BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 13 2013, pages 1593–1599
doi:10.1093/bioinformatics/btt258

Genome analysis

Advance Access publication May 6, 2013

Evidence for the dissemination of cryptic non-coding RNAs transcribed from intronic and intergenic segments by retroposition

Yoonsoo Hahn

Department of Life Science, Research Center for Biomolecules and Biosystems, Chung-Ang University, Seoul 156-756, Korea

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Insertion of DNA segments is one mechanism by which genomes evolve. The bulk of genomic segments are now known to be transcribed into long and short non-coding RNAs (ncRNAs), promoter-associated transcripts and enhancer-templated transcripts. These various cryptic ncRNAs are thought to be dispersed in the human and other genomes by retroposition.

Results: In this study, I report clear evidence for dissemination of cryptic ncRNAs transcribed from intronic and intergenic segments by retroposition. I used highly stringent conditions to find recently retroposed ncRNAs that had a poly(A) tract and were flanked by target site duplication. I identified 73 instances of retroposition in the human, mouse, and rat genomes (12, 36 and 25 instances, respectively). The inserted segments, in some cases, served as a novel exon or promoter for the associated gene, resulting in novel transcript variants. Some disseminated sequences showed sequence conservation across animals, implying a possible regulatory role. My results indicate that retroposition is one of the mechanisms for dispersion of ncRNAs. I propose that these newly inserted segments may play a role in genome evolution by potentially functioning as novel exons, promoters or enhancers.

Contact: yoonsoo.hahn@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 20, 2013; revised on April 30, 2013; accepted on May 1, 2013

1 INTRODUCTION

Insertion of DNA segments into a genome is one of the mechanisms underlying genome evolution. Inserted DNAs include genomic segments derived from other locations in the genome (Linardopoulou *et al*., 2005), repetitive elements (Kim and Hahn, 2011), organellar chromosome fragments, such as nuclear mitochondrial DNAs (Richly and Leister, 2004), and retroviruses (Doxiadis *et al*., 2008). Duplication of genomic segments is common in eukaryotic genomes. These duplications are produced either by DNA-mediated or RNA-mediated mechanisms. DNA-mediated duplication includes interchromosomal recombination and tandem duplication by non-allelic homologous recombination or non-homologous end-joining (Conrad and Hurles, 2007).

Many DNA insertion events are mediated by a retroposition mechanism, which involves transcription of genomic segments into RNA molecules, reverse transcription of these RNAs to cDNAs and insertion of the cDNA segments into new genomic locations (Marques *et al*., 2005). Retroposition of RNAs of various origins has contributed abundant innovations for genome evolution (Brosius, 1999, 2003). The most common retroposed DNA segments are retrotransposable elements, such as Alu, SVA, L1, MIR and endogenous retroviruses; the inserted segments can be exapted as novel exons for protein-coding genes (Krull *et al*., 2007; Lev-Maor *et al*., 2003), act as alternative promoters of nearby genes or induce *de novo* transcription (Kim and Hahn, 2010, 2011; van de Lagemaat *et al*., 2003) and contribute to regulatory elements, such as enhancers (Nishihara *et al*., 2006).

Retroposition of fully or partially processed protein-coding genes is a well-known mechanism for gene duplication, and results in generation of either non-functional retropseudogenes or functional retrocopies (retrogenes) of the original parental genes (Baertsch *et al*., 2008; Marques *et al*., 2005). Functional RNA molecules, such as small nucleolar RNAs (snoRNAs) and small Cajal body-specific RNAs (scaRNAs), are also to be subject to retroposition and have been reported to generate novel sno/scaRNAs during evolution (Schmitz *et al*., 2008; Vitali *et al*., 2003; Weber, 2006; Zemann *et al*., 2006).

It is now known that the bulk of human and other eukaryotic genomes, other than protein-coding genes or RNA genes, is transcribed pervasively (Birney *et al*., 2007; Brosius, 2005; Jacquier, 2009; Salta and De Strooper, 2012). Although a lot of these transcripts, which are collectively known as non-coding RNAs (ncRNAs) or untranslated RNAs (utRNAs), were thought to be simply transcriptional noise, some of them were reported to play important roles in genome regulation (Berretta and Morillon, 2009; Lakhotia, 2012; Wilusz *et al*., 2009). Enhancer regions have been reported to be bidirectionally transcribed to produce a distinct class of ncRNAs, enhancer-templated ncRNAs or eRNAs, which are involved in transcriptional control of nearby genes (Kim *et al*., 2010; Wang *et al*., 2011).

I hypothesized that various cryptic ncRNAs that were pervasively transcribed from intronic or intergenic segments would be subject to retroposition. There are many genomic duplicons that originated from intronic or intergenic regions in the human and other genomes. However, it is difficult to determine whether

these duplicons are generated by RNA-mediated retroposition or DNA recombination mediated by non-homologous end-joining. Retrogenes and retropseudogenes are easy to identify because they usually lack introns compared with their source genes. Retroposed intronic or intergenic segments, in contrast, would not show any splicing features, but would still exhibit other signatures of retroposition, namely, a poly(A) tract and target site duplication (TSD).

In this study, I devised a bioinformatics procedure to identify retroposed ncRNAs originating from intronic or intergenic segments in the human, mouse and rat genomes. The signature of retroposition such as the poly(A) tract decays over time (Grandi *et al.*, 2013). Loss of the retroposition signature would make it difficult to judge whether a duplicon originated by retroposition or DNA-mediated recombination. Therefore, in this study, I used highly stringent conditions to detect only recently retroposed ncRNAs.

To collect clear evidence for retroposition of ncRNAs, I designed a procedure to identify only recently retroposed segments that retained the following retroposition signatures: (i) strong sequence similarity between the source and the insert copies; (ii) a poly(A) tract at the end of the insert; and (iii) a TSD at both boundaries of the insert. By using these criteria, I identified 73 recent retroposition events of ncRNAs in the human, mouse and rat genomes. Furthermore, I comprehensively explored and discussed the impact of these retroposition events on nearby genes.

2 METHODS

2.1 Datasets and bioinformatics tools

Genome sequences and bulk annotation data for the human (hg19), mouse (mm9) and rat (rn4) were downloaded from the University of California Santa Cruz (UCSC) Genome Browser database (ftp://hgdownload.cse.ucsc.edu) in July 2012. Additional annotation information was accessed at the UCSC Genome Browser web server (http://genome.ucsc.edu) (Kuhn *et al.*, 2013). HOMER package version 3.16 (http://biowhat.ucsd.edu/homer/) was used to annotate the source and the target regions (Heinz *et al.*, 2010). Pairwise alignments of DNA sequences were generated using ALIGN software in the FASTA package version 20u66 (http://iubio.bio.indiana.edu/soft/iubionew/molbio/align/search/fasta/). In-house ad hoc PERL scripts were written and used to manipulate data files.

2.2 Identification of retroposed intronic and intergenic segments

To identify clear evidence for retroposition of ncRNAs, I developed a procedure to collect genomic sequences that showed distinct signatures of a recent retroposition event (Fig. 1). First, potential poly(A) tail sequences, which were listed as '(A)n' or '(T)n' repeats in the 'RepeatMasker' track of the UCSC Genome Browser database, were collected. There were 26 294, 55 836 and 29 105 '(A)n' and '(T)n' repeats in the human, mouse and rat genomes, respectively. Then, 'chainSelf' data, comprising segmental duplications of chromosomes, were analyzed to collect duplicons that were close (<10 nt apart) together with a putative poly(A) tail. Because I only wanted to collect recent retroposition events, only those duplicons that met the following conditions were selected: length difference <100 nt; a 'normalized score' of 'chainSelf' data >80; and the poly(A) tract associated with only one pair of duplicons. After this step, there were 142, 683 and 389 duplicon-associated poly(A) tracts in the human, mouse and rat genomes, respectively.

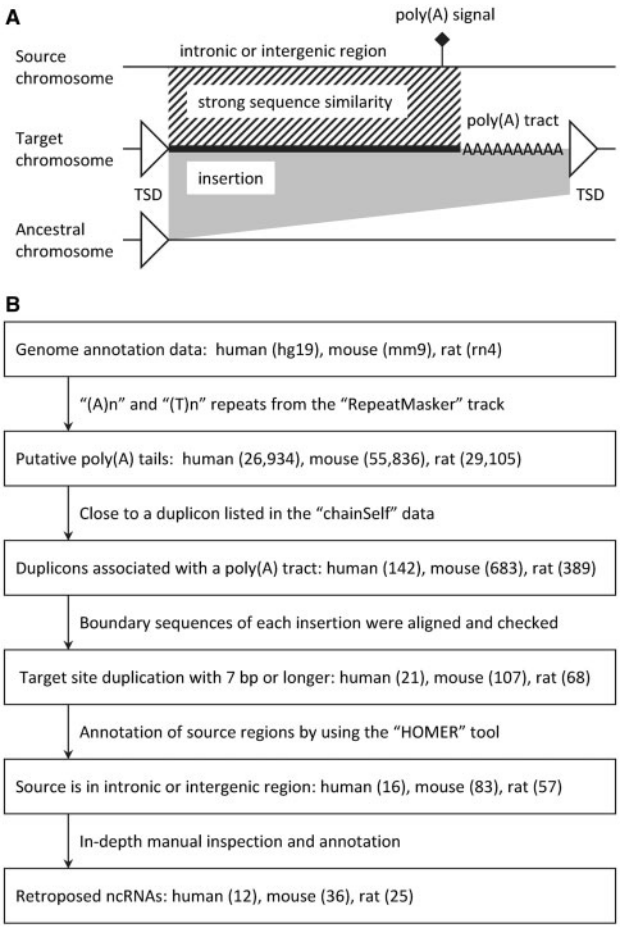

Fig. 1. Schematic representation of retroposition signatures and the procedure I used to identify recently retroposed ncRNAs. (**A**) Signatures of recently retroposed ncRNAs include a strong sequence similarity between the source and the insert (hatched box), a poly(A) signal in the source (diamond lollipop), a poly(A) tract in the insert and a target site duplication (TSD) (triangles). The gray trapezium indicates an insertion event compared with a hypothetical ancestral chromosome; this was assessed by comparison of orthologous regions of other species. (**B**) The procedure for identification of recently retroposed ncRNAs used in this study is depicted, with the number of data collected at each step indicated

Next, I checked whether the inserted sequence was flanked by a TSD. Boundary sequences, defined as 60 nt from each end of an insert centered at the boundary position, were extracted from the chromosome sequences. The two boundary sequences of each insert were aligned using the ALIGN program. Those cases with a TSD of 7 nt or longer were selected. After this step, there were 21, 107 and 68 TSD-flanked retroposition candidates in the human, mouse and rat genomes, respectively.

To select cases for which the source was intronic or intergenic regions, the source region of each case was annotated using the HOMER package. If the source was annotated as 'exon', '3′-UTR' or 'TTS', where TTS stands for transcription termination site, the case was discarded. In contrast, if the source was annotated as 'intron', 'intergenic' or 'promoter-TSS', where TSS stands for transcription start site, then the case was retained. After this step, there were 16, 83 and 57 final candidates from the human, mouse and rat genomes, respectively.

As the final step, I performed an in-depth manual inspection of the candidates. Those cases that met the following conditions were discarded: the source and the insert were two different retrogenes derived from a

single parental gene; more than half of the source sequence was masked as repetitive elements; the poly(A) tract was not derived from a poly(A) tail but a microsatellite sequence, such as '(AAAAG)n' (Subramanian *et al.*, 2003). Finally, 12, 36 and 25 recently retroposed ncRNAs were identified in the human, mouse and rat genomes, respectively. The source and target regions were annotated based on the information available at the UCSC Genome Browser web server.

3 RESULTS

3.1 Identification of retroposed ncRNAs in the human, mouse and rat genomes

To collect strong evidence for retroposition of ncRNAs, I developed and applied a procedure to identify genomic duplicons based on distinct signatures of recent retroposition events (Fig. 1). Genomic duplicons with a poly(A) tail at one end and flanked by a TSD of 7 nt or longer were collected. Then, those cases for which the source was intronic or intergenic segment were collected.

I identified 73 retroposed ncRNAs (12, 36 and 25, respectively) in the human, mouse and rat genomes (Supplementary Table S1). Detailed information on the retroposed segments is provided in Supplementary Dataset S1.

Analysis of sequences around the insertion break points showed the 5′-TTTT/A-3′ motif (Supplementary Fig. S1), which is a canonical sequence for the LINE-1 endonuclease target sites (Cost and Boeke, 1998). This observation suggests that the retroposition was LINE-1 endonuclease dependent.

The source of the retroposed segments included intronic regions of known genes probably derived from primary transcripts, promoters or 5′ flanking regions of known genes and intronic or intergenic regions potentially transcribed as a result of a cryptic promoter or a nearby repetitive element. A survey of the expressed sequence tag (EST) and RNA-seq data available at the UCSC Genome Browser revealed clear evidence for the transcription of the source segments. In eight cases, there were unspliced ESTs that spanned the source region. And almost all the human and mouse cases had at least one RNA-seq read that aligned to the source segment (Supplementary Table S2). Some retroposed sequences showed strong sequence conservation across animals, indicating a possible regulatory role.

Twelve cases were identified in the human genome (Supplementary Table S1). Nine of these were human-specific insertions that occurred in the human genome after the human–chimpanzee divergence. The source of the retroposed ncRNAs was intronic (eight cases) or 5′ flanking regions (one case) of known genes or intergenic regions (three cases). In seven cases, the inserted segment was associated with a known gene, including transcript variants of the *RPBJ* (ID H004), *CCNB3* (ID H011) and *TBC1D8B* (ID H012) genes.

Thirty-six recently retroposed ncRNAs were identified in the mouse genome (Supplementary Table S1). In 13 cases, the source was sequence associated with a known gene, whereas in the remaining 23 cases, the source was intergenic. In 21 cases, the inserted segment was present in a known gene; in one case, namely, the *Dab1* gene (ID M011), the inserted segment functions as an internal exon of a variant transcript.

I identified 25 recently retroposed ncRNAs in the rat genome (Supplementary Table S1). In 14 cases, the source was derived from a genomic region associated with a known gene. The 15 inserted segments were associated with a known gene. Two of these segments are part of the 3′ untranslated region (UTR) of the *Trim16* (ID R018) and *Wrb* (ID R019) genes, respectively.

3.2 Novel exons produced by the retroposed ncRNAs

DNA insertion in a gene could result in exaptation of the inserted segment as an exon, resulting in novel transcripts and protein variants (Kim and Hahn, 2011). Some of the retroposed ncRNAs identified in this study now function as novel exons in the associated gene, resulting in novel transcript variants.

The human *TBC1D8B* gene (ID H012) contains a retroposed ncRNA in intron 11 (Fig. 2). Within the inserted segment, there is an activated splice acceptor. A cryptic poly(A) site in the adjacent genomic region, which is preceded by two consecutive canonical poly(A) signals, was also activated on insertion. This resulted in generation of the 12th or terminal exon of the *TBC1D8B* transcript variant 2 (NCBI accession number NM_198881). The source of the retroposed ncRNA was the antisense transcript of *EBF1*. This retroposition occurred in the human genome after the human–chimpanzee divergence; hence, *TBC1D8B* transcript variant 2 is human specific, as described previously (Kim and Hahn, 2011).

In the case of the mouse *Dab1* gene (ID M011), ncRNA insertion produced a novel alternative internal exon, the 10th exon of a *Dab1* transcript variant (NCBI accession number Y08380) (Supplementary Fig. S2). The splice acceptor and donor sites of this exon are within the insert and in the adjacent region, respectively. It was suggested that this transcript variant would produce a 45 kDa isoform of the Dab1 protein in mouse cells (Howell *et al.*, 1997). However, the inserted exon contains a premature termination codon, which might make this transcript variant susceptible to nonsense-mediated mRNA decay (Amrani *et al.*, 2006).

The middle of the rat *Trim16* gene (ID R018) 3′-UTR contains a retroposed segment, the entire region of which is transcribed as part of the 3′ UTR (Supplementary Fig. S3). The retroposed ncRNA originates from the *Btf3* gene 5′ flanking region, which is divergently transcribed from the *Btf3* gene promoter. The *Btf3* gene has a CpG island promoter that may have bidirectional promoter activity, in common with many other CpG island promoters (Yang and Elnitski, 2008). Similarly, the 3′-end of the rat *Wrb* gene (ID R019) transcript contains a retroposed segment originating from an intron of the *C7* gene. This insertion extended the 3′-UTR of the *Wrb* gene by ~1 kb.

3.3 Promoters induced or affected by retroposed ncRNA

Some newly inserted DNA segments can induce *de novo* transcription of adjacent genomic segments (Kim and Hahn, 2010, 2011). Some of the retroposed ncRNAs identified in this study induce transcription of associated genes, producing novel transcript variants. An example is the human *RBPJ* gene (ID H004), where the novel transcript variant starts within the insert (Fig. 3). Retroposition occurred in intron 1 of the *RBPJ* gene transcript variant 4 (NCBI accession NM_203284). An EST (NCBI accession number DC397179) starts within this retroposed ncRNA, and the first exon of EST DC397179 is entirely derived from the insert. The second exon, which contains a start codon, and the last coding exons are shared with *RBPJ* transcript variant 4.

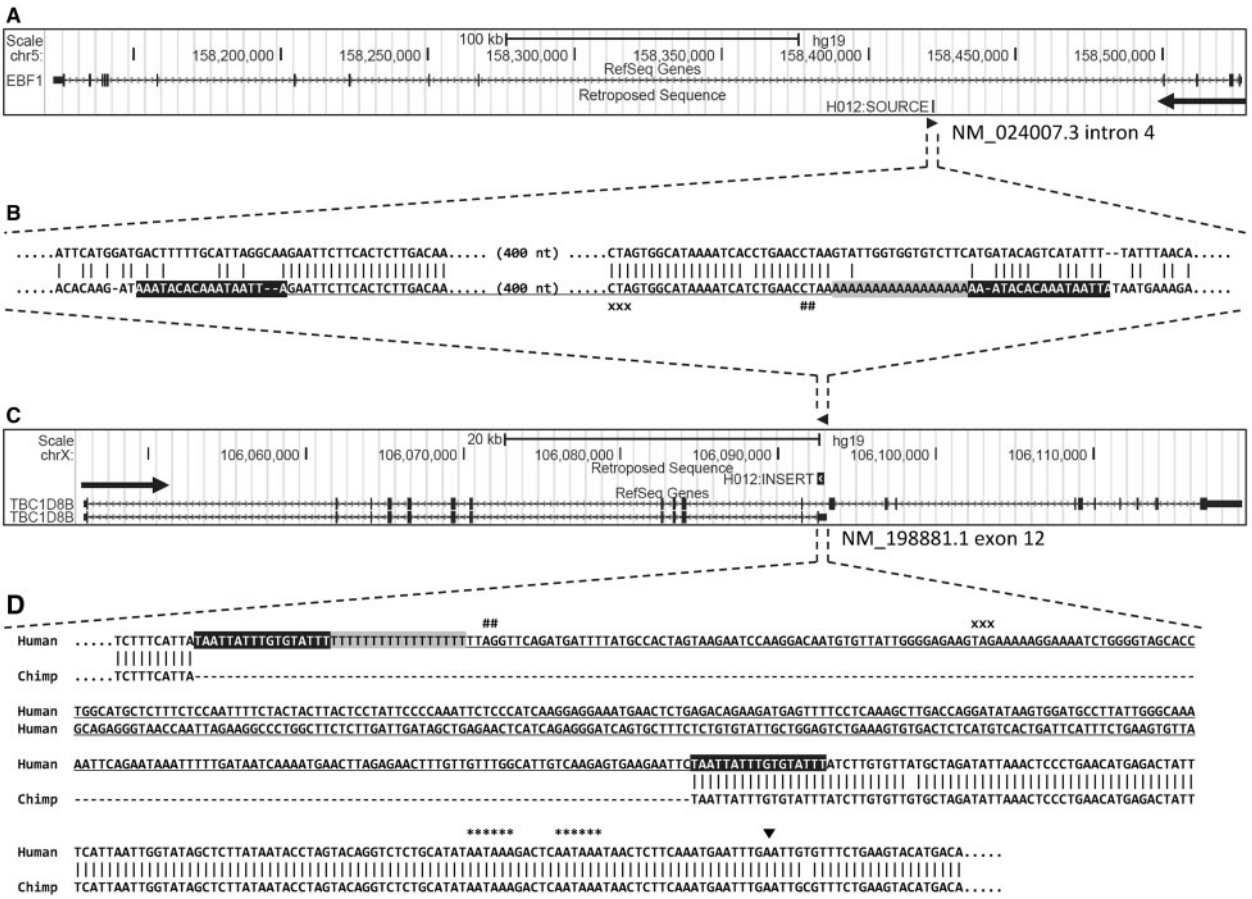

Fig. 2. The human *TBC1D8B* gene (ID H012) in which the retroposed ncRNA serves as the terminal exon of transcript variant 2. (**A**) The source is located within intron 4 of the human *EBF1* gene. The retroposed ncRNA was transcribed from the opposite strand of the *EBF1* gene. The transcription directions of the *EBF1* gene and the retroposed ncRNA are marked by big and small arrows, respectively. (**B**) Alignment of the source (top) and the target (bottom) sequences is shown. (**C**) The *TBC1D8B* gene locus contains the retroposed ncRNA. This insert is present in *TBC1D8B* variant 2 (RefSeq accession NM_198881) as terminal exon 12. (**D**) Alignment of the human and the chimpanzee *TBC1D8B* gene is shown. The inserted sequence is underlined. The poly(A) tail and the TSD at the target site are shown on gray and black background, respectively. The splice acceptor (AG), the stop codon (TAG), the two consecutive poly(A) signals (AATAAA) and the poly(A) site of *TBC1D8B* transcript variant 2 are marked with two number signs, three x marks, six asterisks and an upside-down triangle, respectively

Therefore, the novel transcript variant DC397179 will produce the same protein as the ancestral transcript variant 4. The source region is the short intergenic region of the *U2AF2* and *EPN1* genes, which are arranged in a tail-to-head manner: the 3′ flanking region of *U2AF2* and the 5′ flanking/promoter region of *EPN1*. However, the retroposed sequence is from the opposite strand of these two genes, indicating divergent transcription from the *EPN1* gene promoter, which may have bidirectional promoter activity. The ncRNA was inserted in the opposite orientation of *RBPJ*, so that the *EPN1* promoter-derived DNA is in the same direction as *RBPJ*. This insertion is shared by humans and chimpanzees, but it is not present in gorillas or other primates, indicating that the insertion occurred in the common ancestor of humans and chimpanzees.

If a retroposition event occurs close to the transcription start site of a gene, the insertion may affect transcription. In the human *CCNB3* gene (ID H011), the transcription start position for two RefSeqs, namely, NCBI accession numbers NM_033031 and MN_033670, is ~170 nt downstream of a retroposed ncRNA (Supplementary Fig. S4). Interestingly, there are two *CCNB3* ESTs, NCBI accession numbers DB071849 and DB063071, for which the promoter is situated ~58-kb upstream of the currently annotated *CCNB3* promoter. This observation raises the possibility that the promoter for these two ESTs could be the ancestral one, and that the promoter for the RefSeq transcripts is a novel promoter that was created by insertion. Otherwise, the insertion occurred in the 5′ flanking region of the promoter of the RefSeq transcripts. Currently, there is insufficient data to determine which scenario is correct. This insertion is present in humans, chimpanzees and gorillas, but not in orangutans or other primates, indicating that this insertion occurred in the common ancestor of African great apes.

3.4 Retroposed ncRNAs that are highly conserved across animals

Interestingly, 19 cases of retroposed ncRNAs identified in this study showed strong sequence conservation across animals,

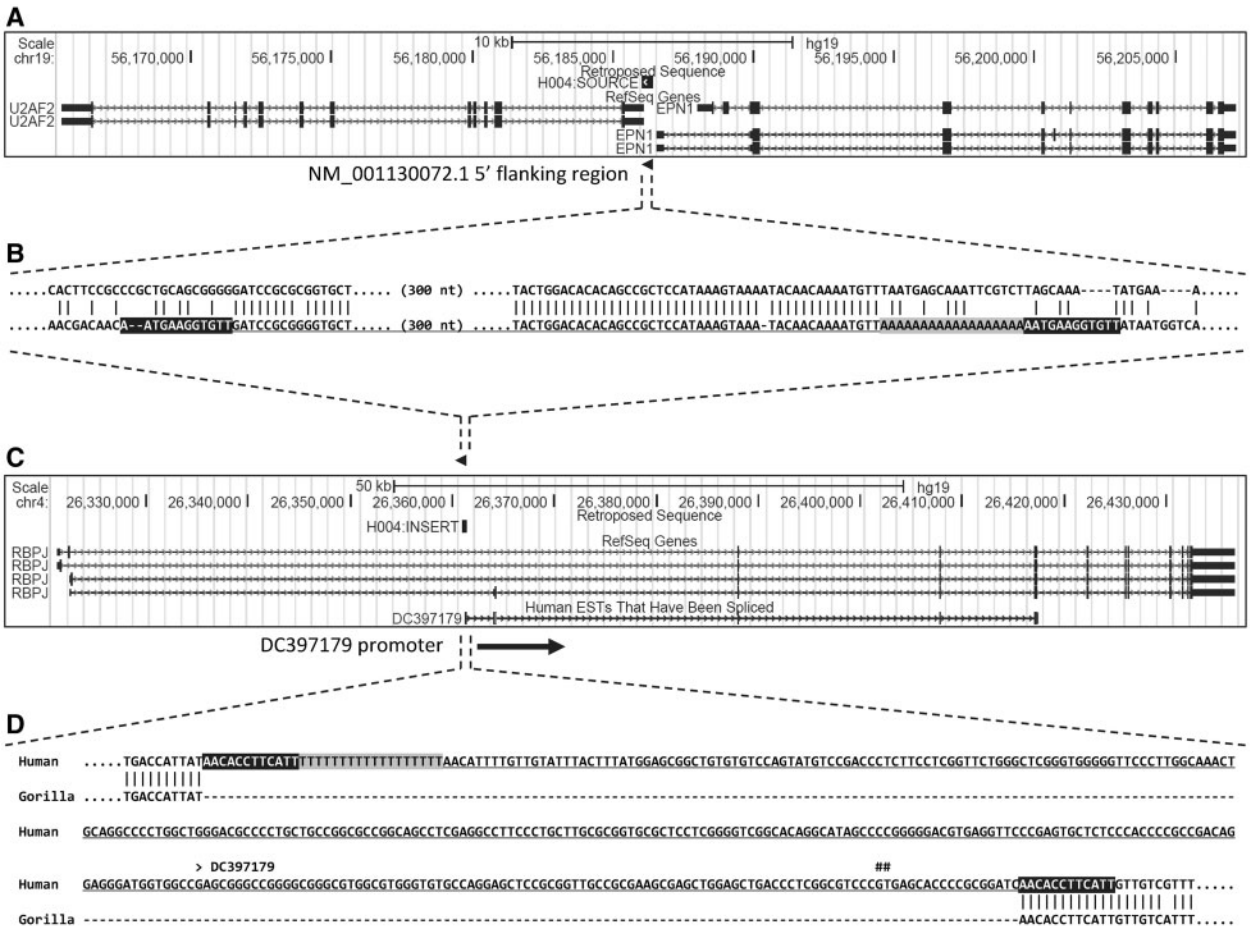

Fig. 3. The retroposed ncRNA induces transcription of the human *RPBJ* gene (ID H004). (**A**) The source is located in the intergenic region of the human *U2AF2* and *EPN1* genes. The retroposed ncRNA was transcribed in the reverse orientation compared with these two genes. (**B**) Alignment of the source (top) and target (bottom) sequences is shown. (**C**) The *RBPJ* gene locus contains retroposed ncRNA, within which EST DC397179 starts. (**D**) Alignment of the human and the gorilla *RBPJ* gene is shown. The inserted sequence is underlined. The poly(A) tail and the TSD at the target site are shown on gray and black background, respectively. The start position of EST DC397179 and the splice donor (GT) are marked with a greater than sign and two number signs, respectively

suggesting that they may play a regulatory role (see Supplementary Table S1 for more information). An example is the rat *Lnx2* gene (ID R020) shown in Figure 4. The source ncRNA is an intergenic region on chromosome 10 that exhibits strong sequence conservation across amniotes, including the chicken. Many conserved sequence elements have been reported to be actively transcribed into RNA molecules (Kapranov *et al*., 2007; Salta and De Strooper, 2012). For example, enhancer regions are transcribed to produce eRNAs, which are involved in transcriptional regulation of nearby genes (Kim *et al*., 2010; Wang *et al*., 2011). Therefore, there is a strong chance that conserved regions with regulatory potential could be mobilized by retroposition and that retroposed copies might affect the regulation of adjacent genes.

4 DISCUSSION

This report presents clear evidence for dissemination of cryptic ncRNAs by retroposition in the human, mouse and rat genomes. Because retroposed ncRNAs have no spliced-out segments, the gradual loss of retroposition signatures, such as a poly(A) tract and a TSD make them practically indistinguishable from DNA-mediated segmental duplicons. In this study, therefore, I used highly stringent conditions to collect 73 recent retroposition cases. Considering the rapid shortening of poly(A) tracts of retroposed transcripts (Grandi *et al*., 2013) and the decay of retroposition signatures over time, there are likely many more 'aged' retroposed ncRNAs that I did not detect using my screening procedure. There is also a distinct class of retroposons that lack poly(A) tails named tailless retropseudogenes (Schmitz *et al*., 2004), which also could not be detected in this study. Therefore, it is highly likely that the dissemination of ncRNAs is more common than observed here.

It has been shown that the bulk of the eukaryotic genome is pervasively transcribed (Birney *et al*., 2007; Jacquier, 2009). Several distinct classes of ncRNAs have been observed. Enhancer elements have been reported to be bidirectionally transcribed to produce eRNAs in human neuronal cells (Kim *et al*.,

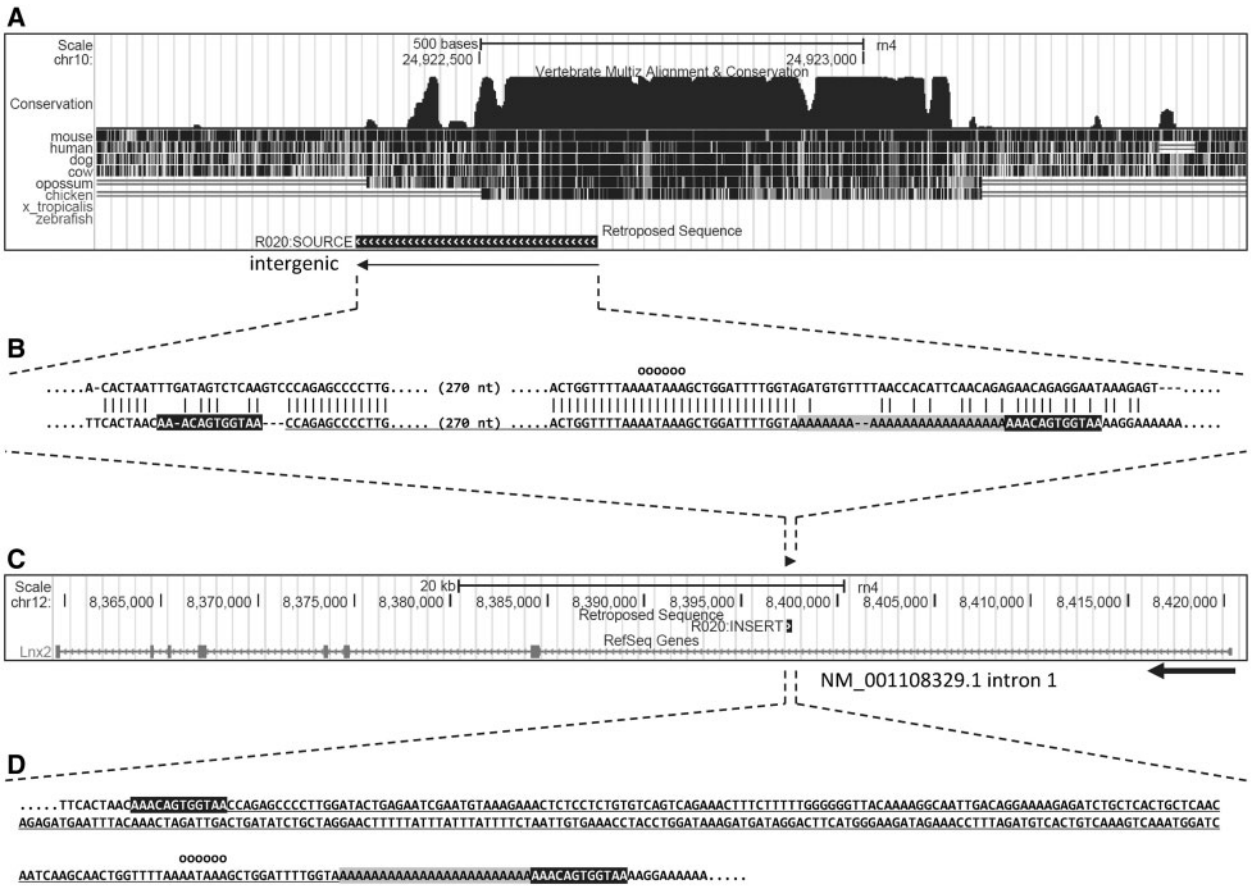

Fig. 4. Evolutionary conservation of the retroposed ncRNA in the rat *Lnx2* gene (ID R020). (**A**) The retroposed ncRNA is an intergenic region from rat chromosome 10 and is a part of a highly conserved region in mammals and chicken. (**B**) The alignment between the source (top) and the target (bottom) sequences is shown. (**C**) Retroposition occurred into intron 1 of the rat *Lnx2* gene. (**D**) The target region is shown. The inserted sequence is underlined. The poly(A) tail and the TSD at the target site are shown on grey and black background, respectively. The putative poly(A) signal for the source transcript is marked with six o marks

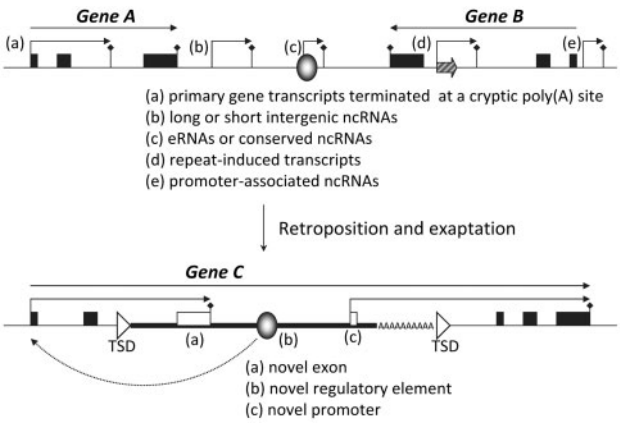

Fig. 5. Potential effects of retroposed cryptic ncRNAs on the insertion region are depicted

2010; Wang *et al*., 2011). Stable intronic sequence RNAs (sisRNAs) were detected in the nucleus of *Xenopus tropicalis* oocytes (Gardner *et al*., 2012). Polyadenylated promoter upstream transcripts are produced upstream of active human promoters (Preker *et al*., 2008). Many other types of promoter-associated RNAs have also been described (Taft *et al*., 2009). Short polyadenylated RNAs transcribed from highly conserved regions in the human genome have been reported (Kapranov *et al*., 2007). These ncRNAs, especially when polyadenylated, can serve as substrates for LINE-1 reverse transcriptase. Therefore, it is highly probable that these distinct classes of ncRNAs are disseminated in the genome by retroposition.

In this study, I identified the source of four retroposed ncRNAs inserted into the human *RBPJ* gene (ID H004, Fig. 3), mouse *Dab1* gene (ID M011, Supplementary Fig. S2), rat *Abcb1b* gene (ID R012) and rat *Trim16* gene (ID R018, Supplementary Fig. S3) as originating from the 5′ region of known genes, indicating that these were promoter-associated transcripts (Preker *et al*., 2008; Taft *et al*., 2009). The human *RBPJ* gene case is particularly interesting because the retroposed promoter-associated ncRNA retains promoter activity, generating a novel transcript variant of the *RBPJ* gene (Fig. 3).

Nineteen retroposed ncRNAs were highly conserved across animals: an example is the rat *Lnx2* gene case (ID R020, Fig. 4). A possible source of these retroposed sequences is short polyadenylated RNAs transcribed from conserved genomic

regions (Kapranov *et al*., 2007). Another possibility is that these sequences were derived from eRNAs (Kim *et al*., 2010; Wang *et al*., 2011), although there is no direct evidence that these regions have enhancer activities.

The steps for retroposition of ncRNAs are likely to be identical or similar to those of retroposition of retro(pseudo)-genes and repetitive elements (Fig. 5). First, the transcript is polyadenylated at an authentic or cryptic poly(A) site. Then, the polyadenylated RNA molecule is recognized by the enzymatic machinery of LINE-1 retrotransposons and inserted into a new location, generating a TSD during the insertion process. The inserted segments can serve as novel exons or promoters to produce novel transcript variants and novel protein isoforms. If the retroposed ncRNA originates from a regulatory element, such as an enhancer, it could potentially play a role in genome regulation.

In summary, I developed a procedure to identify recently retroposed ncRNAs in the human, mouse and rat genomes and provided clear evidence for retroposed ncRNAs. I strongly argue that the dissemination of ncRNAs by retroposition is common. The inserted segments can create novel transcript variants of the associated genes and play a role in genome regulation.

Funding: Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology [NRF-2012R1A1B3001513].

Conflict of Interest: none declared

REFERENCES

Amrani,N. *et al*. (2006) Early nonsense: mRNA decay solves a translational problem. *Nat. Rev. Mol. Cell Biol*., **7**, 415–425.

Baertsch,R. *et al*. (2008) Retrocopy contributions to the evolution of the human genome. *BMC Genomics*, **9**, 466.

Berretta,J. and Morillon,A. (2009) Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep*., **10**, 973–982.

Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Brosius,J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**, 115–134.

Brosius,J. (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica*, **118**, 99–116.

Brosius,J. (2005) Waste not, want not-transcript excess in multicellular eukaryotes. *Trends Genet*., **21**, 287–288.

Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. *Nat. Genet*., **39**, S30–S36.

Cost,G.J. and Boeke,J.D. (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, **37**, 18081–18093.

Doxiadis,G.G. *et al*. (2008) Impact of endogenous intronic retroviruses on major histocompatibility complex class II diversity and stability. *J. Virol*., **82**, 6667–6677.

Gardner,E.J. *et al*. (2012) Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. *Genes Dev*., **26**, 2550–2559.

Grandi,F.C. *et al*. (2013) LINE-1-derived poly(A) microsatellites undergo rapid shortening and create somatic and germline mosaicism in mice. *Mol. Biol. Evol*., **30**, 503–512.

Heinz,S. *et al*. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Howell,B.W. *et al*. (1997) Mouse disabled (mDab1): a Src binding protein implicated in neuronal development. *EMBO J*., **16**, 121–132.

Jacquier,A. (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet*., **10**, 833–844.

Kapranov,P. *et al*. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.

Kim,D.S. and Hahn,Y. (2010) Human-specific antisense transcripts induced by the insertion of transposable element. *Int. J. Mol. Med*., **26**, 151–157.

Kim,D.S. and Hahn,Y. (2011) Identification of human-specific transcript variants induced by DNA insertions in the human genome. *Bioinformatics*, **27**, 14–21.

Kim,T.K. *et al*. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.

Krull,M. *et al*. (2007) Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res*., **17**, 1139–1145.

Kuhn,R.M. *et al*. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform*., **14**, 144–161.

Lakhotia,S.C. (2012) Long non-coding RNAs coordinate cellular responses to stress. *Wiley Interdiscip. Rev. RNA*, **3**, 779–796.

Lev-Maor,G. *et al*. (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science*, **300**, 1288–1291.

Linardopoulou,E.V. *et al*. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, **437**, 94–100.

Marques,A.C. *et al*. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*., **3**, e357.

Nishihara,H. *et al*. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res*., **16**, 864–874.

Preker,P. *et al*. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, **322**, 1851–1854.

Richly,E. and Leister,D. (2004) NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol*., **21**, 1081–1084.

Salta,E. and De Strooper,B. (2012) Non-coding RNAs with essential roles in neurodegenerative disorders. *Lancet Neurol*., **11**, 189–200.

Schmitz,J. *et al*. (2004) A novel class of mammalian-specific tailless retropseudogenes. *Genome Res*., **14**, 1911–1915.

Schmitz,J. *et al*. (2008) Retroposed SNOfall-a mammalian-wide comparison of platypus snoRNAs. *Genome Res*., **18**, 1005–1010.

Subramanian,S. *et al*. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol*., **4**, R13.

Taft,R.J. *et al*. (2009) Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle*, **8**, 2332–2338.

van de Lagemaat,L.N. *et al*. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet*., **19**, 530–536.

Vitali,P. *et al*. (2003) Identification of 13 novel human modification guide RNAs. *Nucleic Acids Res*., **31**, 6543–6551.

Wang,D. *et al*. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390–394.

Weber,M.J. (2006) Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet*., **2**, e205.

Wilusz,J.E. *et al*. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev*., **23**, 1494–1504.

Yang,M.Q. and Elnitski,L.L. (2008) Diversity of core promoter elements comprising human bidirectional promoters. *BMC Genomics*, **9 (Suppl. 2)**, S3.

Zemann,A. *et al*. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res*., **34**, 2676–2685.