

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,500

Open access books available

176,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter

Deciphering the Hidden Language of Long Non-Coding RNAs: Recent Findings and Challenges

Assaf C. Bester

Abstract

Long non-coding RNAs (lncRNAs) are crucial non-coding RNA genes involved in diverse cellular processes. However, the mechanisms underlying their emergence and functions remain incompletely understood. A major challenge in the field is to understand how lncRNA sequences affect their function. In recent years, comprehensive genetic and genomic studies have started to unfold the function of lncRNAs through their interactions, cellular organization, and structure. This comprehensive review delves into the intricate interplay between lncRNA sequences and their functional implications. Unlike other RNA types, lncRNAs exhibit a complex syntax, employing diverse functional elements such as protein recognition and miRNA binding sites, repeat elements, secondary structures, and non-canonical interactions with RNA and DNA binding proteins. By unraveling the hidden language that governs the function and classification of lncRNAs, we aim to shed light on the underlying principles shaping their diverse functions. Through a detailed examination of the intricate relationship between lncRNA sequences and their biological effects, this review offers insights into the sequences underlying lncRNA functionality. Understanding the unique sequence characteristics and functional elements employed by lncRNAs has the potential to advance our knowledge of gene regulation and cellular processes, providing a foundation for the development of novel therapeutic strategies and targeted interventions.

Keywords: lncRNA, ncRNA, RNA regulation, RNA motifs, RNA binding proteins

1. Introduction

Long non-coding RNAs (lncRNAs) represent a substantial group of non-protein-coding genes (PCGs) in the human genome, with at least 20,000 unique genes [1]. These transcripts typically exceed 200 nucleotides (nt) in length, possess minimal or no open-reading frame (ORF) potential, and do not belong to any other recognized RNA groups.

lncRNAs can be classified based on their genetic context in relation to PCGs (intergenic, divergent, intronic, and antisense), or their mechanism of action, functioning either locally (cis) or remotely from the transcription site (trans). However, unlike other types of RNAs (e.g., mRNA, miRNA, rRNA, tRNA, and

snoRNA) where the function is clearly embodied in the sequence, the sequence-to-function relationship of lncRNAs is not as clear. This complexity forms the core focus of this manuscript.

Some lncRNAs function in a sequence-dependent manner, interacting specifically with other macromolecules, including RNA, DNA, and proteins (e.g., NORAD, NEAT1, and MALAT1). Others may rely solely on their processing, such as the initiation of their transcription (e.g., PVT1 [2]) or their splicing (e.g., Blustr [3]). In these instances, the specific sequence of the mature transcript is not essential. These lncRNAs often regulate PCGs in close proximity (cis-regulation) or overlap with DNA regulatory elements, such as enhancers (**Figure 1**) [4, 5].

While lncRNAs are characterized by having minimal or no ORF potential, some have been found to translate into functional peptides [6, 7]. However, lncRNAs can function through various mechanisms, which can vary depending on the cell type and context.

Despite the annotation of thousands of lncRNAs, our understanding of how the lncRNA sequence influences its function remains nascent. Advances in gene editing and molecular biology technologies have facilitated the identification of functional lncRNAs and their partners [8].

Like mRNAs, many lncRNAs have introns, can undergo post-transcriptional capping and polyadenylation, and are transcribed by RNA polymerase II. However, while our understanding of mRNA biology allows us to predict mRNA function based on their sequence, this is not the case with lncRNAs; little is known about the rules shaping lncRNA sequences and how they relate to their function.

In recent years, the language of lncRNAs has begun to unfold through experimental and computational approaches. In this review, we will delve into these challenges, discuss the current state of knowledge, and outline potential future directions to further unravel the mystery of lncRNA sequences and their functional implications.

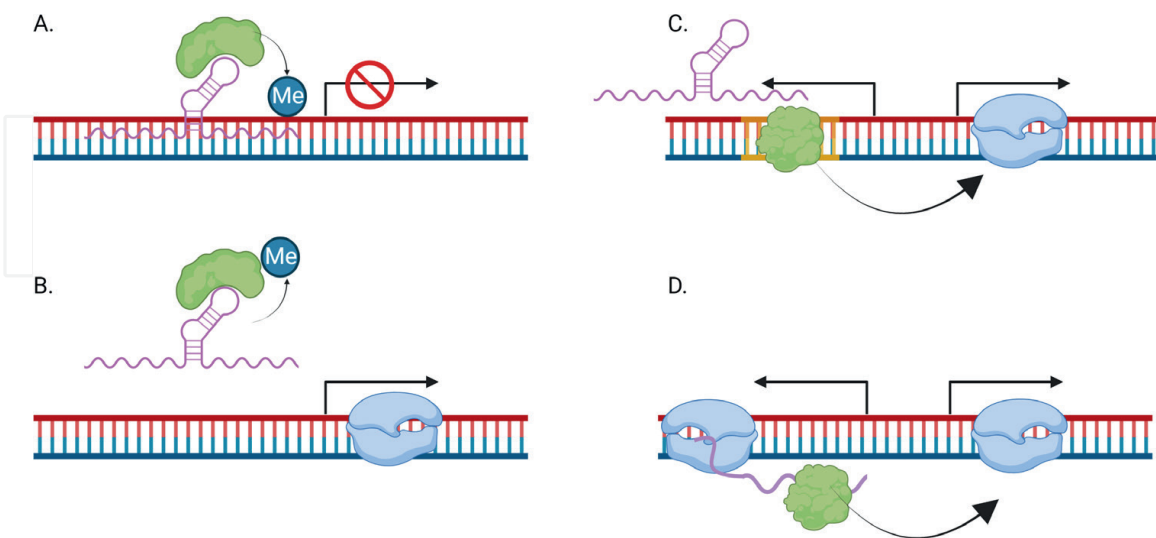


Figure 1. Mechanism of function for lncRNAs. (A) Interaction and recruitment of proteins that hearten their function (e.g., Xist). (B) Interaction with proteins to shield other interactions (e.g., lncPRESS1). (C) “by-product” of open chromatin in regulatory elements (e.g., PVT1). (D) RNA-based regulatory elements (e.g., Blustr) (created with BioRender.com).

2. Sequence-based classification of lncRNAs and mRNAs

lncRNAs and mRNAs share several characteristics, including the presence of exons and introns, and the processes of splicing, capping, and polyadenylation. Interestingly, lncRNAs have been hypothesized to be infant mRNAs that have yet to gain their coding potential [9]. However, the factors that shape their sequences differ significantly.

The sequences of protein-coding genes (PCGs) primarily serve to hold information in their open-reading frames (ORFs), which is influenced by the frequency of specific codons. PCGs often contain conserved and modular functional domains that can be duplicated across multiple transcripts, resulting in stronger sequence constraints for PCGs compared to lncRNAs.

Several *in silico* approaches have been developed to distinguish between PCGs and lncRNAs based on transcript sequences. These methods can also provide insights into the forces that shape lncRNA sequences. For instance, a high frequency of the GC dinucleotide is a characteristic of higher organisms and is associated with effective splicing [10]. lncRNAs, which typically have fewer exons and weaker splicing efficiency, exhibit lower GC content, making GC dinucleotide composition a useful classifier for distinguishing between lncRNAs and PCGs [11].

Conversely, the TA dinucleotide is strictly regulated in PCGs due to its association with stop codons (TAG, TAA, and TGA), contributing to a higher GC content in PCGs. While the frequency of CG dinucleotides across the genome is low due to their unique regulatory role in DNA methylation, CGs are more enriched in PCGs compared to lncRNAs [12].

Additional sequence-based strategies employ short (2–8 nt) sequential fragments (k-mers) to differentiate between lncRNAs and PCGs. Differences in k-mer frequency between transcript groups may represent functional domains in PCGs or lncRNAs, but they may also mirror single-nucleotide preferences [13–15].

However, PCG transcripts also include significant untranslated regions (UTRs) flanking the ORF. These UTRs, similar to lncRNAs, can interact with RNA-binding proteins (RBPs) and microRNAs (miRNAs). Additionally, many lncRNAs have the potential to encode short ORFs, complicating the classification between PCGs and lncRNAs.

Interestingly, a subset of lncRNAs has been found to harbor short ORFs that can be translated into functional peptides. These lncRNAs, often referred to as “micropeptide-encoding RNAs” or “small peptide-encoding RNAs,” challenge the traditional definition of lncRNAs and add another layer of complexity to our understanding of the non-coding transcriptome [16, 17].

Therefore, unexpectedly, ORFs may appear as an important feature of lncRNAs, adding a new dimension to our understanding of these complex molecules.

3. lncRNAs with functional low-complexity sequences

Recent studies have provided valuable insights into the functional roles of lncRNAs that possess low-complexity sequences. Through high-throughput perturbation experiments and correlation analyses of gene expressions, it has been consistently demonstrated that the majority of functional lncRNAs act in a cis-regulatory

manner, exerting their regulatory effects on nearby protein-coding genes [18–21]. While multiple models have been proposed to explain this regulation, it is evident that interactions between lncRNAs and mRNAs, chromatin, or proteins often rely on low-sequence specificity.

One prominent function of lncRNAs is their involvement in epigenetic regulation and chromatin accessibility. Numerous studies have identified lncRNAs that interact with the Polycomb repressive complex 2 (PRC2), influencing its activity on nearby genes [22]. PRC2, a core complex involved in histone 3 lysine 27 methylation (H3K27me3), exhibits unique DNA interactions in *Drosophila* through specific Polycomb response elements (PREs). However, similar motifs have not been identified in mammalian DNA, indicating that PRC2 recruitment to chromatin involves interactions with other mediators. Notably, well-characterized lncRNAs such as X-inactive-specific transcript (XIST) and HOTAIR have been found to interact with PRC2, suggesting their crucial role in PRC2 localization [23–25]. XIST, a conserved lncRNA, consists of six domains (A-F) with tandem repeats. The GC-rich B and E repeats have been identified as essential for PRC2 recruitment and subsequent X chromosome silencing [24].

PRC2 exhibits differential interactions with the four ribonucleotides, displaying a higher affinity for G-rich sequences [26]. Although a clear RNA-binding motif for PRC2 remains elusive, advancements in techniques such as denaturing crosslinking and immunoprecipitation (dCLIP) have enabled the identification of more specific RNA-binding motifs within PRC2 [27]. However, it is important to note that current evidence suggests the existence of partial non-specific interactions between PRC2 and lncRNAs. Functionally, lncRNAs guide PRC2 to specific chromatin regions, facilitating epigenetic changes both in cis and trans. Cis interactions between PRC2 and lncRNAs act as checkpoints for chromatin silencing, while lncRNA-PRC2 interactions may function as decoys, preventing the transcriptional silencing of active genes and reducing the catalytic function of PRC2 [28].

Another model describing interactions with low-sequence specificity involves the recruitment of transcription factors (TFs) during lncRNA metabolism. In such cases, the function of an lncRNA depends on the transcription and processing of its first intron, irrespective of the sequence of exons and introns within the gene.

Notably, studies have demonstrated the essential role of splicing the first intron of the mouse lncRNA 1319, also known as the bivalent locus (*Sfmbt2*) or *Blustr*, in regulating the transcription of the protein-coding gene *Sfmbt2* [3]. Hence, the metabolism of *Blustr* governs the regulation of *Sfmbt2*.

Moreover, research has highlighted the significance of splicing and the recruitment of the spliceosome in the regulation of enhancers, which may overlap with lncRNAs [29, 30]. These findings suggest that many lncRNAs contribute to the regulation of enhancer activity. However, it is crucial to acknowledge that the function of an lncRNA can vary depending on the specific context.

Overall, the functional roles of lncRNAs with low-complexity sequences encompass diverse mechanisms, including the cis-regulation of nearby genes, interactions with PRC2 and other proteins, recruitment of transcription factors, and involvement in enhancer regulation. These roles underscore the complex and multifaceted nature of lncRNA functions in gene expression and epigenetic regulation.

Intriguingly, recent research has demonstrated the formation of 3D membrane-less structures called “high-concentration territories” through interactions involving lncRNAs, DNA, and proteins. These interactions, which do not rely on stringent sequence recognition, significantly shape the local environment surrounding lncRNA

transcription sites and contribute to the regulation of neighboring gene expressions [31]. This model is particularly attractive in explaining the function of lncRNAs in shaping the local chromatin environment, as they interact with transcription factors and chromatin remodeling agents based on general sequence perforations.

4. lncRNA localization motifs

Long non-coding RNAs (lncRNAs) and messenger RNAs (mRNAs) share several common features. Both are transcribed by RNA polymerase II, undergo splicing, and are often modified with 5' caps and poly(A) tails. However, their functional locations differ. While mRNAs operate in the cytoplasm and are rapidly exported from the nucleus, many lncRNAs function within the nucleus, necessitating their retention and inefficient export to the cytoplasm. The sequences affecting lncRNA localization have been a major focus of research in the recent years.

The export of RNA molecules longer than 200 nucleotides, including mRNA and lncRNAs, is a tightly regulated process that relies on the formation of a messenger ribonucleoprotein (mRNP) complex. This mRNP formation is intertwined with mRNA/lncRNA processing and is contingent upon its success.

Two key components of the transport machinery are the transcription-and-export complex (TREX) and the nuclear transcription factor X-Box Binding 1 (NFX1). A recent study delineated two distinct RNA export pathways, each favoring different transcript sequences and characteristics. TREX is involved in the export of multi-exonic, C/G-rich transcripts, while NFX1 interacts with A/U-rich transcripts that have fewer and longer exons, with focal high C/G segments. Moreover, NXF1 typically interacts with m6A methylated RNA sites. Both export pathways can target mRNAs and lncRNAs. However, given that lncRNA transcripts tend to be shorter, with fewer exons and lower C/G content, NXF1 is suggested to be the more dominant pathway for lncRNA export [32].

RNA processing is intricately linked with its export to the cytoplasm. Studies have shown that overall splicing efficiency is a strong predictor of RNA localization for both lncRNAs and mRNAs [33, 34]. Key splicing proteins, such as UAP56, interact with nuclear export complexes to enhance export efficiency. However, splicing is not a prerequisite for RNA export [35]. For instance, single-exon transcripts like NORAD are enriched in the cytoplasm, indicating that endogenous intron-less transcripts can be efficiently exported [35].

For transcripts with introns, proper RNA splicing is critical for their stability and export. Recent findings suggest that lncRNA processing is less efficient than that of mRNA due to the extensive use of non-canonical splicing motifs, such as GC-AG, which are associated with less competent splicing compared to the canonical GT-AG motifs [36, 37]. Consequently, lncRNAs tend to be less stable and fail to accumulate in the cytoplasm. However, intron retention can also generate stable and functional lncRNAs. For example, in the case of lncRNA TUG1, intron retention has led to the nuclear enrichment of otherwise cytoplasmic lncRNAs [38]. Therefore, the use of non-canonical and weak splicing motifs may affect lncRNA stability and nuclear export, thereby altering their localization.

The use of non-canonical splicing sites may also influence the polyadenylation process. All mRNA molecules, and most lncRNAs, undergo polyadenylation. This process involves the identification of an A[A/U]UAAA hexamer and a GU/U-rich sequence located approximately 20 nucleotides downstream, followed by RNA

cleavage and non-template-dependent synthesis of a poly(A) tail [39]. Interestingly, alternative sequences may also participate in the polyadenylation of mRNAs and lncRNAs [40]. Similar to splicing, polyadenylation occurs during transcription and plays a pivotal role in RNA stability and localization. A connection between polyadenylation and TREX has been reported, with TREX mutations leading to the accumulation of bulk poly(A) RNAs [41]. As most stable lncRNA transcripts are polyadenylated, they are subject to the export machinery.

Two of the most extensively studied lncRNAs, MALAT1 and NEAT1_2, exhibit a unique sequence involved in lncRNA processing. These single-exon lncRNAs are processed by the ribonuclease, RNase P, which is known for its role in tRNA maturation by cleaving the leader sequence at the 5'-end of pre-tRNA. Both MALAT1 and NEAT1_2 generate tRNA-like structures at their 3'-ends that are cleaved by RNase P [42]. However, the unprocessed ends of these transcripts could lead to their rapid degradation by a 3'-to-5' exonuclease. To protect their 3'-ends, MALAT1 and NEAT1_2 generate a triple-helix structure, composed of two U-rich motifs and an A-rich tract [43]. Currently, this remains a unique sequence feature of these two human lncRNAs, as no other lncRNA has been found to undergo RNase P processing or to have such a unique triple-helix structure. However, structures resembling this triple-helix have been identified in lncRNAs of other genomes, such as the polyadenylated nuclear (PAN) RNA, produced by Kaposi's sarcoma-associated herpesvirus (KSHV). Hence, MALAT1 and NEAT1 represent unique structures of lncRNAs that favor their nuclear retention. However, while some lncRNAs are enriched in the nucleus due to inefficient processing or unique features, other lncRNAs contain specific sequences that actively retain them in the nucleus.

Comprehensive studies using computational and molecular tools have identified additional motifs that influence lncRNA localization. Using a reporter transcript, which by default exports to the cytoplasm, these studies cloned fragments of nuclear-enriched lncRNAs (and mRNAs). Using thousands of gene fragments, these studies identified sequences affecting nuclear localization. These include a fragment of 42 nt Alu [44]. Further analysis refined the motif and identified a core sequence of GCCUCCC. More generally, these studies identify a general tendency toward C-rich sequences that interact with nuclear proteins, such as hnRNPK, SLTM, and SNRNP70 [44, 45]. Nevertheless, both studies highlight the finding that the RNA localization signal is a combination of multiple sequences and is context-dependent. Other high-throughput studies identified the CAGGUGAGU motif, which interacts with U1 snRNA [46], as well as the importance of other transposable elements in transcript localization [47, 48]. These experimental findings underscore the complexity of RNA localization but also pave the way for a better understanding of the relationship between RNA sequence and function.

5. Motif-based lncRNA-protein interactions

While some lncRNA-protein interactions exhibit low-sequence specificity, other RNA-protein complexes rely on interactions between RNA-binding proteins (RBPs) and a sequence-specific motif. RBPs play a critical role in the RNA lifecycle by regulating their processing, localization, stability, and function. When lncRNA was discovered, it was suggested that these transcripts function as scaffolds for RBPs [49]. RBPs' recognition of motifs of 4–6 nt long is done with a certain flexibility and is context-dependent (**Figure 2**).

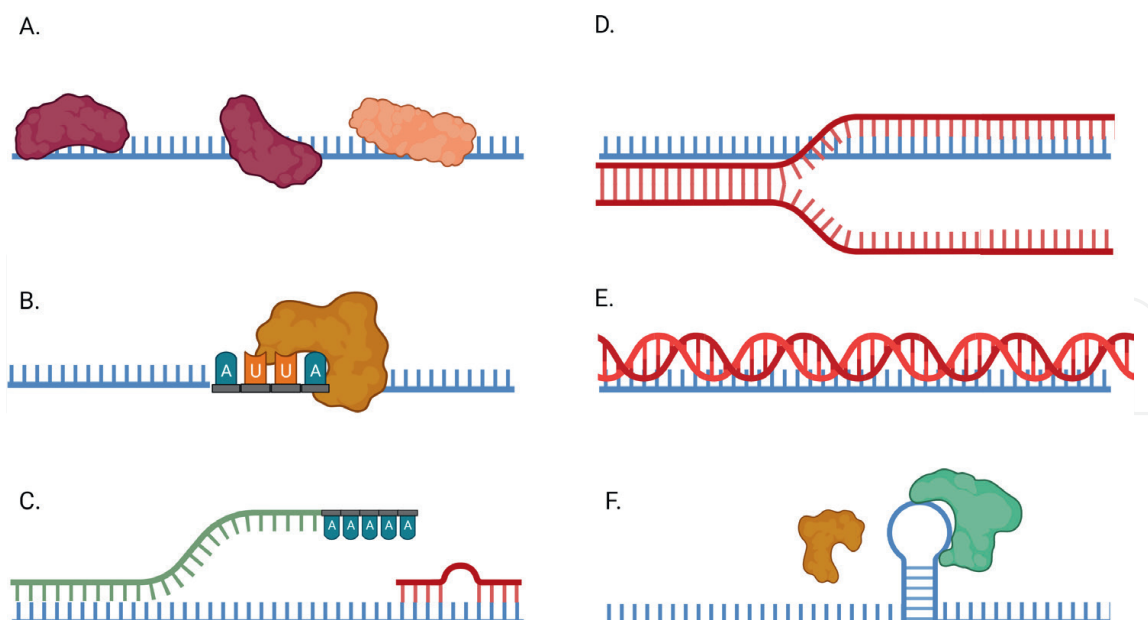


Figure 2. *lncRNAs can interact with different partners. (A) Non-sequence-specific interactions with proteins. (B) Interaction with RBPs based on motifs. (C) Interactions with mRNA or miRNA. (D) Interaction with DNA (R-loop) (E) interaction with DNA (DNA:DNA:RNA triplex). (F) Self-interactions to generate secondary structures affecting interaction with other molecules. (created with BioRender.com).*

The identification of protein-RNA interactions can be performed through RNA pulldown, followed by immunoblotting or mass spectrometry analysis. Alternatively, the pulldown of a suspected protein, followed by quantitative polymerase chain reaction (qPCR) or next-generation sequencing (NGS), can identify which RNA molecules interact with a known protein of interest. However, identifying these interactions is not trivial. The low expression of most lncRNAs, and the vulnerability of RNA molecules to degradation, limits the sensitivity of *in situ* pulldown approaches. On the other hand, *in vitro* approaches may identify non-physiological interactions because of inconsistencies in stoichiometry, localization, and competing interactions. Furthermore, as mentioned above, some proteins are “sticky,” and they do not have specific binding sites, but rather interact with sequence or structural motifs on the transcript [50, 51]. While these interactions may be functional, they may also result in misleading interpretations.

NORAD is a highly abundant, cytoplasmic-enriched, conserved, single-exon lncRNA. Functionally, at the cellular level, NORAD knockdown/–out leads to chromosomal instability and to aging-related phenotypes in mouse models. Pulldown experiments using biotinylated NORAD fragments and cell extract identified PUMILIO homolog 2 (PUM2) as the most enriched interacting protein. These findings were further supported by a sequence analysis that identified at least 15 conserved PUMILIO response elements (PREs) [52]. Further studies demonstrated that NORAD can generate membrane-less bodies in the cytoplasm. This RNA-protein structure generates non-conserved protein-protein interactions that can sponge PUMILIO by overcoming naive protein-RNA stoichiometry. However, NORAD interacts with many other proteins; for example, the RBP SAM68 interacts functionally with UAAA motifs in NORAD [53]. Interestingly, a different pulldown approach, based on *in situ* crosslinking, led to a different list of NORAD-interacting proteins. This analysis highlights the role of nuclear NORAD in DNA replication and genome stability through its direct interaction with the RNA-binding motif protein X-linked (RBMX) [54].

Surprisingly, even though RBPs play a critical role in lncRNA function, only a few specific interaction motifs have been identified. The in-depth characterization of the specific sequences involved in these interactions will provide a further understanding of lncRNA function.

6. The role of secondary structures in lncRNA interactions

Single-stranded RNA (ssRNA) is energetically unstable, folding into secondary and tertiary structures that play a pivotal role in the function of many lncRNAs. These structures generate new binding sites or prevent interactions by altering the context or the availability of motifs. RNA secondary structures are primarily dictated by Watson–Crick base pairing, but non-canonical base pairing is also prevalent, thereby increasing the functional repertoire of RNA [55–57].

An early example of the significance of secondary RNA structures is the lncRNA maternally expressed gene 3 (MEG3), a tumor suppressor that regulates the expressions of P53-downstream genes. Structure and functional analyses identified three structural motifs on MEG3, where the deletion of two affected its function. Interestingly, the function of MEG3 can be restored simply by replacing these deleted motifs with a hybrid RNA of different sequences that can generate the same structures. Therefore, in this case, the structure, rather than the sequence itself, is critical for RNA function [58].

Another example is the lncRNA steroid receptor RNA activator (SRA), which interacts with chromatin modifiers and plays a key role in regulating the expressions of hundreds of genes. Interestingly, SRA is assembled by four domains, whose secondary structure is conserved across 45 vertebrates. However, the sequence itself is much less conserved, indicating that the secondary structure is the driving force in SRA function [59].

Recent studies have shown that dynamic changes in secondary structures can regulate the availability of RNA-binding motifs to their RNA-binding proteins (RBPs). Most motifs are short; therefore, they can be randomly distributed across transcripts. However, higher-order structures affect the availability and affinity of such motifs in protein interactions. It was shown that in a synthetic stem-loop structure, the human U1A (SNRPA) protein interacts better when its motif is located on the loop rather than on the stem part [60]. This way, certain sequences are protected from unwanted interactions. However, the RNA secondary structure is dynamic. Using *in situ* hybridization followed by high-resolution structure analysis, it was shown that the secondary structure of the lncRNA non-coding RNA activated by DNA damage (NORAD) can change under stressful conditions, affecting its biological function [61]. This suggests a new level of lncRNA regulation.

RNA secondary structures may interact directly with DNA-binding proteins. While RBPs usually recognize and interact with single-strand RNA (ssRNA), DNA-binding proteins interact with double-strand sequences. Furthermore, lncRNA secondary structures can function as decoys or activators of transcription factors (TFs) and DNA damage response proteins that identify double-strand DNA. For example, TFs interact with specific DNA sequences that have different chemical/physical characteristics. However, it was shown that some lncRNA sequences can generate double-strand RNA, which can interact with TFs. lncRNA growth arrest-specific 5 (Gas5) can generate a stem-loop structure that mimics the DNA glucocorticoid response elements found in promoters of its target genes. Under physiological

conditions of nutrient deprivation, the Gas5 expression is elevated, and it competes with the endogenous sites of glucocorticoid in the promoters to regulate the expressions of metabolic and apoptotic genes [62, 63].

Recent research has further underscored the significance of RNA secondary structures in the regulation of chromatin-associated proteins. A notable example is the interaction between DNMT1, a DNA methyltransferase, and lncRNAs. Multiple studies have shown that DNMT1 can interact with lncRNAs to regulate the expression of specific genes; however, the mode of interaction was not fully clear [64–66]. Recently, it was shown that DNMT1 exhibits a strong and specific affinity for GU-rich RNAs that form a pUG-fold, a noncanonical G-quadruplex [67]. This interaction is not merely incidental; pUG-fold-capable RNAs were found to inhibit DNMT1 activity by preventing the binding of hemimethylated DNA. This suggests that the secondary structure of RNA can directly influence the activity of chromatin-associated proteins, further emphasizing the critical role of RNA secondary structures in the regulation of protein function and activity.

More recently, it has been discovered that besides interacting with DNA, the DNA damage response proteins Ku70/80 can also interact with the secondary structures of lncRNAs. These interactions have been shown to be functional, increasing the activity of such proteins and the efficiency of the DNA damage response [68, 69]. lncRNAs can also interact with TFs and regulate their function. The secondary structure of the mouse lncRNA Braveheart generates a stable stem-loop structure, protecting ssRNA and functioning as a binding site for the nucleic acid binding protein (CNBP/ZNF9) TF [70]. However, it has been recently revealed that these TFs might not interact with lncRNAs through their canonical DNA-binding domain. The TF signal transducer and activator of transcription 3 (STAT3) was found to interact directly with the lncRNA NORAD. However, this interaction is not mediated by the STAT3 DNA-binding domain but rather by aptamer-like interactions with the SH2 domain. This suggests that NORAD does not function as a decoy but rather as a scaffold for the proper function and localization of STAT3 [50].

In conclusion, the higher-order structures of lncRNAs play a crucial role in their function. Furthermore, these secondary structures have recently emerged as promising therapeutic targets for inhibitory small molecules [71]. However, the methods for studying RNA secondary structures are still in their early stages. As the field continues to develop, we anticipate shedding lighter on the important aspects of lncRNA biology and its therapeutic potential.

7. lncRNA interaction with DNA (R-loop) and DNA:DNA:RNA triplexes

R-loops are RNA:DNA hybrids that can form during transcription when the newly synthesized RNA interacts with the DNA template (cis). R-loops play key roles in replication, transcription, and genome integrity. They are enriched near the promoters/transcription start sites (TSSs) of highly expressed genes and near transcription termination sites, suggesting that R-loops play a functional role in transcription regulation [72].

While theoretically, RNA:DNA hybrids can be generated simply by displacement and Watson-Crick hybridization during transcription, R-loop formation is regulated by proteins known to be involved in homologous recombination during DNA damage repair [73]. Importantly, this active R-loop formation mechanism can lead to the hybridization of RNA molecules that share complementary sequences but are

transcribed from spatially distinct sites (trans). However, R-loops that are adverse byproducts of transcription must be resolved, typically by RNase H or the Fanconi anemia (FA) pathway; otherwise, these R-loops may interfere with transcription and DNA replication (**Figure 2D**).

Antisense lncRNAs can form R-loops with their sense protein-coding genes (PCGs) and can regulate their expression in cis. For instance, lncRNA ANRASSF1 is transcribed in an antisense orientation to the PCG RASSF1. ANRASSF1 transcription results in the formation of an R-loop, which recruits PRC2 and promotes the epigenetic silencing of RASSF1. Conversely, other R-loop forming lncRNAs are associated with increased transcription. The lncRNA VIM-AS1 is a divergent transcript with PCG vimentin (VIM). The transcription of PCG VIM is regulated by the methylation of a CpG island, and it has been shown that the antisense transcription overlapping the TSS generates an R-loop that inhibits the interaction of NF- κ B [74]. Genome-wide studies show that R-loops frequently form at sites of antisense transcription, close to the promoter and TSS, especially at sites enriched in C/G nucleotides. These interactions correlate with low DNA methylation and active promoters [75].

Another significant type of lncRNA interaction with DNA is mediated by the generation of triplexes (DNA:DNA:RNA) between pyrimidine (T and C)-rich RNAs and purine (A and G)-rich DNA *via* Hoogsteen-type hydrogen bonding (**Figure 2E**). The mouse lncRNA Fendrr recruits the epigenetic modulators PRC2 and TrxG/MLL to the promoters of developmental genes. It has been shown that Fendrr interacts with the promoters of the *Foxf1* gene in cis and *Pitx2* in trans. *In vitro*, these interactions are sensitive to digestion by RNase V1 but not RNase H, which specifically cleaves the DNA:RNA duplex, indicating that Fendrr is involved in triplex (DNA:DNA:RNA) interactions [76]. Another example is lncRNA KHPS1, which is transcribed in antisense orientation to enhancer RNA (eRNA) SPHK1 and generates a DNA:RNA triplex to regulate the expression of SPHK1 mRNA. Interestingly, the incorporation of 7-deaza-purine nucleotides to the triplex-forming region (TFR), which prevents the formation of Hoogsteen base pairing and DNA:RNA triplex formation, abolishes the function of the RNA. Furthermore, by replacing the TFR of KHPS1 with MEG3 [77] and generating a chimeric lncRNA, it was shown that TFRs interact with their RNA partners in an interchangeable manner [78].

Moreover, an analysis of lncRNA-chromatin interactions, based on experimental data, suggests that lncRNAs have cis and trans targets, and lncRNAs regulate multiple genes that can be clustered into specific pathways [77, 79, 80]. Genome-wide analysis indicates that the DNA:RNA triplex is a major mechanism of lncRNA function *in vivo*. An interesting finding based on these analyses is the mechanistic regulatory function of lncRNA HIF1 α -AS1 at multiple genomic sites by recruiting the silencing complex under specific physiological contexts [81]. In addition, computational modeling based on chromatin fractionation and an analysis of the dynamic of the dissociation of transcripts from their transcription sites show that lncRNAs overlapping enhancer regions tend to harbor longer before releasing to the nucleoplasm [82]. This may indicate their role in chromatin regulation, directly coupled with their transcription through DNA:RNA interactions. These findings pave the way for a better understanding of these DNA:RNA interactions.

In summary, R-loops and DNA:RNA triplexes constitute important regulatory mechanisms for transcription initiation and termination. In many cases, this is mediated by cis-antisense transcription, which generates unstable RNA. However, accumulating evidence indicates that bona fide lncRNAs can generate R-loops and a DNA:RNA triplex in cis and trans to affect chromatin modification and accessibility.

8. lncRNA interaction with miRNA

MicroRNAs (miRNAs) constitute a significant class of non-coding RNAs (ncRNAs) that modulate post-transcriptional regulation by fine-tuning mRNA translation and stability. miRNAs interact with the miRNA-induced silencing complex (miRISC) and guide the miRISC to its miRNA recognition element (MRE) target based on Watson–Crick base pairing. Notably, in animal cells, most miRNA:MRE interactions are incomplete and include multiple mismatches, which makes the prediction of physiological miRNA targets a challenging task due to the vast number of potential interactions.

Experimental data suggest that miRNAs primarily target mRNA at their 3' untranslated regions (UTRs) and that miRISC predominantly functions in the cytoplasm, although several studies have demonstrated nuclear function under specific physiological conditions [83, 84]. miRNAs can target mRNAs, as well as lncRNAs, pseudogene-derived transcripts, and circular RNA [85]. As a result, all transcripts compete for the same pool of miRISCs. Consequently, changes in the level of a specific transcript may affect the availability of miRNAs for other targets. This mechanism is described as sponging or competitive endogenous RNA (ceRNA), and it is suggested as a mechanism of action for numerous lncRNAs [86].

While the sponging/ceRNA concept provides an appealing framework to predict lncRNA:mRNA post-transcriptional regulation, accurate prediction is challenging due to the complexity of the synergistic networks and the lack of physiological data required to establish reliable variables and constants [87–91]. Therefore, while growing evidence indicates the attractiveness of sponging as a mechanism and computational predictions suggest a unified language that can explain the function of many lncRNAs, the prediction and validation of specific ceRNA cross-talk remain challenging.

9. Modular sequences in lncRNA and the role of repeat elements

Protein-coding genes (PCGs) are characterized by functional domains, which are modular units with distinct structures and functions [92, 93]. The question of whether long non-coding RNAs (lncRNAs) are similarly structured by modular sequences remains largely unexplored. Transposable elements (TEs), which are enriched in lncRNAs, are proposed to serve as functional domains (**Figure 3**) [93].

TEs are sequences with the ability to duplicate (retroviral oriented) or relocate (non-retroviral elements) within the genome. They are abundant in the human genome, with certain types of TEs specifically enriched in lncRNAs [94]. Previously regarded as parasitic, non-functional elements, TEs are now acknowledged as vital

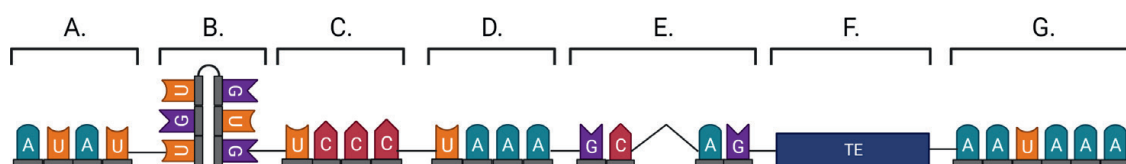


Figure 3. Sequence motifs in lncRNAs. (A) lncRNA tends to be enriched with a/U. (B) Simple repeats and microsatellites. (C) C-rich sequences involved in nuclear retention. (D) Recognition motifs for proteins and miRNAs. (E) Wick splicing signal. (F) Transposon elements. (G) Polyadenylation and other mRNA motifs. (created with BioRender.com).

to the architecture and function of the genome, driving evolution through their high mobility and consequent rapid genomic sequence changes.

It is estimated that around 80% of human lncRNAs incorporate at least one TE sequence, with approximately 30% of lncRNA sequences being TE-related. TE sequences contribute to the architecture and function of lncRNAs by providing splicing sites, binding domains for RNA-binding proteins (RBPs), and domains for RNA:RNA and RNA:DNA interactions. The insertion of Alu TEs can promote the formation of new exons (a process known as exonization). TEs also contribute sequences for poly(A) signals, as well as sequences affecting lncRNA transcription and localization. The high content of TEs in lncRNAs can promote secondary structures (inverted repeats) that are substrates for A-to-I RNA editing.

As previously discussed, the SIRLOIN motif found in Alu TE elements [44, 95] and sequences from the L1PA16, L2b, MIRb, and MIRc TEs [47] play a role in promoting RNA nuclear retention and the formation of nuclear bodies. Interestingly, when a SINE TE from murine Malat1 was removed using CRISPR-Cas9, lncRNAs were dispersed from the nuclear speckles to the cytoplasm, suggesting a role for TE in its localization and organization. While the SINEB1 TE element is murine-specific, functional sequences can be found in human MALAT1, arguably through convergent evolution [96]. Similarly, the deletion of TE sequences within NEAT1 led to abnormal speckle organization and function [97].

Repetitive elements can also lead to lncRNA-mRNA duplexing. Evidence suggests that these interactions can be functional in mRNA translation and stability [98–102]. However, these interactions are limited by secondary structures, transcript abundance, and localization.

Overall, these repeats play a functional role by interacting with RBPs, RNA, and chromatin, and they demonstrate a proposed mechanism for lncRNA evolution that is based on TEs as functional domains of lncRNAs, termed repeat insertion domains of lncRNAs (RIDLs) [93].

While many lncRNAs contain TEs, which are generated by ancient independent elements, other repetitive sequences are simpler in their nature. Tandem repeats (TRs) and local repeats (LRs) are low-complexity sequences. While TRs appear in multiple loci across the genome, LRs are generally found in a specific locus. Although these sequences are frequent in introns and intergenic DNA, some are located in exons. LRs are enriched in lncRNAs compared to PCGs, and they are suggested to be involved in interactions with RBPs (**Figure 3**). The exonic repeating RNA domain (RRD) in lncRNA functional intergenic repeating RNA element (FIRRE) is a 156-bp repeating sequence that was shown to interact with hnRNPU and to play a critical role in the nuclear retention of the transcript [103].

In addition, lncRNA sequence analyses identified an enrichment of short tandem repeat sequences that can interact with RBPs. For example, pyrimidine-rich noncoding transcript (PNCTR) is a 10-Kb-long lncRNA transcribed by RNA pol I. PNCTR is composed of more than two thousand PTBP1-binding motifs. PNCTR interacts with PRBP1 to generate nuclear bodies called perinucleolar compartments (PNCs). The sequestration and antagonization of PRBP1 by PNCTR play a functional role in splicing regulation [104].

In conclusion, repeat elements, whether they are TEs, TRs, or LRs, play a significant role in the structure and function of lncRNAs. They contribute to the interaction with RBPs, RNA, and chromatin, and they provide a mechanism for lncRNA evolution. These findings pave the way for a better understanding of the role of repeat elements in lncRNAs.

10. Unraveling lncRNA function through sequence analysis

RNA sequences carry the blueprint of their function. While experimental approaches often focus on function first and subsequently search for related sequences, computational approaches primarily compare sequences to identify the functional motifs.

The assumption that functional elements within a sequence are modular allows for the evaluation of lncRNA sequence similarity not merely based on linear alignment, but rather on the accumulation of shared motifs. The linear sequence can be deconstructed into constituent short sequences (K-mers). The distance between any two transcripts can then be described by the correlation between their K-mers, an approach known as Sequence Evaluation from K-mers Representation (SEEKR) [105]. This method has successfully clustered together lncRNAs with known similar functions and differentiated between activating and repressing lncRNAs. Interestingly, SEEKR analysis found that repressor lncRNAs, such as XIST, were enriched in G/C K-mers, while activator lncRNAs, like HOTTIP, clustered together due to the enrichment of A/U nucleotides. Furthermore, K-mer distribution can predict lncRNA subcellular localization, with K-mers enriched in the nucleus found to harbor motifs corresponding with proteins known to affect RNA retention.

Sequence conservation through evolution is a hallmark of functionality. However, many lncRNAs appear to evolve in a species-restricted manner. While most lncRNAs do not show sequence similarity, many reveal syntenic similarities between distinct species. This suggests that lncRNAs maintain their cis-regulatory function without obvious sequence similarities. Furthermore, lncRNAs can change their linear sequence but maintain similar structure and function. While SEEKR looks for the enrichment of multiple K-mers in a non-consecutive order, an analysis of lncRNAs, such as NORAD, identified short RBP recognition motifs in which not only the sequence was conserved, but the sequential order was also maintained through evolution [53]. The LncLOOM framework predicts lncRNA functional motifs based on the assumptions of short motif conservation and motif order.

Collectively, advancements in RNA biology have enabled the development of analytical approaches dedicated to the analysis of lncRNAs. Yet, fundamental questions remain unanswered. Do lncRNAs share a common syntax? Are lncRNAs developed mainly in the context of a specific locus or function? How do the second and third dimensions of lncRNAs function in the syntax of lncRNAs? How can we divide lncRNAs into functional subclasses based on their sequence? How can we predict the function of lncRNAs based on their sequence? The answers to these questions will not only improve our understanding of lncRNA biology but may also open opportunities for novel therapeutic approaches.

11. Conclusions

Identifying a functional sequence in a specific lncRNA remains a challenging task due to the inherent biological nature of lncRNAs and the lack of robust methods. Many of the functional sequences discussed in this review are short and have low sequence restrictions, and their function is highly context-dependent. RNA-binding protein (RBP) and miRNA recognition elements are relatively short and can be found distributed by “chance” along multiple sequences. Furthermore, lncRNAs are enriched in transposable elements (TEs) and other repeat elements that can

theoretically interact with many RNA and chromatin sites. Therefore, the function of a sequence depends not only on the context but also on its subcellular localization and abundance [31, 106, 107]. Moreover, recent studies have shown that lncRNA expression is highly dynamic due to its long transcriptional bursts [108], thereby increasing the spatial-temporal function of lncRNAs.

The advancement in experimental techniques has also paved the way for a better understanding of lncRNA sequence-to-function relationships. High-throughput experiments using an expression system with various putative functional sequences enable the identification of motifs that affect RNA nuclear retention. Additionally, CRISPR-based deletions of endogenous sequences or truncated exogenous constructs help to underline functional domains in lncRNAs, especially when combined with evidence from pulldown and phenotypic experiments.

However, the identification of functional domains through phenotypic screening is limited by the complexity of the experimental design and the interpretation of output phenotypes. Therefore, computational approaches, in conjunction with experimental data, will play a crucial role in our understanding of the language of lncRNAs. Sequence analysis shows that lncRNAs differ from protein-coding genes (PCGs), not only due to the lack of substantial open-reading frame (ORF) potential, but also in their nucleotide composition and repetitive elements [13], indicating that lncRNAs function under unique sequence constraints.

Our understanding of the biology and function of lncRNAs is gradually unfolding. This process goes hand in hand not only with developments in analytic and experimental approaches but also with the acknowledgment of lncRNAs' roles in cell function and genome architecture. The role of repeat elements in the structure and function of lncRNAs, as well as the interactions between lncRNAs and proteins or miRNAs, is emerging as significant areas of study.

As we continue to unravel the mysteries of lncRNAs, we anticipate that these non-coding RNAs will provide novel insights into cellular function and offer new avenues for therapeutic intervention. Future research should focus on overcoming the current challenges in identifying functional sequences in lncRNAs, further exploring the potential applications of lncRNAs, and filling the gaps in our current understanding of lncRNAs. The questions posed at the beginning of this chapter remain partially answered, and further research is needed to fully understand the common syntax, functional subclasses, and predictive function of lncRNAs based on their sequence.

Acknowledgements

This work was funded by the Israeli Science Foundation (ISF; grant no. 2228/19) and the Israel Cancer Association (ICA; grant no. 20230069).

Conflict of interest

The authors declare no conflict of interest.

IntechOpen

IntechOpen


Author details

Assaf C. Bester

Department of Biology, Technion-Israel Institute of Technology, Haifa, Israel

*Address all correspondence to: bestera@technion.ac.il

IntechOpen

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Frankish A et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*. 2019;**47**(D1):D766-D773. DOI: 10.1093/nar/gky955
- [2] Cho SW et al. Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell*. 2018;**173**(6):1398-1412.e22. DOI: 10.1016/j.cell.2018.03.068
- [3] Engreitz JM et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*. 2016;**539**(7629):452-455. DOI: 10.1038/nature20149
- [4] Gil N, Ulitsky I. Production of spliced long noncoding RNAs specifies regions with increased enhancer activity. *Cell Systems*. 2018;**7**(5):537-547.e3. DOI: 10.1016/j.cels.2018.10.009
- [5] Ali T, Grote P. Beyond the RNA-dependent function of lncRNA genes. *ELife*. 2020;**9**:e60583. [Published 2020 Oct 23] DOI: 10.7554/eLife.60583
- [6] Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*. 2015;**4**:e08890. DOI: 10.7554/eLife.08890
- [7] Matsumoto A et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*. 2017;**541**(7636):228-232. DOI: 10.1038/nature21034
- [8] Hazan J, Bester AC. CRISPR-based approaches for the high-throughput characterization of long non-coding RNAs. *Non-coding RNA*. 2021;**7**(4):79. DOI: 10.3390/ncrna7040079
- [9] An NA, Zhang J, Mo F, et al. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nature Ecology and Evolution*. 2023;**7**:264-278. DOI: 10.1038/s41559-022-01925-6
- [10] Amit M et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports*. 2012;**1**(5):543-556. DOI: 10.1016/j.celrep.2012.03.013
- [11] Haerty W, Ponting CP. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA*. 2015;**21**(3):333-346. DOI: 10.1261/rna.047324.114
- [12] Ulveling D, Dinger ME, Francastel C, Hubé F. Identification of a dinucleotide signature that discriminates coding from non-coding long RNAs. *Frontiers in Genetics*. 2014;**5**:316. DOI: 10.3389/fgene.2014.00316
- [13] Wen J, Liu Y, Shi Y, Huang H, Deng B, Xiao X. A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. *BMC Bioinformatics*. 2019;**20**(1):469. DOI: 10.1186/s12859-019-3039-3
- [14] Li A, Zhang J, Zhou Z. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*. 2014;**15**:311. DOI: 10.1186/1471-2105-15-311
- [15] Hu L, Xu Z, Hu B, Lu ZJ. COME: A robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Research*. 2017;**45**(1):e2. DOI: 10.1093/nar/gkw798
- [16] Anderson DM et al. A micropeptide encoded by a putative long noncoding

- RNA regulates muscle performance. *Cell*. 2015;**160**(4):595-606. DOI: 10.1016/j.cell.2015.01.009
- [17] Wright BW, Yi Z, Weissman JS, Chen J. The dark proteome: translation from noncanonical open reading frames, 2022. doi: 10.1016/j.tcb.2021.10.010.
- [18] Bester AC et al. An integrated genome-wide CRISPRa approach to functionalize lncRNAs in drug resistance. *Cell*. 2018;**173**(3):649-664.e20. DOI: 10.1016/j.cell.2018.03.052
- [19] Liu SJ et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*. 2017;**355**:eaah7111. DOI: 10.1126/science.aah7111
- [20] Joung J et al. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nature Protocols*. 2017;**12**(4):828-863. DOI: 10.1038/nprot.2017.016
- [21] Derrien T et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*. 2012;**22**(9):1775-1789. DOI: 10.1101/gr.132159.111
- [22] Khalil AM et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;**106**(28):11667-11672. DOI: 10.1073/pnas.0904715106
- [23] Colognori D, Sunwoo H, Kriz AJ, Wang CY, Lee JT. Xist deletional analysis reveals an interdependency between Xist RNA and polycomb complexes for spreading along the inactive X. *Molecular Cell*. 2019;**74**(1):101-117.e10. DOI: 10.1016/j.molcel.2019.01.015
- [24] Dixon-McDougall T, Brown CJ. Independent domains for recruitment of PRC1 and PRC2 by human XIST. *PLoS Genetics*. 2021;**17**(3):e1009123. DOI: 10.1371/journal.pgen.1009123
- [25] Bousard A et al. The role of Xist-mediated polycomb recruitment in the initiation of X-chromosome inactivation. *EMBO Reports*. 2019;**20**(10):e48019. DOI: 10.15252/embr.201948019
- [26] Wang X et al. Targeting of polycomb repressive complex 2 to RNA by short repeats of consecutive guanines. *Molecular Cell*. 2017;**65**(6):1056-1067.e5. DOI: 10.1016/j.molcel.2017.02.003
- [27] Rosenberg M et al. Motif-driven interactions between RNA and PRC2 are rheostats that regulate transcription elongation. *Nature Structural & Molecular Biology*. 2021;**28**(1):103-117. DOI: 10.1038/s41594-020-00535-9
- [28] Kaneko S, Son J, Bonasio R, Shen SS, Reinberg D. Nascent RNA interaction keeps PRC2 activity poised and in check. *Genes & Development*. 2014;**28**(18):1983-1988. DOI: 10.1101/gad.247940.114
- [29] Gil N, Ulitsky I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews. Genetics*. 2020;**21**(2):102-117. DOI: 10.1038/s41576-019-0184-5
- [30] Tan JY, Biasini A, Young RS, Marques AC. Splicing of enhancer-associated lincRNAs contributes to enhancer activity. *Life Science Alliance*. 2020;**3**(4):1-12. DOI: 10.26508/LSA.202000663
- [31] Quinodoz SA et al. RNA promotes the formation of spatial compartments in the nucleus. *Cell*. 2021;**184**(23):5775-5790.e30. DOI: 10.1016/j.cell.2021.10.014

- [32] Zuckerman B, Ron M, Mikl M, Segal E, Ulitsky I. Gene architecture and sequence composition underpin selective dependency of nuclear export of long RNAs on NXF1 and the TREX complex. *Molecular Cell*. 2020;**79**(2):251-267.e6. DOI: 10.1016/j.molcel.2020.05.013
- [33] Wu KE, Parker KR, Fazal FM, Chang HY, Zou J. RNA-GPS predicts high-resolution RNA subcellular localization and highlights the role of splicing. *RNA*. 2020;**26**(7):851-865. DOI: 10.1261/rna.074161.119
- [34] Zuckerman B, Ulitsky I. Predictive models of subcellular localization of long RNAs. *RNA*. 2019;**25**(5):557-572. DOI: 10.1261/rna.068288.118
- [35] Lei H, Dias AP, Reed R. Export and stability of naturally intronless mRNAs require specific coding region sequences and the TREX mRNA export complex. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;**108**(44):17985-17990. DOI: 10.1073/pnas.1113076108
- [36] Krchnáková Z et al. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Research*. 2019;**47**(2):911-928. DOI: 10.1093/nar/gky1147
- [37] Abou Alezz M, Celli L, Belotti G, Lisa A, Bione S. GC-AG introns features in long non-coding and protein-coding genes suggest their role in gene expression regulation. *Frontiers in Genetics*. 2020;**11**:488. DOI: 10.3389/fgene.2020.00488
- [38] Dumbović G et al. Nuclear compartmentalization of TERT mRNA and TUG1 lncRNA is driven by intron retention. *Nature Communications*. 2021;**12**(1):3308. DOI: 10.1038/s41467-021-23221-w
- [39] Richter JD. Cytoplasmic polyadenylation in development and beyond. *Microbiology and Molecular Biology Reviews*. 1999;**63**(2):446-456. DOI: 10.1128/MMBR.63.2.446-456.1999
- [40] Li X-Q, Du D. Motif types, motif locations and base composition patterns around the RNA polyadenylation site in microorganisms, plants and animals. *BMC Evolutionary Biology*. 2014;**14**:162. DOI: 10.1186/s12862-014-0162-7
- [41] Katahira J, Okuzaki D, Inoue H, Yoneda Y, Maehara K, Ohkawa Y. Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. *Nucleic Acids Research*. 2013;**41**(14):7060-7072. DOI: 10.1093/nar/gkt414
- [42] Wilusz JE, Freier SM, Spector DL. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell*. 2008;**135**(5):919-932. DOI: 10.1016/j.cell.2008.10.012
- [43] Wilusz JE, JnBaptiste CK, Lu LY, Kuhn C-D, Joshua-Tor L, Sharp PA. A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(a) tails. *Genes & Development*. 2012;**26**(21):2392-2407. DOI: 10.1101/gad.204438.112
- [44] Lubelsky Y, Ulitsky I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature*. 2018;**555**(7694):107-111. DOI: 10.1038/nature25757
- [45] Zhang B, Gunawardane L, Niazi F, Jahanbani F, Chen X, Valadkhan S. A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Molecular and Cellular Biology*. Jun 2014;**34**(12):2318-2329. DOI: 10.1128/MCB.01673-13

- [46] Yin Y et al. U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature*. 2020;**580**(7801):147-150. DOI: 10.1038/s41586-020-2105-3
- [47] Carlevaro-Fita J, Johnson R. Global positioning system: Understanding long noncoding RNAs through subcellular localization. *Molecular Cell*. 2019;**73**(5):869-883. DOI: 10.1016/j.molcel.2019.02.008
- [48] Ron M, Ulitsky I. Context-specific effects of sequence elements on subcellular localization of linear and circular RNAs. *Nature Communications*. 2022;**13**(1):2481. DOI: 10.1038/s41467-022-30183-0
- [49] Poramba-Liyanage DW et al. Inhibition of transcription leads to rewiring of locus-specific chromatin proteomes. *Genome Research*. 2020;**30**(4):635-646. DOI: 10.1101/gr.256255.119
- [50] Dvir S et al. Uncovering the RNA-binding protein landscape in the pluripotency network of human embryonic stem cells. *Cell Reports*. 2021;**35**(9):109198. DOI: 10.1016/j.celrep.2021.109198
- [51] Zhang Z, Sun W, Shi T, Lu P, Zhuang M, Liu JL. Capturing RNA-protein interaction via CRUIS. *Nucleic Acids Research*. 2020;**48**(9):e52-e52. DOI: 10.1093/nar/gkaa143
- [52] Lee S et al. Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell*. 2016;**164**(1-2):69-80. DOI: 10.1016/j.cell.2015.12.017
- [53] Tichon A, Perry RBT, Stojic L, Ulitsky I. SAM68 is required for regulation of pumilio by the NORAD long noncoding RNA. *Genes & Development*. 2018;**32**(1):70-78. DOI: 10.1101/gad.309138.117
- [54] Munschauer M et al. The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature*. 2018;**561**(7721):132-136. DOI: 10.1038/s41586-018-0453-z
- [55] Yang SY et al. Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nature Communications*. 2018;**9**(1):4730. DOI: 10.1038/s41467-018-07224-8
- [56] Lyu K, Chow EY-C, Mou X, Chan T-F, Kwok CK. RNA G-quadruplexes (rG4s): Genomics and biological functions. *Nucleic Acids Research*. 2021;**49**(10):5426-5450. DOI: 10.1093/nar/gkab187
- [57] Simko EAJ et al. G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. *Nucleic Acids Research*. 2020;**48**(13):7421-7438. DOI: 10.1093/nar/gkaa475
- [58] Zhang X et al. Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: Isoform structure, expression, and functions. *Endocrinology*. 2010;**151**(3):939-947. DOI: 10.1210/en.2009-0657
- [59] Novikova IV, Hennelly SP, Sanbonmatsu KY. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Research*. 2012;**40**(11):5034-5051. DOI: 10.1093/nar/gks071
- [60] Komatsu KR, Taya T, Matsumoto S, Miyashita E, Kashida S, Saito H. RNA structure-wide discovery of functional interactions with multiplexed RNA motif library. *Nature Communications*. 2020;**11**(1):6275. DOI: 10.1038/s41467-020-19699-5

- [61] Ziv O, Farberov S, Lau JY, Miska E, Kudla G, Ulitsky I. Structural features within the NORAD long noncoding RNA underlie efficient repression of Pumilio activity. *bioRxiv*. 2021. DOI: 10.1101/2021.11.19.469243
- [62] Parsonnet NV, Lammer NC, Holmes ZE, Batey RT, Wuttke DS. The glucocorticoid receptor DNA-binding domain recognizes RNA hairpin structures with high affinity. *Nucleic Acids Research*. 2019;**47**(15):8180-8192. DOI: 10.1093/nar/gkz486
- [63] Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Science Signaling*. 2010;**3**(107):ra8. DOI: 10.1126/scisignal.2000568
- [64] Somasundaram S et al. The DNMT1-associated lincRNA DACOR1 reprograms genome-wide DNA methylation in colon cancer 06 biological sciences 0604 genetics. *Clinical Epigenetics*. 2018;**10**(1):1-15. DOI: 10.1186/S13148-018-0555-3/FIGURES/10
- [65] Di Ruscio A et al. DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature*. 2013;**503**(7476):371-376. DOI: 10.1038/nature12598
- [66] Merry CR et al. DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer. 2015. doi: 10.1093/hmg/ddv343
- [67] Jansson-Fritzberg LI et al. DNMT1 inhibition by pUG-fold quadruplex RNA. *RNA*. 2023;**29**(3):346-360. DOI: 10.1261/RNA.079479.122/-/DC1
- [68] Haemmig S et al. Long noncoding RNA SNHG12 integrates a DNA-PK-mediated DNA damage response and vascular senescence. *Science Translational Medicine*. 2020;**12**(531):eaaw1868. DOI: 10.1126/scitranslmed.aaw1868
- [69] Unfried JP et al. Long noncoding RNA NIHCOLE promotes ligation efficiency of DNA double-Strand breaks in hepatocellular carcinoma. *Cancer Research*. 2021;**81**(19):4910-4925. DOI: 10.1158/0008-5472.CAN-21-0463
- [70] Xue Z et al. A G-rich motif in the lincRNA Braveheart interacts with a zinc-finger transcription factor to specify the cardiovascular lineage. *Molecular Cell*. 2016;**64**(1):37-50. DOI: 10.1016/j.molcel.2016.08.010
- [71] Childs-Disney JL, Yang X, Gibaut QMR, Tong Y, Batey RT, Disney MD. Targeting RNA structures with small molecules. *Nature Reviews. Drug Discovery*. 2022;**21**(10):736-762. DOI: 10.1038/s41573-022-00521-4
- [72] Aguilera A, García-Muse T. R Loops: From transcription Byproducts to threats to genome stability. *Molecular Cell*. Apr 2012;**46**(2)115-124. DOI: 10.1016/j.molcel.2012.04.009
- [73] Wahba L, Gore SK, Koshland D. The homologous recombination machinery modulates the formation of RNA-DNA hybrids and associated chromosome instability. *eLife*. Jun 2013;**2**:e00505. DOI: 10.7554/eLife.00505
- [74] Boque-Sastre R et al. Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;**112**(18):5785-5790. DOI: 10.1073/pnas.1421197112
- [75] Sanz LA et al. Prevalent, dynamic, and conserved R-loop structures associate with specific Epigenomic signatures in mammals. *Molecular Cell*.

- 2016;**63**(1):167-178. DOI: 10.1016/j.molcel.2016.05.032
- [76] Grote P et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Developmental Cell*. 2013;**24**(2):206-214. DOI: 10.1016/j.devcel.2012.12.012
- [77] Mondal T et al. MEG3 long noncoding RNA regulates the TGF- β pathway genes through formation of RNA-DNA triplex structures. *Nature Communications*. 2015;**6**:7743. DOI: 10.1038/ncomms8743
- [78] Blank-Giwojna A, Postepska-Igielska A, Grummt I. lncRNA KHPS1 activates a poised enhancer by triplex-dependent recruitment of Epigenomic regulators. *Cell Reports*. 2019;**26**(11):2904-2915.e4. DOI: 10.1016/j.celrep.2019.02.059
- [79] Zhang G et al. Comprehensive analysis of long noncoding RNA (lncRNA)-chromatin interactions reveals lncRNA functions dependent on binding diverse regulatory elements. *The Journal of Biological Chemistry*. 2019;**294**(43):15613-15622. DOI: 10.1074/jbc.RA119.008732
- [80] Sentürk Cetin N, Kuo C-C, Ribarska T, Li R, Costa IG, Grummt I. Isolation and genome-wide characterization of cellular DNA:RNA triplex structures. *Nucleic Acids Research*. 2019;**47**(5):2306-2321. DOI: 10.1093/nar/gky1305
- [81] Leisegang MS, Bains JK, Seredinski S, et al. HIF1 α -AS1 is a DNA:DNA:RNA triplex-forming lncRNA interacting with the HUSH complex. *Nature Communications*. 2022;**13**:6563. DOI: 10.1038/s41467-022-34252-2
- [82] Lin Y, Li C, Xiong W, Fan L, Pan H, Li Y. ARSD, a novel ER α downstream target gene, inhibits proliferation and migration of breast cancer cells via activating hippo/YAP pathway. *Cell Death & Disease*. 2021;**12**(11):1042. DOI: 10.1038/s41419-021-04338-8
- [83] Castanotto D et al. A stress-induced response complex (SIRC) shuttles miRNAs, siRNAs, and oligonucleotides to the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*. 2018;**115**(25):E5756-E5765. DOI: 10.1073/pnas.1721346115
- [84] La Rocca G, Cavalieri V. Roles of the Core components of the mammalian miRISC in chromatin biology. *Genes (Basel)*. 2022;**13**(3):414. DOI: 10.3390/genes13030414
- [85] Hansen TB et al. Natural RNA circles function as efficient microRNA sponges. *Nature*. 2013;**495**(7441):384-388. DOI: 10.1038/nature11993
- [86] Thomson DW, Dinger ME. Endogenous microRNA sponges: Evidence and controversy. *Nature Reviews. Genetics*. 2016;**17**(5):272-283. DOI: 10.1038/nrg.2016.20
- [87] Broderick JA, Zamore PD. Competitive endogenous RNAs cannot alter microRNA function in vivo. *Molecular Cell*. 2014;**54**(5):711-713. DOI: 10.1016/j.molcel.2014.05.023
- [88] Bosson AD, Zamudio JR, Sharp PA. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Molecular Cell*. 2014;**56**(3):347-359. DOI: 10.1016/j.molcel.2014.09.018
- [89] Denzler R, Agarwal V, Stefano J, Bartel DP, Stoffel M. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Molecular Cell*.

2014;**54**(5):766-776. DOI: 10.1016/j.molcel.2014.03.045

[90] Denzler R, McGeary SE, Title AC, Agarwal V, Bartel DP, Stoffel M. Impact of MicroRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Molecular Cell*. 2016;**64**(3):565-579. DOI: 10.1016/j.molcel.2016.09.027

[91] Jens M, Rajewsky N. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nature Reviews. Genetics*. 2015;**16**(2):113-126. DOI: 10.1038/nrg3853

[92] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature*. 2012;**482**(7385):339-346. DOI: 10.1038/nature10887

[93] Johnson R, Guigó R. The RIDL hypothesis: Transposable elements as functional domains of long noncoding RNAs. *RNA*. 2014;**20**(7):959-976. DOI: 10.1261/rna.044560.114

[94] Kapusta A et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics*. 2013;**9**(4):e1003470. DOI: 10.1371/journal.pgen.1003470

[95] Chillón I, Pyle AM. Inverted repeat Alu elements in the human lincRNA-p21 adopt a conserved secondary structure that regulates RNA function. *Nucleic Acids Research*. 2016;**44**(19):9462-9471. DOI: 10.1093/nar/gkw599

[96] Nguyen TM et al. The SINEB1 element in the long non-coding RNA Malat1 is necessary for TDP-43 proteostasis. *Nucleic Acids Research*. 2020;**48**(5):2621-2642. DOI: 10.1093/nar/gkz1176

[97] Yamazaki T et al. Functional domains of NEAT1 architectural lincRNA induce Paraspeckle assembly through phase separation. *Molecular Cell*. 2018;**70**(6):1038-1053.e7. DOI: 10.1016/j.molcel.2018.05.019

[98] Schein A, Zucchelli S, Kauppinen S, Gustincich S, Carninci P. Identification of antisense long noncoding RNAs that function as SINEUPs in human cells. *Scientific Reports*. 2016;**6**:33605. DOI: 10.1038/srep33605

[99] Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*. 2011;**470**(7333):284-288. DOI: 10.1038/nature09701

[100] Fasolo F et al. The RNA-binding protein ILF3 binds to transposable element sequences in SINEUP lincRNAs. *The FASEB Journal*. 2019;**33**(12):13572-13589. DOI: 10.1096/fj.201901618RR

[101] Kim YK, Furic L, Parisien M, Major F, DesGroseillers L, Maquat LE. Stauf1 regulates diverse classes of mammalian transcripts. *The EMBO Journal*. 2007;**26**(11):2670-2681. DOI: 10.1038/sj.emboj.7601712

[102] Kretz M et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*. 2013;**493**(7431):231-235. DOI: 10.1038/nature11661

[103] Hacısuleyman E, Shukla CJ, Weiner CL, Rinn JL. Function and evolution of local repeats in the Firre locus. *Nature Communications*. 2016;**7**:11021. DOI: 10.1038/ncomms11021

[104] Yap K, Mukhina S, Zhang G, Tan JSC, Ong HS, Makeyev EV. A short tandem repeat-enriched RNA assembles a nuclear compartment to control

alternative splicing and promote cell survival. *Molecular Cell*. 2018;**72**(3):525-540.e13. DOI: 10.1016/j.molcel.2018.08.041

[105] Kirk JM et al. Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*. 2018;**50**(10):1474-1482. DOI: 10.1038/s41588-018-0207-8

[106] Elguindy MM, Mendell JT. NORAD-induced Pumilio phase separation is required for genome stability. *Nature*. 2021;**595**(7866):303-308. DOI: 10.1038/s41586-021-03633-w

[107] Cabili MN et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biology*. 2015;**16**:20. DOI: 10.1186/s13059-015-0586-4

[108] Johnsson P et al. Transcriptional kinetics and molecular functions of long noncoding RNAs. *Nature Genetics*. 2022;**54**(3):306-317. DOI: 10.1038/s41588-022-01014-1