University of Mazandaran

Journal of Genetic Resources
J Genet Resour 2018;4 (1): 56-64
http://sc.journals.umz.ac.ir
doi: 10.22080/jgr.2018.14224.1103

Iranian Biology Society

Phylogenetic Analysis of Three Long Non-coding RNA Genes: *AK082072*, *AK043754* and *AK082467*

Farzane Amirmahani[1*] and Kiarash Jamshidi Goharrizi[2]

[1] Department of Biology, Faculty of Science, University of Isfahan, Isfahan, Iran
[2]Department of Plant Breeding, Yazd Branch, Islamic Azad University, Yazd, Iran

[*]Corresponding author: farzanemahani@yahoo.com

Received: 15 October 2017 Accepted: 11 January 2018

Abstract

Now, it is clear that protein is just one of the most functional products produced by the eukaryotic genome. Indeed, a major part of the human genome is transcribed to non-coding sequences than to the coding sequence of the protein. In this study, we selected three long non-coding RNAs namely *AK082072*, *AK043754* and *AK082467* which show brain expression and local region conservation among vertebrates. Thus, the sequences of these genes are appropriate for phylogenetic analysis. In order to evaluate the evolutionary and molecular trend of lncRNAs in vertebrates, phylogenetic analysis and natural selection process were analyzed during evolution. The nucleotide sequences of selected long non-coding RNAs from different vertebrates were aligned and the phylogenetic trees were constructed using Neighbor Joining method with maximum sequence differences of 0.75. Our analysis of nucleotide sequences to find closely evolved organisms with high similarity by NCBI-BLAST tools and MEGA7 showed that the selected sequence of *AK082072* in human and *M. fascicularis* (macaque) were placed into the same cluster and they may originate from a common ancestor. In addition, the human sequence of *AK082467* and *AK043754* had the closest similarity with cow. Also, bioinformatic analysis showed that the dN/dS ratio is lower than 1 for all three genes which demonstrates purifying selection for the longest predicted ORF of each lncRNA. Together, these results indicate that lncRNAs act as regulatory genes that have important roles in development.

Key words: Natural selection; Long non-coding RNA; Phylogenetic tree; Common ancestor; Development

Introduction

Whereas only about 1.06% of the human genome encodes protein (Church *et al*., 2009), at least four times that amount is transcribed to non-protein coding transcripts (Bertone *et al*., 2005). It is likely that many ncRNAs develop simply from transcriptional 'noise'. If so, their sequence and transcription might be expected not to be conserved outside of restricted phyletic lineages (Chodroff *et al*., 2010). Long non-coding RNAs (lncRNAs) that have 200 nt to 100 kb length and do not show any evidence of being translated to protein have manifested as key regulators of important biological processes (Jannat Alipoor *et al*., 2017) and played a role in development and differentiation (Klattenhoff *et al*., 2013; Kretz *et al*., 2013). Most lncRNAs in each species did not show any detectable homology with lncRNAs in other species, demonstrating rapid turnover of lncRNA repertoires, as also showed by others (Necsulea *et al*., 2014; Washietl *et al*., 2014). Upon this backdrop of high turnover, many lncRNAs are conserved between various vertebrates showing their functional potency. Generally, genomic sequences of lncRNAs exhibit decreased substitution and insertion/deletion rates in comparison to expected random rates (Marques and Ponting, 2009; Necsulea *et al*., 2014). Moreover, lncRNA transcripts show distinct tissue-specific expression and lower mutation rate showing that they are subject to significant purifying selection. Rapid transcriptional output of lncRNAs is found to impact lineage-specific emergence or invisibility of them (Kutter *et al*., 2012) and the lower expression level of lncRNAs may be associated with their rapid rate of evolution (Managadze *et al*., 2011).

Recently, it has been demonstrated that three long non-coding RNAs namely AK082072, AK043754 and AK082467 demonstrate pronounced evolutionary limitation within their putative promoter region and across exon-intron boundaries, generally. Many of these lncRNA loci may be included in the cis regulation of adjacent protein-coding transcription factor genes (Valadkhan and Nilsen, 2010). In addition, some of the first orthologs present between vertebrates show conservation of brain expression (Chodroff *et al*., 2010). Due to the limited transcription of these lncRNAs to the developing nervous system in distantly relevant vertebrates, the transcripts could play important roles in neurogenesis and neuronal differentiation in specific parts of the developing telencephalon. Although determining whether expression of AK082072 transcriptionally regulates Mef2C , a gene involved in autism and intellectual disability phenotypes, requires detailed investigations (Le Meur *et al*., 2010).

Nowadays, there are many developments in the field of primate evolution. Furthermore, it is clear that phylogenomics would be a main challenging approach for re-analyzing species to determine the degrees of differences between these great creatures. With the growing understanding of the significance of some lncRNAs in different biological pathways, there is a great interest in the perception of their evolution and in using comparative genomics to study their functional determinants (Ulitsky, 2016). Therefore, the purpose of the present study was to determine the evolutionary relationships of the nucleotide sequences of three lncRNAs, AK082072, AK082467, and AK043754 and their selection procedure during evolution. Our observations provide the first investigation of comparative genomics of these lncRNAs. In this research, we have shown the evolutionary view of these genes to find the closest organism to human by the orthologous of them, which can be instructive with regard to their role in human biology.

Material and methods

LncRNA selection

We selected three lncRNAs namely *AK082072*, *AK082467*, and *AK043754* based on previous study which have high overlap with phastCons-predicted conserved elements that express in embryonic or neonatal brain according to the origin of the cDNA library from which they were recognized. They are transcribed from the mouse genome regions whose sequence aligns to vertebrate genome sequences from species at least as distantly associated as chicken, with nucleotide identity more than 80% at some intervals (Chodroff *et al*., 2010).

Phylogenetic analysis

The sequences of three lncRNAs with accession numbers: *AK082072*, *AK082467*, and *AK043754* and their ortholougs among vertebrates were taken from NCBI database. In this research bioinformatics programs such as NCBI-BLAST and MEGA7 software were applied for sequence similarity search. In addition, they were utilized for local alignments, for example, the maximal regions of high similarity among the query sequence and the database sequences. The fast nucleotide Megablast was applied as the BLAST tool, because it could compare a query to closely related sequences, and when the target percentage identity was 95% or more it could be better utilized (Zhang *et al*., 2000). In this regard, very similar sequences were chosen for alignment. Therefore, the BLAST results were applied for phylogenetic tree construction using definite methods. Furthermore, fast minimum Evolution and Neighbor Joining tools were utilized for the evaluation of the data (Desper and Gascuel, 2004; Saitou and Nei, 1987). The Maximum sequence differences of 0.75 were utilized and the Maximum sequence differences larger than 0.5 were considered as precise for grouping of sequence as determined by NCBI. Also, pairwise distances and the probability of substitution (r) from one base to another were computed by MEGA7 software. These analyses were conducted using the Maximum Composite Likelihood and Tamura-Nei model (Tamura and Nei, 1993; Tamura *et al*., 2004). Evaluating the nucleotide changes that alter amino acid sequences (dN) into those that do not affect amino acid sequences (dS) of predicted ORFs is useful in analyzing natural selection and was done by HIV sequence database (http://www.hiv.lanl.gov) (Korber, 2000). Also, JBrowse

(https://bioinf.eva.mpg.de/jbrowse/) was used to study conservation and copy number of lncRNA genes in Neanderthal genome (Skinner *et al*., 2009).

Results

Comparative analysis of three lncRNA genes

The complete cDNA sequences of different species as introduced in the materials and methods section were aligned (Fig.1). The comparative results from the present research demonstrated that human selected sequence of *AK082072* and *AK082467* had the closest similarity with *M. fascicularis* (macaque) and *Bos taurus* (cow), respectively and they may come from the same ancestor (Fig. 2A and 2C). According to the results of BLASTn, alignments of *AK082072* human sequence share approximately 67% identity with mouse ortholog. This observation is confirmed in our phylogenic three where the main cluster of human (*Homo sapiens*) and *M. musculus* (mouse) were located near each other. In addition, MEGA7 analysis demonstrated that human (*Homo sapiens*) and *M. fascicularis* (macaque) main cluster of *AK082072* was close to that of *M. musculus* (mouse). But, cDNA of *Lupus familiaris* (dog) was far from those of human, *M. fascicularis* (macaque), and *M. musculus* (mouse) (Fig. 2A). Whereas in *AK043754*, the main cluster of *M. musculus* (mouse) and human (*Homo sapiens*) were far from each other but the human (*Homo sapiens*) and *Sus scrofa* (cow) clusters are close together (Fig. 2B). In addition, pairwise distances estimate the evolutionary divergence between Sequences (Table 1). The probability of substitution (r) from one base to another is shown in Table 2. Rates of transitional substitutions are higher than transversional ones in AK043754 although in two other genes this ratio was lower. The values of the dN/dS ratio were 0.81, 0.77 and 0.78 for *AK082072*, *AK082467*, and *AK043754* respectively which demonstrated purifying selection for predicted ORFs. Furthermore, in all Neanderthal genomes sequenced, we found partly conservation and single copy number similar to human and other primates. Our analysis showed that there were some single nucleotide variants (SNVs) throughout lncRNA genes across Neanderthal genomes (Fig. 3). [All data are not shown].

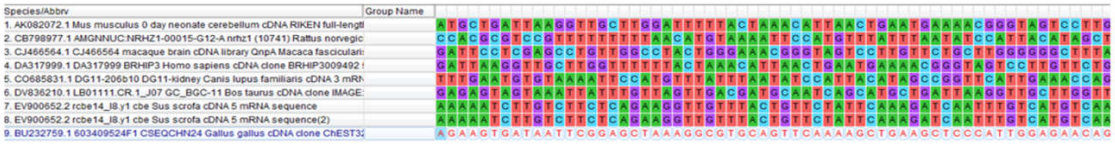
Table 1. Estimates of evolutionary divergence between sequences: All positions containing gaps and missing data were eliminated.

A: *AK082072*							
AK082072							
CB798977	1.389						
CJ466564	0.284	1.473					
DA317999	0.212	1.434	0.092				
CO685831	1.305	0.307	1.482	1.414			
DV836210	0.795	1.775	0.761	0.665	1.869		
EV900652	0.589	1.316	0.473	0.480	1.357	1.277	
BU232759	0.862	1.740	1.026	0.909	1.827	1.116	1.436
B: *AK043754*							
AK043754							
BF565173	0.102						
DB326634	1.170	1.245					
CO886535	1.115	1.083	1.541				
EW186118	1.386	1.208	2.778	1.770			
C: *AK082467*							
AK082467							
BF397583	1.354						
DA347802	2.400	1.408					
CB447323	2.558	1.573	1.019				
BI405055	2.749	1.591	2.808	2.220			
CO586030	1.385	1.225	1.460	1.680	1.595		

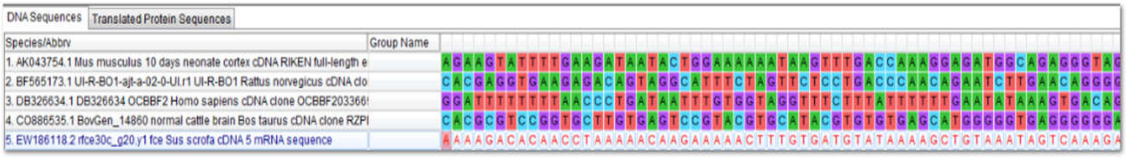
Table 2. Maximum composite likelihood estimate of the pattern of nucleotide substitution.

		AK043754		
From/To	**A**	**T**	**C**	**G**
A	-	6.2449	4.7524	13.2083
T	4.496	-	14.2165	3.9649
C	4.496	18.6814	-	3.9649
G	14.9776	6.2449	4.7524	-
		AK082072		
From/To	**A**	**T**	**C**	**G**
A	-	5.4098	4.3493	15.2061
T	5.8157	-	11.5934	5.0370
C	5.8157	14.4200	-	5.0370
G	17.5570	5.4098	4.3493	-
		AK082467		
From/To	**A**	**T**	**C**	**G**
A	-	6.3256	3.3256	12.4733
T	6.7442	-	9.4382	4.6745
C	6.7442	17.9523	-	4.6745
G	17.9963	6.3256	3.3256	-

A



B



C

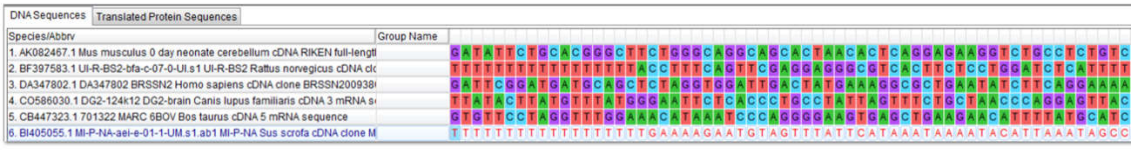
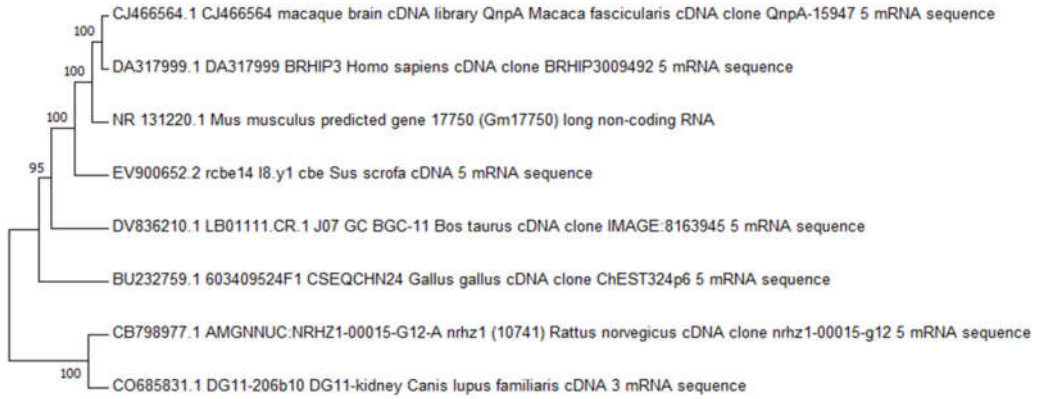
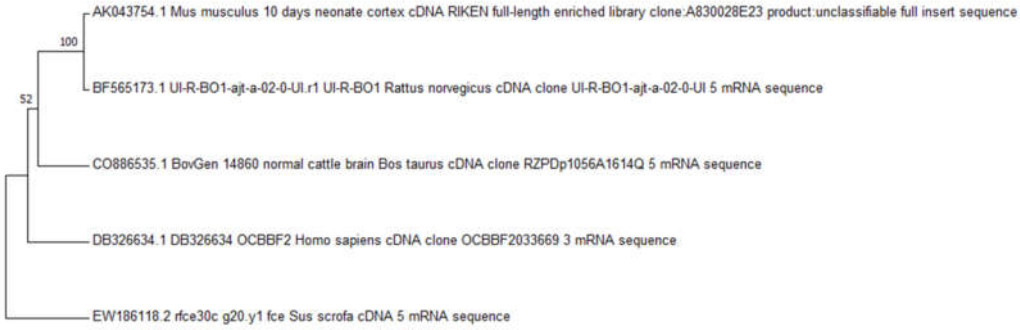


Fig. 1. Multiple sequence alignment. Alignment of lncRNA genes from collected nucleotide sequences of different species: (A) *AK082072*; (B) *AK043754*; (C) *AK082467*.

A



B



C

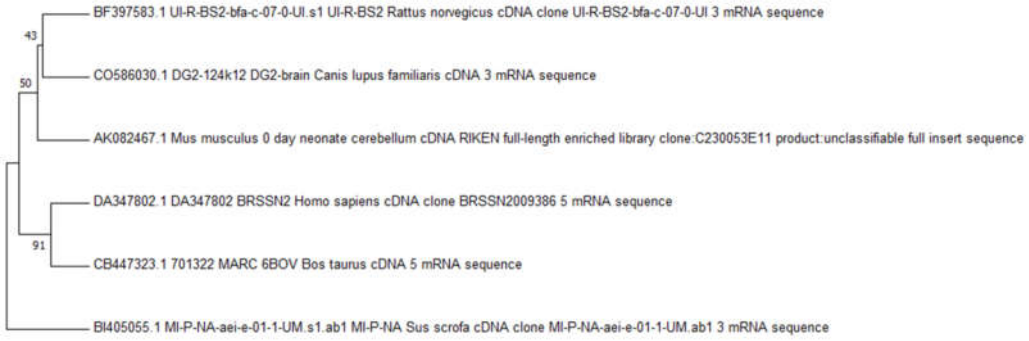


Fig. 2. Phylogenetic tree of the three lncRNA genes. The numbers at each node are the bootstrap support values obtained by maximum likelihood: (A) *AK082072*; (B) *AK043754*; (C) *AK082467*.

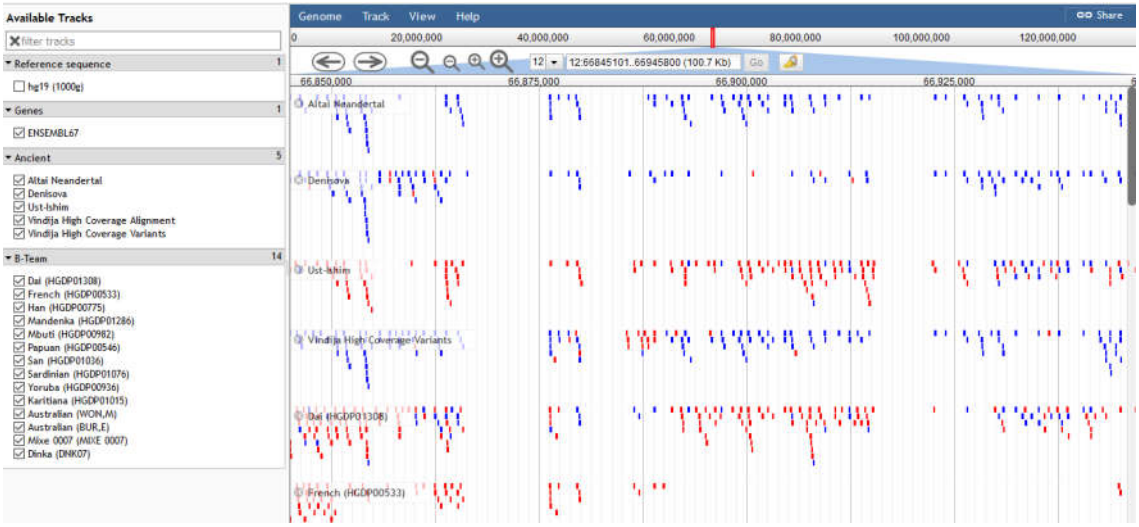


Fig. 3. Conservation of *AK082072* gene: The single copy and conservation of *AK082072* gene across Neanderthal genomes using Neanderthal genomes database (https://bioinf.eva.mpg.de/jbrowse/). Blue and red dots refer to homozygous and heterozygous variants, respectively.

Discussion

In this study a comparative research on the divergence of three conserved lncRNAs was carried out. Phylogenetic trees demonstrate the evolutionary relationships of nucleotide sequences of selected lncRNA genes. The numbers at each node are the bootstrap support values obtained by maximum likelihood. Pairwise distance matrix was used to estimate the evolutionary divergence between sequences.

The patterns of nucleotide conservation for these lncRNA loci demonstrated higher conservation near exon boundaries (Chodroff *et al*., 2010). In this regard, these lncRNA loci differ from protein-coding genes, markedly, that typically include more distributed uniformly and potent conservation within exons (Chinwalla *et al*., 2002). Less limitation within the central portions of exons may demonstrate the insertion of large transposable element sequences, that are generally free of selective limitation within exons of lncRNA in early eutherian evolution (Lunter *et al*., 2006).

In accordance with the multi-species genome sequence alignment, all transcripts use a conserved 5' donor site. In contrast, only the mammalian transcripts utilize the predicted 3' acceptor site and terminate after the predicted poly (A) signal, immediately (Chodroff *et al*., 2010). This is consistent with previous studies that amniote species had at least 70% nucleotide identity restricted to the 3'end (approximately 500 bp) demonstrating that this locus has evolved extremely rapidly after divergence from other vertebrates or originated within the amniote lineage. Also, *AK082467* orthologs in human and cow show >70% sequence identity over their proximal promoters, first exons, and 5' splice donor sites (Chodroff *et al*., 2010).

The three selected lncRNA loci have elements which are generally associated with protein-coding genes. These are GT-AG donor-acceptor splice sites, polyadenylation signals, and chromatin marks in their putative promoter regions. The putative core promoter regions are under greater evolutionary limitation than lncRNA exonic sequences, generally (Carninci, 2007; Marques and Ponting, 2009).

The results of the substitution percentage of the nucleotide sequences showed high rates of pyrimidine substitution for *AK043754* gene which is due to the cytosine methylation. The substitution rates decrease in comparison with expected random rates which is consistent with previous studies (Marques and Ponting, 2009; Necsulea *et al*., 2014). The results of the dN/dS ratio are a useful and highly effective method for recognizing the natural selection process during evolution of genes. If it is higher than 1, it shows positive selection, equal to 1 represents neutral selection and lower than 1 indicates purifying selection. Although it remains possible that the lncRNAs encode short peptides, there is a negative selection on their protein coding capacity as the dN/dS ratio was lower than 1 for all studied lncRNA genes.

Specifically, of the > 10,000 recently annotated human lncRNAs, ~ 100 have homologs in fish, ~ 300 in non-mammalian vertebrates, and more than a thousand have sequence-similar counterparts in other mammals (Hezroni *et al*., 2015). Most of the lncRNAs which are conserved only in mammals, including *XIST*, *HOTAIR*, and *NORAD* have established functions (Augui *et al*., 2011;Lee *et al*., 2016; Li *et al*., 2013; Tichon *et al*., 2016). One presumption is that these lncRNAs are conserved outside of mammals, but the sequence similarity is so low that it is no longer identifiable in contemporary species. Indeed, the number of positionally conserved pairs of mammalian and non-mammalian lncRNAs is actually higher than expected (Amaral *et al*., 2016; He *et al*., 2015; Hezroni *et al*., 2015) and the variations between the numbers of observed and the expected syntenic pairs between mammals and other vertebrates is greater than the number of pairs with sequence similarity (Hezroni *et al*., 2015). But, these variations are small in comparison with the number of lncRNAs which do not have indictable homologs outside mammals, and thus it is likely that most of the lncRNAs observed between mammals are innovations of them (Hezroni *et al*., 2017). Similar to previous studies, our results show evidence of purifying selection in proximal promoter regions than in the transcripts themselves. The observed sequence conservation in promoter regions in addition to the expression and transcription of selected lncRNA genes demonstrate that these genes have important functions among different vertebrates. Due to the limited transcription of these genes to the developing nervous system in related vertebrates, the transcripts could play important roles in neuronal differentiation and neurogenesis in particular sections of the developing telencephalon (Chodroff *et al*., 2010).

Conclusion

In recent years, many phylogenetic studies have been conducted on protein coding genes, but the evolutionary studies of non-coding RNA genes have been considered less. Despite limited conservation of lncRNA genes in comparison with small RNAs or protein coding genes, many of them have local regions that are conserved between different species. In this study, three long non-coding RNAs that have conserved promoter regions and brain expression were studied to assess the evolutionary process and find the closest organism to human by orthologous of them. In addition, tissue specific expression and lower rate of base substitution in comparison with protein coding genes show that they are subject to considerable purifying selection which was confirmed by computing the dN/dS ratio. It seems that lncRNAs have high spatio-temporal specificity and rapid turnover during the evolution which suggest that these long non-coding RNAs are as regulatory genes and have important roles in specific organisms.

References

Amaral PP, Leonardi T, Han N, Vire E, Gascoigne DK, Arias-Carrasco R, Buscher M, Zhang A, Pluchino S, Maracaja-Coutinho V. 2016. Genomic positional conservation identifies topological anchor point (tap) RNAs linked to developmental loci. *Genome Biol* 19: 32.

Augui S, Nora EP, Heard E. 2011. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat Rev Genet* 12: 429-442.

Bertone P, Gerstein M, Snyder M. 2005. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res* 13: 259-274.

Carninci P. 2007. Constructing the landscape of the mammalian transcriptome. *J Exp Biol* 210: 1497-1506.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520-562.

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnár Z, Ponting CP. 2010. Long non-coding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol* 11: R72.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* 7:e1000112.

Desper R, Gascuel O. 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol Biol Evol* 21: 587-598.

He Y, Ding Y, Zhan F, Zhang H, Han B, Hu G, Zhao K, Yang N, Yu Y, Mao L. 2015. The conservation and signatures of lincRNAs in Marek's disease of chicken. *Sci Rep* 5: 15184.

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of long non-coding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep* 11: 1110-1122.

Hezroni H, Perry RB-T, Meir Z, Housman G, Lubelsky Y, Ulitsky I. 2017. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol* 18: 162.

Jannat Alipoor F, Asadi MH, Torkzadeh-Mahani M. 2017. LncRNA Miat promotes proliferation of cervical cancer cells and acts as an anti-apoptotic factor. *J Genet Resour*. 3(2): 80-86.

Klattenhoff CA, Scheuermann JC, Surface LE, Bradley RK, Fields PA, Steinhauser ML, Ding H, Butty VL, Torrey L, Haas S. 2013. Braveheart, a long non-coding RNA required for cardiovascular lineage commitment. *Cell* 152: 570-583.

Korber B. 2000. HIV Signature and sequence variation analysis. In: Computational analysis of HIV molecular sequences, Allen G and Gerald H (eds). Dordrecht, Kluwer Academic Publishers, Netherlands.

Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 493: 231-235.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long non-coding RNAs and the evolution of gene expression. *PLoS genet* 8: e1002841.

Le Meur N, Holder-Espinasse M, Jaillard S, Goldenberg A, Joriot S, Amati-Bonneau P, Guichet A, Barth M, Charollais A, Journel H. 2010. MEF2C haploinsufficiency caused by either microdeletion of the 5q14. 3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J Med Genet* 47: 22-29.

Lee S, Kopp F, Chang T-C, Sataluri A, Chen B, Sivakumar S, Yu H, Xie Y, Mendell JT. 2016. Non-coding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell* 164: 69-80.

Li L, Liu B, Wapinski OL, Tsai M-C, Qu K, Zhang J, Carlson JC, Lin M, Fang F, Gupta RA. 2013. Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell Rep* 5: 3-12.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLOS Comput Biol* 2: e5.

Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. 2011. Negative correlation between expression level and evolutionary rate of long intergenic non-coding RNAs. *Genome Biol Evol* 3: 1390-1404.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long non-coding RNAs: modest conservation and incompleteness. *Genome Biol* 10: R124.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505: 635.

Ponjavic J, Oliver PL, Lunter G, Ponting CP. 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* 5: e1000617.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406-425.

Skinner ME1, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. *Genome Res* 19: 1630-1638.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10: 512-526.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci* 101: 11030-11035.

Tichon A, Gil N, Lubelsky Y, Solomon TH, Lemze D, Itzkovitz S, Stern-Ginossar N, Ulitsky I. 2016. A conserved abundant cytoplasmic long non-coding RNA modulates repression by Pumilio proteins in human cells. *Nat Commun* 7: 12209.

Ulitsky I. 2016. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* 10: 601.

Valadkhan S, Nilsen TW. 2010. Reprogramming of the non-coding transcriptome during brain development. *J Biol* 9: 5.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long non-coding RNAs in six mammals. *Genome Res* 24: 616-628.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7: 203-214.