Received: 1 September 2021 | Revised: 19 November 2021 | Accepted: 4 December 2021

DOI: 10.1002/wrna.1708

ADVANCED REVIEW

WIREs RNA WILEY

Discovering functional motifs in long noncoding RNAs

Caroline Jane Ross | **Igor Ulitsky**

Biological Regulation and Molecular Neuroscience, Weizmann Institute of Science, Rehovot, Israel

Correspondence
Igor Ulitsky, Biological Regulation and Molecular Neuroscience, Weizmann Institute of Science, Rehovot 76100, Israel.
Email: igor.ulitsky@weizmann.ac.il

Funding information
German-Israel foundation for Scientific Research and Development, Grant/Award Number: I-1455-417.13/2018; Israel Science Foundation, Grant/Award Numbers: 2406/18, 852/19; MOST-PRC program; The EU Joint Programme - Neurodegenerative Disease Research (JPND ERA-Net localMND)

Edited by: Jeff Wilusz, Editor-in-Chief

Abstract

Long noncoding RNAs (lncRNAs) are products of pervasive transcription that closely resemble messenger RNAs on the molecular level, yet function through largely unknown modes of action. The current model is that the function of lncRNAs often relies on specific, typically short, conserved elements, connected by linkers in which specific sequences and/or structures are less important. This notion has fueled the development of both computational and experimental methods focused on the discovery of functional elements within lncRNA genes, based on diverse signals such as evolutionary conservation, predicted structural elements, or the ability to rescue loss-of-function phenotypes. In this review, we outline the main challenges that the different methods need to overcome, describe the recently developed approaches, and discuss their respective limitations.

This article is categorized under:

RNA Evolution and Genomics > Computational Analyses of RNA

RNA Interactions with Proteins and Other Molecules > Protein-RNA Interactions: Functional Implications

Regulatory RNAs/RNAi/Riboswitches > Regulatory RNAs

KEYWORDS

computational biology, long noncoding RNA, RNA-protein interactions

1 | INTRODUCTION

Over the past decade, progress in high throughput sequencing and genome-wide transcriptome profiling have revealed an unprecedented landscape of long noncoding RNAs (lncRNAs) in animal and plant genomes (J. B. Brown et al., 2014; Cabili et al., 2011; Clark & Mattick, 2011; Guttman et al., 2009; Iyer et al., 2015; Necsulea et al., 2014; Sarropoulos et al., 2019). Only a fraction of these lncRNAs have been studied and those have been attributed with regulatory functions in various cell processes and occasionally implicated in human disease (Guo et al., 2019; Hezroni et al., 2019; Liu et al., 2021; Statello et al., 2020). LncRNA synthesis is similar to that of mRNAs in that they are transcribed by RNA polymerase II (Pol II), undergo splicing and are capped and polyadenylated. The classification of lncRNAs is largely dependent on exclusion criteria and typically encompasses transcripts of at least 200 nucleotides (nt) that do not exhibit potential to encode functional proteins and do not belong to any other well-defined group of noncoding RNAs, such as ribosomal RNA (rRNA). Thus, the currently annotated lncRNAs comprise a hash of thousands of genes with diverse properties, possible functions, and mechanisms for which the principles of subclassification have not yet been elucidated. A striking difference between lncRNA genes and coding genes, is the rapid turnover of lncRNAs during evolution. This has underpinned earlier assumptions that the majority of lncRNAs may not be functional. However, studies on specific lncRNAs have shown that function can be maintained across distantly related species, even in cases where

the sequences have substantially diverged (reviewed in Ulitsky, 2016). As the evidence continues to mount that a sizable number of lncRNAs are relevant to human disease, it is becoming more important to develop methods that can rapidly distinguish which lncRNAs are likely to be functional. The understanding of how lncRNA functions have been maintained across sequences that have extensively changed is fundamental to deriving principles for the large-scale classification of these genes into functional groups. This classification will allow us to transfer functional knowledge across lncRNAs with similar properties, and will guide the design of experimental studies to systematically probe their numerous biological mechanisms.

The first studies that delved into mapping the homology of lncRNA sequences relied on alignment-based algorithms, which were successful in comparison of closely related species. However, these algorithms assume the equivalence between homology and nucleotide sequence similarity, which is problematic for the comparison of lncRNA sequences that have largely diverged. Nevertheless, comparisons of lncRNAs between human and other mammals suggest that lncRNA sequences often contain short conserved patches that evolve under purifying selection and that these patches are disguised by their surrounding sequence context which evolves rapidly and thus seems to have limited contribution to function. This model is attractive because it is sufficient for supporting interactions of lncRNAs with RNA binding proteins (RBPs) or microRNAs (miRNAs), which has previously been shown to modulate lncRNA function and to be regulated by them (Bitetti et al., 2018; Kleaveland et al., 2018; Ulitsky et al., 2011; Xue et al., 2016). More recent studies have expanded on this idea, and implemented a combination of computational and experimental approaches to uncover functional elements in an array of lncRNA genes. This concept has also underpinned the recent development of novel alignment-free approaches which have advanced the capability to identify and compare orthologous lncRNAs in more distal species.

In this review, we summarize the recent advancements in the computational and experimental approaches to uncover functional motifs in lncRNA genes. We begin with a brief description of the challenges that have hindered lncRNA annotation across diverse species and the comparative analysis of their sequences, with a particular focus on considerations for the unbiased identification of lncRNA orthologs in distal species. We then discuss the fundamental traits of lncRNA evolution that have been uncovered from early studies, and go on to describe how these traits have been leveraged to develop new algorithms that present more sensitive approaches for the functional classification of lncRNAs. Last, we give an overview of the experimental methods that have been developed to validate the functional importance of sequence and structural motifs in lncRNA sequences.

2 | THE MAIN CHALLENGES THAT HAVE STALLED COMPARATIVE GENOMICS OF lncRNAs

Functional conservation is a well-defined evolutionary constraint that has been successfully leveraged to annotate protein coding genes (PCGs) and other noncoding RNAs (Bartel, 2009; Michel & Westhof, 1990; Woese et al., 1980). However, the application of comparative genomics to lncRNAs has been hindered by two major obstacles. The first obstacle stems from the rapid evolution of lncRNA sequences. Conventional algorithms that were developed for PCGs are based on multiple sequence alignments (MSAs). These can be derived from whole genome alignments (WGAs), but those require extensive and pairwise-significant sequence similarity between the studied genomes, and much or all of the sequence of a given lncRNA will often not have informative MSAs. Such methods thus perform poorly in identifying and comparing orthologous lncRNAs that lack long continuous regions that are highly constrained at the sequence level. The analysis of *CHASERR* lncRNA sequences from distantly related species presents an example of this problem (Figure 1a). *CHASERR* is transcribed in close proximity to the transcription start site (TSS) of *CHD2* and acts in *cis* to regulate expression levels of *CHD2* (Rom et al., 2019). This regulatory mechanism has been shown to be essential for mouse viability (Rom et al., 2019). Analysis of RNA-seq data clearly indicates that *CHASERR* has syntenic counterparts that are transcribed in species as far as zebrafish (>400 million years of evolution; Figure 1a), however BLASTN does not detect significant alignment (E-value <0.001) between human *CHASERR* and species beyond amniotes (Ross et al., 2021). An alternative to using MSAs is to directly compare sequences of lncRNAs across species by pairwise sequence alignment. However, the number of lncRNAs that have been annotated remains very limited in most vertebrate species. This is a consequence of both incomplete genome sequences or partial annotations of protein-coding genes, and the limited accuracy of algorithms for reconstruction of transcripts from RNA-seq data. In some cases this reduces the ability to identify orthologous lncRNAs, while in other cases it can lead to the false identification of a lncRNA that appears to stand alone, but actually aligns with a region of a PCG that has not been fully annotated

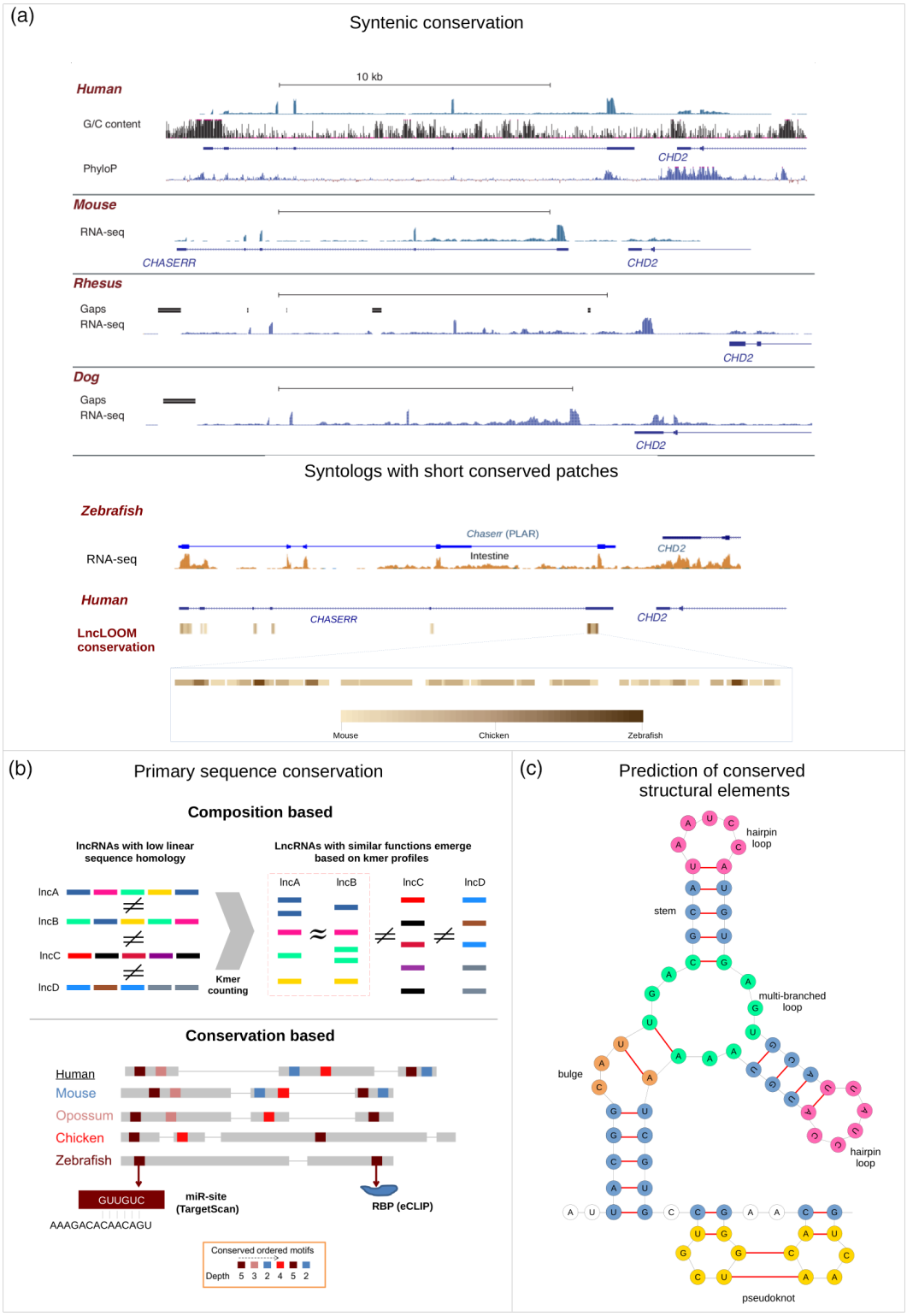

FIGURE 1 Dimensions of lncRNA evolution that have been leveraged for comparative sequence analysis. (a) Genomic locus of *Chaserr* across vertebrate species. RNA-seq data obtained from HPA and SRA. PhyloP scores indicate low sequence conservation, despite synteny. Genomic gaps in respective species and high GC content of the region are shown to highlight potential challenges in identification of lncRNA orthologs in this region. Zebrafish *Chaserr* reconstructed by PLAR. Short conserved motifs identified from lncLOOM analysis of Chaserr sequences from 16 vertebrate species (ordered from Human to Zebrafish). Graded brown color indicates conservation per number of species. (b) Computational approaches to identify functional motifs in lncRNA sequences: A composition-based approach (as used in Seekr) vs a conservation based approach (as used in lncLOOM). (c) Schematic illustration of common secondary structures in RNA sequences

(discussed in Chen et al., 2016). The issue can be partially addressed by using programs such as Trinity (Grabherr et al., 2011) to reconstruct the transcriptome de novo. However, these tools also have their limitations, especially in the reconstruction of full transcripts from short RNA-seq reads (Hölzer & Marz, 2019). These shortcomings have left substantial parts of the transcriptomes from many vertebrate species unexplored. This has been a major incentive to develop more sensitive algorithms that can confidently predict orthologous genes by directly comparing transcripts to annotated lncRNAs from a query species (discussed below). Another option to overcome the challenges in d§e novo transcriptome assembly lies in long-read sequencing technologies, which have advanced substantially in recent years (Amarasinghe et al., 2020; Oikonomopoulos et al., 2020) and have already been used in transcriptome assembly in recent studies (Müller et al., 2021; Y. H. Sun et al., 2021). Although there are still several limitations related to the high error rate of long-read sequencing, we expect that it will become more applicable to transcriptome assembly as the technologies continue to advance and improved analysis methods for error correction are developed.

Faced with these challenges, the first studies that began to explore the evolutionary trajectories of lncRNAs were based on aligning (with BLASTN, BLASTZ, and LastZ) lncRNAs within clades of closely related species, particularly across mammals (Bu et al., 2015; Carninci & Hayashizaki, 2007; Kutter et al., 2012; Washietl et al., 2014). Although these studies were limited in that they could miss parts of the conservation because of the challenges mentioned above, they shed light on sequence and structural characteristics that are shared between lncRNAs. These characteristics can now be leveraged to establish more sensitive computational approaches for comparison between distantly related species (discussed in section 3). Importantly, these studies also established standards for the analysis of lncRNA sequences and revealed fundamental considerations that are important to avoid ascertainment bias of orthologous lncRNA identification (Hezroni et al., 2015).

2.1 | Considerations for the unbiased comparison of orthologous lncRNA genes

The majority of studies have investigated lncRNA evolution by projecting human lncRNA sequences across WGAs to identify candidate loci where the DNA sequence is conserved and also transcribed in distal species (Bu et al., 2015; Hezroni et al., 2015; Kutter et al., 2012; Necsulea et al., 2014; Washietl et al., 2014). These studies have largely focused on long intergenic noncoding RNAs (lincRNAs), which are lncRNAs that do not overlap PCGs, and are therefore easier to identify. In comparison to coding genes, lncRNA expression is notably more tissue-specific (Cabili et al., 2011) and can also vary significantly between individuals within a population (Kornienko et al., 2016). Taking this into account, the comprehensive catalogs of lncRNA transcripts for various species were typically reconstructed from RNA-seq data from multiple tissues. There are two conflicting caveats in the WGA-based approach. From one side, lncRNA conservation may be underestimated if the lncRNAs in other species are not detected due to the spatiotemporal nature of lncRNA expression (Cabili et al., 2011; Chodroff et al., 2010; Morán et al., 2012; X.-Q. Zhang et al., 2017) or because reconstruction algorithms fail to properly annotate them (e.g., a lncRNA transcript is merged to a PCG next to it, and not annotated as a lncRNA). This may be partially circumvented by carefully comparing expression levels from multiple tissue types or combining RNA-seq data with 3P-seq data (Hezroni et al., 2015). As an alternative, Necsulea et al., 2014 increased their estimates of the evolutionary age of lncRNA families by including species for which, according to a probabilistic likelihood, the absence of transcription could be attributed to read coverage or exonic length. Although this improved the sensitivity, it presents potential bias in the other direction, as the assumption that if a lncRNA is transcribed in some species, all sequences homologous to it in other species are also transcribed, is often too strong. Instead, it is possible that genomic loci may be conserved due to other constraints. This potential bias is illustrated through close inspection of the *Sox21* loci in distal vertebrate species. In humans, the 20 kb region surrounding *Sox21* contains three lincRNAs, *SOX21-AS1*, *linc-SOX21-B*, and *linc-SOX21-C*, that are expressed and overlap DNA sequences that are alignable to other mammals, while *linc-SOX21-B* even overlaps a region that is alignable to zebrafish. Despite this conservation between the DNA sequences, Hezroni et al. did not detect transcription of either *linc-SOX21-B* or *linc-SOX21-C* in any of the other species that were studied, including two primates. Rather, the significant alignment of human *linc-SOX21-B* to other genomes was attributed to its overlap with a highly conserved brain and neural tube enhancer (VISTA, element hs488; Visel et al., 2007). With this scenario in mind, it is advised to first reconstruct lncRNAs independently in each species and then subsequently compare the RNA sequences to each other.

It is now generally accepted that many lncRNAs are lineage specific and do not have recognizable homologues in distant species (Carninci & Hayashizaki, 2007; Church et al., 2009; Hezroni et al., 2015; Necsulea et al., 2014; Okazaki et al., 2002; Paralkar et al., 2014). This is strongly substantiated by studies that examined lncRNAs expressed in specific

tissues across closely related mammalian species (Chen et al., 2016; Kutter et al., 2012; Morán et al., 2012; Mustafi et al., 2013). For example, Kutter et al. found that only 60% of the lncRNAs that are expressed in mouse liver are also expressed in rat liver, while only 27% are also expressed in human liver (Kutter et al., 2012). As expected, the conservation of lncRNAs decreases dramatically across longer evolutionary distances. One of the first transcriptome-wide comparisons between lincRNA expression in zebrafish and mammals found that only 29 out of 567 (∼5%) lincRNAs that were identified in zebrafish had detectable putative orthologs in mammals (Ulitsky et al., 2011). A subsequent study explored lncRNA conservation beyond vertebrates by comparing 16 vertebrate species and sea urchin (Hezroni et al., 2015). Although hundreds of putative orthologs were detected beyond mammals, only 99 lincRNA genes could be traced to the last common ancestor of tetrapods and teleost fish, and no significant homology was detected between vertebrates and sea urchin. Although this number of putative orthologs between human and fish was substantially less than the 171 lncRNAs reported by Necsulea et al. (potentially due to the combination of factors previously discussed) the overall consensus is that homology between mammals and fish is detectable for <5% of human lncRNAs. These observations mark out two possible roadmaps for fruitful comparative genomic analysis. One possibility, if the majority of lncRNAs did arise de novo after major speciation events, is that we can use more intermediate evolutionary distances or a phylogenetic tree-based approach to carefully compare lncRNAs between more closely-related species. Alternatively, it is also plausible that many more lncRNAs are deeply conserved and their sequences contain spasmodic similarity that is not statistically detectable by alignment-based algorithms. If this is the case, it is expected that many lncRNAs will have synthetic counterparts in distal species, but without significant sequence similarity. Indeed, such lncRNAs have been observed (Amaral et al., 2018; Bryzghalov et al., 2020; Hezroni et al., 2015; Ponjavic et al., 2009; Ulitsky et al., 2011). For example, Hezroni et al. found no significant sequence homology, but over 2000 human lincRNAs had putative homologues with only positional conservation (referred to as syntologs) in the sea urchin genome, which was ∼600 more than the number expected by chance, evaluated by randomly placing lncRNAs in the sea urchin genome (Amaral et al., 2018; Bryzghalov et al., 2020; Hezroni et al., 2015; Ponjavic et al., 2009; Ulitsky et al., 2011). Such cases are worth exploring further as they will elucidate particularly short conserved motifs that are potentially easy to study further experimentally. Recent advances in computational methods for the comparative analysis of lncRNAs have therefore focused on alignment-free methods to uncover subtle motifs that have been purified in conserved lncRNAs from distant species or are shared between lncRNAs that perform similar functions in the same species. In the next section we provide examples of how the following dimensions: (1) syntenic conservation across species, (2) primary sequence conservation, and (3) the conservation of elements that adopt shared secondary structures, have been used to increase the sensitivity of lncRNA identification and motif discovery.

3 | COMPUTATIONAL APPROACHES THAT LEVERAGE SEQUENCE FEATURES OBSERVED IN CONSERVED IN lncRNAs

3.1 | Syntenic conservation increases the sensitivity of orthologous lncRNA detection in distal species

Although synteny alone is not informative for functional motif discovery, it is a fundamental property that can be used to identify and construct databases of putatively orthologous lncRNAs, and these datasets can be used as input for subsequent motif analysis (Table 1), which is especially useful in cases where significant sequence similarity is not detectable. Syntenic conservation was recently used to characterize a new subgroup of lncRNAs in mammals which are positioned at chromatin loop anchor points and the borders of topologically associating domains (TADs; Amaral et al., 2018). These lncRNAs were named topological anchor point RNAs (tapRNAs) and were very often associated with developmental genes with which they are co-expressed. Motif analysis, performed by the direct alignment of human and mouse tapRNAs, showed that they are enriched with binding sites of transcription factors (TFs) and zinc-finger proteins such as the chromatin organizer CTCF. Synteny has also been used to construct extensive resources of putative orthologous lincRNA sequences that are now available for future studies. For instance, thousands of full-length orthologous lincRNAs were generated using the PLAR (lncRNA annotation from RNA-seq data) pipeline by comparison of lncRNAs across vertebrates and sea urchin (Hezroni et al., 2015). Another good resource for obtaining lncRNA sequences is SyntDB a database of syntenic lncRNAs that are conserved across 11 primate species (Bryzghalov et al., 2020). The identification of lncRNAs that are positionally conserved has also been implemented as a key step in publicly available pipelines that have been developed for the identification of orthologous lncRNA genes. Below we

TABLE 1 Databases and pipelines for the retrieval of lncRNA homologues across species

Resource		Species analyzed	Description	Website	References
Pipeline	PLAR	17 vertebrates and sea urchin	Pipeline for reconstruction and of lncRNA transcripts from RNA-seq from multiple tissues and 3P-seq data	http://webhome.weizmann.ac.il/home/igoru/PLAR	(Hezroni et al., 2015)
	Slncky	29 mammals	lncRNA discovery tool based on RNA-seq from multiple species. Orthologs identified by synteny	https://slncky.umassmed.edu/ https://slncky.github.io/	(Chen et al., 2016)
	LincOFinder	Vertebrates and invertebrates	lncRNA discovery tool based on annotated orthologous PCGs and RNA-seq from multiple species. lncRNA orthologs identified by microsynteny cluster analysis	https://github.com/cherrera1990/LincOFinder	(Herrera-Úbeda et al., 2019)
	lncEVO	5 mammals	Pipeline for identification, reconstruction and comparison of conserved lncRNAs based on RNA-seq from multiple species	https://gitlab.com/spirit678/lncrna_conservation_nf	(Bryzghalov et al., 2021)
Database	Evo-devo	6 mammals and chicken	Catalogs of lncRNAs identified from RNA-seq in seven major developmental organs from early organogenesis to adulthood	https://apps.kaessmannlab.org/lncRNA_app/	(Sarropoulos et al., 2019)
	SyntDB	12 primates	lncRNAs identified from RNA-seq from multiple tissues. Implements *slncky* to identify syntologs across species	http://syntdb.amu.edu.pl	(Bryzghalov et al., 2020)
	RNAcentral	>50 species	Collection of 44 ncRNA databases including NONCODE, LncBOOK, LNCipedia, and lncRNAdb	https://rnacentral.org	(RNAcentral Consortium, 2021)
Study	Kutter et al.	Mouse and rat species	lncRNA catalogs constructed from RNA-seq and supported by H3K4me3-bound (ChIPseq) DNA data	(Kutter et al., 2012)	
	Washietl et al.	6 mammals	lncRNA catalogs constructed from RNA-seq from multiple tissues	(Washietl et al., 2014)	
	Necsulea et al.	11 vertebrates	lncRNA catalogs constructed from RNA-seq from multiple tissues	(Necsulea et al., 2014)	

describe how two such methods, *slncky* (Chen et al., 2016) and LincOfinder (Herrera-Úbeda et al., 2019), use syntenic conservation to increase the sensitivity of lncRNA identification.

The *slncky* pipeline is based on two steps, both of which rely on syntenic conservation. First, *slncky* aligns putative lncRNAs to syntenic noncoding transcripts from different species in order to evaluate coding potential. If an alignment is detected, slncky then interrogates putative open reading frames (ORF) that are longer than 30 nt and conserved in both species. For each ORF, *slncky* computes the ratio of nonsynonymous to synonymous mutations (dN/dS) and only considers the putative lncRNA as protein-coding if the ORF has a significantly low dN/dS ratio. Second, *slncky* increases the alignment score between two putative lncRNA orthologs by expanding the alignment to include the highly conserved sequences of their flanking protein-coding genes. Specifically, once *slncky* has defined syntenic regions that contain a noncoding transcript, it performs a second alignment of only the area 150,000 nt upstream and downstream of the syntenic region. To minimize false positive alignments (that can often arise from aligning repetitive elements), *slncky* aligns each lncRNA to shuffled intergenic sequences and determines an empirical 5% threshold for classifying significant alignment scores. Using this approach, the authors were able to identify 1466 out of 1521 (>95%) of orthologous lncRNAs previously reported between human and mouse and also identify a further 121 pairs (8%) of the homologous human-mouse lncRNAs that were previously classified as species-specific. They also found that 18% of lncRNAs that are expressed in mammalian pluripotent cells were likely present prior to the divergence between rodents and primates (Chen et al., 2016). Although *slncky* is useful, it is largely dependent on the quality of pairwise WGAs to project lncRNA expression to loci in any other species and may therefore have limitations in comparing more distant species.

The use of synteny to identify orthologous lncRNAs was recently expanded to establish another pipeline, LincOFinder, that was used to infer putatively conserved lncRNAs between human and amphioxus (Herrera-Úbeda et al., 2019). Instead of relying on WGAs to define syntenic regions, LincOFinder creates two lists of genes: one list contains all the genes in the reference species sorted by genomic position and the second list contains all corresponding orthologs of protein-coding genes in the species being interrogated. These orthologs are identified by using known sets of orthologous families or helper programs such as Orthofinder (Emms & Kelly, 2015). Sets of candidate genes (that may potentially define a microsyntenic region) in the interrogated species are then selected if they are orthologs of genes that neighbor a lincRNA in the reference species. The coordinates of these candidate genes form the input to a UPGMA hierarchical clustering algorithm which identifies microsyntenic clusters that contain the orthologous genes that are sufficiently close together. Each of these clusters are then scanned for expression of a lincRNA in the interrogated species. Using this approach, the authors were able to identify 16 lincRNAs putatively conserved between human and amphioxus, including *HOTAIRM1* which is located in the anterior part of the Hox cluster across several vertebrate lineages (Gardner et al., 2015; H. Yu et al., 2012).

3.2 | Computational frameworks focused on short motifs

Although the primary sequences of lncRNAs are not well conserved, multiple studies have identified short conserved motifs using alignment-based approaches (Chureau et al., 2002; Hezroni et al., 2015; Jin et al., 2021; Ulitsky et al., 2011) and experimental techniques (Ilik et al., 2013; Quinn et al., 2014, 2016) to uncover short functional domains in lncRNAs that are evolutionary conserved. For example, in Hezroni et al., 2015 the direct comparison of RNA sequences using BLASTN identified short patches that were alignable in lincRNAs that were conserved in human and one of 15 other vertebrates. Although short conserved patches could be detected as far as shark, the length of these patches significantly decreased in fish species with averages less than 100 bases, and approaching the lower limit of lengths that can be detected by BLASTN. To overcome limitations in the alignment, the study also characterized motif enrichment across lncRNAs in each species by counting the number of occurrences of all possible 6mers in exonic sequences compared to the number of random occurrences in shuffled sequences with preserved dinucleotide frequencies. This approach is similar to the n-gram metrics that were systematically evaluated against alignment-based approaches for the analysis of lncRNA sequences in Noviello et al. (2018), where each n-gram would comprise the six consecutive nucleotides in each 6mer. From their analysis, Hezroni et al. identified 31 motifs that were enriched in at least 12 species, thus capturing motifs enriched in lncRNAs in both mammals and fish. A substantial fraction of these motifs corresponded to exonic splicing enhancers (ESEs), were purine-rich, or composed combinations of CUG and CAG which form binding sites for the splicing factors CUG-BP and Muscleblind. This together with additional studies (Gil & Ulitsky, 2018; Haerty & Ponting, 2015; Schuler et al., 2014; J. Y. Tan et al., 2020) provides evidence that the

subsequences that control the processing of lncRNA transcripts are functionally important and under purifying selection, hinting that in many cases the biological function of the lncRNA may be incidental to the process of its maturation.

In addition to mapping motif enrichment across lncRNAs that are evolutionary conserved, motif enrichment has also been described across more focused sets of lncRNAs that can be potentially considered as coming from the same gene family. Gil and Ulitsky (2018) explored the differences in sequence composition and chromatin landscape between enhancer regions that transcribed lncRNAs and enhancers that transcribed much shorter and less stable enhancer-RNA (eRNA) molecules. They found that, in comparison to enhancers that only produce eRNAs, enhancers that harbor lncRNA transcripts have heightened activity and are significantly enriched with motifs recognized by specific RBPs, particularly splicing factors that are required for the maturation of the lncRNA. They inferred, from analyzing expression data in which several splicing factors were knocked down, that this heightened enhancer activity is driven by the maturation of the lncRNAs. This was further evidenced in a similar study that reported the significant enrichment of splicing-associated motifs in multi-exonic lincRNAs that are produced from enhancers compared to single-exonic counterparts (Tan et al., 2020). Likewise, their analysis also showed that enhancers that are associated with multi-exonic lncRNAs have a 2.5-fold increase in enhancer activity. Both studies also reported that lncRNA producing enhancers are enriched with binding sites for proteins involved in chromatin remodeling and loop formation, specifically CTCF. To achieve their results these studies used motif analysis tools such as AME (McLeay & Bailey, 2010), FIMO (Grant et al., 2011), and TOMTOM (Gupta et al., 2007) that scan sequences for known motifs and are not reliant on sequence alignments (Table 2).

There are also now several examples where the mature lncRNA product is functional, and in many of these cases the functionality can be attributed to short conserved motifs that correspond to miRNA binding sites or RBPs that confer post-transcriptional function. One well known example is *Cyrano (OIP5-AS1* in human), which harbors an extensively paired site to miR-7, which is required for degradation of miR-7 (Kleaveland et al., 2018; Ulitsky et al., 2011). The degradation of miR-7 enables CDR1as (a circular RNA that has many binding sites to miR-7) to accumulate in the cytoplasm of neurons, where it potentially regulates neuronal activity (Kleaveland et al., 2018). The miR-7 binding site in *Cyrano* is contained within a deeply conserved stretch of 67 nucleotides, which is the only region that is significantly alignable between human (8862 nt) and zebrafish (4630 nt) by BLASTN. Another well-known example is NORAD, a highly conserved cytoplasmic lncRNA that antagonizes the repression of mRNA levels of genes involved in cell division by sequestering Pumilio proteins (Elguindy et al., 2019; Kopp et al., 2019; S. Lee et al., 2016; Tichon et al., 2016). Human NORAD contains at least 17 Pumilio recognition elements (PREs) that bind PUM1 and PUM2 proteins (Tichon et al., 2016).

Collectively the findings in these studies support the notion that lncRNA function often requires only short conserved patches that are under selection, and that the remainder of the sequence can tolerate extensive changes. This concept has been applied in recent frameworks, specifically developed to compare noncoding sequences (Table 2). Below we discuss two of these frameworks, namely SEEKR (*SE*quence *E*valuation from *K*mer *R*epresentation; Kirk et al., 2018) and LncLOOM (Ross et al., 2021) that use alternative implementations of n-grams to home in on short motifs (or kmers) that underlie lncRNA function (Figure 1b).

Proteins that function via similar mechanisms adopt similar folds and bear specific domains that can be described by statistical models of amino acid sequences. This enabled the large-scale classification of protein families within and across species. In contrast, lncRNAs that function in similar ways often lack detectable significant sequence homology. For example, *Xist* and *Kcnq1ot1* are both known to recruit the Polycomb Repressor Complex to repress their target genes in *cis* (Lee & Bartolomei, 2013). However, the lack of sequence homology between them made it difficult to predict this shared mode of action from their linear sequences (Kirk et al., 2018). To overcome this limitation, SEEKR uses a statistical approach to compare profiles of kmer abundance between different lncRNA sequences. Kmer profiles are derived for each lncRNA sequence by counting the occurrences of 4-, 5-, and 6mers, and these profiles are then normalized by the length of the sequence and compared with Pearson correlation. The authors used SEEKR to compare human and mouse lncRNAs with putatively related functions. In both species, SEEKR was able to cluster well-known *cis*-repressive lncRNAs (including: *XIST*, *TSIX, KCNQ1OT1*, *UBE3A-ATS*, *ANRIL/CDKN2B-AS1*, and *Airn*) that were characterized by a high abundance of AU-rich kmers, separately from *cis*-activating lncRNAs (including: *PCAT6*, *HOTTIP*, *LINC00570*, *DBE-T*, and *HOTAIRM1*) that had a significant enrichment of GC-rich kmers. Overall, their results show that kmer-based quantitation can infer related functions in lncRNAs, irrespective of their low sequence homology.

In an alternative approach, we recently developed LncLOOM, a new framework that identifies combinations of short motifs that are found in the same order in putatively homologous sequences from different species. To identify

TABLE 2 Computational tools that have been used for comparative analysis of lncRNA sequences

Program	Algorithm description	Recommended dataset for lncRNA sequence analysis	Example of study where applied
BLASTN	Local alignment-based discovery of segments with statistically significant similarity	Closely related species	Identification of short conserved patches in lncRNAs across vertebrates (Hezroni et al., 2015)
MEME	Alignment-based motif discovery based on expectation–maximization (EM) algorithm	Can be used across intermediate evolutionary distances	Discovery of a functional conserved motif in Nron lncRNA, that binds to E3 ubiquitin ligase CUL4B to regulate ERα stability. (Jin et al., 2021)
AME	Statistical comparison of differential motif enrichment between sequences, based on a linear regression model	To compare groups of lncRNAs that have similar functions in the same or closely related species	Characterization of motifs that are enriched in lncRNAs that are transcribed from enhancer regions, and dictate higher enhancer activity. (Gil & Ulitsky, 2018)
TOMTOM	Statistical comparison of one or more DNA/RNA motifs against a database of known motifs	Functional annotation of motifs discovered in lncRNA sequences that have similar functions	Characterization of transcription factor binding sites that are enriched in multi-exonic lincRNAs that are produced from enhancers, compared to single-exon counterparts. (Tan et al., 2020)
FIMO	Statistical tool to locate specific motifs within a DNA/RNA sequence	To locate motifs of interest in a set of lncRNA sequences	
Seekr	Statistical comparison of kmer composition profiles across different lncRNA sequences	To identify and compare groups lncRNAs that have similar functions in the same species	Differential characterization of kmer profiles in *cis*-activating and *cis*-repressive lncRNAs in human and mouse. (Kirk et al., 2018)
LncLOOM	Implementation of the longest common subsequence problem using a graph-based approach and integer linear programming to identify conserved combinations of motifs that appear in the same order in sequences from different species	To compare lncRNA sequences that are orthologous across large evolutionary distances. Ideally, the set of sequences should represent species that monotonically increase in evolutionary distance	Discovery of deeply conserved, novel functional elements in *Chaserr* lncRNA that were experimentally validated to regulate CHD2 expression. (Ross et al., 2021)
TDF (Triplex Domain Finder)	Statistical ranking, based on motif over-representation, of lncRNA domains that are likely to interact with specific DNA targets, such as protomers of differentially expressed genes upon lncRNA perturbation	To map DNA binding domains in a single lncRNA sequence, for which differentially expressed genes are known from experimental knockdown of the lncRNA	Identification of known triple helices of lncRNAs *Fenderr*, *HOTAIR*, *MEG3*, and prediction of *GATA6-AS* triple helices that modulate cardiac mesoderm differentiation. (Kuo et al., 2019)
DeepLncRNA	Deep learning algorithm that classifies localization of lncRNAs based on sequence composition	To predict the localization of a specific lncRNA sequence, based on specific sequence motifs	Large-scale prediction of the localization of human and primate lncRNAs. (Gudenas & Wang, 2018)
iLoc-lncRNA	SVM is a machine-learning algorithm based on the statistical learning theory that classifies localization of lncRNAs based on octamer composition lncRNA sequences	To predict the localization of a specific lncRNA sequence, based on specific sequence motifs	Large-scale prediction of the localization of lncRNAs from the RNALocate database. (Su et al., 2018)

conserved motif combinations, LncLOOM builds a directed graph from a set of homologous sequences that are ideally ordered from species with a monotonically increasing evolutionary distance with respect to a query sequence. Each sequence is modeled as a layer of kmers, where each kmer represents a node in the graph and identical nodes in consecutive layers are connected by edges. Motif discovery is then performed using integer linear programming to find long nonintersecting paths in the graph. Essentially, LncLOOM aims to identify the longest common subsequence (Maier, 1978) in a set of homologous sequences. The LncLOOM approach is underpinned by the assumption that the linear order of short subsequences has been conserved across long evolutionary distances. Although the order of kmers may not always be important for function, LncLOOM can capture combinations of kmers where the order has been maintained because the functionality of the conserved elements is aided by sequence or structural context in the longer RNA molecule. More importantly, it is unlikely that large combinations of short conserved kmers that appear in the same order were independently gained during evolution. Rather, it is more plausible that these elements are remnants of their rapidly evolving sequences and have been evolutionarily selected because they are important to the function of the lncRNA. From a computational standpoint, the constraint of the linear order of kmers increases the power of motif discovery because an ordered set of *k*-mers is much less likely to appear by chance than any one of its possible permutations, and so while the presence of each of the kmers or even their combination might not be statistically significant, the addition of the conserved order constraint enables discovery of significant conservation. To demonstrate the power of LncLOOM, the framework was used to analyze the sequences of *Cyrano*, *Chaserr*, and *Libra* (a zebrafish lncRNA that is homologous to the *Nrep* mRNA in mammals) across vertebrate species. In *Cyrano*, LncLOOM identified a set of nine ordered kmers that are conserved in 17 vertebrate species from human to zebrafish, seven of which were also conserved to elephant shark. In addition to known miRNA and protein binding sites, LncLOOM identified novel functional elements in *Chaserr* that were conserved throughout vertebrates and were experimentally validated to regulate CHD2 expression (Ross et al., 2021).

3.3 | Comparison of lncRNA secondary structures

RNA secondary structures are known to mediate RNA-protein and RNA–RNA interactions. They have been shown to be critical to the biogenesis and function of many ncRNAs such as miRNAs, small nucleolar RNAs (snoRNAs), and transfer RNAs (tRNAs; Bhartiya & Scaria, 2016; Parisien et al., 2013). They also contribute to the stability, processing, and localization of mRNA molecules as in the case of riboswitches, terminal stem loops and 3′ stem loops in histone mRNAs (Mignone et al., 2002; Singh & Singh, 2019; Svoboda & Di Cara, 2006; D. Tan et al., 2013). Due to their functional importance, such elements are often constrained during evolution more than their flanking linear RNA sequences. For instance, *TERC*, the RNA component of the telomerase complex, is structurally very conserved across vertebrates despite sharing only 60% sequence identity between human and mouse (Seemann et al., 2017; Wang et al., 2016). Likewise, we expect that the low sequence conservation of lncRNAs across vertebrates does not preclude the existence of conserved structural motifs that underpin their function (Diederichs, 2014; Guttman & Rinn, 2012; Wutz et al., 2002). Indeed, there is some evidence from experimentally solved structures of conserved secondary structure elements, including in *XIST* (Pintacuda et al., 2017; Smola et al., 2016), *MALAT1* (Brown et al., 2014; McCown et al., 2019), *Cyrano* (Jones et al., 2020), *MEG3* (Uroda et al., 2019; Zhang et al., 2010), and *COOLAIR* (Hawkes et al., 2016). We describe the proposed contributions of these structures to lncRNA function in more detail in section 4, when we consider experimental approaches to uncover functional motifs.

From a computational perspective, structural motif discovery in lncRNA sequences is a major challenge (Figure 1C). Although many methods have been developed to predict the secondary structure of RNA sequences (reviewed in B. Yu et al., 2020), the accuracy of these algorithms for predicting the full structures of long RNAs is limited. The most popular programs rely on minimum free energy (MFE) calculations, for which the accuracy of predicted base pairs has been reported to be as low as 40% for RNAs longer than 500 nt (Doshi et al., 2004; Lorenz et al., 2016). This is mainly due to an incomplete understanding of the thermodynamic parameters that govern RNA molecular interactions and stability; such that any small change to the thermodynamic models that underpin MFE calculations, can generate significantly different optimal structures (Rogers et al., 2017; Schroeder, 2018). This becomes particularly relevant in the case of longer RNAs where the number of possible folds is enormous, and many possible optimal structures can be generated with no significant differences in their MFE scores. Another caveat in this approach is that the model assumes that the optimal structure of the RNA molecule corresponds to a global free energy minimum. This assumption does not account for the co-transcriptional folding of RNA transcripts (Watters et al., 2016; Yu et al., 2021),

which may result in the RNA adopting local folds such that the kinetics of the molecule differ from the global free energy minimum of the whole RNA. Another conceptual problem is that even random RNA sequences can adopt highly stable structures in cells (Schultes et al., 2005), and so the presence of a particular structure cannot usually be used as evidence for its functional importance. Similarly, random sequences have multiple predicted plausible folds, that may comprise a similar number of base pairs and have similar stability (Rivas & Eddy, 2000). This highlights the challenge in deciphering which RNA molecules harbor biologically relevant structures as the presence of a structural motif in an RNA sequence from one species (determined either experimentally or predicted) does not provide evidence that the motif is functional. To infer functional importance, it is necessary to uncover statistical evidence that the motif has been evolutionarily conserved beyond phylogenetic expectation, that is, that the functional constraints are imposed on the structural motif and not the primary sequence in which the motif is embedded (reviewed in Rivas, 2021). Covariance models are the gold standard for such analysis (Eddy & Durbin, 1994; Griffiths-Jones et al., 2003) and have been instrumental in predicting the structure of other ncRNAs, such as rRNA (Gutell et al., 2002). However, opinions differ on the extent to which covariation statistically supports predicted secondary structures in lncRNA sequences (Rivas, 2021). Here, we briefly discuss the handful of studies that have used covariance based approaches to predict conserved structural motifs in lncRNAs.

Covariance models have been implemented in several computational tools that predict RNA secondary structure including: R-scape (Rivas et al., 2017) and frequently used programs such as INFERNAL (Nawrocki et al., 2009), CMFinder (Yao et al., 2005), EvoFold (Pedersen et al., 2006), and others (reviewed in Tahi et al., 2017). To circumvent the inaccurate prediction of the secondary structure of longer RNA molecules, studies have predicted local structures of shorter putative functional regions within lncRNAs such as *NORAD* (Tichon et al., 2016, 2018), *Cyrano* (Jones et al., 2020), and *XIST* (Wutz et al., 2002). These studies used programs such as EvoFold that uses a phylogenetic stochastic context-free grammar (phylo-SCFG) to model coevolving base pairs within the structural motifs; ScanFoldScan (Andrews et al., 2018) that predicts local structures across an RNA sequence and then searches the RFam database for conservation matches; or CMFinder (Yao et al., 2005) that uses covariance models to improve the accuracy of the predicted structural motifs. Although these programs provide evidence of motif conservation, they do not statistically test if the level of covariance is significant, such that the motif is conserved beyond random expectation. Nonetheless, CMFinder was used by Seemann et al. (2017) to screen vertebrate genomes for putatively conserved RNA structures (CRSs). The authors reported CRSs in several lncRNAs including *XIST* and *MALAT1* and noted that the density of CRSs decreases from the 5′ to the 3′ end of lncRNAs. On the other hand, R-scape provides a method that quantitatively tests whether covariance supports the presence of a conserved RNA secondary structure (Rivas et al., 2017). R-scape analysis detected no statistically significant support for the secondary structures determined experimentally in lncRNAs *HOTAIR*, *SRA*, and *XIST*. A later study proposed that the covariance support of evolutionary conserved structures in *HOTAIR*, and other lncRNAs, can be increased by extending the depth of the alignment beyond mammals to increase variation or adjusting the default parameters of R-scape to scan shorter window sizes or use different statistical metrics (Tavares et al., 2019). However, this has since been rebutted by evidence that the alignments of *HOTAIR* and *SRA* do have sufficient variation to detect covariations, yet still lack signals of evolutionary conserved structures (Rivas et al., 2020). All in all, the analysis of conserved structural motifs in lncRNA sequences remains a major challenge with very few positive controls that can be used for better method development.

The best approach may be to consider structural elements in combination with other features as a means to enhance the sensitivity of lncRNA classification. Along these lines, Quinn et al. (2016) used a combination of synteny, microhomology, and conserved secondary structural elements to identify *roX1* and *roX2* lncRNAs across diverse *Drosophila* species. In the past, homology between these genes has remained undetected because the similarity between their sequences is comparable to that of random sequences. In an earlier study, Tycowski et al. (2012) presented a structure-based bioinformatic screen that successfully identified conserved expression and nuclear retention elements (ENEs) in the genome of diverse viral genomes. The ENE, a structured element that is composed of a stem-loop structure with an asymmetric internal U-rich loop, was first identified in the genome of Kaposi's sarcoma-associated herpesvirus (Conrad & Steitz, 2005). The element is located ~120 nt upstream of the polyadenylation site (PAS) of PAN lncRNA, and plays an essential role in stabilizing PAN lncRNA through triple-helix formation with its poly(A) tail, thereby preventing the initiation of RNA decay (Conrad et al., 2006; Mitton-Fry et al., 2010). The conserved ENEs identified by Tycowski et al. (2012), were indeed located in putative lncRNAs that were transcribed from intergenic regions of the viral genomes.

4 | EXPERIMENTAL STUDIES OF FUNCTIONAL ELEMENTS IN lncRNA GENES

In parallel to the computational predictions, functional elements can also be identified and studied in detail experimentally, as we discuss next.

4.1 | Systematic dissection of lncRNA sequences that dictate subcellular localization

To further understand the mechanisms that drive lncRNA function, several studies have used massively parallel reporter assays to systematically interrogate lncRNA sequences to decipher elements that dictate their function (Figure 2a). For example, to identify elements in lncRNAs and mRNAs that can force nuclear localization, Lubelsky and Ulitsky cloned thousands of short fragments that tiled the exons of human lncRNAs and 3′UTRs of selected mRNAs into the untranslated region of an otherwise cytoplasmic mRNA. By comparing reporter RNA from cytoplasmic and nuclear fractions of cells that were transfected with the library, they attributed nuclear accumulation to a short C-rich sequence (42 nt) derived from an Alu repeat, named SIRLOIN. Using RNA immunoprecipitation (RIP) with an

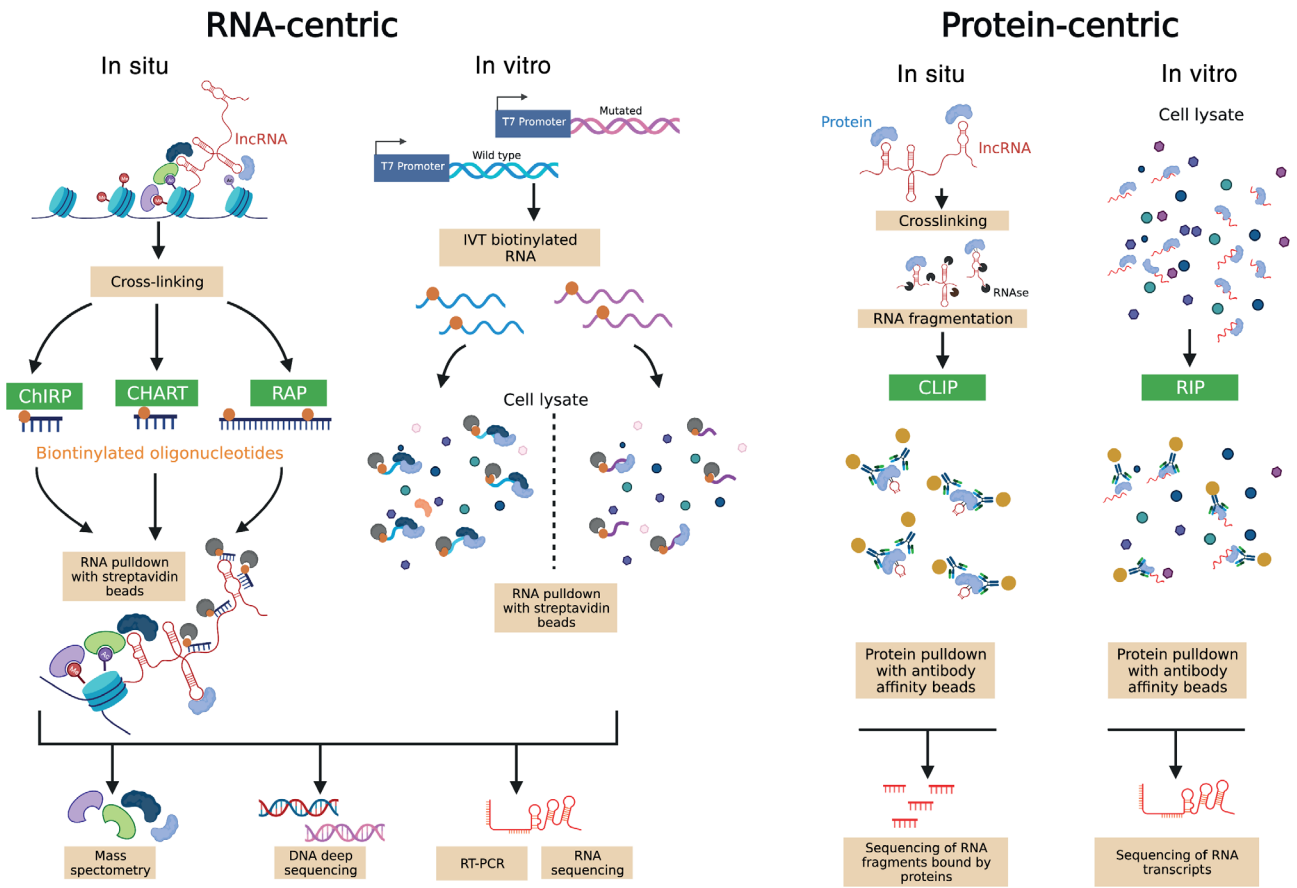

FIGURE 2 Experimental approaches that have been used to investigate the functional importance of conserved motifs in lncRNAs. (a) A high throughput tilling assay to identify sequence elements that modulate RNA localization. (b) RNA-centric approaches (left) can be used to identify proteins, RNA, or DNA bound to an RNA of interest. Cross-linking approaches allow for *in situ* discovery of RNA interaction partners, through pull down with biotinylated antisense probes specific to the RNA of interest. In vitro transcription of biotinylated RNA molecules of wild-type and mutated fragments can be used to investigate interactions with conserved regions within the RNA molecule. Protein-centric approaches (right) identify protein:RNA interactions with a protein of interest, based on protein pulldown with antibody affinity beads. (c) Perturb-and-rescue experiments. Left: Endogenous rescue with fragments that encode functional domains of a lncRNA transcript, following knockdown of the endogenous express lncRNA using CRISRPi. Right: Cross-species rescue of post-transcriptional perturbation of lncRNA transcript using ASOs that target conserved regions

hnRNPK-specific antibody, they further showed that SIRLOIN drives nuclear localization by recruiting hnRNPK (Lubelsky & Ulitsky, 2018). In a follow-up study, the authors investigated the mode of action of SIRLOIN at higher resolution by using a suite of massively parallel RNA assays and libraries that contained thousands of sequence variants to pinpoint the regions within the SIRLOIN element that are essential for the nuclear retention of lncRNA transcripts (Lubelsky et al., 2021). Although SIRLOIN contains multiple CCC elements that bind hnRNPK, their findings suggest that nuclear retention is solely dependent on hnRNPK binding to the GCCUCCC element that is precisely positioned in the SIRLOIN core (Lubelsky et al., 2021). In an independent study, Shukla et al. also used a tiling-reporter assay to characterize sequence elements that dictated the localization of lncRNA transcripts. They identified a shorter C-rich motif (15 nt), along with 108 other RNA elements that increased nuclear localization of their cytoplasmic reporter (Shukla et al., 2018). As an alternative to the tiling strategy, a recent study introduced mutREL (RNA elements for subcellular localization by sequencing), which is a high-throughput method coupled with random mutagenesis to identify motifs that specify localization (Yin et al., 2020). In this approach, DNA fragments of candidate genes are PCR amplified from their genomic DNA and then cloned into GFP-reporters which are then stably integrated into the genome of cells. RNA is then isolated and sequenced from the chromatin, nucleoplasm, and cytoplasm fractions. The random mutagenesis is achieved by using an error-prone PCR (McCullum et al., 2010). The authors used mutREL to investigate lncRNAs *Malat1*, *Neat1*, *NXF1-IR*, and uncovered an RNA motif that recognizes the U1 small nuclear ribonucleoprotein (snRNP) to promote lncRNA-chromatin retention (Yin et al., 2020).

4.2 | Affinity assays uncover lncRNA motifs that facilitate lncRNA:protein, lncRNA:RNA, and lncRNA:genome interactions

Many of the proposed mechanisms for lncRNAs have been uncovered from affinity-based assays that can also be used to attribute binding to a protein, DNA, or RNA to specific regions within lncRNA molecules. To date, an array of versatile methods have been developed to map interactions with RNA molecules (reviewed in Ramanathan et al., 2019). These methods can be broadly categorized into RNA-centric methods which identify interacting partners bound to an RNA of interest; or protein-centric methods which identify RNA molecules bound to a protein of interest. In this review we give selected examples of how some of these assays have been used to investigate the molecular mechanism of certain lncRNAs, several more examples have been described in (Constanty & Shkumatava, 2021).

RNA-centric methods include an array of hybridization methods that capture RNA interactions with protein, RNA and DNA (Figure 2b). These methods have been widely used to explore the modes of action of many lncRNAs (reviewed in Cao et al., 2019). In 2011 two main methods were developed to investigate the in vivo genomic binding sites of chromatin associated lncRNA, namely: ChIRP (Chromatin isolation by RNA purification; Chu et al., 2011, 2012) and CHART (capture hybridization analysis of RNA targets; M. D. Simon, 2013; M. D. Simon et al., 2011). The methods are similar in that they use biotin-conjugated 20–25mer DNA probes to purify a lncRNA of interest RNA and determine its associated chromatin fraction by DNA deep sequencing. Subsequently, RAP (RNA antisense purification), a method that increases specificity by using longer DNA probes, was developed (Engreitz et al., 2015; Engreitz et al., 2013). All three methods have been successfully used to establish high resolution maps of genomic binding *roX2* (Chu et al., 2011; Simon et al., 2011) and *Xist* (Engreitz et al., 2015; Engreitz et al., 2013) lncRNAs across the X chromosome. Using ChiRP, Chu et al. (2011) also showed that *HOTAIR* preferentially binds genomic regions containing a GA-rich motif. More recently, several high throughput methods have been developed to systematically map RNA:chromatin interactions on a genome-wide scale (reviewed in Kato & Carninci, 2020). Additionally, RNA and protein that is enriched from purification by these methods can also be isolated and subjected to RNA and protein analysis. For example RAP-RNA was used to show that *Malat1* interacts with many nascent pre-mRNAs as a means to localize to the chromatin at active genes (Engreitz et al., 2014). While a combination of RAP and mass spectrometry revealed that *Xist* interacts directly with SHARP, which also interacts with the SMRT corepressor to activate HDAC3 deacetylation activity on chromatin (McHugh et al., 2015). In a recent study, RAP was used to characterize the functional mechanism of ADEPTR, a lncRNA recently found to be necessary for activity-dependent changes in synaptic transmission and structural plasticity of dendritic spines (Grinman et al., 2021): ADEPTR contains a conserved sequence that binds to actin-scaffolding regulators AnkB and Spnt1, which are then transported to distal process through ADEPTR interactions with the motor protein Kif2A. Although cross-linking methods have proved to be useful, they can also be technically challenging due to the inefficiency of RNA-pulldowns, in particular for RNAs that are not abundant. As an alternative, the high-throughput method incPRINT, was recently developed to screen protein interactions with an RNA of interest in

cells (Graindorge et al., 2019). This is achieved by using a library of tagged proteins with a target RNA that is tethered to a luciferase detector. The method enables the characterization of in-cell RNA-interacting proteomes, as well as the mapping of protein bindings across different regions of long RNA transcripts.

In addition to studying the interactomes of endogenously expressed lncRNAs, affinity-based RNA pulldown methods that are based on in vitro transcription (IVT) of biotinylated target can also be used to characterize interacting proteomes of lncRNAs (reviewed in Cao et al., 2019). The method is advantageous in that it is fast and can be used to enrich RBPs associated with unabundant target RNA. However, the approach also has limitations as in vitro transcribed RNA may not have the same modifications or adopt the same structure as cellular RNA, and they may encounter in lysates proteins that are not normally found in their proximity in cells, for example, a nuclear lncRNA may bind strongly to cytoplasmic proteins in lysates, but not in cells. Nevertheless, IVT affinity-based RNA pulldown has been widely used to characterize lncRNA:protein interactions (Huang et al., 2017; Noh et al., 2016; Unfried et al., 2021). We recently used an IVT approach, combined with mutagenesis, to interrogate the functional importance of short conserved elements in the last exon of *Chaserr* (Ross et al., 2021). Following incubation with cell lysate, WT and mutated transcripts were pulled down with streptavidin beads and interacting proteins were identified by mass-spectrometry, and comparison of the identified interactomes homed in on proteins bound specifically to the conserved sites in *Chaserr*.

Protein-centric methods include well-established assays such as RIP and cross-linking and immunoprecipitation (CLIP) to identify endogenous RNAs that interact with a protein of interest (Figure 2b). RIP is an effective assay to discover RNA molecules that interact with specific proteins, however it is limited in that it does not map the precise regions of the RNA molecule that form the interaction. On the other hand, CLIP provides a method to map the nucleotide resolution of the binding sites that interact with specific proteins (Ule et al., 2003). Many variants of CLIP have now been developed (Hafner et al., 2021) and have been extensively used to understand the molecular mechanisms of lncRNA function, particularly in lncRNAs that are associated with human disease (reviewed in Jonas et al., 2020 and J.-M. Carter et al., 2021). Much of the experimental data that has been generated from CLIP experiments in recent years, has been integrated into public repositories such as GEO (Barrett et al., 2013; Edgar et al., 2002), CLIPdb (Yang et al., 2015), and starBASE (Li et al., 2014; Yang et al., 2010). Large-scale projects such as ENCODE have mapped the binding sites of hundreds of RBPs in certain human cell lines (Van Nostrand et al., 2020). Publicly available CLIP data has also been integrated in computational frameworks for lncRNA analysis: for instance, eCLIP data from ENCODE has been incorporated into LncLOOM to annotate evolutionarily conserved motifs that correspond to binding sites of RBPs (Ross et al., 2021).

4.3 | Discovery of functional structured RNA elements in lncRNAs

Generally, it is still unclear to what extent secondary structure is important to lncRNA function (Rivas, 2021; Ulitsky & Bartel, 2013). As it stands, the functional importance of structural elements have been experimentally studied in relatively few lncRNAs (Table 3). Indeed, these studies have provided key insights into how structural elements modulate the function and stability of some lncRNAs. One lncRNA where structure has proved to be relevant is *MEG3*. Based on computational prediction, Zhang et al. reported that the *MEG3* lncRNA contains three structural motifs (termed M1, M2, and M3), two of which were experimentally validated to be required for p53 activation by *MEG3* (Zhang et al., 2010). When the primary sequence of M2 was replaced by an entirely unrelated sequence that artificially folded into a similar structure, the transcript retained the functions of both p53 activation and growth suppression, in comparison to almost complete loss of function when the motif was deleted. More recently, chemical probing and atomic force microscopy revealed that the p53-activating core of *MEG3* comprises two evolutionary conserved distal motifs that interact by base complementarity to form pseudoknots (Uroda et al., 2019). Notably, one of these motifs overlapped the M2 motif identified by Zhang et al. They showed that point mutations that break the long-range interactions between these motifs disrupt the *MEG3* architecture and severely impair p53 activation, although the mechanism by which this occurs still needs to be elucidated. In addition to mediating the function of lncRNA molecules, structural elements have also been found to be essential to lncRNA stability. As an example, RNA crystallization of the 3′-end of MALAT1 revealed a triple helix structure that protects it from RNase degradation (Brown et al., 2014).

Multiple methods have been developed to determine the secondary structure of RNA molecules in vitro and in cells (reviewed in Zampetaki et al., 2018). So far, the majority of structural elements found in lncRNA sequences have been solved using a combination of chemical and enzymatic probing such as: SHAPE-seq, DMS-seq, and PARS. To overcome

TABLE 3 Examples of structural elements that have been supported by experimental data

LncRNA	Description	Method	References
MEG3	The p53-activating core comprises two evolutionary conserved distal motifs that interact to form pseudoknots	SHAPE, probing with 1M7, NMIA, 1M6, and DMS Hydroxyl radical footprinting Atomic force microscopy	(Uroda et al., 2019)
XIST	XIST A-region contains repeated stem-loop structures that recruit the PRC2 complex	Probing with RNAse V1 and DMS Targeted Structure-Seq	(Fang et al., 2015; Maenner et al., 2010)
	XIST A-repeat inter-repeat duplexes that span across the X-chromosome to facilitate the assembly SPEN	PARIS	(Lu et al., 2016)
CYRANO	Cloverleaf structure (100 nt) that contains the highly complementary site to miR-7	3S shotgun approach, SHAPE, probing with 1M7, NMIA, and S1 nuclease	(Jones et al., 2020)
SRA	Complex structural organization consisting of four domains that contain multiple loops and helical structures	SHAPE, probing with DMS and RNase V1	(Novikova et al., 2012)
HOTAIR	Four independently-folded highly structured domains	SHAPE, probing with DMS and terbium	(Somarowthu et al., 2015)
COOLAIR	Evolutionarily conserved complex structure with multi-helix junctions	3S shotgun approach, SHAPE, probing with 1M7	(Hawkes et al., 2016)
Braveheart	Adopts a modular fold that contains a short asymmetric G-rich internal loop that binds to CNBP	SHAPE, probing with 1M7 and DMS	(Xue et al., 2016)
ROX1 and *ROX2*	Evolutionarily conserved domains that contain tandem stem-loops, in which the roXbox motif is embedded	SHAPE and parallel analysis of RNA structure (PARS) analysis	(Ilik et al., 2013)
MALAT1	3′-end forms triple helix structures that stabilize the lncRNA	RNA crystallization	(Brown et al., 2014)
	3′-end contains evolutionary conserved tRNA-like structures	SHAPE, probing with DMS	(Zhang et al., 2017)
PAN	*Cis-acting* elements MRE and ENE are organized within a branched secondary structure comprised of three domains that contain multiple hairpin loops	SHAPE-mutational profiling (SHAPE-MaP)	(Sztuba-Solinska et al., 2017)

the challenges imposed from the length of lncRNA sequences Novikova et al. (2013) introduced the (3S) shotgun approach, which consists of two steps to determine subdomains that adopt modular folds in the context of full-length RNA structure. In the first step, the entire lncRNA sequence is chemically probed using SHAPE and DMS. Next, the sequence is divided into overlapping segments which are then probed individually to determine local structural domains that are subsequently compared to their folds obtained from the full-length sequence. This approach has been successfully applied to several lncRNAs including *Cyrano* (Jones et al., 2020), *COOLAIR* (Hawkes et al., 2016), and *Braveheart* (Xue et al., 2016). In particular, the structural study of *Braveheart* revealed key insights into its mode of action. It was found that *Braveheart* adopts a modular secondary structure and contains a 5′ asymmetric G-rich internal loop (AGIL) that recruits and antagonizes the zinc-finger protein CNBP, a TF that represses cardiac differentiation, to promote cardiac fate (Xue et al., 2016). More recent technologies have now advanced our capability to determine RNA structures in living cells and thereby map long-range RNA interactions across the transcriptome (Lu et al., 2016; Ziv et al., 2018). One such approach, PARIS (psoralen analysis of RNA interactions and structures) showed that *XIST* A-repeats form complex inter-repeat duplexes that span many kilobases across the X chromosome. The *XIST* A-repeats coordinate the assembly of SPEN, an *Xist*-binding transcriptional repressor that is required for X chromosome inactivation (XCI), across the X chromosome (Lu et al., 2016). The functional importance of structural motifs in RNA has also been demonstrated in viral genomic RNA (vRNA) and viral mRNA (reviewed in Ferhadian et al., 2018; Fernández-Sanlés et al., 2017). For example, in vivo enzymatic probing revealed that the efficient replication and infectivity of Influenza A virus (IAV) is dependent on the formation of stable structural motifs in the viral mRNAs (Simon et al., 2019), while conserved G-quadruplex structures have been reported in IAV vRNA (Tomaszewska et al., 2021).

Similarly, SHAPE analysis revealed conserved structural motifs in the RNA genome and ORF of hepatitis C virus that contribute to viral fitness and regulate specific stages in the life cycle of the virus (Mauger et al., 2015; Pirakitikulr et al., 2016).

Overall experimental studies corroborate the role of structural elements in lncRNA function, however some features may be worth exploring in future studies. First, previous analysis of the sequence and context preferences of RBPs has shown that RBPs often bind low-complexity RNA motifs and their specificity is influenced by the surrounding sequence context including secondary structures and their flanking nucleotides (Dominguez et al., 2018). Therefore, to fully understand how the functions of lncRNAs are encoded in their sequences, we also need to determine the factors that control the functionality of structured domains that are positioned in precise locations within their lncRNA molecules. Second, there is evidence to suggest that the structures of lncRNA molecules are dynamic and adopt different conformations across different subcellular compartments and/or cell states, subject to modifications or interactions with chromatin, proteins, or other RNA molecules (Sun et al., 2019). This kind of structural dynamics could dictate temporally variable roles of the same lncRNA, or nonmutually exclusive functions that are performed in different compartments of the cell. What may also be of interest is the relationship between structural changes that occur due to species-specific factors and the diversification of lncRNA function. Substantial levels of species-specific structural divergence has been previously reported. For example, the structures of small RNAs that are highly conserved between human and mouse undergo both shared and unique conformational changes (Sun et al., 2019).

4.4 | Perturb-and-rescue validates motifs that mediate regulatory activity of specific lncRNAs

Perturbation and rescue experiments provide a powerful approach to validate the function of conserved motifs and have been included in multiple studies to further understand how selected lncRNAs perform their function (Figure 2c). The Mendell Lab recently used an array of comparative genetic rescue experiments to interrogate two alternative pathways that have been proposed for *NORAD* function (Elguindy et al., 2019), a lncRNA that is required for genome stability in mammals (Lee et al., 2016). As mentioned earlier in this review, several studies have now found that *NORAD* acts in the cytoplasm as a binding decoy for PUM1 and PUM2 proteins, which in turn reduces the repression of PUM targets including key regulators of mitosis, DNA repair, and DNA replication genes (Elguindy et al., 2019; Kopp et al., 2019; Lee et al., 2016; Tichon et al., 2016, 2018). This mechanism is strongly supported by the high number of PREs located in the *NORAD* lncRNA. Alternatively, Munschauer et al. proposed that the regulation of genomic stability by *NORAD* is mediated through *NORAD*:RBMX interaction that facilitates the assembly of a ribonucleoprotein (RNP) complex in the nucleus (Munschauer et al., 2018), which includes proteins such as Topoisomerase I (TOP1) that are critical for genome maintenance. Similarly, this mechanism is supported by an RBMX binding site that spans the first 898 nt (15%) of the *NORAD* sequence. To directly interrogate the importance of PUM1/PUM2 and RBMX binding for *NORAD* function, Elguindy et al. generated a series of mutant *NORAD* constructs lacking either PUM or RBMX binding sites; as well as truncated constructs of domains that contain multiple PREs or the RBMX binding site. These constructs were integrated as potential endogenous rescues into the *AAVS1/PPP1R12C* locus of HCT116 cells. Following the depletion of endogenous *NORAD* transcripts using CRISPR interference (CRISPRi), it was found that the PRE-mutant construct was not able to rescue the increase in aneuploid cells, while the construct lacking only the RBMX binding site fully restored genome stability. Furthermore, no increase in aneuploid cells was observed after knockdown of *NORAD* in cells that expressed the truncated fragment containing one PRE-rich domain (referred to as the ND4 domain). This demonstrated the ability of the ND4 domain to function independently from the full *NORAD* transcript. In contrast to ND4, expression of the 5′ fragment comprising only the RBMX binding site had no rescue activity. These results by Elguindy et al. (2019) strongly suggest that *NORAD* interaction with PUM is critical to maintaining genome stability, while its interaction with RBMX may be dispensable for this function.

In the study of lncRNAs that have low sequence homology between species, the demonstration that loss of function of a lncRNA in one species can be rescued by the exogenous expression of the homologue from different species is a compelling approach to ascertain that the function of the lncRNA may indeed be encoded in more subtle sequence features that are common to both orthologs. This concept was illustrated by a comprehensive analysis of *roX1* and *roX2* lncRNAs that correlated changes in their binding potency and their functionality in diverse Drosophilid species across 40 million years of evolution (Quinn et al., 2016). Although *roX1* and *roX2* functionality is conserved across species, the sequence identity between orthologs is similar to that of random sequences. Likewise, *roX1* and *roX2* differ greatly in

sequence and size, yet they initially appeared to be functionally redundant in *D. melanogaster* and disruption of either lncRNA results in failed dosage compensation and male-specific lethality (Meller & Rattner, 2002), though more recent studies found some hierarchy in their functional importance (Valsecchi et al., 2021). The functionality of *roX* lncRNAs has been attributed to repeats of the short roXbox motif (8 nt), that is embedded in tandem stem-loop structures across their sequences (Ilik et al., 2013; Seung-Won et al., 2008). Quinn et al. used genomic occupancy maps in four species to show that the targeting of the *roX* lncRNAs to the X chromosome is conserved, however their precise binding sites differ considerably across the species. Nonetheless, when *roX* lncRNAs from other species were introduced into *roX*-null *D. melanogaster*, they were able to modestly rescue ~20% of the *roX*-null *D. melanogaster* male mutants by binding to *D. melanogaster* high-affinity sites (HASs; Quinn et al., 2016). Importantly, the modest cross-species rescue was consistent with prior reports that rescue efficiency decreases with increasing evolutionary distance (Seung-Won et al., 2008), suggesting that functionality decreases as the sequence similarity degrades but is still maintained at a lower level across evolution. Notably, rescue efficiency was substantially improved by chimeric constructs with an increased number of stem-loops. Furthermore, when the genomic occupancy maps of *roX1* and *roX2* were compared within the same species, it was found that in certain species, including *D. virilis* and *D. busckii*, the redundancy between the two lncRNAs had degenerated and *roX1* binding was less potent. Interestingly, genetic rescue with engineered constructs that contained varying numbers of the repetitive stem-loop structure from different ortholog species, confirmed that the decreasing potency of *roX1* in these species correlated with loss of stem-loops and roXbox motifs. These results suggest that *D. virilis* and *D. busckii roX1* lncRNAs have vestigial function due to loss of key sequence elements, while the preservation of these elements in in *D. melanogaster* has maintained the overall roX1-roX2 functional redundancy (Quinn et al., 2016). As another example, a cross-species rescue was also used to characterize the function of *TERMINATOR* and *PUNISHER*, two lncRNAs that play a critical role in cardiovascular development across vertebrates (Kurian et al., 2015). Both lncRNAs are conserved across vertebrates and PhyloP analysis revealed relatively short regions (250–500 bp) conserved across their sequences. Loss-of-function experiments in zebrafish embryos using morpholino antisense oligonucleotides (MOs) against the conserved regions and splice sites in *terminator* compromised development at the gastrulation stage and resulted in >70% lethality, while MO injections targeting *punisher* resulted in severe defects in branching and vessel formation. Both morphant phenotypes observed in zebrafish embryos were sufficiently rescued upon coinjection with the respective human lncRNA sequences of *TERMINATOR* and *PUNISHER*. A similar approach was also previously used to investigate if *Cyrano* lncRNA has conserved function in human and zebrafish embryonic development, given that short patches of conservation were interspersed across their sequences (Ulitsky et al., 2011).

Another powerful way to ascertain the essential function of a specific sequence domain, is to rescue the perturbed phenotype by introducing a similar domain encoded from an alternative gene that performs a similar function. This approach was recently used to establish an evolutionary link between XCI by *Xist* lncRNA and RNA-mediated endogenous retrovirus (ERV) silencing (Carter et al., 2020). The authors show that SPEN also interacts with structural motifs in ERV RNA to recruit chromatin silencing machinery to loci from which ERVs are transcribed. These structural motifs are similar to the A-repeats in *Xist* lncRNA, which are derived from TEs and required for the recruitment of Spen and subsequent gene silencing in *cis* (Giorgetti et al., 2016; Wutz et al., 2002). Using CRISPR-Cas9, they show that knock-in of an ERV-derived Spen binding site into an A-repeat deficient *Xist* is sufficient to rescue interaction with Spen and restores strict local gene silencing in *cis*. Overall their findings suggest that *Xist* mediates XCI by repurposing antiviral chromatin silencing machinery and that *Xist* acquired its ability to silence genes in *cis* from TEs derived from ancient ERVs. This confirms previous suggestions that the A-repeat sequence is derived from the insertion and duplication of an ERV (Elisaphenko et al., 2008), and sheds light on how ERVs have impacted the evolution of functional lncRNAs.

5 | CONCLUSION

Progress in understanding the biology of individual lncRNA genes has given rise to specific concepts of what types of functional elements are expected to be found in lncRNA genes. These have been the basis of computational tools looking for short conserved elements, for evidence of selection on secondary structures, and/or for repeated occurrences of short elements. Still the number of lncRNAs for which there is convincing evidence of a specific mode of action and the sequences underlying it remains very low, and they correspond to a small minority of lncRNAs with reported functions. Since this is now the main frontier in lncRNA research, we expect that many additional functional domains within lncRNAs will be characterized in the near future, leading to both the refinement of the computational and experimental tools developed so far, and catalyzing development of new methodologies. These will likely also be able to

go beyond discovery of individual domains and decode combinations of functional elements that co-exist within the same lncRNAs, as well as hierarchical grouping of functional domains found in different lncRNAs into families with different granularity levels, similar to the situation in the research of domains in proteins. Once such tools become available, they will enable large-scale prediction of lncRNA mode of action and possibly even functions. Since validation of predicted functional aspects is typically easier than their de novo experimental discovery, these tools are likely to play instrumental roles in a leap forward of research programs focused on lncRNA biology.

ACKNOWLEDGMENT

The authors thank Yoav Lubelsky and members of the Ulitsky laboratory for helpful discussions and comments on the manuscript.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Caroline Jane Ross: Conceptualization (equal); visualization (lead); writing – original draft (lead); writing – review and editing (equal). **Igor Ulitsky:** Conceptualization (equal); funding acquisition (lead); supervision (lead); writing – review and editing (lead).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Caroline Jane Ross https://orcid.org/0000-0003-4354-6372
Igor Ulitsky https://orcid.org/0000-0003-0555-6561

RELATED WIREs ARTICLES

Evolutionary clues in lncRNAs
Cytoplasmic functions of long noncoding RNAs
Long noncoding RNAs in mammalian cells: what, where, and why?

REFERENCES

Amaral, P. P., Leonardi, T., Han, N., Viré, E., Gascoigne, D. K., Arias-Carrasco, R., Büscher, M., Pandolfini, L., Zhang, A., Pluchino, S., Maracaja-Coutinho, V., Nakaya, H. I., Hemberg, M., Shiekhattar, R., Enright, A. J., & Kouzarides, T. (2018). Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biology*, *19*(1), 32.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30.

Andrews, R. J., Roche, J., & Moss, W. N. (2018). ScanFold: An approach for genome-wide discovery of local RNA structural elements—Applications to Zika virus and HIV. *PeerJ*, *6*, e6136.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Research*, *41*(Database issue), D991–D995.

Bartel, D. P. (2009). MicroRNAs: Target recognition and regulatory functions. *Cell*, *136*(2), 215–233.

Bhartiya, D., & Scaria, V. (2016). Genomic variations in non-coding RNAs: Structure, function and regulation. *Genomics*, *107*(2–3), 59–68.

Bitetti, A., Mallory, A. C., Golini, E., Carrieri, C., Carreño Gutiérrez, H., Perlas, E., Pérez-Rico, Y. A., Tocchini-Valentini, G. P., Enright, A. J., Norton, W. H. J., Mandillo, S., O'Carroll, D., & Shkumatava, A. (2018). MicroRNA degradation by a conserved target RNA regulates animal behavior. *Nature Structural & Molecular Biology*, *25*(3), 244–251.

Brown, J. A., Bulkley, D., Wang, J., Valenstein, M. L., Yario, T. A., Steitz, T. A., & Steitz, J. A. (2014). Structural insights into the stabilization of MALAT1 noncoding RNA by a bipartite triple helix. *Nature Structural & Molecular Biology*, *21*(7), 633–640.

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., ... Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*, *512*(7515), 393–399.

Bryzghalov, O., Makałowska, I., & Szcześniak, M. W. (2021). lncEvo: Automated identification and conservation study of long noncoding RNAs. *BMC Bioinformatics*, *22*(1), 59.

Bryzghalov, O., Szcześniak, M. W., & Makałowska, I. (2020). SyntDB: Defining orthologues of human long noncoding RNAs across primates. *Nucleic Acids Research*, *48*(D1), D238–D245.

Bu, D., Luo, H., Jiao, F., Fang, S., Tan, C., Liu, Z., & Zhao, Y. (2015). Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. *Science China. Life Sciences*, *58*(8), 787–798.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–1927.

Cao, M., Zhao, J., & Hu, G. (2019). Genome-wide methods for investigating long noncoding RNAs. *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie*, *111*, 395–401.

Carninci, P., & Hayashizaki, Y. (2007). Noncoding RNA transcription beyond annotated genes. *Current Opinion in Genetics & Development*, *17*(2), 139–144.

Carter, A. C., Xu, J., Nakamoto, M. Y., Wei, Y., Zarnegar, B. J., Shi, Q., Broughton, J. P., Ransom, R. C., Salhotra, A., Nagaraja, S. D., Li, R., Dou, D. R., Yost, K. E., Cho, S.-W., Mistry, A., Longaker, M. T., Khavari, P. A., Batey, R. T., Wuttke, D. S., & Chang, H. Y. (2020). Spen links RNA-mediated endogenous retrovirus silencing and X chromosome inactivation. *eLife*, *9*, e54508. https://doi.org/10.7554/eLife.54508

Carter, J.-M., Ang, D. A., Sim, N., Budiman, A., & Li, Y. (2021). Approaches to identify and characterise the post-transcriptional roles of lncRNAs in cancer. *Non-Coding RNA*, *7*(1), 19. https://doi.org/10.3390/ncrna7010019

Chen, J., Shishkin, A. A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J. H., Regev, A., & Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biology*, *17*, 19.

Chodroff, R. A., Goodstadt, L., Sirey, T. M., Oliver, P. L., Davies, K. E., Green, E. D., Molnár, Z., & Ponting, C. P. (2010). Long noncoding RNA genes: Conservation of sequence and brain expression among diverse amniotes. *Genome Biology*, *11*(7), R72.

Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., & Chang, H. Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Molecular Cell*, *44*(4), 667–678.

Chu, C., Quinn, J., & Chang, H. Y. (2012). Chromatin isolation by RNA purification (ChIRP). *Journal of Visualized Experiments: JoVE*, *61*, 3912. https://doi.org/10.3791/3912

Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., Bult, C. J., Agarwala, R., Cherry, J. L., DiCuccio, M., Hlavina, W., Kapustin, Y., Meric, P., Maglott, D., Birtle, Z., Marques, A. C., Graves, T., Zhou, S., Teague, B., ... Mouse Genome Sequencing Consortium. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, *7*(5), e1000112.

Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P., & Duret, L. (2002). Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Research*, *12*(6), 894–908.

Clark, M. B., & Mattick, J. S. (2011). Long noncoding RNAs in cell biology. *Seminars in Cell & Developmental Biology*, *22*(4), 366–376.

Conrad, N. K., Mili, S., Marshall, E. L., Shu, M.-D., & Steitz, J. A. (2006). Identification of a rapid mammalian deadenylation-dependent decay pathway and its inhibition by a viral RNA element. *Molecular Cell*, *24*(6), 943–953.

Conrad, N. K., & Steitz, J. A. (2005). A Kaposi's sarcoma virus RNA element that increases the nuclear abundance of intronless transcripts. *The EMBO Journal*, *24*(10), 1831–1841.

Constanty, F., & Shkumatava, A. (2021). lncRNAs in development and differentiation: From sequence motifs to functional characterization. *Development*, *148*(1), dev182741. https://doi.org/10.1242/dev.182741

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics: TIG*, *30*(4), 121–123.

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., Yeo, G. W., Graveley, B. R., & Burge, C. B. (2018). Sequence, structure, and context preferences of human RNA binding proteins. *Molecular Cell*, *70*(5), 854–867.e9.

Doshi, K. J., Cannone, J. J., Cobaugh, C. W., & Gutell, R. R. (2004). Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, *5*, 105.

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, *22*(11), 2079–2088.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210.

Elguindy, M. M., Kopp, F., Goodarzi, M., Rehfeld, F., Thomas, A., Chang, T.-C., & Mendell, J. T. (2019). PUMILIO, but not RBMX, binding is required for regulation of genomic stability by noncoding RNA NORAD. *eLife*, *8*, e48625. https://doi.org/10.7554/eLife.48625

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogozin, I. B., Nesterova, T. B., Brockdorff, N., & Zakian, S. M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS One*, *3*(6), e2521.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*, 157.

Engreitz, J., Lander, E. S., & Guttman, M. (2015). RNA antisense purification (RAP) for mapping RNA interactions with chromatin. *Methods in Molecular Biology*, *1262*, 183–197.

Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K., & Guttman, M. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, *341*(6147), 1237973.

Engreitz, J. M., Sirokman, K., McDonel, P., Shishkin, A. A., Surka, C., Russell, P., Grossman, S. R., Chow, A. Y., Guttman, M., & Lander, E. S. (2014). RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell*, *159*(1), 188–199.

Fang, R., Moss, W. N., Rutenberg-Schoenberg, M., & Simon, M. D. (2015). Probing Xist RNA structure in cells using targeted structure-Seq. *PLoS Genetics*, *11*(12), e1005668.

Ferhadian, D., Contrant, M., Printz-Schweigert, A., Smyth, R. P., Paillart, J.-C., & Marquet, R. (2018). Structural and functional motifs in influenza virus RNAs. *Frontiers in Microbiology*, *9*, 559.

Fernández-Sanlés, A., Ríos-Marco, P., Romero-López, C., & Berzal-Herranz, A. (2017). Functional information stored in the conserved structural RNA domains of Flavivirus genomes. *Frontiers in Microbiology*, *8*, 546.

Gardner, P. P., Fasold, M., Burge, S. W., Ninova, M., Hertel, J., Kehr, S., Steeves, T. E., Griffiths-Jones, S., & Stadler, P. F. (2015). Conservation and losses of non-coding RNAs in avian genomes. *PLoS One*, *10*(3), e0121797.

Gil, N., & Ulitsky, I. (2018). Production of spliced long noncoding RNAs specifies regions with increased enhancer activity. *Cell Systems*, *7*(5), 537–547.e3.

Giorgetti, L., Lajoie, B. R., Carter, A. C., Attia, M., Zhan, Y., Xu, J., Chen, C. J., Kaplan, N., Chang, H. Y., Heard, E., & Dekker, J. (2016). Structural organization of the inactive X chromosome in the mouse. *Nature*, *535*(7613), 575–579.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Graindorge, A., Pinheiro, I., Nawrocka, A., Mallory, A. C., Tsvetkov, P., Gil, N., Carolis, C., Buchholz, F., Ulitsky, I., Heard, E., Taipale, M., & Shkumatava, A. (2019). In-cell identification and measurement of RNA-protein interactions. *Nature Communications*, *10*(1), 5317.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. (2003). Rfam: An RNA family database. *Nucleic Acids Research*, *31*(1), 439–441.

Grinman, E., Nakahata, Y., Avchalumov, Y., Espadas, I., Swarnkar, S., Yasuda, R., & Puthanveettil, S. V. (2021). Activity-regulated synaptic targeting of lncRNA ADEPTR mediates structural plasticity by localizing Sptn1 and AnkB in dendrites. *Science Advances*, *7*(16), eabf0605. https://doi.org/10.1126/sciadv.abf0605

Gudenas, B. L., & Wang, L. (2018). Prediction of LncRNA subcellular localization with deep learning from sequence features. *Scientific Reports*, *8*(1), 16385.

Guo, J., Liu, Z., & Gong, R. (2019). Long noncoding RNA: An emerging player in diabetes and diabetic kidney disease. *Clinical Science*, *133*(12), 1321–1339.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*, *8*(2), R24.

Gutell, R. R., Lee, J. C., & Cannone, J. J. (2002). The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, *12*(3), 301–310.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., & Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, *458*(7235), 223–227.

Guttman, M., & Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature*, *482*(7385), 339–346.

Haerty, W., & Ponting, C. P. (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA*, *21*, 320–332. https://doi.org/10.1261/rna.047324.114

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., & Zavolan, M. (2021). CLIP and complementary methods. *Nature Reviews Methods Primers*, *1*(1), 1–23.

Hawkes, E. J., Hennelly, S. P., Novikova, I. V., Irwin, J. A., Dean, C., & Sanbonmatsu, K. Y. (2016). COOLAIR antisense RNAs form evolutionarily conserved elaborate secondary structures. *Cell Reports*, *16*(12), 3087–3096.

Herrera-Úbeda, C., Marín-Barba, M., Navas-Pérez, E., Gravemeyer, J., Albuixech-Crespo, B., Wheeler, G. N., & Garcia-Fernàndez, J. (2019). Microsyntenic clusters reveal conservation of lncRNAs in chordates despite absence of sequence conservation. *Biology*, *8*(3), 61. https://doi.org/10.3390/biology8030061

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of Transcriptomes in 17 species. *Cell Reports*, *11*, 1110–1122. https://doi.org/10.1016/j.celrep.2015.04.023

Hezroni, H., Tov Perry, R. B., & Ulitsky, I. (2019). Long noncoding RNAs in development and regeneration of the neural lineage. *Cold Spring Harbor Symposia on Quantitative Biology*, *84*, 165–177.

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, *8*(5), giz039. https://doi.org/10.1093/gigascience/giz039

Huang, M., Hou, J., Wang, Y., Xie, M., Wei, C., Nie, F., Wang, Z., & Sun, M. (2017). Long noncoding RNA LINC00673 is activated by SP1 and exerts oncogenic properties by interacting with LSD1 and EZH2 in gastric cancer. *Molecular Therapy: The Journal of the American Society of Gene Therapy*, *25*(4), 1014–1026.

Ilik, I. A., Quinn, J. J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R. C., Luscombe, N., Backofen, R., Chang, H. Y., & Akhtar, A. (2013). Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in drosophila. *Molecular Cell*, *51*(2), 156–173.

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., & Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, *47*(3), 199–208.

Jin, F., Li, J., Zhang, Y.-B., Liu, X., Cai, M., Liu, M., Li, M., Ma, C., Yue, R., Zhu, Y., Lai, R., Wang, Z., Ji, X., Wei, H., Dong, J., Liu, Z., Wang, Y., Sun, Y., & Wang, X. (2021). A functional motif of long noncoding RNA Nron against osteoporosis. *Nature Communications*, *12*(1), 1–15.

Jonas, K., Calin, G. A., & Pichler, M. (2020). RNA-binding proteins as important regulators of long non-coding RNAs in cancer. *International Journal of Molecular Sciences*, *21*(8), 2969. https://doi.org/10.3390/ijms21082969

Jones, A. N., Pisignano, G., Pavelitz, T., White, J., Kinisu, M., Forino, N., Albin, D., & Varani, G. (2020). An evolutionarily conserved RNA structure in the functional core of the lincRNA Cyrano. *RNA*, *26*(9), 1234–1246.

Kato, M., & Carninci, P. (2020). Genome-wide technologies to study RNA–chromatin interactions. *Non-Coding RNA*, *6*(2), 20.

Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., & Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, *50*(10), 1474–1482.

Kleaveland, B., Shi, C. Y., Stefano, J., & Bartel, D. P. (2018). A network of noncoding regulatory RNAs acts in the mammalian brain. *Cell*, *174*(2), 350–362.

Kopp, F., Elguindy, M. M., Yalvac, M. E., Zhang, H., Chen, B., Gillett, F. A., Lee, S., Sivakumar, S., Yu, H., Xie, Y., Mishra, P., Sahenk, Z., & Mendell, J. T. (2019). PUMILIO hyperactivity drives premature aging of Norad-deficient mice. *eLife*, *8*, 8. https://doi.org/10.7554/eLife.42650

Kornienko, A. E., Dotter, C. P., Guenzl, P. M., Gisslinger, H., Gisslinger, B., Cleary, C., Kralovics, R., Pauler, F. M., & Barlow, D. P. (2016). Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biology*, *17*, 14.

Kuo, C.-C., Hänzelmann, S., Sentürk Cetin, N., Frank, S., Zajzon, B., Derks, J.-P., Akhade, V. S., Ahuja, G., Kanduri, C., Grummt, I., Kurian, L., & Costa, I. G. (2019). Detection of RNA–DNA binding sites in long noncoding RNAs. *Nucleic Acids Research*, *47*(6), e32–e32.

Kurian, L., Aguirre, A., Sancho-Martinez, I., Benner, C., Hishida, T., Nguyen, T. B., Reddy, P., Nivet, E., Krause, M. N., Nelles, D. A., Esteban, C. R., Campistol, J. M., Yeo, G. W., & Belmonte, J. C. I. (2015). Identification of novel long noncoding RNAs underlying vertebrate cardiovascular development. *Circulation*, *131*(14), 1278–1290.

Kutter, C., Watt, S., Stefflova, K., Wilson, M. D., Goncalves, A., Ponting, C. P., Odom, D. T., & Marques, A. C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genetics*, *8*(7), e1002841.

Lee, J. T., & Bartolomei, M. S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell*, *152*(6), 1308–1323.

Lee, S., Kopp, F., Chang, T.-C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y., & Mendell, J. T. (2016). Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell*, *164*(1–2), 69–80.

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., & Yang, J.-H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. In. *Nucleic Acids Research*, *42*(D1), D92–D97. https://doi.org/10.1093/nar/gkt1248

Liu, S. J., Dang, H. X., Lim, D. A., Feng, F. Y., & Maher, C. A. (2021). Long noncoding RNAs in cancer metastasis. *Nature Reviews. Cancer*, *21*(7), 446–460.

Lorenz, R., Wolfinger, M. T., Tanzer, A., & Hofacker, I. L. (2016). Predicting RNA secondary structures from sequence and probing data. *Methods*, *103*, 86–98.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., Mesirov, J. P., Cech, T. R., & Chang, H. Y. (2016). RNA duplex map in living cells reveals higher-order Transcriptome structure. *Cell*, *165*(5), 1267–1279.

Lubelsky, Y., & Ulitsky, I. (2018). Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature*, *555*(7694), 107–111.

Lubelsky, Y., Zuckerman, B., & Ulitsky, I. (2021). High-resolution mapping of function and protein binding in an RNA nuclear enrichment sequence. *The EMBO Journal*, *40*, e106357. https://doi.org/10.15252/embj.2020106357

Maenner, S., Blaud, M., Fouillen, L., Savoye, A., Marchand, V., Dubois, A., Sanglier-Cianférani, S., Van Dorsselaer, A., Clerc, P., Avner, P., Visvikis, A., & Branlant, C. (2010). 2-D structure of the a region of Xist RNA and its implication for PRC2 association. *PLoS Biology*, *8*(1), e1000276.

Maier, D. (1978). The complexity of some problems on subsequences and supersequences. *Journal of the ACM*, *25*(2), 322–336.

Mauger, D. M., Golden, M., Yamane, D., Williford, S., Lemon, S. M., Martin, D. P., & Weeks, K. M. (2015). Functionally conserved architecture of hepatitis C virus RNA genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3692–3697.

McCown, P. J., Wang, M. C., Jaeger, L., & Brown, J. A. (2019). Secondary structural model of human MALAT1 reveals multiple structure–function relationships. *International Journal of Molecular Sciences*, *20*(22), 5610.

McCullum, E. O., Williams, B. A. R., Zhang, J., & Chaput, J. C. (2010). Random mutagenesis by error-prone PCR. *Methods in Molecular Biology*, *634*, 103–109.

McHugh, C. A., Chen, C.-K., Chow, A., Surka, C. F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M. J., Shishkin, A. A., Su, J., Lander, E. S., Hess, S., Plath, K., & Guttman, M. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature*, *521*(7551), 232–236.

McLeay, R. C., & Bailey, T. L. (2010). Motif enrichment analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, *11*, 165.

Meller, V. H., & Rattner, B. P. (2002). The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *The EMBO Journal*, *21*(5), 1084–1091.

Michel, F., & Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology*, *216*(3), 585–610.

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, *3*(3), REVIEWS0004.

Mitton-Fry, R. M., DeGregorio, S. J., Wang, J., Steitz, T. A., & Steitz, J. A. (2010). Poly(A) tail recognition by a viral RNA element through assembly of a triple helix. *Science*, *330*(6008), 1244–1247.

Morán, I., Akerman, I., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., Arnes, L., Nakić, N., García-Hurtado, J., Rodríguez-Seguí, S., Pasquali, L., Sauty-Colace, C., Beucher, A., Scharfmann, R., van Arensbergen, J., Johnson, P. R., Berry, A., Lee, C., Harkins, T., ... Ferrer, J. (2012). Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metabolism*, *16*(4), 435–448.

Müller, T., Boileau, E., Talyan, S., Kehr, D., Varadi, K., Busch, M., Most, P., Krijgsveld, J., & Dieterich, C. (2021). Updated and enhanced pig cardiac transcriptome based on long-read RNA sequencing and proteomics. *Journal of Molecular and Cellular Cardiology*, *150*, 23–31.

Munschauer, M., Nguyen, C. T., Sirokman, K., Hartigan, C. R., Hogstrom, L., Engreitz, J. M., Ulirsch, J. C., Fulco, C. P., Subramanian, V., Chen, J., Schenone, M., Guttman, M., Carr, S. A., & Lander, E. S. (2018). Publisher correction: The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature*, *563*(7733), E32.

Mustafi, D., Kevany, B. M., Bai, X., Maeda, T., Sears, J. E., Khalil, A. M., & Palczewski, K. (2013). Evolutionarily conserved long intergenic non-coding RNAs in the eye. *Human Molecular Genetics*, *22*(15), 2992–3002.

Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, *25*(10), 1335–1337.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., & Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, *505*(7485), 635–640.

Noh, J. H., Kim, K. M., Abdelmohsen, K., Yoon, J.-H., Panda, A. C., Munk, R., Kim, J., Curtis, J., Moad, C. A., Wohler, C. M., Indig, F. E., de Paula, W., Dudekula, D. B., De, S., Piao, Y., Yang, X., Martindale, J. L., de Cabo, R., & Gorospe, M. (2016). HuR and GRSF1 modulate the nuclear export and mitochondrial localization of the lncRNA RMRP. *Genes & Development*, *30*(10), 1224–1239.

Noviello, T. M. R., Di Liddo, A., Ventola, G. M., Spagnuolo, A., D'Aniello, S., Ceccarelli, M., & Cerulo, L. (2018). Detection of long non–coding RNA homology, a comparative study on alignment and alignment–free metrics. *BMC Bioinformatics*, *19*(1), 407.

Novikova, I. V., Dharap, A., Hennelly, S. P., & Sanbonmatsu, K. Y. (2013). 3S: Shotgun secondary structure determination of long non-coding RNAs. *Methods*, *63*(2), 170–177.

Novikova, I. V., Hennelly, S. P., & Sanbonmatsu, K. Y. (2012). Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Research*, *40*(11), 5034–5051.

Oikonomopoulos, S., Bayega, A., Fahiminiya, S., Djambazian, H., Berube, P., & Ragoussis, J. (2020). Methodologies for transcript profiling using long-read technologies. *Frontiers in Genetics*, *11*, 606.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schönbach, C., Gojobori, T., Baldarelli, R., ... RIKEN Genome Exploration Research Group Phase I & II Team. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, *420*(6915), 563–573.

Paralkar, V. R., Mishra, T., Luan, J., Yao, Y., Kossenkov, A. V., Anderson, S. M., Dunagin, M., Pimkin, M., Gore, M., Sun, D., Konuthula, N., Raj, A., An, X., Mohandas, N., Bodine, D. M., Hardison, R. C., & Weiss, M. J. (2014). Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood*, *123*(12), 1927–1937.

Parisien, M., Wang, X., & Pan, T. (2013). Diversity of human tRNA genes from the 1000-genomes project. *RNA Biology*, *10*(12), 1853–1867.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., & Haussler, D. (2006). Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology*, *2*(4), e33.

Pintacuda, G., Young, A. N., & Cerase, A. (2017). Function by structure: Spotlights on Xist long non-coding RNA. *Frontiers in Molecular Biosciences*, *4*, 90.

Pirakitikulr, N., Kohlway, A., Lindenbach, B. D., & Pyle, A. M. (2016). The coding region of the HCV genome contains a network of regulatory RNA structures. *Molecular Cell*, *62*(1), 111–120.

Ponjavic, J., Oliver, P. L., Lunter, G., & Ponting, C. P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genetics*, *5*(8), e1000617.

Quinn, J. J., Ilik, I. A., Qu, K., Georgiev, P., Chu, C., Akhtar, A., & Chang, H. Y. (2014). Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nature Biotechnology*, *32*(9), 933–940.

Quinn, J. J., Zhang, Q. C., Georgiev, P., & Ilik, I. A. (2016). Rapid evolutionary turnover underlies conserved lncRNA–genome interactions. *Genes Dev 30*(2), 191–207. http://genesdev.cshlp.org/content/30/2/191.short

Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019). Methods to study RNA–protein interactions. *Nature Methods*, *16*(3), 225–234.

Rivas, E. (2021). Evolutionary conservation of RNA sequence and structure. *WIREs RNA*, *12*(5), e1649. https://doi.org/10.1002/wrna.1649

Rivas, E., Clements, J., & Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, *14*(1), 45–48.

Rivas, E., Clements, J., & Eddy, S. R. (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, *36*(10), 3072–3076.

Rivas, E., & Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, *16*(7), 583–605.

RNAcentral Consortium. (2021). RNAcentral 2021: Secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research*, *49*(D1), D212–D220.

Rogers, E., Murrugarra, D., & Heitsch, C. (2017). Conditioning and robustness of RNA Boltzmann sampling under thermodynamic parameter perturbations. *Biophysical Journal*, *113*(2), 321–329.

Rom, A., Melamed, L., Gil, N., Goldrich, M. J., Kadir, R., Golan, M., Biton, I., Perry, R. B.-T., & Ulitsky, I. (2019). Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nature Communications*, *10*(1), 5092.

Ross, C. J., Rom, A., Spinrad, A., Gelbard-Solodkin, D., Degani, N., & Ulitsky, I. (2021). Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biology*, *22*(1), 29.

Sarropoulos, I., Marin, R., Cardoso-Moreira, M., & Kaessmann, H. (2019). Developmental dynamics of lncRNAs across mammalian organs and species. *Nature*, *571*(7766), 510–514.

Schroeder, S. J. (2018). Challenges and approaches to predicting RNA with multiple functional structures. *RNA*, *24*(12), 1615–1624.

Schuler, A., Ghanbarian, A. T., & Hurst, L. D. (2014). Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Molecular Biology and Evolution*, *31*(12), 3164–3183.

Schultes, E. A., Spasic, A., Mohanty, U., & Bartel, D. P. (2005). Compact and ordered collapse of randomly generated RNA sequences. *Nature Structural & Molecular Biology*, *12*(12), 1130–1136.

Seemann, S. E., Mirza, A. H., Hansen, C., Bang-Berthelsen, C. H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C. T., Pociot, F., Nielsen, H., Tommerup, N., Ruzzo, W. L., & Gorodkin, J. (2017). The identification and functional annotation of RNA structures conserved in vertebrates. *Genome Research*, *27*(8), 1371–1383.

Seung-Won, P., Kuroda, M. I., & Yongkyu, P. (2008). Regulation of histone H4 Lys16 acetylation by predicted alternative secondary structures in roX noncoding RNAs. *Molecular and Cellular Biology*, *28*(16), 4952–4962.

Shukla, C. J., McCorkindale, A. L., Gerhardinger, C., Korthauer, K. D., Cabili, M. N., Shechner, D. M., Irizarry, R. A., Maass, P. G., & Rinn, J. L. (2018). High-throughput identification of RNA nuclear enrichment sequences. *The EMBO Journal*, *37*(6), e98452. https://doi.org/10.15252/embj.201798452

Simon, L. M., Morandi, E., Luganini, A., Gribaudo, G., Martinez-Sobrido, L., Turner, D. H., Oliviero, S., & Incarnato, D. (2019). In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Research*, *47*(13), 7003–7017.

Simon, M. D. (2013). Capture hybridization analysis of RNA targets (CHART). *Current Protocols in Molecular Biology*. 2013, Chapter 21, Unit 21 21.25.

Simon, M. D., Wang, C. I., Kharchenko, P. V., West, J. A., Chapman, B. A., Alekseyenko, A. A., Borowsky, M. L., Kuroda, M. I., & Kingston, R. E. (2011). The genomic binding sites of a noncoding RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20497–20502.

Singh, N. N., & Singh, R. N. (2019). How RNA structure dictates the usage of a critical exon of spinal muscular atrophy gene. *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms*, *1862*(11–12), 194403.

Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M., & Weeks, K. M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(37), 10322–10327.

Somarowthu, S., Legiewicz, M., Chillón, I., Marcia, M., Liu, F., & Pyle, A. M. (2015). HOTAIR forms an intricate and modular secondary structure. *Molecular Cell*, *58*(2), 353–361.

Statello, L., Guo, C.-J., Chen, L.-L., & Huarte, M. (2020). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews. Molecular Cell Biology*, *22*(2), 96–118.

Su, Z.-D., Huang, Y., Zhang, Z.-Y., Zhao, Y.-W., Wang, D., Chen, W., Chou, K.-C., & Lin, H. (2018). iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*, *34*(24), 4196–4204.

Sun, L., Fazal, F. M., Li, P., Broughton, J. P., Lee, B., Tang, L., Huang, W., Kool, E. T., Chang, H. Y., & Zhang, Q. C. (2019). RNA structure maps across mammalian cellular compartments. *Nature Structural & Molecular Biology*, *26*(4), 322–330.

Sun, Y. H., Wang, A., Song, C., Shankar, G., Srivastava, R. K., Au, K. F., & Li, X. Z. (2021). Single-molecule long-read sequencing reveals a conserved intact long RNA profile in sperm. *Nature Communications*, *12*(1), 1361.

Svoboda, P., & Di Cara, A. (2006). Hairpin RNA: A secondary structure of primary importance. *Cellular and Molecular Life Sciences: CMLS*, *63*(7–8), 901–908.

Sztuba-Solinska, J., Rausch, J. W., Smith, R., Miller, J. T., Whitby, D., & Le Grice, S. F. J. (2017). Kaposi's sarcoma-associated herpesvirus polyadenylated nuclear RNA: A structural scaffold for nuclear, cytoplasmic and viral proteins. *Nucleic Acids Research*, *45*(11), 6805–6821.

Tahi, F., Du T Tran, V., & Boucheham, A. (2017). In silico prediction of RNA secondary structure. *Methods in Molecular Biology*, *1543*, 145–168.

Tan, D., Marzluff, W. F., Dominski, Z., & Tong, L. (2013). Structure of histone mRNA stem-loop, human stem-loop binding protein, and 3′hExo ternary complex. *Science*, *339*(6117), 318–321.

Tan, J. Y., Biasini, A., Young, R. S., & Marques, A. C. (2020). Splicing of enhancer-associated lincRNAs contributes to enhancer activity. *Life science Alliance*, *3*(4), e202000663. https://doi.org/10.26508/lsa.202000663

Tavares, R. C. A., Pyle, A. M., & Somarowthu, S. (2019). Phylogenetic analysis with improved parameters reveals conservation in lncRNA structures. *Journal of Molecular Biology*, *431*(8), 1592–1603.

Tichon, A., Gil, N., Lubelsky, Y., Havkin Solomon, T., Lemze, D., Itzkovitz, S., Stern-Ginossar, N., & Ulitsky, I. (2016). A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nature Communications*, *7*, 12209.

Tichon, A., Perry, R. B.-T., Stojic, L., & Ulitsky, I. (2018). SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. *Genes & Development*, *32*(1), 70–78.

Tomaszewska, M., Szabat, M., Zielińska, K., & Kierzek, R. (2021). Identification and structural aspects of G-Quadruplex-forming sequences from the influenza a virus genome. *International Journal of Molecular Sciences*, *22*(11), 6031. https://doi.org/10.3390/ijms22116031

Tycowski, K. T., Shu, M.-D., Borah, S., Shi, M., & Steitz, J. A. (2012). Conservation of a triple-helix-forming RNA stability element in noncoding and genomic RNAs of diverse viruses. *Cell Reports*, *2*(1), 26–32.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., & Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science*, *302*(5648), 1212–1215.

Ulitsky, I. (2016). Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nature Reviews. Genetics*, *17*(10), 601–614.

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: Genomics, evolution, and mechanisms. *Cell*, *154*(1), 26–46.

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, *147*(7), 1537–1550.

Unfried, J. P., Marín-Baquero, M., Rivera-Calzada, Á., Razquin, N., Martín-Cuevas, E. M., de Bragança, S., Aicart-Ramos, C., McCoy, C., Prats-Mari, L., Arribas-Bosacoma, R., Lee, L., Caruso, S., Zucman-Rossi, J., Sangro, B., Williams, G., Moreno-Herrero, F., Llorca, O., Lees-Miller, S. P., & Fortes, P. (2021). Long noncoding RNA NIHCOLE promotes ligation efficiency of DNA double-strand breaks in hepatocellular carcinoma. *Cancer Research*, *81*, 4910–4925. https://doi.org/10.1158/0008-5472.CAN-21-0463

Uroda, T., Anastasakou, E., Rossi, A., Teulon, J.-M., Pellequer, J.-L., Annibale, P., Pessey, O., Inga, A., Chillón, I., & Marcia, M. (2019). Conserved Pseudoknots in lncRNA MEG3 are essential for stimulation of the p53 pathway. *Molecular Cell*, *75*(5), 982–995.e9.

Valsecchi, C. I. K., Basilicata, M. F., Georgiev, P., Gaub, A., Seyfferth, J., Kulkarni, T., Panhale, A., Semplicio, G., Manjunath, V., Holz, H., Dasmeh, P., & Akhtar, A. (2021). RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature*, *589*(7840), 137–142.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., ... Yeo, G. W. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, *583*(7818), 711–719.

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Research*, *35*(Database issue, D88–D92.

Wang, Y., Yesselman, J. D., Zhang, Q., Kang, M., & Feigon, J. (2016). Structural conservation in the template/pseudoknot domain of vertebrate telomerase RNA from teleost fish to human. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(35), E5125–E5134.

Washietl, S., Kellis, M., & Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Research*, *24*(4), 616–628.

Watters, K. E., Strobel, E. J., Yu, A. M., Lis, J. T., & Lucks, J. B. (2016). Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature Structural & Molecular Biology*, *23*(12), 1124–1131.

Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J. J., & Noller, H. F. (1980). Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Research*, *8*(10), 2275–2293.

Wutz, A., Rasmussen, T. P., & Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics*, *30*(2), 167–174.

Xue, Z., Hennelly, S., Doyle, B., Gulati, A. A., Novikova, I. V., Sanbonmatsu, K. Y., & Boyer, L. A. (2016). A G-rich motif in the lncRNA Braveheart interacts with a zinc-finger transcription factor to specify the cardiovascular lineage. *Molecular Cell*, *64*(1), 37–50.

Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., & Qu, L.-H. (2010). starBase: A database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Research*, *39*(suppl_1), D202–D209.

Yang, Y.-C. T., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., & Lu, Z. J. (2015). CLIPdb: A CLIP-seq database for protein-RNA interactions. *BMC Genomics*, *16*, 51.

Yao, Z., Weinberg, Z., & Ruzzo, W. L. (2005). CMfinder—A covariance model based RNA motif finding algorithm. *Bioinformatics*, *22*(4), 445–452.

Yin, Y., Lu, J. Y., Zhang, X., Shao, W., Xu, Y., Li, P., Hong, Y., Cui, L., Shan, G., Tian, B., Zhang, Q. C., & Shen, X. (2020). U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature*, *580*(7801), 147–150.

Yu, A. M., Gasper, P. M., Cheng, L., Lai, L. B., Kaur, S., Gopalan, V., Chen, A. A., & Lucks, J. B. (2021). Computationally reconstructing cotranscriptional RNA folding from experimental data reveals rearrangement of non-native folding intermediates. *Molecular Cell*, *81*(4), 870–883.e10.

Yu, B., Lu, Y., Zhang, Q. C., & Hou, L. (2020). Prediction and differential analysis of RNA secondary structure. *Quantitative Biology (Beijing, China)*, *8*(2), 109–118.

Yu, H., Lindsay, J., Feng, Z.-P., Frankenberg, S., Hu, Y., Carone, D., Shaw, G., Pask, A. J., O'Neill, R., Papenfuss, A. T., & Renfree, M. B. (2012). Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC Genomics*, *13*, 251.

Zampetaki, A., Albrecht, A., & Steinhofel, K. (2018). Long non-coding RNA structure and function: Is there a link? *Frontiers in Physiology*, *9*, 1201. https://doi.org/10.3389/fphys.2018.01201

Zhang, B., Mao, Y. S., Diermeier, S. D., Novikova, I. V., Nawrocki, E. P., Jones, T. A., Lazar, Z., Tung, C.-S., Luo, W., Eddy, S. R., Sanbonmatsu, K. Y., & Spector, D. L. (2017). Identification and characterization of a class of MALAT1-like genomic loci. *Cell Reports*, *19*(8), 1723–1738.

Zhang, X., Rice, K., Wang, Y., Chen, W., Zhong, Y., Nakayama, Y., Zhou, Y., & Klibanski, A. (2010). Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: Isoform structure, expression, and functions. *Endocrinology*, *151*(3), 939–947.

Zhang, X.-Q., Wang, Z.-L., Poon, M.-W., & Yang, J.-H. (2017). Spatial-temporal transcriptional dynamics of long non-coding RNAs in human brain. *Human Molecular Genetics*, *26*(16), 3202–3211.

Ziv, O., Gabryelska, M. M., Lun, A. T. L., Gebert, L. F. R., Sheu-Gruttadauria, J., Meredith, L. W., Liu, Z.-Y., Kwok, C. K., Qin, C.-F., MacRae, I. J., Goodfellow, I., Marioni, J. C., Kudla, G., & Miska, E. A. (2018). COMRADES determines in vivo RNA structures and interactions. *Nature Methods*, *15*(10), 785–788.

How to cite this article: Ross, C. J., & Ulitsky, I. (2022). Discovering functional motifs in long noncoding RNAs. *Wiley Interdisciplinary Reviews: RNA*, e1708. https://doi.org/10.1002/wrna.1708