of SINEs and LTR elements (**Figures 3A** and **4A**). However, there is a significant suppression of LINEs in exonic and promoter regions, in both human and mouse (**Figures 3A** and **4A**). This effect cannot be explained by fluctuations of the base composition in different gene regions because there are no significant compositional differences between exons, introns, and promoter regions for human and mouse lincRNA genes (results not shown). The same trend was observed for different lincRNA sets (Kapusta et al., 2013) suggesting that re-distribution of TEs is a general property of mammalian lincRNA genes. Furthermore, similar tendency is observed in promoter sequences of protein-coding genes (**Figures 3B** and **4B**), the overall lower abundance of TEs notwithstanding. Conceivably, when the smaller SINEs are inserted into functionally important parts of genes, they typically exert a milder deleterious effect than the larger LINEs and LTR elements and accordingly, are more often fixed in the course of evolution.

Higher Frequency of Ancient Transposable Element-Derived Sequences in Promoters and Exons Compared to Introns

Evolutionary conservation of TEs is likely to reflect molecular domestication of the respective elements (Jordan et al., 2003; Feschotte, 2008; Jurka, 2008; Bourque, 2009; Sinzelle et al., 2009). We analyzed the fraction of ancient mobile elements in different regions of lincRNA genes (**Figure 5**). A significantly higher abundance of ancient TEs ($P < 10^{-5}$ according to the Fisher exact test) was detected in exons and especially in promoter regions compared to introns (**Figure 5**). This finding is consistent with the hypothesis that TEs, in some cases, may perform novel functions in the host organisms (Makalowski, 2000; Jordan et al., 2003). The excess of ancient TEs was more pronounced in human compared to mouse lincRNA genes (**Figure 5**), possibly reflecting differences in evolutionary processes in rodents and primates although a bias caused by technical problems with the detection of 5′-ends of human lincRNA sequences cannot be ruled out (Kutter et al., 2012). We searched the putative promoter regions of lincRNA genes for the presence of TATA boxes and found a substantially elevated frequency of TATA-like sequences in the region −25 to −35 (Figure S3 in Supplementary Material). Given that a similar distribution is observed in many well-annotated human protein-coding genes (Yang et al., 2007), these observations suggest acceptable accuracy of 5′-end identification in lincRNA genes. The fractions of TATA-containing promoters are similar for protein-coding genes (10–25%) (Yang et al., 2007; Anish et al., 2009) and the analyzed sets of lincRNA genes (19–30%; Table S1 in Supplementary Material). The higher frequency of ancient TEs in promoters and exons compared to introns (**Figure 5**; Table S2 in Supplementary Material) suggests that many lincRNA genes emerged before the split of primates and rodents, and that TEs contributed to the origin of these ancient lincRNAs. This finding is consistent with recent observations that 60–70% of the lincRNAs genes are conserved between human and mouse (Kutter et al., 2012; Managadze et al., 2013), and with the observed

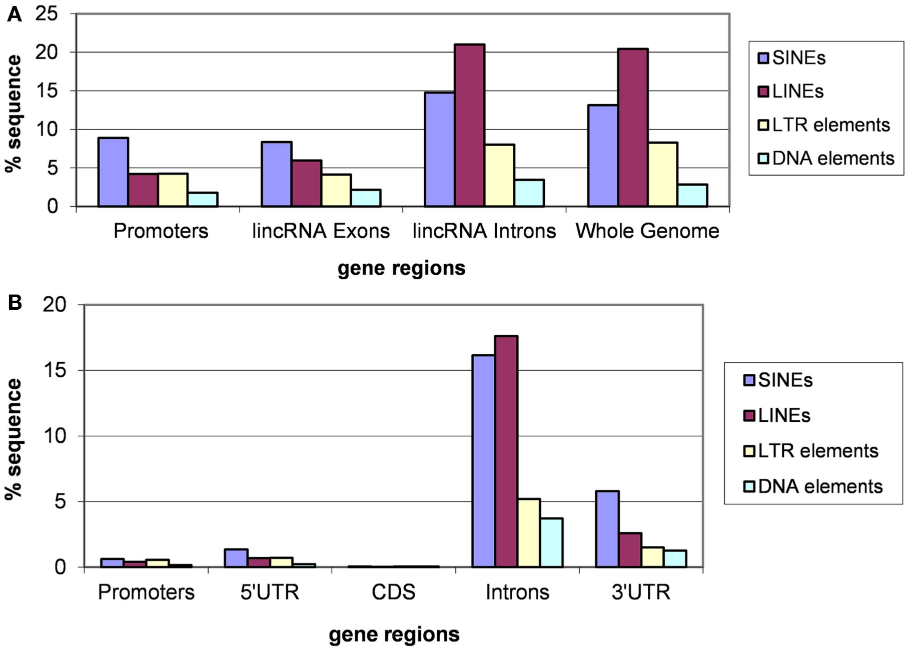

FIGURE 3 | Fractions (proportions) of human lincRNA gene regions (concatenated promoters, exons, and introns) and the whole genome sequence (A) and protein-coding gene regions (B) occupied by sequences derived from different types of TEs. Differences for pairwise comparisons "SINEs vs. LINEs" and "LTRs vs. LINEs" are statistically significant for both classes of genes ($P < 10^{-5}$ according to the Fisher exact test; the raw counts of nucleotides in TES vs. the raw counts of nucleotides in TE-free regions was used as the input for 2 × 2 contingency tables).

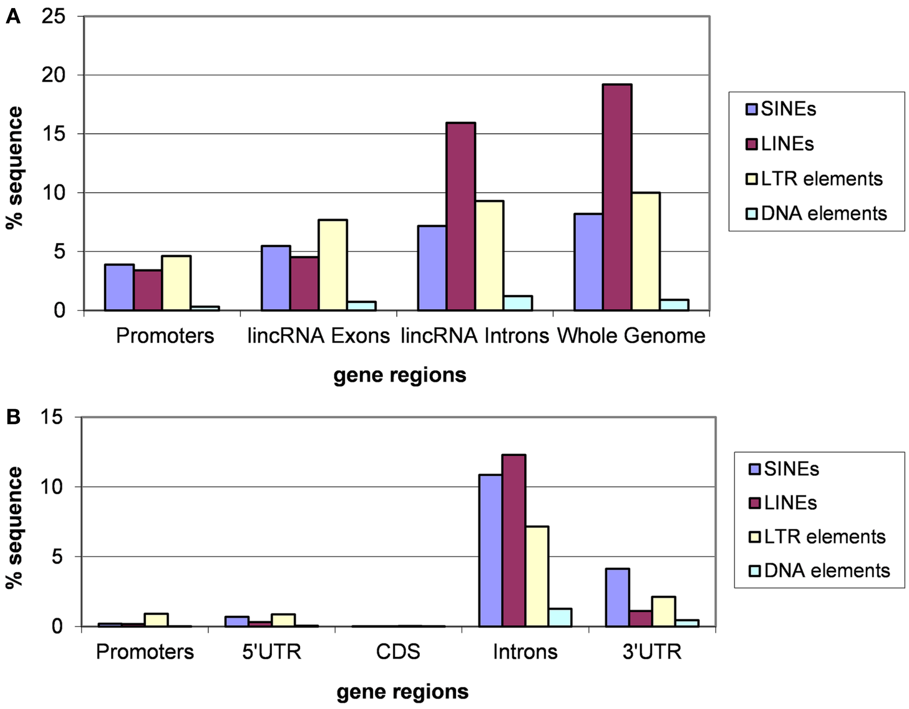

FIGURE 4 | Fractions (proportions) of mouse lincRNA gene regions (concatenated promoters, exons, and introns) and the whole genome sequence [the data for the whole genome sequence were from Waterston et al. (2002)] (A) and protein-coding gene regions (B) occupied by sequences derived from different types of TEs. Differences for pairwise comparisons "SINEs vs. LINEs" and "LTRs vs. LINEs" are statistically significant for both classes of genes ($P < 10^{-5}$ according to the Fisher exact test; the raw counts of nucleotides in TES vs. the raw counts of nucleotides in TE-free regions was used as the input for 2 × 2 contingency tables).

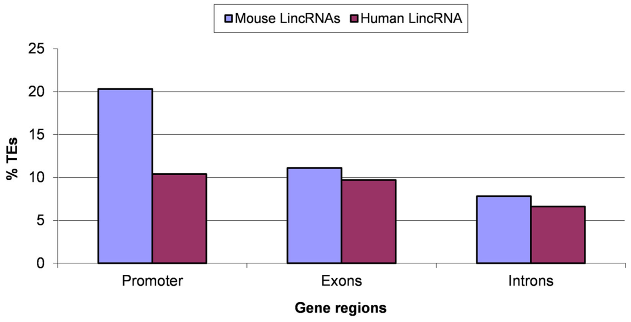

FIGURE 5 | Ancient transposable elements (TEs) in putative promoter regions, exons, and introns of lincRNA genes. A TE was considered ancient if the alignment between human–mouse orthologous TE sequences was longer than 100 bp and contained <5% insertions/deletions (the stringent threshold). "% TEs" stands for the fraction (proportion) of ancient TEs. Results for the relaxed threshold (the alignment between human–mouse orthologous TE sequences was longer than 100 bp and contained no more than 25% insertions/deletions) are shown in the Table S2 in Supplementary Material. Differences between pairwise comparisons "promoters vs. introns" and "exons vs. introns" are statistically significant ($P < 10^{-5}$ according to the Fisher exact test; the raw counts of ancient TES vs. the raw counts of lineage-specific TES was used as the input for 2 × 2 contingency tables).

higher conservation of lincRNA promoter regions compared to exons (Elisaphenko et al., 2008; Kapusta et al., 2013; Johnson and Guigo, 2014).

Discussion

The staggering evolutionary success of TEs in eukaryotes is often attributed to their ability to out replicate the host genomes in which they reside, as opposed to any selective advantage that they might provide to their hosts. Indeed, it has been shown that TEs can spread within and among genomes even in the face of a selective cost to the host (Hickey, 1982). Hence, the selfish DNA concept of TEs focuses on the parasitic nature of these elements and emphasizes the deleterious effects of transposition as well as the negligible evolutionary benefit that TEs provide to their hosts (Orgel et al., 1980; Gould and Vrba, 1982). However, the sheer abundance of TEs in the genome, as well as the variety of mutation effects induced by their mobility, suggest that they might, in some cases, be exapted (Gould and Vrba, 1982) or domesticated (Miller et al., 1999), to serve the evolutionary interests of the host (Makalowski, 2000; Jordan et al., 2003). Indeed, multiple lines of evidence indicate that the presence of TEs can result in host adaptation by shaping and reshaping the genome in many different ways (Smit, 1996; Makalowski, 2000; Rogozin et al., 2000; Kidwell and Lisch, 2001; Deininger and Batzer, 2002; Jordan et al., 2003).

The TEs comprise at least half of the mammalian genomes, and in particular, are found in most lincRNAs [this study and Kelley and Rinn (2012)]. Here, we demonstrate that TEs substantially contribute to the evolution of lincRNAs and their promoter regions. Although the densities of TES in these regions are much lower than those in introns, ostensibly, due to the purifying selection that affects functional regions, the contributions of TEs to the evolution of these regions is substantially greater than in the respective regions of protein-coding genes. The higher density of TES in the exons of lincRNAs compared to protein-coding exons appears to reflect the much lower level of functional constraint characteristic of the former (Ponjavic et al., 2007; Managadze et al., 2011). The promoters of lincRNA appear to similarly enjoy greater plasticity and flexibility compared to the promoters of protein-coding genes.

Thus, TE insertion is an important factor that affects lincRNA evolution and biological function. An analysis of TEs in human lincRNAs revealed that the TES composition in lincRNA genes significantly differs from genomic averages: LINEs and SINEs are depleted whereas LTR retrotransposons are enriched (Kelley and Rinn, 2012). The TES occur in biased positions and orientations at lincRNA transcription start sites suggesting a functional role in lincRNA transcriptional regulation (Kelley and Rinn, 2012). In many cases, lincRNAs devoid of TES are expressed at higher levels than lincRNAs containing TES in all tested tissues and cell lines (Kelley and Rinn, 2012). Thus, it has been suggested that TES divide lincRNAs into classes and have contributed to lincRNA evolution and function by conferring tissue-specific expression from extant transcriptional regulatory signals (Kelley and Rinn, 2012). Here, we add another facet to these observations by showing that the promoter regions of lincRNAs are specifically enriched for ancient TES. This finding indicates that not only have many lincRNA genes evolved before the radiation of primates and rodents but also that at least some features of their regulation were already established at that time through TE insertion.

The possibility that some lincRNA genes encode short peptides that are translated, perhaps in a tissue-specific manner, is the subject of an ongoing debate (Brosius and Tiedge, 2004; Mattick and Makunin, 2006; Dinger et al., 2008; Makalowska et al., 2010; Carvunis et al., 2012; Chew et al., 2013). It is extremely hard to rule out such a role for a fraction of purported lincRNAs. A recent peptidomics study demonstrated that most annotated lincRNAs do not generate stable protein (peptide) products (Banfai et al., 2012). Furthermore, ribosomal profiling of lincRNAs suggests that ribosomal engagement with lincRNAs is likely to be regulatory (Chew et al., 2013). The presence of ORFs in the analyzed lincRNA data sets had been analyzed before using different approaches (Managadze et al., 2011, 2013). Importantly, removal of ORF-containing lincRNAs did not affect the conclusions of both studies (Managadze et al., 2011, 2013). The much higher abundance of TES in lincRNA compared to 5′UTR and protein-coding regions of mRNAs is consistent with the low frequency of protein-coding regions in the analyzed data sets.

It has been proposed that lncRNAs are organized into combinations of discrete functional domains, but the nature of these domains and their identification remain elusive (Guttman and Rinn, 2012). Insertion of TEs and exaptation of TES could represent an important route of evolution of the domain structure of lncRNAs. More specifically, Johnson and Guigo (2014) have proposed that exonic TES comprise functional domains of lncRNAs that they dubbed repeat insertion domains of LncRNAs (RIDLs). A growing number of RIDLs have been experimentally identified whereby lncRNA TES function as RNA-, DNA-, and protein-binding domains/motifs (Elisaphenko et al., 2008; Kelley and Rinn, 2012; Grote and Herrmann, 2013; Holdt et al., 2013; Johnson and Guigo, 2014). These examples are likely to reflect a more general phenomenon of exaptation and/or domestication during lncRNA evolution whereby TES are employed as DNA-, RNA-, and protein-binding domain/motifs (Johnson and Guigo, 2014). The RIDL hypothesis has the potential to explain how functional evolution can keep pace with the fast evolution observed in many lncRNA genes (Johnson and Guigo, 2014). The findings on the distribution of TES across different regions of lincRNA genes, the higher occurrence of TES in lincRNA promoters and exons compared to introns, and significant correlations between the content of TES and evolutionary rate presented here appear to be compatible with the RIDL hypothesis. More specifically, even if a substantial fraction of TES are not fixed in lincRNA exons and promoter regions, those TES that are fixed tend to persist in the genome longer than intronic TES. Moreover, given the near ubiquity of recognizable TES in lincRNA genes, TE mapping can be a useful approach for characterization of lincRNAs and possibly even prediction of their functions. The correlations between the content of TES and various features of lincRNA genes described here could be useful for the characterization of lincRNA functions.

Conclusion

The results of the present analysis, along with several previous studies, indicate that TEs have contributed to the evolution of many if not most mammalian lincRNAs. Whereas the density of TES in the introns of lincRNA genes is about the same as in introns of protein-coding genes exons, and promoters of lincRNAs are markedly enriched in TES compared to the counterparts in protein-coding genes. This high prevalence of TES reflects the relatively weak evolutionary constraints on lincRNA genes and itself appears to contribute to the plasticity and functional diversification of lincRNAs. Furthermore, the distribution of TE types in the functional regions of lincRNA genes significantly differs from that in introns (or whole genomes), conceivably, because the smaller SINEs that encode no proteins are more suitable for exaptation than the larger, protein-coding LINEs. The prodigious exaptation of TE could account, at least in part, for the functionality of many lincRNAs despite their rapid evolution.

Acknowledgments

We thank Joshua Cherry, Jean Thierri-Mieg, Kira Makarova, and Yuri Wolf for useful discussions. *Funding*: This research was supported in part by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health and the National Institutes of Health Intramural Research Program of the National Eye Institute (US Department of Health and Human Services), Italian Ministry of Education, University and Research (MIUR) Flagship "InterOmics" (PB05), HIRMA (RBAP11YS7K) and by the "MIMOMICS" European projects.

Supplementary Material

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/fbioe.2015.00071/abstract

References

Amaral, P. P., Dinger, M. E., and Mattick, J. S. (2013). Non-coding RNAs in homeostasis, disease and stress responses: an evolutionary perspective. *Brief. Funct. Genomics* **12**, 254–278. doi:10.1093/bfgp/elt016

Anish, R., Hossain, M. B., Jacobson, R. H., and Takada, S. (2009). Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements. *PLoS ONE* **4**:e5103. doi:10.1371/journal.pone.0005103

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W. E. Jr., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646–1657. doi:10.1101/gr.134767.111

Beltran, M., Puig, I., Pena, C., Garcia, J. M., Alvarez, A. B., Pena, R., et al. (2008). A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev.* **22**, 756–769. doi:10.1101/gad.455708

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246. doi:10.1126/science.1103388

Boccaccio, C., Deschatrette, J., and Meunier-Rotival, M. (1990). Empty and occupied insertion site of the truncated LINE-1 repeat located in the mouse serum albumin-encoding gene. *Gene* **88**, 181–186. doi:10.1016/0378-1119(90)90030-U

Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.* **19**, 607–612. doi:10.1016/j.gde.2009.10.013

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762. doi:10.1101/gr.080663.108

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., et al. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515–526. doi:10.1016/0092-8674(92)90519-I

Brosius, J. (1999). Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107**, 209–238. doi:10.1023/A:1004018519722

Brosius, J., and Tiedge, H. (2004). RNomenclature. *RNA Biol.* **1**, 81–83. doi:10.4161/rna.1.2.1228

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563. doi:10.1126/science.1112014

Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., et al. (2012). Proto-genes and de novo gene birth. *Nature* **487**, 370–374. doi:10.1038/nature11184

Centonze, D., Rossi, S., Napoli, I., Mercaldo, V., Lacoux, C., Ferrari, F., et al. (2007). The brain cytoplasmic RNA BC1 regulates dopamine D2 receptor-mediated transmission in the striatum. *J. Neurosci.* **27**, 8885–8892. doi:10.1523/JNEUROSCI.0548-07.2007

Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G., and Rowley, J. D. (2005). Human antisense genes have unusually short introns: evidence for selection for rapid transcription. *Trends Genet.* **21**, 203–207. doi:10.1016/j.tig.2005.02.003

Chew, G. L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140**, 2828–2834. doi:10.1242/dev.098343

Deininger, P. L., and Batzer, M. A. (2002). Mammalian retroelements. *Genome Res.* **12**, 1455–1465. doi:10.1101/gr.282402

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789. doi:10.1101/gr.132159.111

Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M., and Mattick, J. S. (2009). NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.* **37**, D122–D126. doi:10.1093/nar/gkn617

Dinger, M. E., Pang, K. C., Mercer, T. R., and Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**:e1000176. doi:10.1371/journal.pcbi.1000176

Drummond, D. A., and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352. doi:10.1016/j.cell.2008.05.042

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653–1655. doi:10.1126/science.1126316

Duret, L., and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68–74. doi:10.1093/oxfordjournals.molbev.a026239

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogozin, I. B., Nesterova, T. B., Brockdorff, N., et al. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS ONE* **3**:e2521. doi:10.1371/journal.pone.0002521

Espinoza, C. A., Goodrich, J. A., and Kugel, J. F. (2007). Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. *RNA* **13**, 583–596. doi:10.1261/rna.310307

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571. doi:10.1038/ng.368

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth. Enzymol.* **266**, 418–427. doi:10.1016/S0076-6879(96)66026-1

Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P., and Kohtz, J. D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20**, 1470–1484. doi:10.1101/gad.1416106

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405. doi:10.1038/nrg2337

Glazko, G. V., Zybailov, B. L., and Rogozin, I. B. (2012). Computational prediction of polycomb-associated long non-coding RNAs. *PLoS ONE* **7**:e44878. doi:10.1371/journal.pone.0044878

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86. doi:10.1186/gb-2010-11-8-r86

Goodrich, J. A., and Kugel, J. F. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* **7**, 612–616. doi:10.1038/nrm1946

Gould, S. J., and Vrba, S. (1982). Exaptation – a missing term in the science of form. *Paleobiology* **8**, 4–14.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* **5**, 578–590. doi:10.1093/gbe/evt028

Grote, P., and Herrmann, B. G. (2013). The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol.* **10**, 1579–1585. doi:10.4161/rna.26165

Guttman, M., and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346. doi:10.1038/nature10887

Hickey, D. A. (1982). Selfish DNA: a sexual*ly-transmitted* nuclear parasite. *Genetics* **101**, 519–531.

Hirota, K., Miyoshi, T., Kugou, K., Hoffman, C. S., Shibata, T., and Ohta, K. (2008). Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**, 130–134. doi:10.1038/nature07348

Holdt, L. M., Hoffmann, S., Sass, K., Langenberger, D., Scholz, M., Krohn, K., et al. (2013). Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet.* **9**:e1003588. doi:10.1371/journal.pgen.1003588

Johnson, R., and Guigo, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20**, 959–976. doi:10.1261/rna.044560.114

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. doi:10.1016/S0168-9525(02)00006-9

Jurka, J. (2008). Conserved eukaryotic transposable elements and the evolution of gene regulation. *Cell. Mol. Life Sci.* **65**, 201–204. doi:10.1007/s00018-007-7369-3

Kapitonov, V. V., and Jurka, J. (2003). A novel class of SINE elements derived from 5S rRNA. *Mol. Biol. Evol.* **20**, 694–702. doi:10.1093/molbev/msg075

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488. doi:10.1126/science.1138341

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., et al. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**:e1003470. doi:10.1371/journal.pgen.1003470

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496. doi:10.1093/nar/gkh103

Kazazian, H. H. Jr., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., and Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166. doi:10.1038/332164a0

Kelley, D., and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107. doi:10.1186/gb-2012-13-11-r107

Kidwell, M. G., and Lisch, D. R. (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**, 1–24. doi:10.1554/0014-3820(2001)055[0001:PTEPDA]2.0.CO;2

Kolesnikov, N. N., and Elisafenko, E. A. (2010). Comparative organization and the origin of noncoding regulatory RNA genes from X-chromosome inactivation center of human and mouse. *Genetika* **46**, 1386–1391. doi:10.1134/S1022795410100200

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**, 2229–2235. doi:10.1101/gr.1589103

Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., et al. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8**:e1002841. doi:10.1371/journal.pgen.1002841

Loeb, D. D., Padgett, R. W., Hardies, S. C., Shehee, W. R., Comer, M. B., Edgell, M. H., et al. (1986). The sequence of a large L1Md element reveals a tandemly repeated 5' end and several features found in retrotransposons. *Mol. Cell. Biol.* **6**, 168–182.

Loewer, S., Cabili, M. N., Guttman, M., Loh, Y. H., Thomas, K., Park, I. H., et al. (2010). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.* **42**, 1113–1117. doi:10.1038/ng.710

Makalowska, I., Rogozin, I. B., and Makalowski, W. (2010). Genome evolution. *Adv. Bioinformatics* **2010**, 643701. doi:10.1155/2010/643701

Makalowski, W. (2000). Genomic scrap yard: how genomes utilize all that junk. *Gene* **259**, 61–67. doi:10.1016/S0378-1119(00)00436-4

Managadze, D., Lobkovsky, A. E., Wolf, Y. I., Shabalina, S. A., Rogozin, I. B., and Koonin, E. V. (2013). The vast, conserved mammalian lincRNome. *PLoS Comput. Biol.* **9**:e1002917. doi:10.1371/journal.pcbi.1002917

Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A., and Koonin, E. V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* **3**, 1390–1404. doi:10.1093/gbe/evr116

Mariner, P. D., Walters, R. D., Espinoza, C. A., Drullinger, L. F., Wagner, S. D., Kugel, J. F., et al. (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* **29**, 499–509. doi:10.1016/j.molcel.2007.12.013

Martens, J. A., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature* **429**, 571–574. doi:10.1038/nature02538

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., and Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666–670. doi:10.1038/nature05519

Martin, S. L. (2006). The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. *J. Biomed. Biotechnol.* **2006**, 45621. doi:10.1155/JBB/2006/45621

Mattick, J. S., and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* **15**, R17–R29. doi:10.1093/hmg/ddl046

Mattick, J. S., Taft, R. J., and Faulkner, G. J. (2010). A global view of genomic information – moving beyond the gene and the master regulator. *Trends Genet.* **26**, 21–28. doi:10.1016/j.tig.2009.11.002

Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., and Mattick, J. S. (2008). Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 716–721. doi:10.1073/pnas.0706729105

Miller, W. J., McDonald, J. F., Nouaud, D., and Anxolabehere, D. (1999). Molecular domestication – more than a sporadic episode in evolution. *Genetica* **107**, 197–207. doi:10.1023/A:1004070603792

Munroe, S. H., and Lazar, M. A. (1991). Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J. Biol. Chem.* **266**, 22083–22086.

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., et al. (2008). The air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717–1720. doi:10.1126/science.1163802

Ng, S. Y., Lin, L., Soh, B. S., and Stanton, L. W. (2013). Long noncoding RNAs in development and disease of the central nervous system. *Trends Genet.* **29**, 461–468. doi:10.1016/j.tig.2013.03.002

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573. doi:10.1038/nature01266

Orgel, L. E., Crick, F. H., and Sapienza, C. (1980). Selfish DNA. *Nature* **288**, 645–646. doi:10.1038/288645a0

Osato, N., Suzuki, Y., Ikeo, K., and Gojobori, T. (2007). Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* **176**, 1299–1306. doi:10.1534/genetics.106.069484

Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., et al. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* **32**, 232–246. doi:10.1016/j.molcel.2008.08.022

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1–5. doi:10.1016/j.tig.2005.10.003

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17**, 556–565. doi:10.1101/gr.6036807

Ponting, C. P., and Belgard, T. G. (2010). Transcribed dark matter: meaning or myth? *Hum. Mol. Genet.* **19**, R162–R168. doi:10.1093/hmg/ddq362

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641. doi:10.1016/j.cell.2009.02.006

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166. doi:10.1146/annurev-biochem-051410-092902

Robinson, R. (2010). Dark matter transcripts: sound and fury, signifying nothing? *PLoS Biol.* **8**:e1000370. doi:10.1371/journal.pbio.1000370

Rogozin, I. B., Mayorov, V. I., Lavrentieva, M. V., Milanesi, L., and Adkison, L. R. (2000). Prediction and phylogenetic analysis of mammalian short interspersed elements (SINEs). *Brief. Bioinformatics* **1**, 260–274. doi:10.1093/bib/1.3.260

Schuler, A., Ghanbarian, A. T., and Hurst, L. D. (2014). Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* **31**, 3164–3183. doi:10.1093/molbev/msu249

Sinzelle, L., Izsvak, Z., and Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell. Mol. Life Sci.* **66**, 1073–1093. doi:10.1007/s00018-009-8376-3

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743–748. doi:10.1016/S0959-437X(96)80030-X

Temin, H. M. (1985). Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* **2**, 455–468.

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550. doi:10.1016/j.cell.2011.11.055

Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., et al. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of polycomb group complexes. *Nat. Genet.* **36**, 1296–1300. doi:10.1038/ng1467

Van Bakel, H., and Hughes, T. R. (2009). Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic. Proteomic.* **8**, 424–436. doi:10.1093/bfgp/elp037

Wang, H., Iacoangeli, A., Lin, D., Williams, K., Denman, R. B., Hellen, C. U., et al. (2005). Dendritic BC1 RNA in translational control mechanisms. *J. Cell Biol.* **171**, 811–821. doi:10.1083/jcb.200506006

Wang, J., Gong, C., and Maquat, L. E. (2011). Control of myogenesis by rodent SINE-containing lncRNAs. *Genes Dev.* **27**, 793–804. doi:10.1101/gad.212639.112

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562. doi:10.1038/nature01262

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M., and Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389**, 52–65. doi:10.1016/j.gene.2006.09.029

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.