Retrotransposons and non-protein coding RNAs

Tobias Mourier and Eske Willerslev

Advance Access publication date 3 September 2009

Abstract

Retrotransposons constitute a significant fraction of mammalian genomes. Considering the finding of widespread transcriptional activity across entire genomes, it is not surprising that retrotransposons contribute to the collective RNA pool. However, the transcriptional output from retrotransposons does not merely represent spurious transcription. We review examples of functional RNAs transcribed from retrotransposons, and address the collection of non-protein coding RNAs derived from transposable element sequences, including numerous human microRNAs and the neuronal BC RNAs. Finally, we review the emerging understanding of how retrotransposons themselves are regulated by small RNAs.

Keywords: transposable elements; RNA genes; gene regulation

INTRODUCTION

Retrotransposons are genetic sequences that proliferate within genomes, and whose replication cycles include a reverse transcription step in which an RNA sequence is reverse transcribed into DNA. Broadly, retrotransposons can be divided into three groups, Long Terminal Repeat (LTR) elements, non-LTR elements (or LINEs, for Long INterspersed Elements) and Short INterspersed Elements (SINEs) (Figure 1) [1, 2].

LTR elements are large retrotransposons flanked by two LTR sequences. Transcription by polymerase II starts in the first LTR and ends in the second LTR. The central region contains one or two open reading frames (ORFs) encoding the products required for transposition [3]. Through recombination between two flanking LTR sequences, a solitary LTR sequence may be formed. The vast majority of human LTR sequences exist as solitary LTRs [4]. LTR elements are closely related to retroviruses and are widespread among eukaryotic genomes [5]. The human endogenous retrovirus, HERV-K shows human-specific insertions as well as within-human polymorphisms, suggesting very recent activity of this LTR element [6, 7].

LINE elements are not flanked by LTR sequences. Full-length members of the predominant mammalian LINE, L1, contain 2 ORFs. The first ORF encodes a nucleic acid binding protein [1, 8], and the ORF2 protein contains reverse transcriptase and endonuclease activity [9]. L1 elements contain an internal RNA polymerase II promoter in their 5′ end [9], as well as an antisense promoter that is able to transcribe neighbouring genes [10]. Furthermore, transcription start sites cluster in the 3′ end of L1 elements suggesting the existence of a non-canonical promoter [11]. LINEs are found in all major eukaryotic lineages, and recently have been described in Giardia lamblia, a protozoan believed to be one of the most primitive eukaryotes [12]. The L1 LINE has existed in the mammalian lineage for ~150 million years [4], and is currently the only active LINE in the human genome. Recent L1 activity has resulted in L1 polymorphisms between human populations [13, 14]. Large differences are found between mammalian genomes regarding the number and nature of their LINE elements. The human genome contains around 150 full-length L1 elements with intact ORFs [15] of which a minor fraction are believed to be responsible
for the majority of retrotransposition events [16]. Although the human and mouse genomes contain comparable numbers of L1 elements [4, 17], ten times as many full-length L1 elements with intact ORFs are found in the mouse genome [15], and retrotransposition rates appear to be considerably higher in inbred mice strains than in humans [18]. In contrast, LINEs have recently become extinct in some mammalian species [19, 20].

SINEs are short non-autonomous elements without ORFs, whose retrotransposition is presumably mediated by LINE elements [21, 22]. Some SINEs show similarity to LINE sequences in their 3' end, suggesting a mean to effectively interact with the enzymes provided by LINE elements [23, 24]. As different RNA transcripts may be targets for LINE-mediated retrotransposition, the populations of SINEs diverge extensively between mammalian genomes [25, 26]. The predominant primate SINE, Alu, contains a weak internal polymerase III promoter [27] that often is not sufficient to initiate transcription by itself [28]. Estimates of Alu activity suggest one new Alu insertion in approximately every 20 human births [29].

Although retrotransposons are widespread among eukaryotic genomes, a few recorded examples of retrotransposon-deprived genomes exist (e.g. [30, 31]). In a typical mammalian genome retrotransposon sequences make up 30–40% of the total amount of DNA [4, 17, 32]. Given the observation of widespread transcription throughout mammalian genomes [33, 34], it is perhaps not surprising if retrotransposons contribute significantly to the RNA pool of a mammalian cell. However, the RNA contribution from retrotransposons does not only represent simple spurious transcription. We here review a number of examples in which retrotransposon-derived RNAs are highly regulated in expression and of functional importance to the cell.

Analysis of retrotransposon transcription

Due to their intrinsic sequence similarity, large-scale transcriptome analyses are often biased against transposable elements, either in terms of experimental setup or analytical procedures [34, 35]. For example, whole-genome tiling arrays are usually restricted to the non-repetitive parts of the genome, and repetitive elements are often masked in genomic sequences before transcripts are mapped onto them. Therefore, transcriptional studies of retrotransposons are often derived from more specific techniques: reports of transcriptional activity from retrotransposons are
available from cDNA sequencing studies [36–38], from EST database searches [39], and from targeted experiments taking advantage of common sequence features of the retrotransposons [40]. Circumstantial evidence for retrotransposon transcriptional activity comes from the indications of apparent selection against transcriptionally active retrotransposons near protein-coding genes [41–43]. Technological advances in unbiased sequencing procedures are beginning to shed light on retrotransposon transcription. Using sequencing-by-synthesis technology, Lister and colleagues found increased transcriptional activity from transposable elements (albeit DNA transposons) in Arabidopsis methylation mutants [44]. More recently, a large-scale transcription tagging analysis revealed extensive—but regulated—transcription initiated within mammalian retrotransposons [11]. Still, novel sequencing techniques produce sequence reads of restricted lengths limiting the resolution by which reads can be assigned to individual retrotransposon loci. Furthermore, in developing the sequencing techniques, the obtained read lengths are likely to be optimized mainly for purposes such as assembly of sequence reads against a reference sequence and for the assignment of sequence reads to genes of interest. Hence, studies of the dynamics and transcriptional activity of retrotransposons may continue to face significant challenges despite current technological advances.

EXAMPLES OF ADAPTIVE RETROTRANSPOSON TRANSCRIPTION

Heat shock regulation by SINE RNAs

Heat shock increases the abundance of Alu RNAs transiently in human cell lines [26, 45]. More than a decade ago, it was shown that over-expression of Alu RNA increased the expression of a reporter gene construct [46]. Later, Hässler and Strub showed that Alu RNA stimulated the initiation of protein translation in vitro [47]. Yet, this is not the only level at which gene expression is regulated by Alu RNA: during heat shock, Alu RNAs (that are up-regulated) bind RNA polymerase II and inhibit the transcription of housekeeping genes [48]. A similar phenomenon is observed in mouse, although here involving RNA from B2, a prevalent rodent SINE [49, 50]. Remarkably, these two SINEs are unrelated, with Alus being derived from a 7SL RNA gene [27], and B2 SINEs derived from a tRNA sequence [51]. The heat shock response is a defence against cellular stresses, in which transcriptional activity of most genes are shut down and a repertoire of heat shock genes are up-regulated (which coincidentally is activated by another non-protein coding RNA) [52, 53]. Hence, RNAs from retrotransposons are essential parts in this cellular response in both mouse and humans [54], although the RNAs themselves are unrelated. One might speculate that the RNA gene that binds and inhibits RNA polymerase II can be changed relatively rapidly during evolution, and that an essential criterion for such an RNA is that it is transcribed by RNA polymerase III and abundant in the genome. A more general connection between heat shock and RNA polymerase III transcribed SINEs is supported by the finding that silk worm SINEs are up-regulated under cell stress [55], and the presence of a RNA polymerase III transcript in Tetrahymena thermophila during heat shock [56].

Retrotransposons and antisense transcription

The extent of naturally occurring antisense transcription has been highlighted by transcriptome data mapping studies [57]. One study found that transposable elements contribute considerably to the antisense transcription of human genes [58]. Interestingly, antisense transcription appears to be more prevalent from older elements, such as MIR and L2 (old SINE and LINE families, respectively) [58], although it is possible that the higher sequence similarity between younger elements introduces a bias in this analysis. The presence of natural antisense transcription is hypothesized to contribute to the regulation of the sense strand gene [57]. This has recently been confirmed for the human SLC4A8 and IFT172 genes, in which an intronic LTR sequence situated in the antisense orientation to the gene generates RNAs that modulate the mRNA levels of the genes [59].

The chromatin status around the mouse growth hormone (GH) gene is defined by boundaries created by the bi-directional transcription of a B2 SINE element [60]. However, this is a result of the transcriptional process itself rather than a functional role of the B2 transcripts [60], so in this case the effect of the retrotransposons does not appear to involve non-protein coding RNAs.
**Retrotransposon RNAs in human diseases**

Elevated expression levels of retrotransposon RNAs have been reported in a number of human diseases. However, it is not clear if the elevated activity signifies a functional response or merely represents a secondary effect associated with the imposed cellular stress. For example, infection with the HIV retrovirus results in increased transcription levels of endogenous retroviruses (i.e. LTR retrotransposons) [61]. It has nevertheless been suggested that polymerase III transcripts are directly implicated in human cancers by affecting cell transformation [62], and concurrently, increased expression of Alu RNA is observed in carcinoma tissues [63]. More generally, a number of human diseases have been linked to insertions of retrotransposons and recombination facilitated by retrotransposons [64, 65], or to the effect of proteins expressed by retrotransposons [66]. As none of these cases are directly linked to non-protein coding RNAs, these studies are not included in this review.

**RNA GENES DERIVED FROM RETROTRANSPOSONS**

Retrotransposons and retrotranspositional activity generate genomic variation in various ways. Retrotransposons carry functional sequences along with them that upon insertion may regulate their novel genomic context [67]. Furthermore, the reverse transcriptase activity provided by retrotransposons may generate processed pseudogenes by incorporating messenger-RNA information into the genome [22, 68]. Finally, individual retrotransposons may themselves be recruited—or exapted—to novel genetic functions [69, 70]. Here, we briefly review two examples of RNA genes derived from retrotransposons.

**BC RNAs**

BC200 (Brain Cytoplasmic RNA 200-nucleotides) is a small non-protein coding RNA that arose from a monomeric Alu and is found in the genomes of Anthropoidea (monkeys and apes) [71]. The existence of numerous BC200 pseudogenes suggests that the gene to some extent has retained transpositional activity [71, 72]. BC200 is specifically transcribed in neurons by polymerase III, and is actively transported to the dendrites [73]. In rodents, a similarly short (152 nucleotides) RNA, termed BC1, is also expressed specifically in neurones and transported to the dendrites where it—similarly to BC200—represses translation at the level of initiation [74–76]. Yet, in contrast to BC200, the BC1 gene is derived from a tRNA sequence [77]. Despite their different evolutionary history and the absence of apparent sequence similarity, BC1 and BC200 adopt similar secondary structures [78, 79]. The functional similarity between the BC RNAs is further underlined by the fact that in transgenic mice, BC200 is expressed in brain areas comparable to those in human [80].

The BC RNAs hence provide another example of apparent convergent exaptations where similar functions are carried out by different retrotransposon-derived genes in primates and rodents. This only underlines the dynamic nature of retrotransposons and the functional flexibility they provide to the host genome. There is no reason to expect that such phenomena are peculiar to rodents and primates, and future functional studies of model organism genomes are likely to reveal a plethora of such evolutionary tinkering mediated by retrotransposons.

**MicroRNAs**

MicroRNAs are short (~21–23 nucleotides) RNA sequences that regulate gene expression. Mature microRNAs are processed from a stem-loop precursor of around 70 nucleotides, which in turn may be excised from a longer primary transcript. By complementary base-pairing with mRNAs, the microRNAs may initiate the destruction of the mRNA or halt its translation (for reviews on microRNAs see e.g. [81, 82]). To date, more than 600 human microRNAs are known [83], a number that is likely to continue to increase. The insertion of multiple related retrotransposons may result in similar sequences residing in close proximity to each other, and in opposite orientations. Such sequence arrangements are—if transcribed—likely candidates for the production of stem-loop RNA structures that may potentially generate novel microRNAs. By searching microRNAs for retrotransposon sequence, Smalheiser and Torvik found that 13 mammalian microRNAs were derived from transposable elements [84]. In a later study (hence using a much larger number of microRNA sequences), 55 experimentally characterized human microRNAs were found to be derived from transposable elements (including DNA transposons) [85]. It should be stressed that the expression of
a retrotransposon-derived sequence that can adopt a stem-loop RNA structure cannot be taken as evidence for a \textit{bona fide} functional microRNA (see ref. [86, 87] for current guidelines on microRNA annotation). Yet, any microRNA derived from retrotransposon sequences is likely to display sequence similarity to related retrotransposons, which may then act as target sequences [88]. As retrotransposons are prevalent in the untranslated regions of mammalian messenger-RNAs, these may be seen as ideal candidate sequences for the generation of microRNAs and their targets, allowing a rapid evolution of regulatory networks [85]. Beside providing the raw sequence material for the generation of novel microRNAs, retrotransposons appear to have diversified a primate-specific family of microRNAs by facilitating ectopic recombination events [89].

**RNAs IN CONTROLLING RETROTRANSPOSON ACTIVITY**

Within the last decade, the involvement of RNA genes in regulatory processes has become more and more evident. Small RNAs form complexes with members of the Argonaute protein superfamily to repress gene activity [90]. Members of the Ago subfamily of Argonaute associate with microRNAs (above) and small interfering RNAs (siRNAs, 20–25 nucleotides long). MicroRNAs are encoded from distinct genomic loci as precursors forming a stem-loop structure that are then processed to mature microRNAs. In contrast, siRNAs are formed from double-stranded RNAs that may result from independent transcription of complementary sequences or strands. The Argonaute proteins of the Piwi subfamily form complexes with slightly larger (24–30 nucleotides) RNAs, called piRNAs for Piwi-interacting RNAs. In \textit{Drosophila}, the majority of siRNAs and piRNAs are targeting transposable elements. Transposon-targeting piRNAs are mainly found in the \textit{Drosophila} germ line, whereas transposon-targeting siRNAs are detectable in somatic tissue [91–93]. Interestingly, \textit{Drosophila} piRNAs are transferred from the mother to the progeny germ line, hence providing an epigenetic repression of transposon activity to the offspring [94]. In mammals, piRNAs were reported in mouse testis, although strikingly, retrotransposon sequences were underrepresented among the piRNAs [95, 96]. However, a later study revealed the presence of retrotransposon-matching siRNA in early stage (pre-pachytene) spermatocytes [97]. Both piRNAs and siRNAs corresponding to retrotransposon sequences are found in mouse oocytes [98, 99]. Retrotransposon activity was increased significantly when knocking out the proteins associated with either piRNAs or siRNAs, suggesting that both types of RNA-induced inhibition pathways are targeting retrotransposons in mouse oocytes [98]. In a more detailed experiment, it was found that siRNAs targeting LINEs were derived directly from bidirectional transcription of the LINE elements themselves [100].

Hence, small RNAs targeting retrotransposons are constraining the propagation of elements during early development in both invertebrates and mammals. In fact, the regulatory systems involving small RNAs allegedly evolved as a response to invading viruses and transposable elements [101]. The RNA-based repressor systems appear to predominantly be active in germ line or during early development. It is less well known what specific restrictions against retrotransposon activity are at work in somatic tissues. Retrotranspositional activity of human LINEs in rodent neuronal stem cells suggests that somatic activity may generate genetic differences between cells [102]. An observation that is further supported by the finding that members of the APOBEC3 protein family—which inhibit the mobilization of LINEs—are not expressed in human brains but in all other tissues examined [103].

**CONCLUDING REMARKS**

Retrotransposons provide a powerful mean for combining genetic components with novel genomic context, and throughout eukaryotic evolution retrotransposons have contributed to the generation of genomic novelties. Furthermore, retrotransposons are highly dynamic by nature, hence ensuring a continuous generation of genomic variation upon which selection can act.

Most novel insertions are likely to be deleterious or at best neutral. Yet, when the right component is combined with the right context, a retrotransposon (or more likely, a part thereof) may undergo exaptation and its genomic sequence or the derived RNA will expand the functional repertoire of the host. In this regard, the multitude of retrotransposons in a genome may be regarded as a simple hallmark of
the numerous unsuccessful (i.e. non-exapted) insertions. However, in some cases the sheer abundance of retrotransposons is likely to be a prerequisite for exaptation: origin of regulatory networks may be facilitated by the presence of a large number of highly similar sequences in the genome. In case of the heat shock response facilitated by SINE RNAs, this response should presumably be rapid and deliver effector sequences in large numbers. It is plausible that the multitude of SINE loci is crucial for exerting such an effect. Hence, exaptation will not only be dependent on a single fortuitous insertion, but also on the fact that insertions have taken place over the course of millions of years.

The recent in-depth study of 1% of the human genome surprisingly revealed that many functional sequences show little conservation across mammalian genomes [33]. The authors suggested: ‘the possibility of a large pool of neutral elements that are biologically active but provide no specific benefit to the organism. This pool may serve as a ‘warehouse’ for natural selection, potentially acting as the source of lineage-specific elements and functionally conserved but non-orthologous elements between species’ [33]. Considering the properties of retrotransposons, it seems that they should be essential contributors to such a pool, both in terms of DNA sequences and of RNA genes.

### Key Points
- Eukaryotic genomes are scattered with retrotransposon sequences.
- Cases of retrotransposon transcripts being of functional importance to the host are known.
- BC RNAs and several microRNAs are derived from retrotransposons.
- Small RNAs are important regulators of retrotransposon activity in eukaryotes.

### Acknowledgements
The authors thank Daniel Jeffares for critical reading of the article.

### FUNDING
Lundbeck Foundation grant [j.nr. 324/06] (to T.M.). Work at the Ancient DNA & Evolution Group is funded by the Villum Kann Rasmussen Foundation, the Danish Natural Science Research Council, and the Danish National Research Foundation.

### References


69. Brosius J. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 1999;238:115–34.


