Biochimica et Biophysica Acta xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Review

Evolutionary conservation of long non-coding RNAs; sequence, structure, function ☆

Per Johnsson [a], Leonard Lipovich [b,c], Dan Grandér [a], Kevin V. Morris [d,e,*]

[a] Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden
[b] Department of Neurology, Wayne State University School of Medicine, Detroit, MI, USA
[c] Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, MI, USA
[d] Biotechnology and Biomedical Sciences, The University of New South Wales, Sydney, Australia
[e] Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 6 June 2013
Received in revised form 15 October 2013
Accepted 21 October 2013
Available online xxxx

Keywords:
Long non-coding RNA
Antisense RNA
Conservation
Secondary structure
Polypurines
Epigenetic

ABSTRACT

Background: Recent advances in genomewide studies have revealed the abundance of long non-coding RNAs (lncRNAs) in mammalian transcriptomes. The ENCODE Consortium has elucidated the prevalence of human lncRNA genes, which are as numerous as protein-coding genes. Surprisingly, many lncRNAs do not show the same pattern of high interspecies conservation as protein-coding genes. The absence of functional studies and the frequent lack of sequence conservation therefore make functional interpretation of these newly discovered transcripts challenging. Many investigators have suggested the presence and importance of secondary structural elements within lncRNAs, but mammalian lncRNA secondary structure remains poorly understood. It is intriguing to speculate that in this group of genes, RNA secondary structures might be preserved throughout evolution and that this might explain the lack of sequence conservation among many lncRNAs.
Scope of review: Here, we review the extent of interspecies conservation among different lncRNAs, with a focus on a subset of lncRNAs that have been functionally investigated. The function of lncRNAs is widespread and we investigate whether different forms of functionalities may be conserved.
Major conclusions: Lack of conservation does not imbue a lack of function. We highlight several examples of lncRNAs where RNA structure appears to be the main functional unit and evolutionary constraint. We survey existing genomewide studies of mammalian lncRNA conservation and summarize their limitations. We further review specific human lncRNAs which lack evolutionary conservation beyond primates but have proven to be both functional and therapeutically relevant.
General significance: Pioneering studies highlight a role in lncRNAs for secondary structures, and possibly the presence of functional "modules", which are interspersed with longer and less conserved stretches of nucleotide sequences. Taken together, high-throughput analysis of conservation and functional composition of the still-mysterious lncRNA genes is only now becoming feasible.

1. Introduction

Studies using the recent technical advances in genomewide platforms have revealed the human genome to be vastly more complex than previously anticipated. While only ~1.2% of the human genome encodes for protein-coding genes [1], it is becoming increasingly apparent that the large majority of the human genome is transcribed into non-protein-coding RNAs (ncRNAs) [2,3]. Thousands of long ncRNAs (lncRNAs) have been identified, but very few have been assigned any function. The lack of functional studies and in many cases absence of evolutionary conservation have raised concerns about the importance of lncRNAs; some argue they are nothing more than transcriptional noise [4]. However, recent reports show thousands of lncRNAs being evolutionarily conserved [5], though not to the same extent as many protein-coding genes [6]. While the transcripts of lncRNAs appear less conserved than protein-encoding mRNAs, the promoter regions of lncRNAs are often just as conserved as the promoters of many protein-coding genes [3,7]. Furthermore as they are RNAs their conservation may be found in functional interactions with proteins and other RNAs, in contrast to the conservation of specific sequence stretches. Functional equivalency of lncRNAs that appear to lack conservation across species may be feasible thanks to the chemical properties of nucleotides and protein interaction affinities.

The function of RNA is indeed widespread; mRNAs encode proteins, rRNA and tRNA are in involved in translation, and microRNAs act by

* Corresponding author.
E-mail address: kmorris@unsw.edu.au (K.V. Morris).

0304-4165/$ – see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagen.2013.10.035

RNA:RNA interactions to modulate mRNA function. In contrast to microRNAs, almost all of which are post-transcriptional repressors, the diverse functions of lncRNAs include both positive and negative regulations of protein-coding genes, and range from lncRNA:RNA and lncRNA:protein to lncRNA:chromatin interactions [8–11]. Due to this functional diversity, it seems reasonable to presume that different evolutionary constraints might be operative for different RNAs, such as mRNAs, microRNAs, and lncRNAs.

The functional importance of lncRNAs is only now becoming revealed, and to date, of the tens of thousands of metazoan lncRNAs discovered from cDNA libraries and RNAseq data by high-throughput transcriptome projects, only a handful of lncRNAs have been functionally characterized. However, this number has been increasing, with more lncRNAs being found recently to be involved in disease [8,10–13]. Although the large majority of lncRNAs remain to be characterized, there is no longer any doubt that at least some are of functional importance. Yet, the non-conservation conundrum remains: For many lncRNAs already proven functional, poor evolutionary conservation is paradoxical and in stark contrast to the conservation of protein-coding genes.

2. Lack of conservation does not imbue a lack of function

While conservation almost always indicates functionality, lack of sequence conservation does not directly imply the opposite [10,14]. The evidence that supports this statement arises from two vastly different classes of non-protein-coding genomic regions with completely opposite evolutionary properties; ultra conserved regions (UCR), which are highly conserved with near perfect sequence identity across all vertebrates, and human accelerated regions (HAR), which show unusually high sequence diversity between human and chimpanzee.

2.1. Ultra conserved regions

In a study by Bejerano et al., 481 segments longer than 200 nt were identified to have complete conservation among human, rat, and mouse genomes, and most also in chickens and dogs [15]. Some of these UCRs were found within protein-coding sequences (111 of 481), while others were found within introns and "gene deserts". A subsequent study specifically addressed whether these UCRs were transcribed into RNA [16]. There, Calin et al. found that the majority of the UCRs were indeed expressed as RNAs, so called transcribed UCRs (T-UCRs), and intriguingly, demonstrated differential expression in cancer [16]. While the function of the majority of these T-UCRs remains to be elucidated, it is clear that many of them give rise to non-protein-coding transcripts that do not host known small RNAs, and as such are categorized as lncRNAs. Initial reports suggest that some T-UCRs are under microRNA mediated control and also dysregulated in several tumors such as chronic lymphocytic leukemia (CLL) [16] and neuroblastoma [17]. However, further functional studies to elucidate and fully understand the role of T-UCRs remain necessary, in order to definitively determine the mechanistic role of T-UCRs. Additionally, it is imperative that transcriptome datasets from non-human species, including cDNA/EST libraries as well as RNAseq results from the modENCODE Consortium, be used to determine the presence, and the exact genomic structure, of any non-human orthologous T-UCR transcripts as a prerequisite for understanding their RNA secondary structure and hence their function.

2.2. Human accelerated regions

In contrast to T-UCRs, which were found and defined by their high sequence conservation, Pollard et al. used an opposite approach [18]. Instead of looking for highly conserved regions, they identified genomic regions with accelerated rate of nucleotide substitution between human and chimpanzee, with an emphasis on sequences whose substitution rates in evolution prior to the emergence of the human terminal lineage had been lower. Because of the latter property, these sequences were termed "human-accelerated" regions (HAR). A total of 49 [18] and 202 [19] HAR regions were initially identified, of which 96% were localized within non-coding segments [18]. The most divergent of these regions, which had multiple substitutions distinguishing humans and chimpanzees but surprisingly tight sequence conservation between chimpanzees and non-primate species, was named HAR1. HAR1 was identified to be bidirectionally transcribed as part of two longer lncRNAs in a sense–antisense pair: the lncRNA HAR1A (HAR1 forward) on the forward genomic strand, and the lncRNA HAR1B (HAR1 reverse) on the opposite strand. The HAR1 region was found to be 118 nt long, to reside precisely in the exon-to-exon sense-antisense overlap of these two lncRNA genes (whose reference transcripts range from 900 to nearly 3000 nt in length, including the HAR1 118 nt sequence), and to fold into an organized secondary RNA structure whose differences between human and chimpanzee have been biochemically confirmed by independent studies [18,20]. Interestingly, it was suggested that the mutations in the human HAR1 compared to the chimpanzee sequence, stabilized this RNA structure further and were therefore evolutionarily produced through positive selection [20]. Alternatively, this varied secondary structure may be involved in sense–antisense pairing of HAR1B and HAR1A, which are reverse complement and overlapping one another, thus allowing for RNA:RNA pairing and higher ordered secondary structures to form. The HAR1 ncRNA was found to be expressed in developing neocortex early in human embryonic development and to co-localize with Reelin, an important brain protein with functions in schizophrenia and aging. Therefore, the authors speculated whether the increased rate of nucleotide substitutions within this region is of importance for human brain evolution. This example illustrates that poorly-conserved ncRNAs can have specific spatiotemporal gene expression patterns that strongly suggest function, and that major aspects of lncRNA secondary structure can undergo drastic changes during evolutionary events, such as during the emergence of modern humans. Neanderthal and Denisovan genomes, which recently became publicly available, collectively provide an invaluable resource that will allow more precise timing of sequence substitutions concomitant with RNA secondary structure changes within the last 50,000 years of human evolution.

Many more HARs, as well as T-UCRs, remain to be investigated, as improved bioinformatics and high-throughput RNA sequencing approaches make it possible to discover additional rapidly evolving regions and additional evidence of transcriptional activity, respectively. It will be of great interest to gauge the extent to which these regions are transcribed as ncRNAs and the role that these regions may have in cellular function and evolution [19].

3. LncRNAs and secondary structures

The vast majority of post-genomic lncRNA experimental biology has been an observational science, a modern equivalent to Darwin's voyage on *The Beagle*: high-throughput cDNA library construction and next-generation RNA sequencing have provided deep and comprehensive catalogs of lncRNA genes and transcripts, while the inherent bottleneck between the large size of these datasets and the low throughput of experimental validation methods has ensured that functional validation lags far behind. For this reason, only a relatively few lncRNAs have been functionally characterized to date, and even fewer have been investigated for their secondary structure and the interplay between structure and function.

Primary sequence conservation of lncRNA genes, across species, has already been studied genomewide in mammals [21–23]. Jointly, these three studies establish that genomic sequence conservation and gene structure conservation are rare at orthologous and positionally-equivalent lncRNA loci, and that intergenic lncRNAs are subjected to rapid turnover during evolution. The presence and absence of apparently species-specific lncRNAs at orthologous loci in related species, and

Please cite this article as: P. Johnsson, et al., Evolutionary conservation of long non-coding RNAs; sequence, structure, function, Biochim. Biophys. Acta (2013), http://dx.doi.org/10.1016/j.bbagen.2013.10.035