ACTA ABP BIOCHIMICA POLONICA

Vol. 63, No 4/2016
825–833
https://doi.org/10.18388/abp.2016_1354

Regular paper

Retroposition as a source of antisense long non-coding RNAs with possible regulatory functions#

Oleksii Bryzghalov*, Michał Wojciech Szcześniak*✉ and Izabela Makałowska

Department of Integrative Genomics, Institute of Antropology, Adam Mickiewicz University in Poznan, Poznań, Poland#

Long non-coding RNAs (lncRNAs) are a class of intensely studied, yet enigmatic molecules that make up a substantial portion of the human transcriptome. In this work, we link the origins and functions of some lncRNAs to retroposition, a process resulting in the creation of intronless copies (retrocopies) of the so-called parental genes. We found 35 human retrocopies transcribed in antisense and giving rise to 58 lncRNA transcripts. These lncRNAs share sequence similarity with the corresponding parental genes but in the sense/antisense orientation, meaning they have the potential to interact with each other and to form RNA:RNA duplexes. We took a closer look at these duplexes and found that 10 of the lncRNAs might regulate parental gene expression and processing at the pre-mRNA and mRNA levels. Further analysis of the co-expression and expression correlation provided support for the existence of functional coupling between lncRNAs and their mate parental gene transcripts.

Key words: lncRNAs, long non-coding RNAs, retroposition, retrocopies, antisense transcription, RNA:RNA duplexes

Received: 04 June, 2016; revised: 30 June, 2016; accepted: 19 July, 2016; available on-line: 02 November, 2016

INTRODUCTION

In higher eukaryotes, non-coding RNAs, such as miRNAs (microRNAs) and lncRNAs (long non-coding RNAs), represent considerable portions of the transcriptome, with the latter class represented by 28 031 transcripts in humans (Ensembl 83), compared to 79 930 protein-coding transcripts. Some other sources provide even higher numbers of these RNAs, such as NONCODE (Zhao *et al*., 2016), with 141 353 lncRNAs. This abundance of lncRNAs sparked interest in deciphering their functions, origins and evolution. However, the tasks appear to be even more demanding than in the case of protein-coding genes, and as a result, the vast majority of lncRNAs has no biological role assigned. In particular, due to poor evolutionary conservation of their sequences, homology-based functional assignment can be applied to only a small subset of lncRNAs. Additionally, detailed studies of the selected lncRNAs, such as HOTAIR (Tsai *et al*., 2010), ANRIL (Yap *et al*., 2010), and ZEB2-NAT (Beltran *et al*., 2008), indicate a high heterogeneity of their modes of action, making the *in silico* functional studies quite inaccurate. The accumulated data associate lncRNAs with biological processes such as transcription, splicing, translation, protein localization, cell cycle and apoptosis. They have also been linked to a number of human diseases, including cancers. It is possible that a large portion of lncRNAs has no biological role and represent a mere transcriptional noise or that the act of their transcription itself has a biological meaning, rather than their sequence does (Kornienko *et al*., 2013). Regarding the modes of action, a number of scenarios has been proposed, with transcriptional regulation being the best studied and being achieved through several mechanisms, such as promoter modifications, creating a permissive chromatin environment or binding transport factors to inhibit the nuclear localization of specific transcription factors (Kugel & Goodrich, 2012). In contrast to transcription-related mechanisms, little is known about the roles lncRNAs play upon base-pairing with fully or partially complementary mate mRNAs. In that scenario, lncRNAs could affect the stability, processing and expression levels of other transcripts (Geisler & Coller, 2013). One possibility is modulating the pre-mRNA splicing by splice site masking and subsequent blocking of the spliceosome assembly, which requires an extensive complementarity with a regulated pre-mRNA molecule. Such complementarity occurs by definition between the natural *cis* antisense transcripts (*cis*-NATs), but interactions *in trans* are also possible. Several lncRNAs are known to be involved in this type of regulation. For instance, it was shown that NAT's influence the splicing patterns of mRNAs at the neuroblastoma MYC, c-ErbAalpha and ZEB2 loci in mammals (Beltran *et al*., 2008). In the case of neuroblastoma MYC and c-ErbAalpha, this was suggested to be achieved through formation of RNA:RNA duplexes, which then inhibit splicing. At the ZEB2 locus, lncRNA expression inhibits splicing of an intron that contains an internal ribosome entry site (IRES). Translation of ZEB2 relies on this IRES; therefore, expression of the NAT indirectly facilitates expression of ZEB2 protein. In addition to splicing modulation, other regulatory mechanisms triggered by lncRNA:RNA duplexes are possible in humans, and they include adenine to inosine RNA editing at dsRNA regions, mRNA stability control by abrogation of miRNA-induced repression and guiding protein-coding genes to degradation within a Staufen-mediated decay (SMD) pathway (Geisler & Coller, 2013). Recently, the potential

✉e-mail: miszcz@amu.edu.pl
*These authors contributed equally to this work
#Information about a preliminary report on the same subject presented at scientific meetings: Oleksii Bryzghalov, Michał Szcześniak, Izabela Makałowska. Polish Evolutionary Conference 2015, Poznan, Poland. *Potential roles of retroposition-derived lncRNAs in splicing regulation* (poster); book of abstracts, page 36.
#Formely: Department of Bioinformatics, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University in Poznan, Poland.
Abbreviations: lncRNAs, long non-coding RNAs; miRNAs, microRNAs; NATs, natural antisense transcripts; IRES, internal ribosome entry site; SMD, Staufen-mediated decay; hnRNPs, heterogeneous nuclear ribonucleoproteins; asRNAs, antisense RNAs

of lncRNAs to exert regulatory roles through RNA:RNA base-pairings has been assessed for the human transcriptome (Szczesniak and Makałowska, 2016) and for several model plant species (Szczesniak *et al.*, 2016).

In this work, we scanned the human lncRNAs to identify those transcribed in antisense to retroposition-derived copies (retrocopies) of protein-coding genes. In retroposition, an mRNA molecule is reversely transcribed into cDNA, which occasionally becomes inserted into the genome at a random location (Fig. 1A). The resulting new copy (retrocopy) of the so-called parental gene typically is not functional because it lacks the core promoter and other regulatory sequences that would enable its transcription. In some cases, however, retrocopies use upstream promoters, either new ones (exaptation of cryptic promoter sequences) or the ones from other genes; such new genes, called retrogenes, constitute ca. 7.4% of the human gene set. They might evolve functions other than those of parental genes (neofunctionalization), play the same roles but with different spatio-temporal pattern (subfunctionalization) or replace the parental gene (orphan retrogenes) (Ciomborowska *et al.*, 2013). Finally, some of them are transcribed from the antisense strand, resulting in production of long non-coding RNAs, as shown in this study. These lncRNAs are expected to have functions other than the corresponding retrocopies or parental genes due to the lack of sequence similarity in the sense/sense orientation. A key to understanding their functions might be the observation that as a consequence of their origin, these lncRNAs are fully or partially complementary to their parental genes and are thus able to interact with each other at the RNA level. Keeping this in mind, we performed *in silico* base-pairing of the antisense lncRNAs with their parental genes and tried to determine whether these results could be linked to the abovementioned functions of RNA:RNA interactions. We found 10 lncRNAs transcribed in antisense to retrocopies and predicted to modulate processing and expression of their parental genes (Suppl. Table 1 at www.actabp.pl). The subsequent analysis of co-expression, expression correlation and sequence conservation led us to the conclusion that retroposition, already known to be one of the most important processes shaping mammalian genomes, might also contribute to the evolution of antisense lncRNAs and be a key to understanding their biological roles.

MATERIALS AND METHODS

Data download. The GENCODE 24 (Derrien *et al.*, 2012) annotation data for human (*Homo sapiens*), mouse (*Mus musculus*) and chimp (*Pan troglodytes*) were downloaded from the Ensembl release 83 (Herrero *et al.* 2016) using BioMart. To obtain long non-coding RNAs, only sequences classified as *3prime_overlapping ncrna*, *antisense*, *lincRNA*, *macro_lncRNA*, *retained_intron*, *sense_intronic*, or *sense_overlapping* were kept. Retrocopy-associated data were obtained from the RetrogeneDB (Kabza *et al.*, 2014). The retrocopies that are known to exist in Ensembl and have assigned Ensembl gene IDs were mapped to Ensembl release 83 using the biomaRt R package, which enabled access to updated, cross-release information on the genes, including transformation of the genomic coordinates from the human genome version hg19 to hg38. The retrocopy genes that are present only in the retrogeneDB were transformed into hg38 coordinates using the LiftOver tool available at the UCSC Genome Browser website (Speir *et al.*, 2016). Retrocopies that could not be mapped to Ensembl or failed coordinates transformation were eliminated from further steps. As a result, the original set of 4927 human retrocopies from the RetrogeneDB was reduced to 4675 loci (Fig. 2). For gene expression analysis, pre-calculated expression estimates from 153 stranded RNA-Seq libraries (Suppl. Table 2 at www.actabp.pl) were downloaded from ENCODE (ENCODE Project Consortium 2012).

***Ab initio* transcriptome assembly for chimp**. *Pan troglodytes* genome and annotation data in the GTF format were downloaded from Ensembl 83 (Herrero *et al.*, 2016). Nineteen stranded RNA-Seq libraries were downloaded from the Sequence Read Archive database (Kodama *et al.*, 2012) in the FASTQ format (Suppl. Table 3 at www.actabp.pl). The reads were filtered for quality, and adapters were trimmed using Trimmomatic (Bolger *et al.*, 2014). For quality filtering, the following parameters were used: LEADING: 20, TRAILING: 20, SLIDINGWINDOW: 5:20, and MINLEN: 50. Additionally, reads mapping to rRNA sequences were discarded using Bowtie 2 (Langmead & Salzberg, 2012). The processed paired-end reads were then mapped to the chimp genome with HISAT (Kim *et al.*, 2015) using the following settings: -X 1000, --rna-strandness RF, and --phred33, in addition to the splice site data from Ensembl. The resulting SAM file was then converted to the BAM format and sorted with SAMtools (Li *et al.*, 2009). Finally, StringTie (Pertea *et al.*, 2015) was used to assemble the transcriptome using known annotations in the GTF format as a reference. The procedure was repeated for each sequencing library, resulting in 19 GTF files. The files were then merged with Cuffmerge from the Cufflinks suite (Trapnell *et al.*, 2010). Using a custom Python script, transcript sequences in the FASTA format were retrieved from the resulting merged GTF file.

Identification of chimp long non-coding RNAs. The obtained GTF file was compared with known annotations from Ensembl using Cuffcompare (Trapnell *et al.*, 2010), and Cufflinks class codes were assigned to the transcripts. All transcripts with class code "s" were discarded because they are likely to result from mapping errors. For the class codes "=", "j", "c", "e", "o", and "p", the newly assembled transcripts are identical to the known transcripts or share part of their sequence; therefore, we used the available annotations to filter them. Briefly, transcripts belonging to the following categories were removed: *miRNA*, *Mt_rRNA*, *Mt_tRNA*, *protein_coding*, *rRNA*, *snoRNA*, and *snRNA*. Additionally, transcripts shorter than 200 bases were removed to accommodate a commonly used threshold for lncRNA length. Then, BLAST (Altschul *et al.*, 1990) search against *Pan troglodytes* ncRNAs from Ensembl was performed using an E-value threshold of 1e-5, and sequences that showed high similarity to miRNAs, mitochondrial rRNAs, mitochondrial tRNAs, rRNAs, snoRNAs, or snRNAs were discarded. Then, the coding potential of the remaining transcripts was assessed with CNCI (Sun *et al.*, 2013) using -m 50 and -S parameters and with CPC using the default settings (Kong *et al.*, 2007). For both tools, transcripts with an assessed coding potential higher than 0.0 were discarded. The protein-coding potential was also checked with TransDecoder (http://transdecoder.github.io/) in three steps. First, all open reading frames of at least 50 amino acids were identified with TransDecoder.LongOrfs. Then, their similarity to known proteins was checked in two ways. The peptides were subjected to search against Swiss-Prot (UniProt Consortium, 2015) proteins with BLASTP from the BLAST+ package using the following criteria, as suggested on the tool's website:

-max_target_seqs 1, -outfmt 6, -evalue 1e-5. Additionally, the PFAM profile-HMM database was searched with hmmscan from the HMMER-3 package (http://hmmer.org/) to identify common protein domains. In the third step, the TransDecoder.predict utility was run to obtain only high-confidence proteins based on the BLASTP and hmmscan results, as well as a built-in model for protein classification. Sequences that passed all filtering steps and were not recognized as protein-coding by TransDecoder were classified as long non-coding RNAs.

Expression analysis. To identify lncRNAs co-expressed with the corresponding parental genes in humans, expression data from ENCODE at the gene level was used (Suppl. Table 2 at www.actabp.pl). In this analysis, only gene pairs with expression values >0.1 TPM in at least one sample were considered to be co-expressed. Expression correlation analysis was performed in R using the same data and requiring that the Spearman's rank correlation coefficient was >0.6 or <–0.6 and the *p*-value <0.05. Both genes were required to have expression values of 0.2 TPM or higher; otherwise, that particular sample was removed from the correlation testing.

Identification of lncRNA-RNA interactions and their possible functions. The lncRNA interactions with their parental genes were predicted using a recently described strategy that is proven to achieve good performance and is able to identify experimentally validated RNA:RNA duplexes (Szczesniak & Makałowska, 2016). Briefly, it uses *lastal* from the LAST package (Kiełbasa *et al*., 2011) with a custom substitution matrix that allows G:U (wobble) pair consideration. Additionally, a mismatch is scored -6, gap opening -20, and gap extension -8. Using this tool, mRNAs and pre-mRNAs (i.e., unspliced transcripts that contain introns) of parental genes were compared against lncRNAs. The pre-mRNA sequences were modified so that any intronic sequences located more than 250 bases from the 3' or 5' splice sites were masked with *N* characters. Then, to assign potential functions to the identified interactions, we followed a previously proposed methodology (Szczesniak & Makałowska, 2016), which takes the following mechanisms into consideration: splicing regulation through masking splicing signals, abrogation of miRNA-dependent regulation, guiding protein-coding transcripts to the SMD pathway, and triggering mRNA editing events.

Other procedures. To identify antisense lncRNA-retrocopy pairs across human, mouse and chimp genomes, the BEDTools *intersect* utility from the BEDTools suite v.2.16.1 (Quinlan & Hall, 2010) was used with the requirement that at least 25% of an lncRNA sequence is overlapped by a retrocopy in a sense/antisense orientation. Conservation analysis for human pairs of antisense lncRNA-retrocopy overlaps was performed in R using human, chimp and mouse 1-to-1 orthology data from Ensembl as the input. An overlap was considered conserved if the human, antisense-transcribed retrocopy had an ortholog in chimp and/or in mouse. Data plotting was performed with custom R scripts using the following libraries: *plyr*, *ggplot2*, *scales*, and *plotly*.

RESULTS AND DISCUSSION

In this work, we took a closer look at lncRNAs transcribed in antisense to retrocopies, focusing on possible RNA:RNA interactions between them and the corresponding parental genes, both at the mRNA and pre-mRNA levels (Fig. 1B, C). To achieve this, we first

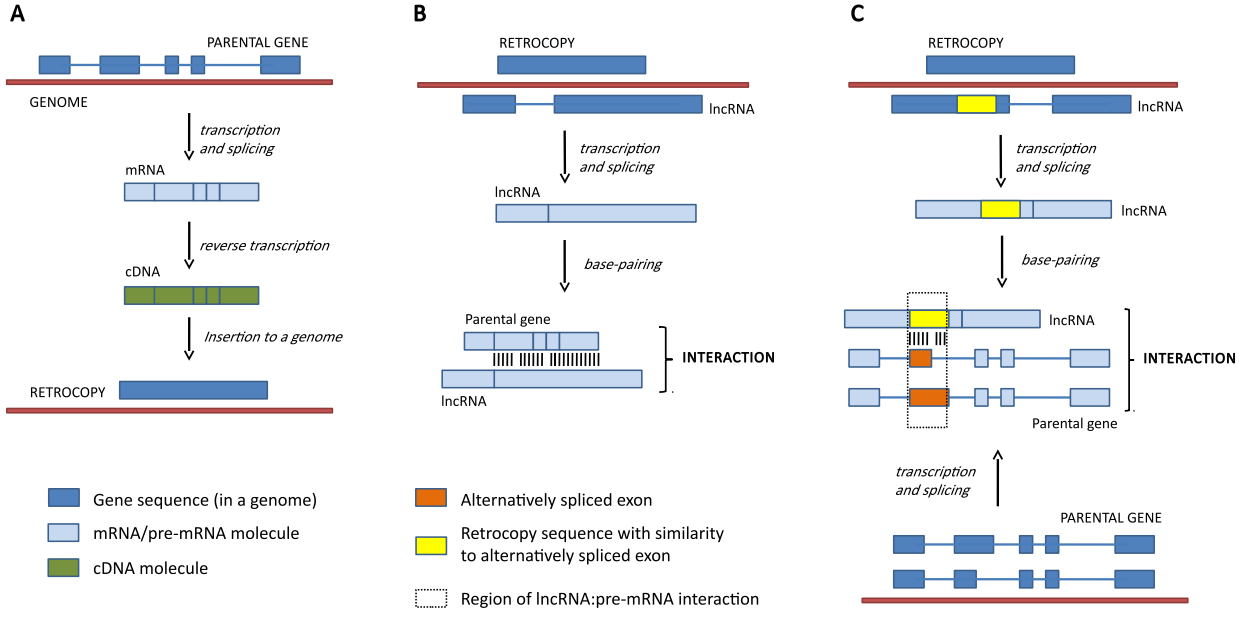

Figure 1. (A) Schematic representation of the retroposition process. (B) Mechanism behind the creation of lncRNA base-pairings with parental genes at the mRNA level. Once a retrocopy is transcribed in the antisense orientation, the resulting lncRNAs share sequence similarity with the parental genes in the sense/antisense orientation, meaning they are able to interact and form RNA:RNA duplexes with possible regulatory implications. (C) Evolutionary mechanism that enables the formation of lncRNA interactions with the pre-mRNAs of parental genes.
A retrocopy is created from one of many splice forms of the parental gene. Its antisense lncRNAs are complementary to the pre-mRNAs of the parental gene. Although retrocopies typically are devoid of introns, some of the retroposition-derived lncRNAs are able to base-pair with intronic parts of the parental gene's pre-mRNAs and mask the intronic splicing signals. This is possible if the process of retroposition and the formation of lncRNA:RNA duplexes engages different splice forms of the parental gene, as shown in the figure.

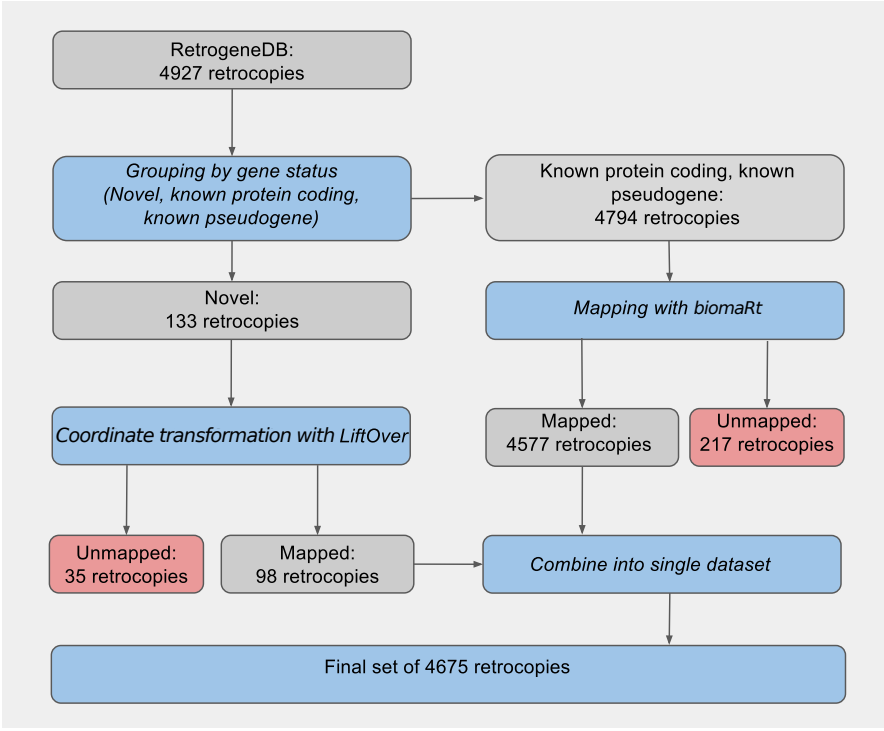

Figure 2. Pipeline for preparation of the dataset with human retrocopies.
Gene ID mapping between Ensembl releases was performed with the biomaRt R package or using LiftOver for retrocopies absent from Ensembl; the resulting two sets of retrocopies were merged into a single dataset.

collected a set of 4,675 human retrocopies from retrogeneDB and 57,145 lncRNAs (25,296 genes) from Ensembl. Using BEDTools *intersect*, we found 58 lncRNAs that were transcribed in antisense to 35 retrocopies. With Ensembl's annotation data for chimp, we found no antisense lncRNA-retrocopy overlaps. We attributed this to the quality of the available data; therefore, we re-annotated the chimp transcriptome, taking advantage of stranded RNA-Seq data available in the NCBI's Sequence Read Archive database (Suppl. Table 3 at www.actabp.pl). Altogether, we identified 167182 transcripts belonging to 101427 genes, including 36010 lncRNA transcripts (14-fold more than in Ensembl). With these new data, we discovered 23 antisense lncRNA-retrocopy overlaps. We also identified 6 antisense overlaps in mouse (Ensembl's annotation data).

Antisense transcripts of retrocopies are quite poorly conserved

The evolutionary conservation of human lncRNA transcription in antisense to retrocopies was tested by comparing human cases with the corresponding chimp and mouse homologs for retrocopies and checking whether there is antisense lncRNA transcription, like in human. We found 8 homologs in mouse and 6 in chimp; however, only one, in mouse, had antisense lncRNAs. We assumed this observation could be partially attributed to poor annotation of lncRNAs in chimp and mouse because Ensembl has only 2586 chimp lncRNAs, as opposed to 28031 for human. To obtain more reliable cross-species comparison, we performed *de novo* assembly of the chimp transcriptome using an extensive set of 19 stranded RNA-Seq libraries, followed by lncRNA identification, which resulted in a set of 36010 lncRNAs. With this new dataset, we found 23 retrocopies with antisense lncRNAs, as opposed to no cases found for the Ensembl data. However, none of them were conserved in humans, showing that antisense transcription of retrocopies is poorly conserved across the analyzed species. This is not surprising because a large fraction of lncRNAs represents species-specific transcripts, and approximately 60–70% are not detectable outside of primates (Necsulea *et al.*, 2014; Washietl *et al.*, 2014; Derrien *et al.*, 2012). However, the poorly resolved orthology relations for retrocopies and their relatively low conservation across species are also a factor: we were able to find 1-to-1 orthologs in chimp and mouse only for ca. 20% of all human retrocopies, which considerably reduced the chances for finding conserved antisense lncRNA-retrocopy pairs. Considering the facts listed above, we manually checked all previously identified 1-to-1 orthologs for human retrocopies with antisense lncRNAs and found a mouse ortholog of the DNAJB8 retrocopy, which also had lncRNAs transcribed in antisense. We used Clustal Omega (Sievers & Higgins, 2014) to align these antisense transcripts with the corresponding human lncRNA and found that the sequence identity is only 46%. Moreover, both mouse antisense RNAs overlap the translated sequence of DNAJB8, while human antisense transcript overlaps only a 5'UTR region (Fig. 3). These observations led us to the startling conclusion that orthologous retrocopies might possess antisense lncRNAs that originated independently and therefore are not orthologous.

Selected antisense lncRNAs show correlation of expression with their parental genes

Considering the lack of conservation of antisense lncRNA-retrocopy pairs, we aimed to provide more support for the supposed functionalities by analyzing the expression values of RNAs that are expected to interact. First, we checked whether they are co-expressed by analyzing human expression data from 153 strand-specific

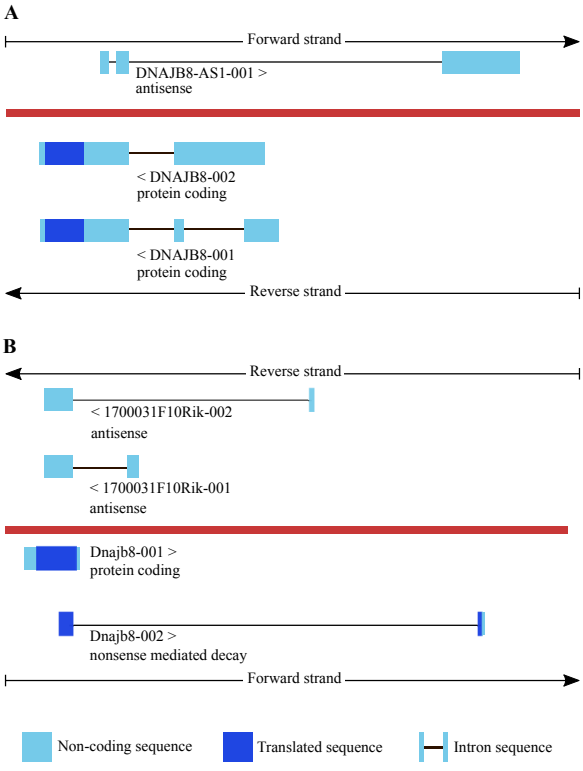

Figure 3. (A) Schematic representation of transcripts for the human DNAJB8 gene (*retro_hsap_54*) and the overlapping antisense lncRNA, DNAJB8-AS1-001. (B) Schematic representation of a 1-to-1 ortholog of DNAJB8 in mouse (Dnajb8) with its antisense lncRNAs: 1700031F10Rik-001 and 1700031F10Rik-002.

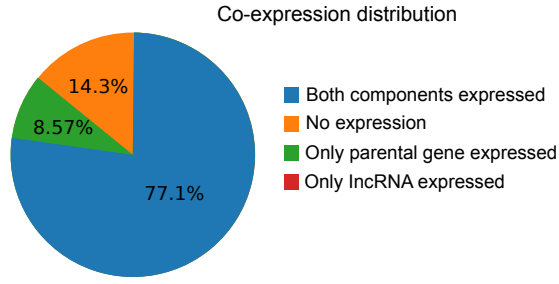

Figure 4. Summary of the co-expression analysis performed for lncRNA-parental gene pairs.

RNA-Seq libraries from ENCODE, and we found that 27 of 35 lncRNA genes are co-expressed with their parental genes (Fig. 4, Suppl. Table 4 at www.actabp.pl). Importantly, two RNAs are expected to be co-expressed if they interact in a cell, but co-expression itself does not imply they base-pair, for instance, they could be expressed in different cellular compartments. We therefore hypothesized that functionally coupled lncRNAs and parental genes should, in addition to being co-expressed, show some level of expression correlation. Therefore, we calculated the Spearman Rho correlation coefficient for co-expressed pairs. Requiring the correlation coefficient to be greater than 0.6 or less than –0.6 with a *p*-value <0.05, we found two pairs with statistically significant positive expression correlation and one pair with negatively correlated expression (Table 1). Using R's uniReg package, isotonic regression models were built for the AC021224.1-HNRNPA1 and RP11-3P17.5-RPL23A correlated pairs and an antitonic regression model was constructed for the RP11-78A19.3-CHMP1A pair (negatively correlated) (Fig. 5A, B and C, respectively). These results provide indirect evidence for the functionality of these three cases. The remaining lncRNA:RNA pairs may not be functional or alternative scenarios could apply, for example: i) the transcripts are co-expressed in a small subset of samples, producing statistically insignificant results for correlation testing, ii) some of the modes of action, such as splicing modulation or triggering mRNA editing events, do not involve changes in gene expression levels; thus, one does not expect to observe (anti-) correlation of expression, and iii) other factors, such as miRNAs and transcription factors, being involved in the regulatory processes.

Functional insights

Next, we identified possible base-pairings between lncRNAs transcribed in antisense to retrocopies and the parental genes using a previously proposed procedure (Szczesniak & Makałowska, 2016). The subsequent analysis of the RNA:RNA duplexes revealed 10 lncRNAs with potential regulatory roles exerted on their parental genes (Suppl. Table 1 at www.actabp.pl), which included stability control (masking miRNA target sites, guiding to the SMD pathway), pre-mRNA processing (modulating alternative splicing) events and mRNA processing (RNA editing). Three previously described examples with statistically significant correlations of expression were among those pairs with possible base-pairings. Therefore, we focused on them in further analysis. These cases include the following parental genes: hnRNPA1, CHMP1A, and RPL23A.

hnRNPA1

hnRNPA1 belongs to the A/B subfamily of ubiquitously expressed heterogeneous nuclear ribonucleoproteins (hnRNPs), RNA-binding proteins that associate with pre-mRNAs in the nucleus and influence pre-mRNA processing, as well as other aspects of mRNA metabolism and transport (Han *et al.*, 2010). It represents one of the most abundant core proteins of hnRNP complexes and plays a key role in the regulation of alternative splicing (Mayeda *et al.*, 1998). Overexpressed hnRNPA1 effectively downregulates the expression of the transcriptional transacti-

Table 1. Summary of antisense lncRNA and parental gene pairs with statistically significant expression correlation.
No. of samples indicates the number of samples with both genes expressed.

Ensembl gene IDs		Gene symbols		Correlation measurement		
lncRNA gene	Parental gene	lncRNA gene	Parental gene	Spearman Rho	p-value	No. of samples
ENSG00000262477	ENSG00000135486	AC021224.1	HNRNPA1	0.77	0.008	11
ENSG00000269888	ENSG00000198242	RP11-3P17.5	RPL23A	0.7	0.0002	23
ENSG00000267165	ENSG00000131165	RP11-78A19.3	CHMP1A	–0.65	0.0008	23

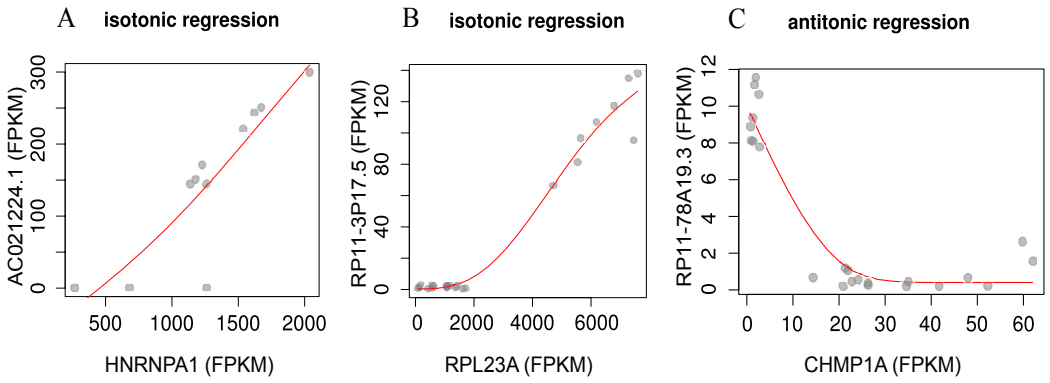

Figure 5. (A) Isotonic regression model of co-expression for AC021224.1 and HNRNPA1. (B) Isotonic regression model of co-expression for RP11-3P17.5 and RPL23A. (C) Antitonic regression model of co-expression for RP11-78A19.3 and CHMP1A.

vator Tat, which in HIV-1 infected cells results in a sharp reduction in the transcription of the viral genome and a 100-fold drop in the production of new HIV-1 virions (Jablonski & Caputi, 2009). As many as 66 retrocopies across the human genome can be found for the hnRNPA1 gene at retrogeneDB (Kabza *et al.*, 2014). One of them, *retro_hsap_1933*, has an antisense transcript ENST00000573479, also known as AC021224.1-201 or NONHSAT058863.2, that is classified as a long non-coding RNA at NONCODE (Zhao *et al.*, 2016). One of the hnRNPA1 splice variants, ENST00000547276, lacks domains necessary for the major functions of hnRNPA1 (Fig. 6B), i.e., those required for alternative splicing activity, stable binding of RNAs and optimal RNA annealing (Mayeda *et al.*, 1994). This isoform, however, plays regulatory roles in HIV-1 splicing and replication. Our bioinformatics predictions link the generation of this splice form to the absence of lncRNA:RNA base-pairing, which normally would lead to masking of the 5' splice site in the 6th intron and emergence of longer isoforms with extended functionality (Fig. 6). Importantly, the lncRNA and parental gene display a statistically significant correlation of expression, with a Spearman Rho coefficient of 0.77 (Fig. 5A), which supports the idea of their functional coupling.

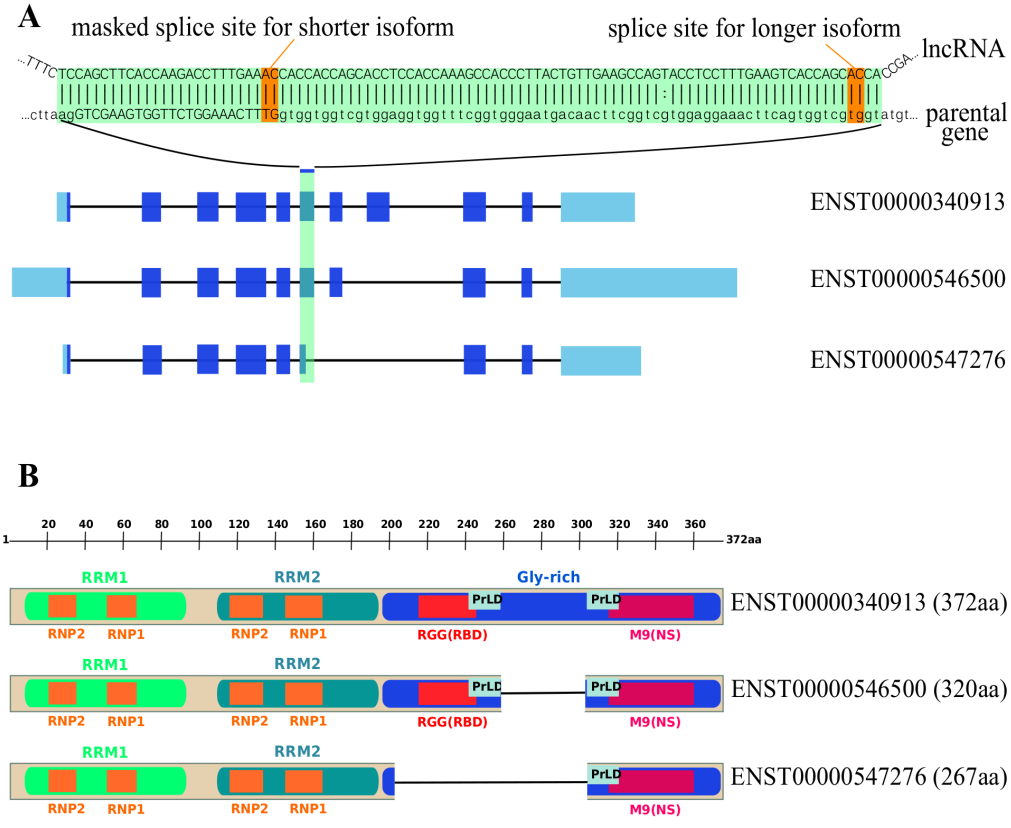

Figure 6. (A) Exon-intron structures of the two main splice forms of HNRNPA1 and a short isoform (ENST00000547276); the predicted base-pairing region between the transcripts and lncRNA (AC021224.1-201) overlaps the 5' splice site in the 6th intron and is marked with a green line. (B) Schematic representation of the hnRNPA1 protein domain structure, corresponding to these three HNRNPA1 transcripts (from Jean-Philippe *et al.*, 2014, modified). The product of ENST00000547276 lacks an RGG-Box and a portion of the prion-like domain.

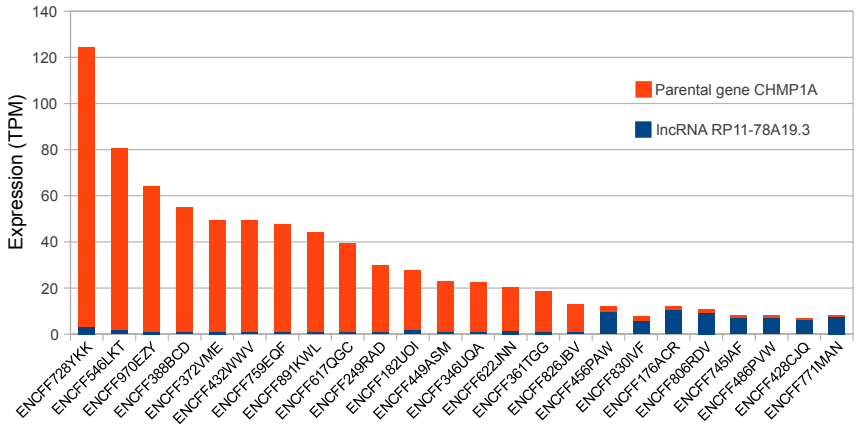

Figure 7. Comparison of RP11-78A19.3 and CHMP1A expression across 24 samples.
Eight samples derived from two cell lines, Epstein-Barr Virus transformed Gm12878 and myelogenous leukemia K562, display a relatively high expression of RP11-78A19.3-001 when compared to the parental gene CHMP1A (plotted to the right), while all the other samples follow a reverse pattern. The expression is provided in transcripts per million.

CHMP1A

CHMP1A has two retrocopies in humans (Kabza *et al.*, 2014). One is *retro_hsap_75* with antisense transcript ENST00000586474 that is also known as RP11-78A19.3-001, and is classified as a long non-coding RNA at NONCODE. CHMP1A encodes a member of the CHMP/Chmp family of proteins, which are involved in multivesicular body sorting of proteins to the interiors of lysosomes (Howard *et al.*, 2001). Overexpression of CHMP1A in cultured cells leads to gene silencing due to interaction with BMI1 transcriptional repressor and the effect on the chromatin structure (Stauffer *et al.*, 2001). Recent studies link CHMP1A to tumor development because the gene is differentially expressed in diverse tumor types (Li *et al.*, 2008; You *et al.*, 2012). For instance, shRNA knockdown of CHMP1A expression in HEK 293T cells results in increased anchorage-independent growth *in vitro* and tumor formation *in vivo* (Li *et al.*, 2008); on the other hand, overexpression of CHMP1A inhibits the growth of pancreatic cancer cells *in vitro* (Li *et al.*, 2008). Moreover, CHMP1A overexpression suppresses the proliferation of renal carcinoma cells *in vitro* and leads to suppressed tumor growth of rat renal carcinoma cells *in vivo*, while inhibition of CHMP1A expression has no effect on tumor cell growth (You *et al.*, 2012).

dsRNA formed by the *CHMP1A* gene and lncRNA may have some functions related to *CHMP1A* activity in tumors. Our expression analysis shows that the antisense RNA (RP11-78A19.3-001) is co-expressed with the parental gene (*CHMP1A*) in 24 of 300 samples, mostly in tumor samples, such as HT1080, A172, SK-MEL-5, and K562. However, with the current knowledge, it is impossible to speculate the relevance of the strong negative correlation between the two genes (Spearman Rho of –0.65 and *p*-value 0.0006; Fig. 7), especially because we were unable to link their RNA:RNA base-pairing to any of the mechanisms considered in this study.

RPL23A

RPL23A encodes a ribosomal protein that is part of the 60S subunit. This gene contains antisense transcripts that mediate downregulation of RPL23A expression in IFN-b-treated cells; they were identified *de novo* in tumor cells and were confirmed by northern blot and RT-PCR assays (Jiang *et al.*, 1997). RPL23A has as many as 68 retrocopies across the human genome (Kabza *et al.*, 2014), and three of them, namely, *retro_hsap_1775*, *retro_hsap_2021*, and *retro_hsap_2874*, have antisense lncRNAs that are co-expressed with RPL23A (Table 2). We found that these three lncRNA transcripts show elevated expression in two cell lines: K562 (derived from erythroleukemia cells) and GM12878 (from normal lymphoblastoid cells). AC016629.3, an lncRNA, shows the highest expression in K562 samples derived from a female patient with chronic myelogenous leukemia (ca. 27 TPM), in contrast to GM12878 and other non-cancer cell lines, where its expression is <2 TPM. Our analysis of lncRNA:RNA duplexes shows that

Table 2. Comparison of the expression values of RPL23A and three lncRNAs expressed in antisense to its retrocopies. Only the samples with the highest RPL23A expression are shown.

Sample ID	Sample	Parental gene	lncRNA genes		
	cell line	RPL23A	RP11--3P17.5	AC016629.3	RP11-264B14.2
ENCFF745IAF	GM12878	6093.71	105.2	1.93	0.39
ENCFF486PVW	GM12878	5863.27	106.75	1.58	0.68
ENCFF830IVF	GM12878	5703.39	72.98	1.81	0.5
ENCFF428CJQ	GM12878	5614.8	104.18	1.72	0.34
ENCFF456PAW	K562	5095.64	74.71	27.66	0.32
ENCFF806RDV	K562	4934.07	85.4	27.53	Not expressed
ENCFF771MAN	K562	4777.03	81.96	27.83	0.97
ENCFF176ACR	K562	4513.05	63.82	28.11	Not expressed

AC016629.3 might mask miRNA target sites in seven splice forms of RPL23A. Another lncRNA, RP11-3P17.5, is co-expressed with RPL23A in 21 samples, more than half of which are cancer-related. The analysis of expression values showed strong correlation between the two genes, with a Spearman Rho of 0.70 and *p*-value of 0.0002. An isotonic regression model was built and is presented in Fig. 5B.

Recent research links functions of antisense lncRNAs to their mate retrocopies and parental genes

A growing body of evidence shows that retrocopies play significant biological roles and are also key players in genome evolution (Szcześniak *et al*., 2012; Ciomborowska *et al*., 2013; Navarro & Galante 2015). A number of them constitute long non-coding RNAs (less than 3% of RetrogeneDB retrocopies have *protein_coding* status in Ensembl and half of them possess premature stop codons and/or frameshifts compared with the coding sequences of their parental genes), making retroposition a significant source of lncRNAs. Recent studies revealed that retrocopies often express antisense RNAs (asRNAs), which are active regulators of their sense counterparts through transcriptional and post-transcriptional mechanisms. They were shown to participate in controlling the promoters and transcription of the retrocopies (Morris *et al*., 2008), while suppression of these asRNAs results in transcriptional activation of the retrocopies (reviewed in Weinberg & Morris, 2013). Finally, lncRNAs transcribed in antisense to the retrocopies might act *in trans* and contribute to regulation of the parental genes. For example, PTEN, a tumor-suppressor gene, is under control of its retrocopy, PTENpg1 (Johnsson *et al*., 2013). PTENpg1 has two antisense RNAs, α and β, which regulate PTEN transcription and the stability of its transcripts. The α isoform functions *in trans* and epigenetically modulates PTEN by recruiting a DNA methyltransferase, while the β isoform interacts with PTENpg1 through RNA:RNA base-pairing, which affects the stability of sense PTENpg1 and thus enables its sponge activity.

A large-scale analysis of antisense lncRNAs in a recent study (Milligan *et al*., 2016) found 2 277 loci containing exon-to-exon overlaps between long non-coding RNAs and pseudogenes. This dataset included retrocopies and other pseudogenes, such as processed and unprocessed pseudogenes from Ensembl. Further analysis of the full-length cDNAs and ESTs that supported 313 pseudogene-lncRNA overlaps indicated that this phenomenon is prevalent. The use of EST/cDNAs as transcriptional evidence represents a conservative approach, and many more cases likely exist. The subsequent comparison of the parental genes of the pseudogenes to all human genes showed enrichment of several ontology categories; however, no insight into the biology behind these findings was provided. In particular, the biological processes and mechanisms that could possibly underlay the hypothesized lncRNA-parental gene associations were not investigated. To the best of our knowledge, we performed this type of assessment for the very first time. Our findings suggest that retroposition-derived, antisense lncRNAs might affect the expression and processing of parental genes in a number of ways, which is supported by the *in silico* base-pairing of the RNA molecules, followed by computational function assignment, co-expression data and, occasionally, correlation of expression and evolutionary conservation.

FINAL REMARKS

In this work, we analyzed the potential roles of retroposition-derived lncRNAs in regulating the expression and processing of the corresponding parental genes. We assumed directionality of the regulatory effect, although the reverse scenario is possible, with lncRNAs being affected by the transcripts of parental genes. Additionally, the antisense lncRNAs could regulate paralogs of the parental genes or any other gene with sufficient sequence similarity, which was not considered in this study. In particular, *in cis* effects are possible between lncRNAs and retrocopies expressed from the opposite strand because antisense transcripts are expected to base-pair easily due to their 100% sequence similarity (in sense/antisense orientation), and they occupy the same genomic loci, which further facilitates contact between the RNA molecules. Importantly, scenarios other than RNA:RNA interactions are possible, including transcription-dependent and transcription-independent mechanisms leading to chromatin remodeling, which has already been the subject of a number of studies (reviewed in Geisler & Coller, 2013; Milligan & Lipovich 2015).

Acknowledgements

This work was supported by the National Science Centre (grant No. 2014/15/D/NZ2/00525 to M.W.S; grant No. 2013/11/B/NZ2/02598 to I.M.), the Foundation for Polish Science (START Scholarship grant editions 2014/2015 and 2015/2016 to M.W.S.), and the KNOW Poznan RNA Centre (grant No. 01/KNOW2/2014).

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410. http://dx.doi.org/S0022-2836(05)80360-2

Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**: 756–769. http://dx.doi.org/10.1101/gad.455708

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170

Ciomborowska J, Rosikiewicz W, Szklarczyk D, Makalowski W, Makalowska I (2013) „Orphan" retrogenes in the human genome. *Mol Biol Evol* **30**: 384–396. http://dx.doi.org/10.1093/molbev/mss235

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775–1789. http://dx.doi.org/10.1101/gr.132159.111

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. http://dx.doi.org/10.1038/nature11247

Geisler S, Coller J (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol* **14**: 699–712. http://dx.doi.org/10.1038/nrm3679

Han SP, Tang YH, Smith R (2010) Functional diversity of the hnRNPs: past, present and perspectives. *Biochem J* **430**: 379–392. http://dx.doi.org/10.1042/BJ20100396

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SM, Amode R, Brent S, Spooner W, Kulesha E, Yates A, Flicek P (2016) Ensembl comparative genomics resources. *Database (Oxford)* **2016**: 10.1093/database/bav096 [doi]. http://dx.doi.org/10.1093/database/bav096

Howard TL, Stauffer DR, Degnin CR, Hollenberg SM (2001) CHMP1 functions as a member of a newly defined family of vesicle trafficking proteins. *J Cell Sci* **114**: 2395–2404

Jablonski JA, Caputi M (2009) Role of cellular RNA processing factors in human immunodeficiency virus type 1 mRNA metabolism,

replication, and infectivity. *J Virol* **83**: 981–992. http://dx.doi.org/10.1128/JVI.01801-08

Jiang H, Lin JJ, Tao J, Fisher PB (1997) Suppression of human ribosomal protein L23A expression during cell growth inhibition by interferon-beta. *Oncogene* **14**: 473–480. http://dx.doi.org/10.1038/sj.onc.1200858

Johnsson P, Ackley A, Vidarsdottir L, Lui WO, Corcoran M, Grander D, Morris KV (2013) A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nat Struct Mol Biol* **20**: 440–446. http://dx.doi.org/10.1038/nsmb.2516

Kabza M, Ciomborowska J, Makalowska I (2014) RetrogeneDB--a database of animal retrogenes. *Mol Biol Evol* **31**: 1646–1648. http://dx.doi.org/10.1093/molbev/msu139

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487–493. http://dx.doi.org/10.1101/gr.113985.110

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360. http://dx.doi.org/10.1038/nmeth.3317

Kodama Y, Shumway M, Leinonen R (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **40** (Database issue): D54-6. http://dx.doi.org/10.1093/nar/gkr854

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35** (Web Server issue): W345-9. http://dx.doi.org/10.1093/nar/gkm391

Kornienko AE, Guenzl PM, Barlow DP, Pauler FM (2013) Gene regulation by the act of long non-coding RNA transcription. *BMC Biol* **11**: 59. http://dx.doi.org/10.1186/1741-7007-11-59

Kugel JF, Goodrich JA (2012) Non-coding RNAs: key regulators of mammalian transcription. *Trends Biochem Sci* **37**: 144–151. http://dx.doi.org/10.1016/j.tibs.2011.12.003

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359. http://dx.doi.org/10.1038/nmeth.1923

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352

Li J, Belogortseva N, Porter D, Park M (2008) Chmp1A functions as a novel tumor suppressor gene in human embryonic kidney and ductal pancreatic tumor cells. *Cell Cycle* **7**: 2886–2893. http://dx.doi.org/10.4161/cc.7.18.6677

Mayeda A, Munroe SH, Caceres JF, Krainer AR (1994) Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins. *EMBO J* **13**: 5483–5495.

Mayeda A, Munroe SH, Xu RM, Krainer AR (1998) Distinct functions of the closely related tandem RNA-recognition motifs of hnRNP A1. *RNA* **4**: 1111–1123.

Milligan MJ, Harvey E, Yu A, Morgan AL, Smith DL, Zhang E, Berengut J, Sivananthan J, Subramaniam R, Skoric A, Collins S, Damski C, Morris KV, Lipovich L (2016) Global intersection of long non-coding RNAs with processed and unprocessed pseudogenes in the human genome. *Front Genet* **7**: 26. http://dx.doi.org/10.3389/fgene.2016.00026

Milligan MJ, Lipovich L (2014) Pseudogene-derived lncRNAs: emerging regulators of gene expression. *Front Genet* **5**: 476. http://dx.doi.org/10.3389/fgene.2014.00476

Morris KV, Santoso S, Turner AM, Pastori C, Hawkins PG (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet* **4**: e1000258. http://dx.doi.org/10.1371/journal.pgen.1000258

Navarro FC, Galante PA (2015) A Genome-Wide Landscape of Retrocopies in Primate Genomes. *Genome Biol Evol* **7**: 2265–2275. http://dx.doi.org/10.1093/gbe/evv142

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**: 635–640. http://dx.doi.org/10.1038/nature12943

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295. http://dx.doi.org/10.1038/nbt.3122

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. http://dx.doi.org/10.1093/bioinformatics/btq033

Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol.* **1079**: 105–116. http://dx.doi.org/10.1007/978-1-62703-646-7_6

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Fujita PA, Eisenhart C, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* **44**: D717–D725. http://dx.doi.org/10.1093/nar/gkv1275

Stauffer DR, Howard TL, Nyun T, Hollenberg SM (2001) CHMP1 is a novel nuclear matrix protein affecting chromatin structure and cell-cycle progression. *J Cell Sci* **114**: 2383–2393

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res* **41**: e166. http://dx.doi.org/10.1093/nar/gkt646

Szczesniak MW, Ciomborowska J, Nowak W, Rogozin IB, Makalowska I (2011) Primate and rodent specific intron gains and the origin of retrogenes with splice variants. *Mol Biol Evol* **28**: 33–37. http://dx.doi.org/10.1093/molbev/msq260

Szczesniak MW, Makalowska I (2016) lncRNA-RNA Interactions across the Human Transcriptome. *PLoS One* **11**: e0150353. http://dx.doi.org/PONE-D-15-35227

Szczesniak MW, Rosikiewicz W, Makalowska I (2016) CANTATAdb: A Collection of Plant Long Non-Coding RNAs. *Plant Cell Physiol* **57**: e8. http://dx.doi.org/10.1093/pcp/pcv201

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–515. http://dx.doi.org/10.1038/nbt.1621

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**: 689–693. http://dx.doi.org/10.1126/science.1192002

UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* **43** (Database issue): D204-12. http://dx.doi.org/10.1093/nar/gku989

Washietl S, Kellis M, Garber M (2014) Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24**: 616–628. http://dx.doi.org/10.1101/gr.165035.113

Weinberg MS, Morris KV (2013) Long non-coding RNA targeting and transcriptional de-repression. *Nucleic Acid Ther* **23**: 9–14. http://dx.doi.org/10.1089/nat.2012.0412

Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ, Zhou MM (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell* **38**: 662–674. http://dx.doi.org/10.1016/j.molcel.2010.03.021

You Z, Xin Y, Liu Y, Sun J, Zhou G, Gao H, Xu P, Chen Y, Chen G, Zhang L, Gu L, Chen Z, Han B, Xuan Y (2012) Chmp1A acts as a tumor suppressor gene that inhibits proliferation of renal cell carcinoma. *Cancer Lett* **319**: 190–196. http://dx.doi.org/10.1016/j.canlet.2012.01.010

Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res* **44**(D1): D203-8. http://dx.doi.org/10.1093/nar/gkv1252