Comparative genomics in the search for conserved long noncoding RNAs

Michal Wojciech Szcześniak, Magdalena Regina Kubiak*, Elżbieta Wanowska* and Izabela Makalowska

Adam Mickiewicz University in Poznań, Faculty of Biology, Institute of Human Biology and Evolution, Uniwersytetu Poznańskiego 6, Poznań 61-614, Poland

Correspondence: Michal Wojciech Szcześniak (miszcz@amu.edu.pl) or Izabela Makalowska (izabel@amu.edu.pl)

Long noncoding RNAs (lncRNAs) have emerged as prominent regulators of gene expression in eukaryotes. The identification of lncRNA orthologs is essential in efforts to decipher their roles across model organisms, as homologous genes tend to have similar molecular and biological functions. The relatively high sequence plasticity of lncRNA genes compared with protein-coding genes, makes the identification of their orthologs a challenging task. This is why comparative genomics of lncRNAs requires the development of specific and, sometimes, complex approaches. Here, we briefly review current advancements and challenges associated with four levels of lncRNA conservation: genomic sequences, splicing signals, secondary structures and syntenic transcription.

Introduction

Long noncoding RNAs (lncRNAs) represent an abundant class of transcripts longer than 200 bases that do not encode proteins. In ENSEMBL 102 [1], there are 82 760 human lncRNAs originating from 28 480 genes; thus, they outnumber protein-coding genes (n=22 803). Despite indisputable progress in accumulating lncRNA-related data, relatively little is known about their functions, with no more than a few thousands of them subjected to detailed functional studies. To date, lncRNAs have been linked to virtually all cellular processes, including transcription, splicing, translation, and the cell cycle [2–6], and have also been implicated in human diseases, such as malignant transformation, while a number of them represent diagnostic and prognostic biomarkers for cancer [7,8]. lncRNAs participate in different steps of gene expression. For instance, many modulate the act of transcription, e.g. through promoter modifications, creating a permissive chromatin environment or binding transport factors to inhibit the nuclear localization of specific transcription factors [9,10].

The heterogeneity in the modes of lncRNA action poses a huge challenge to functional studies and for distinguishing which are *bona fide* functional. Here, comparative genomics, conservation studies in particular, may be helpful. First, evolutionary conservation is often associated with functionality, while the presence of orthologs with established roles makes the predictions stronger. Second, the level and type of lncRNA conservation may be used to assign lncRNAs to a hypothetical functional domain, thus facilitating subsequent functional studies. For instance, conservation of exon sequences points to functions exerted through mature lncRNA molecules, such as lncRNA:RNA base pairings that are associated with RNA processing and stability [11,12]. On the other hand, the evidence accumulated thus far strongly suggests that for most lncRNAs it is not their sequences but the act of transcription that is associated with their biologically relevant functions. This is especially true for so-called antisense lncRNAs [13,14], whose function is to modulate the expression of sense partners. Because of the nature of the underlying lncRNA:protein interactions with low sequence specificity, there are virtually no constraints upon the conservation of sequences of these lncRNAs. They do, however, often exhibit positional conservation; i.e., they occupy the same or neighboring genomic loci in closely related species.

In this review, we consider four dimensions of lncRNA conservation from the perspective of the genome. The abovementioned conservation of primary sequences (Figure 1A) and positional...
conservation of lncRNAs (Figure 1B) are complemented by insight into their splice sites and other splicing signals (Figure 1C), and their secondary structures (Figure 1D): we consider the pros and cons of existing methodologies and their algorithmic challenges, discuss relevant studies and pinpoint some novel or emerging approaches, such as the application of alignment-free metrics and the conservation of promoter sequences.

We are not covering the issue of conservation of lncRNA expression, processing, and tissue-specificity but one should keep in mind this represents another layer of complexity in conservation and functional studies. For example, it has been demonstrated that positionally conserved lncRNAs in human embryonic stem cells are in most cases spliced and then exported to the cytoplasm, while their mouse counterparts are predominantly unspliced and retained in the nucleus, leading to species-specific functions of lncRNAs in pluripotency maintenance [15]. Generally, the most conserved lncRNAs tend to be broadly expressed and they display significantly higher expression levels than those with high evolutionary rates [16,17].

Sequence-level conservation
In recent years, several software programs based on sequence homology have been proposed for use in lncRNA conservation analysis [18–20]. In particular, alignment-based algorithms (e.g., BLASTN or BLAT) are widely used to search for similarities in primary sequences between lncRNAs. Such comparisons are performed utilizing solely lncRNA sequences or, additionally, genomes of interest. For example, Marques et al. compared the substitutions in mouse and rat transcriptional initiation regions and demonstrated that enhancer-associated lncRNAs generally display lower conservation than other classes of lncRNAs [21]. In another study, Hezroni et al. applied BLASTN to create a network of sequence similarities, comparing long intergenic RNAs (lincRNAs; autonomously transcribed lncRNAs that do not overlap annotated coding genes) from sea urchins and 16 vertebrates and demonstrated that most lincRNAs are lineage-specific [18]. A number of studies have revealed that only a small fraction of lncRNAs exhibit primary sequence conservation similar to that of protein-coding genes [22], with some well-studied examples among them. For example, phylogenetic analysis comparing 20 mammals showed high evolutionary conservation of MALAT1 lncRNA [23], known to be a modulator of gene expression and involved in tumor growth and metastasis. Rapid evolution of lncRNAs has resulted in weak sequence conservation not only within animals but also in plants. It has been shown that fewer than 2% of the lncRNAs in A. thaliana are conserved across plants [24]. Although the general conservation level of lncRNAs is very low, there is a special group of lncRNAs dubbed transcribed-ultraconserved regions (T-UCRs) that are encoded by a subset of ultraconserved genomic regions and are highly conserved in humans, rats, and mice [25]. Typically, they are detected as transcribed by microarray experiments and not present in most RNA-seq based reconstructions of polyadenylated transcripts [25]. Recent studies highlight the importance of T-UCRs in the
regulation of genes expression. For example, uc.372 promotes the expression of genes related to lipid synthesis and uptake by binding pri-miR-195/pri-miR-4668 and by suppressing maturation of mir-195/mir-4668 [26].

Because the full sequences of many IncRNAs are poorly preserved, analyses of shorter regions have been proposed, leading to the discovery of short-sequence homologies (microhomologies) [18,27]. It has been shown that these microhomologous regions are repeated within an IncRNA and can act as protein binding sites [28,29]. However, another solution has been proposed by Noviello et al., who successfully applied alignment-free string similarity metrics for the identification of IncRNA homologs based on their promoter sequences [19]. Kirk et al. have also developed an alignment-free approach called SEEKR. They found that IncRNAs with related functions have similar k-mer profiles despite lacking linear homology [30]. Recently, Ross et al. have introduced a novel technique that complements the SEEKR algorithm, LncLOOM. This method efficiently detects both known and novel functional elements in IncRNAs [31]. Interestingly, a growing body of evidence shows that promoter regions in IncRNAs are more conserved than the remainder of their sequences [29,32]. Moreover, Derrien et al. compared the conservation level of different regions of IncRNA genes and observed that promoters have been conserved to almost the same extent as protein-coding genes [33]. In addition, a number of studies have shown that exons of IncRNAs are better conserved than introns [34,35]. On the other hand, Pegueroles and Gabaldón analyzed sets of human IncRNAs, demonstrating that although IncRNA exons are enriched in conserved elements, they are not conserved to a greater extent than introns [36].

Conservation of splicing signals

Because of the overall weak primary sequence conservation, studies of other IncRNA features, such as their gene structures, may shed light on the evolution and functional potential of these molecules. In general, the conservation of splicing patterns has been widely studied, as it refers directly to the preservation of the exon–intron structure, considered a hallmark of functional transcripts. Despite the fact that IncRNAs seem to be less efficiently spliced than pre-mRNAs [37–40], noncoding exons are reported to universally originate from alternative splicing [41], and distinct IncRNA isoforms have the potential to act in a divergent manner [42–44].

The most common approach to splicing conservation studies requires a set of high-quality splice sites in species of interest, multiple genome alignments and, finally, statistical testing to detect and measure conservation. For example, Ponjavic et al. [45] evaluated splice site conservation with a chi-square test to compare signals from interest, multiple genome alignments and, finally, statistical testing to detect and measure conservation. For example, lncRNA isoforms have the potential to act in a divergent manner [42–44].

On the other hand, Pegueroles and Gabaldón analyzed sets of human IncRNAs, demonstrating that although IncRNA exons are enriched in conserved elements, they are not conserved to a greater extent than introns [36].

Conservation of splice sites is commonly assessed using the MaxEntScan program with positions of mouse or human introns as a reference for gene structures. Interestingly, some of them have not been conserved for over 100 million years. However, a substantial fraction of lincRNAs introns are not conserved – a high turnover rate of lincRNAs introns in the Glires lineage was reported [46].

Machine learning or statistical models are implemented in conservation studies as well. For instance, Rose et al. trained a support vector machine model to detect and rate donor and acceptor splice site candidates in multiple sequence alignments [47]. They also used MaxEntScan [48], a tool that scores splice site sequences in terms of their similarity to canonical splice sites. MaxEntScan is based on maximum entropy distribution and is considered to offer the most unbiased approximation for modeling short sequence motifs [49]. Remarkably, it considers dependencies between both nonadjacent and adjacent positions and is used extensively in analyses of IncRNA splice-site conservation [50–52]. One such study was performed on vertebrate IncRNAs [50], in which multiple sequence alignment together with transcriptomics data were utilized to determine the homologous positions in splice sites, leading to the construction of a comparative map of splice sites. The authors showed that in conserved human transcripts, 87% of splice sites are present in other species, including mice, rats, cows, and dogs, yet most of the novel splice sites originated during primate evolution. MaxEntScan has also been used in recent studies on IncRNA splicing conservation in 16 Brassica species [52]. These studies revealed that nearly 18% of lincRNAs display splicing conservation in at least one exon in Brassicaceae plant family members. In comparison with that of vertebrates, the level of lincRNA conservation measured by gene structure is significantly lower in plants.

The splicing conservation may also involve short sequence motifs recognized by proteins associated with splicing regulation. Examples of this motif are purine-rich hexamers that are recognized by serine arginine-rich proteins (SR proteins). These so-called exonic splice enhancers (ESEs) were explored, inter alia, by Schüler et al. [53]; these authors showed that experimentally confirmed human hexamers (ESEs) evolve considerably slower than nonenhancer
sequences in IncRNAs. Another group implemented the RESCUE-ESE algorithm [54] to identify hexamers that are significantly enriched or depleted within exonic sequences relative to their flanking intronic sequences [55]. The authors reported that evolutionary constraints are more concentrated near intron–exon boundaries. Moreover, these regions in IncRNAs contain a high density of ESEs, which are conserved across mammals [55]. Despite the fact that both of the aforementioned studies indicate the action of purifying selection to preserve exonic splicing enhancers within IncRNAs, the biological interpretation is difficult, particularly because splicing of IncRNAs is generally inefficient [37,39,40]. More in-depth insight is provided by experimental evidence showing that IncRNAs are unable to bind SR proteins securely [56] and by studies showing the diverse dynamics of intron excision across IncRNAs [38].

**Secondary structure conservation**

Diversified functions of IncRNAs, including participation in post-transcriptional regulation, are often associated with their ability to recognize and bind molecular targets, where the secondary and tertiary structures of related RNA molecules may play pivotal roles. As a result, studying the structural architecture of IncRNAs may help to determine their exact modes of action. The possibility of these molecules to maintain certain structures was reported, for example by Smith et al. in a genome-wide study of mammalian evolutionarily conserved structures. Smith et al. identified millions of putative functional RNAs, some of which overlap with annotated ncRNAs [57]. Another report pointed to a set of conserved brain IncRNAs being significantly enriched with predicted RNA secondary structures [58]. Furthermore, Pegueroles and Gabaldón [36] observed that positions in paired IncRNA regions that participate in the formation of folds show lower probabilities of having SNPs. Thus secondary structures impose limits on changes to IncRNA sequences (which may lead to degeneration of the structural motifs), suggesting their functional relevance. Yang and Zhang showed, using data from a parallel analysis of RNA structure (PARS), that IncRNAs are substantially less folded than mRNAs [59]. These results were surprising as it is assumed that IncRNAs often require a secondary structure to perform their functions, while the mRNAs do not. However, the authors pointed out that it is difficult to clearly explain the observations since analyses may be biased by insufficient representation of functional IncRNAs, for example. On the other hand, strong secondary structures may not be preferred as they could hinder base-pairings that are at the core of many RNAs functions.

Although the secondary structure conservation of IncRNAs is lower than that of messenger RNAs or ribosomal RNAs [20,59], some functional IncRNAs with experimentally confirmed structural motifs, including *Cyrano* [60], *Xist* [61,62], *MALAT1* [63,64], *SRA* [65] and *GAS5* [66], have been found. For instance, recently published studies on *MALAT1* in vertebrates revealed an evolutionarily conserved core consisting of numerous helices [64]. At the center of the functional core of *Cyrano* is a cloverleaf structure maintained for more than 400 million years, enriched in several protein binding sites and masking a target site for miR-7 [60]. The aforementioned studies were based on a similar approach: secondary structures are experimentally determined by chemical and/or enzymatic probing and then used in comparative sequence analyses. Interestingly, resolved secondary structures may be used in the identification of IncRNAs in other species, as shown in studies on *COOLAIR* [67]; its secondary structure was successfully used to predict the corresponding exons in evolutionarily divergent *Brassicaceae* species.

Computational methods for the detection of conserved secondary structures are quite well established, such as those based on covariance models (CMs), which are able to automatically learn statistical models derived from multiple RNA alignments. CMs are implemented with popular homology search tools, e.g., Infernal [68] or CMfinder [69]. CMfinder was applied in research on vertebrate genomes, which were screened for conserved RNA structures (CRSs) [70]. The authors pinpointed some structurally conserved IncRNAs and noted that CRS density decreases from the 5′ to the 3′ end of IncRNAs. Another tool based on covariance models is R-scape, which initially showed limited utility on eukaryotic IncRNAs, as it found no statistically significant support for the proposed secondary structures in some of the IncRNAs with experimentally verified structures [71]. However, recent studies have shown that R-scape efficiently detects conserved secondary structures when appropriately parameterized [72]. Notably, since CMs need to be trained on sets of homologous sequences that are not always available, attempts have been made to utilize algorithms comparing structures without sequence alignments, e.g., BEAGLE [73], which was used to successfully study the secondary structure similarities in four different *Caenorhabditis* worm species [20].

**Syntenic IncRNAs**

Since it is difficult to identify orthologous IncRNAs by sequence similarity [74,75] or the conservation of secondary structures [71], the evolutionarily conserved position in the genome (synteny) may be particularly useful. Syntenic IncRNAs are found either in the same genomic region across compared species, as determined by whole genome
alignment (WGA), or are found between syntenic protein-coding genes. Syntenic IncRNAs, also referred to as syntologs, represent loci with conserved transcription that typically exhibit little or no sequence similarity. In other words, syntenic transcription may generate RNAs whose expression itself, rather than the sequence, is required for IncRNA function. Indeed, a growing body of evidence shows that for most human IncRNAs, their transcription alone, rather than the production of the mature RNA molecules, is of biological importance. Good examples are natural antisense transcripts (NATs), which originate from the opposite genomic strand compared with their sense partners. It is estimated that as many as 70% of human genes display antisense transcription [13,14], with NATs modulating their expression and processing in a number of ways [76,77]. The prevailing mechanism is the recruitment of complex epigenetic machinery that results in histone modifications and subsequent transcriptional deregulation of target genes [78]. Antisense IncRNAs may also affect expression of their sense counterparts through transcriptional interference (TI) mechanisms. One such example is yeast SER3 protein coding gene being overlapped by the SRG1 IncRNA, whose transcription increases nucleosome density at the SER3 promoter [79]. Antisense IncRNAs may also trigger methylation at GC-rich genomic regions that are often associated with vertebrate genomes. An example is α-thalassemia, where antisense RNA LUC7L represses expression of HBA2 (α-Globin) by triggering methylation of its CpG islands [80]. More functions played by NATs are reviewed elsewhere [76,77]. Examples of syntologs with relatively degenerate sequences are those referred to as topological anchor point RNAs (tapRNAs), whose roles are linked to chromatin looping and topology, especially in the neighborhood of developmental genes [81]. They act by binding the DNA in cis; hence, they maintain complementarity irrespective of how much the counterpart DNA sequence is changed, indicating little or no evolutionary constraints upon their sequences. It should be noted, however, that particular fragments of these transcripts might be under stricter evolutionary constraints. tapRNAs contain conserved sequence domains recognized by transcription factors and RNA-binding proteins with zinc finger domains, leading to their elevated overall sequence conservation compared with most IncRNAs [81].

Although the search for syntenic IncRNAs proved to be an effective way of obtaining orthologous IncRNAs [82,83], there are substantial issues when calling syntologs. In general, whether two IncRNAs found within syntenic genomic blocks are orthologous remains in question, as they tend to occupy nonoverlapping conserved fragments of genomic DNA or are even found in a relatively large, nonconserved genomic region between two syntenic blocks (or two orthologous protein-coding genes). The syntenic IncRNAs might also show conservation limited to short parts of their sequences (e.g. within one exon out of many), which further sows the seeds of doubt whether they should be called orthologs. Inference pipelines, such as slncky [82], also fail to unanimously call IncRNA homologs in cases where more than one IncRNA can be found at neighboring genomic loci or when IncRNA and its genomic neighborhood are enriched with repetitive elements. In these cases, a single orthologous pair is selected based on statistical analysis but with no guarantee that the best solution is provided. For example, to reduce reporting alignments that may be driven by repetitive elements, slncky aligns each IncRNA to the shuffled intergenic regions and seeks to establish a null distribution and to determine the empirical 5% threshold for significant alignment scores. Second, the search for syntologs relies heavily on the quality of WGAs, which project expressed lncRNA loci that correspond to loci in other species. As a result, although the approach is less dependent on evolutionary distance than other sequence-based analyses, it works best with closely related species. Nevertheless, Herrera-Ubeda et al. managed to find syntenic IncRNAs between humans and lancelets using LincOFinder, a new pipeline that identifies conserved lincRNAs between evolutionarily distant species by means of microsynteny analyses [83]. The pipeline considers only the clusters formed by one upstream gene, the lincRNA and one downstream gene and led to the discovery of 32 clusters, with only 16 found to have bona fide orthologous lincRNA after manual inspection.

Notably, an efficient search for syntologous IncRNAs requires high-quality annotations for both species of interest, which are not always available but are essential in different algorithm steps. In particular, protein-coding gene annotations are used to resolve spurious conservation relationships, but the ability to detect IncRNA orthologs relies mostly on extensive annotations of IncRNAs in a given species. As mentioned earlier, available IncRNA datasets are scarce for most species, while model organisms possess multiple and rather poorly overlapping catalogs of IncRNAs, as they are populated with fundamentally different experimental and in silico methods. Keeping this in mind, in our recent research on conserved IncRNAs across primates, we used our custom pipeline built on the basis of slncky, and we assembled custom IncRNA annotations based on data from publicly available RNA-Seq repositories [17]. Our analysis yielded over 78 000 expressed human IncRNAs, and from 2054 to 18 226 were conserved across individual primate species. Interestingly, IncRNAs showing only positional conservation represent as many as 66.78% of the conserved human IncRNAs in the great apes, while syntologs make up only 42.37% of the ultraconserved IncRNAs, defined as those found in all eleven primate species considered in the study.
Current challenges in comparative genomics of IncRNAs

When deciphering the evolution of IncRNAs, the use of comparative genomics is hindered by the poor conservation of sequences and secondary structures. Therefore, specialized computational approaches tailored to the specific characteristics of IncRNA evolution are needed. Some of the most frequently used approaches are briefly described above. The genomic context of IncRNAs represents another layer of complexity. As many of IncRNAs are expressed in antisense to protein-coding loci, are located in introns or represent alternative isoforms of protein-coding genes, it is virtually impossible to investigate their evolution independent of their associated protein-coding genes. This is why most studies are focusing on IncRNAs, but their evolution does not necessarily reflect that of all IncRNAs. Another major obstacle is extensive bias in the quality of annotations between compared species. First, in addition to model organisms, available IncRNA catalogs are quite scarce (compare 82,760 IncRNAs for humans and 1,778 for chimps in ENSEMBL 102), which is often coupled with the lack of fully assembled genomes and diversified RNA-Seq data, both of which are typically required to build comprehensive sets of expressed IncRNAs. On the other hand, model organisms, such as humans, possess highly divergent datasets of IncRNAs across public resources, which largely stems from the various computational criteria being used, in some situations leading to barely overlapping sets of IncRNAs based on the same input data [84]. Additionally, since IncRNAs are most often identified from RNA-Seq data, some IncRNAs with highly specific spatiotemporal expression patterns might remain undiscovered; similarly, nonpolyadenylated IncRNAs that prevalently remain elusive when oligo(dT) protocols for reverse transcription are applied. Reliable analysis of IncRNA conservation, orthologs across species and evolution in general depend on the establishment of standard criteria for their search and annotation, along with the growth of genomic and transcriptomic data for nonmodel organisms.

Summary

- Comparative genomics of IncRNAs typically involves the formulation of dedicated algorithms and the use of specialized software, in addition to those used in previous studies on protein-coding genes.
- From the perspective of genomics, there are four different levels of IncRNA conservation, and in general, IncRNA orthologs are not required to show sequence similarities.
- Comparative genomics of IncRNAs is severely hampered by limited availability of quality annotations, and further progress in the field is dependent on expansion of IncRNA catalogs across species.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

This work was supported by the National Science Centre [grant number 2017/27/N/NZ1/00417 (to E.W.)].

Author Contribution

All authors wrote and edited the manuscript. E.W. and M.R.K. prepared the figure. All authors read and approved the final manuscript.

Abbreviations

CM, covariance model; CRS, conserved RNA structure; ESE, exonic splice enhancer; lincRNA, long intergenic RNA; lncRNA, long noncoding RNA; NAT, natural antisense transcript; T-UCR, transcribed-ultraconserved region; WGA, whole genome alignment.

References


