Role of lncRNAs in health and disease—size and shape matter

Vakul Mohanty, Yesim Gökmen-Polar, Sunil Badve and Sarath Chandra Janga

Advance Access publication date 11 September 2014

Abstract

Most of the mammalian genome including a large fraction of the non-protein coding transcripts has been shown to be transcribed. Studies related to these non-coding RNA molecules have predominantly focused on smaller molecules like microRNAs. In contrast, long non-coding RNAs (lncRNAs) have long been considered to be transcriptional noise. Accumulating evidence suggests that lncRNAs are involved in key cellular and developmental processes. Several critical questions regarding functions and properties of lncRNAs and their circular forms remain to be answered. Increasing evidence from high-throughput sequencing screens also suggests the involvement of lncRNAs in diseases such as cancer, although the underlying mechanisms still need to be elucidated. Here, we discuss the current state of research in the field of lncRNAs, questions that need to be addressed in light of recent genome-wide studies documenting the landscape of lncRNAs, their functional roles and involvement in diseases. We posit that with the availability of high-throughput data sets it is not only possible to improve methods for predicting lncRNAs but will also facilitate our ability to elucidate their functions and phenotypes by using integrative approaches.

Keywords: gene regulation; genome organization; post-transcriptional control; non-coding RNA; transcriptional noise; eukaryotes

INTRODUCTION

Extensive analysis of mammalian transcriptomes using high-throughput RNA sequencing techniques have shown that most of the mammalian genome is transcribed, usually from both strands [1, 2]. This has brought to light a huge disparity between the number of transcripts and protein coding genes encoded in these genomes. It is now known that a large fraction of the transcriptome consists of sequences that have no protein coding abilities, i.e. non-coding RNA (ncRNA) [3–5]. Several small ncRNAs, like microRNA (miRNA), have been studied extensively over the past two decades. However, the class of long non-coding RNA
(lncRNA) commonly defined as ncRNAs with length >200 nt have not been studied in great depth. It is only recently that they are drawing attention of the scientific community owing to their abundance [1, 6], association with myriad of disease phenotypes [7–11] and diverse cellular roles [12, 13]. Although lncRNAs are important in several cellular functions, knowledge about the actual number of functional lncRNA and their mechanisms of action is still limited.

LncRNAs were first identified using large-scale sequencing of full-length cDNA libraries in mouse. In genomic terms, lncRNAs range in size from ~100 to 100 000 bases, can be spliced or unspliced, polyadenylated or non-polyadenylated, nuclear or cytoplasmic, and are usually transcribed by RNA polymerase II and/or III [4, 14]. They lack easily recognizable traits such as sequence or functional features compared with well-studied classes of small RNA such as miRNAs and tRNAs. The inability to distinguish lncRNAs from mRNAs and lack of a firm understanding of their functional mechanisms makes it difficult to identify functional lncRNAs and their targets computationally and experimentally. Currently, very few high-throughput methods are available to dissect the precise functional roles of lncRNAs [15, 16]; most lncRNAs have been studied on a case-to-case basis. It is now clear that lncRNAs play an important role in a number of cellular processes including regulation of gene expression, X-chromosome silencing and chromosomal organization [17–20]. Further, lncRNAs were found to be dysregulated in a number of different kinds of cancers and neurological diseases, although their precise role and contribution towards these conditions remains unclear [7–11]. Another perplexing finding is their contextual ‘dual’ nature, i.e. act as a functional lncRNA or encode a protein [21, 22]. Though recent evidence from GENCODE [4, 5] suggests that this phenomenon may not be widespread and, hence, the duality may be infrequent and limited to a few cellular transcripts [22, 23]. Nevertheless, it does add an additional layer of complexity to the transcriptome and should be a subject of future studies.

In this review, we discuss our current understanding of the genomic organization of lncRNAs, i.e. their genomic occurrence, properties and methods for identifying lncRNAs, followed by a discussion of their functional roles, possible mechanisms of action and impact in causing disease phenotypes in light of the recent large- and small-scale studies.

**GENOMIC ORGANIZATION AND PROPERTIES OF lncRNAs**

LncRNAs are found scattered throughout the genome and have been classified based on their genomic location, or more precisely, based on their proximity to protein coding genes [4, 24]. Though this approach is unable to classify several lncRNAs into any of the defined categories, it is still widely accepted. The classification and description of different categories is illustrated in Figure 1. In addition to these classes, circular forms of ncRNAs (circRNAs) representing a diverse class of endogenous RNAs have been detected in various tissues [25]. Although their existence was reported for few genes in earlier studies (i.e. DCC, ETS-1, SRY, MLL, cytochrome P-450 2C18 and dystrophin), they have been primarily considered as rare RNA types, transcriptional noise or RT-PCR artefacts [26–31]. They are usually non-polyadenylated and are expressed at much lower levels than the normal transcript. Recent deep sequencing of RNA using RNA-Seq...
methodologies has confirmed the existence of circular isoforms in hundreds of human loci in both normal and malignant cells [25, 32–34]. Although circRNAs exhibit a circular structure in common, they come from diverse RNA types including Coding Sequence Exons, antisense, lncRNA and intergenic regions [25, 35, 36]. These circRNAs are highly stable and unlike the linear RNAs, are resistant to RNase R degradation. Furthermore, the order of exons in circRNA can be different from genomic DNA, suggesting that they are created via a non-canonical RNA splicing [25, 32, 34]. Circular RNAs have been shown to be evolutionarily conserved and their splicing patterns for each gene were found to be regulated in a cell-specific manner [32, 33].

The GENCODE consortium within the ENCODE project has reported and in some cases reconfirmed several intriguing properties of lncRNAs [4, 37]. Consistent with prior studies [15, 38], this consortium has shown that lncRNAs are under weak selective constrains and have a level of conservation which is lower than that of protein coding genes but is higher than random sequences [4]. lncRNA also tend to show lower levels of expression when compared with protein coding genes and frequently exhibit tissue-specific and developmental-specific expression [4, 39–43]. This has been argued as an evidence for lncRNAs to unlikely represent transcriptional noise [4, 44]. Additionally, original studies showed that lncRNAs are enriched in the nuclear compartment (rather than cytoplasm), supporting their role in epigenetic mechanisms and modulation of chromatin modification machinery [4] although this is still under debate with a recent study questioning these trends [45].

At the transcriptional level, lncRNAs produce independent transcriptional units and show a strong positive correlation of expression with their neighbouring genes, suggesting a positive coupling with genomically adjacent protein coding regions [4]. Most lncRNAs are composed of at most two exons, so splicing, although present, is less profound when compared with that of mRNA [46]. Splicing of lncRNAs was found to be predominantly co-transcriptional, as is the case with mRNAs [46]. In contrast, recent global studies report that the epigenetic markers of lncRNAs are different from those of mRNAs [47]. In particular, the methylation density of lncRNAs, irrespective of the expression levels is high around their Transcription Start Site (TSS) unlike that of a repressed gene whose methylation density is high in the downstream of its TSS. However, the presence of H3K4me3 and H3K36me3 marks a direct correlation in the activity of the lncRNA, suggesting that these modifications play a key role in their activation. It was also shown that H3K9me3 does not dictate repression in case of lncRNAs unlike H3K27me3 [47].

**METHODS TO PREDICT lncRNAs**

The lack of a definitive set of criteria to distinguish lncRNAs from mRNA has been a significant drawback for confidently predicting lncRNAs from mRNA has been a significant drawback for confidently predicting lncRNAs [1, 4, 48]. Early prediction methods labeled those sequences that exhibited no homology with known protein coding and annotated sequences as a potential lncRNA (Tables 1 and 2 summarize some of the methods and databases for lncRNAs, respectively). The accuracy of such techniques depends on the accuracy of the present annotation of protein coding genes, which often makes them error prone [59].

Another strategy is based on the hypothesis that Open Reading Frame (ORF) of coding RNA sequences is longer than that of the non-coding RNA (ncRNA). To reduce the chances of error, a minimum ORF length was set as a cut-off, above which the RNA sequence is classified as an mRNA. However, there is no consensus as to what this cut-off should be. The FANTOM consortium [1] set its cut-off at 300 nt while the H-invitational consortium [48] set it at 60 nt, which led to a large disparity in the number of lncRNAs estimated. The higher threshold for ORF length in FANTOM project yields large lncRNAs like Xist to be misclassified as mRNA. On the other hand, a lower cut-off in the H consortium misclassifies short peptide coding RNAs as lncRNAs. To overcome such problems, methods that incorporate the conservation of ORF have been employed. These were based on the assumption that coding sequences have greater evolutionary restrictions than lncRNAs. For example, CRITICA [50] and CSTminer [49] make use of this conservation to identify mRNAs and thus by extension, lncRNAs. CRITICA identifies coding sequences by making use of evolutionary conservation and dicodon bias that coding sequences exhibit. After identifying local sequences similar to the query sequence in a DNA database using BLASTN, CRITICA breaks the alignments into triplets and the amino acid encoded by each triplet is
determined. Synonymous changes receive a positive score, while non-synonymous changes incur a negative score. The coding evidence due to the alignment is defined as the logarithm of the probability of finding the given sequence of triplets in a coding frame divided by the probability of finding the sequence in a non-coding frame. Additionally, a statistical score is generated based on dicodon frequencies that are computed in an iterative manner. As these methods do not depend on the accuracy of present annotation of protein coding genes, they are more likely to accurately identify novel lncRNAs. Nevertheless, a possible caveat of these methods could be the fact that some lncRNAs may be evolutionary remnants of protein coding genes [15], and might exhibit persistence of conservation inherent in the parent gene.

The approaches described above are based on the negative characteristics, i.e. identification of features present in mRNAs and presumed to be absent in lncRNAs. Structural methods, on the other hand, look directly at the structural properties under the premise that functional lncRNAs must have an underlying structural similarity even if there is large sequence divergence. Knudsen and Hein [60] developed a method that makes use of phylogenetic information in addition to stochastic context-free grammar (Phylo-SCFGs) to predict secondary structures. Evofold [51] takes multiple sequence alignment and a phylogenetic tree as input and generates a secondary structure and folding potential score. EvoFold uses two phylo-SCFGs, one each from the modelling regions of functional RNAs and non-functional RNAs, to calculate the log-likelihood ratio. Using EvoFold, the authors were able to predict miRNAs with great accuracy. As EvoFold is designed to look for functional RNAs that are conserved in structure and remain in the same genomic context in all vertebrates, it fails to identify fast evolving sequences. Because lncRNAs show a degree of conservation determined. Synonymous changes receive a positive score, while non-synonymous changes incur a negative score. The coding evidence due to the alignment is defined as the logarithm of the probability of finding the given sequence of triplets in a coding frame divided by the probability of finding the sequence in a non-coding frame. Additionally, a statistical score is generated based on dicodon frequencies that are computed in an iterative manner. As these methods do not depend on the accuracy of present annotation of protein coding genes, they are more likely to accurately identify novel lncRNAs. Nevertheless, a possible caveat of these methods could be the fact that some lncRNAs may be evolutionary remnants of protein coding genes [15], and might exhibit persistence of conservation inherent in the parent gene.

The approaches described above are based on the negative characteristics, i.e. identification of features present in mRNAs and presumed to be absent in lncRNAs. Structural methods, on the other hand, look directly at the structural properties under the premise that functional lncRNAs must have an underlying structural similarity even if there is large sequence divergence. Knudsen and Hein [60] developed a method that makes use of phylogenetic information in addition to stochastic context-free grammar (Phylo-SCFGs) to predict secondary structures. Evofold [51] takes multiple sequence alignment and a phylogenetic tree as input and generates a secondary structure and folding potential score. EvoFold uses two phylo-SCFGs, one each from the modelling regions of functional RNAs and non-functional RNAs, to calculate the log-likelihood ratio. Using EvoFold, the authors were able to predict miRNAs with great accuracy. As EvoFold is designed to look for functional RNAs that are conserved in structure and remain in the same genomic context in all vertebrates, it fails to identify fast evolving sequences. Because lncRNAs show a degree of conservation
that is lesser than snoRNAs [38], it is likely that EvoFold would be ineffective in predicting functional lncRNAs. RNAz [53], another structure-based method, measures conservation and thermodynamic stability of a RNA's secondary structure. RNAz makes use of RNAFOLD and RNAALIFOLD for folding a single sequence and a consensus folding of aligned sequences, respectively. A minimum free energy values is calculated for 1000 synthetic sequences (length 50–400 nt) and a Support Vector Machine (SVM) is trained on them to estimate z-scores, which are then used as a binary classifier. Secondary structure-based methods are hampered by the accuracy of the secondary structure prediction methods, which typically do not take non-canonical interactions (such as pseudoknots) into consideration [61, 62]. Although tertiary structures of RNA contain more information pertaining to the functions of a molecule, there is a dearth of solved 3D structures and a lack of computational algorithms that can accurately model these. With improvement in tertiary structure prediction algorithms, it may become possible to use tertiary structures of RNAs and ascertain functionality. Nevertheless, since both EvoFold and RNAz were primarily designed to find relatively short and highly structured RNAs, they are unlikely to be effective in predicting functional lncRNAs especially as there is no compelling evidence indicating that lncRNAs are highly structured.

Yet another group of methods for identification of lncRNAs make use of machine learning algorithms like neural networks (DIANA-EST) [54] and SVMs (CPC) [56]. DIANA-EST [57] makes use of artificial neural networks to access the coding potential of sequences. Recent methods use SVMs, which are supervised learning algorithms that analyse data and identify patterns, to classify RNAs. For example, CONC, a SVM-based method, is based on a set of 180 features including protein properties of potential peptides like peptide length and amino acid compositions. This method uses known data from non-redundant sets of cDNAs (n = 5610) and lncRNAs constructed from NONCODE [58] and RNAdb [63] to train the algorithm. CPC, another SVM-based method, used just six features to achieve equivalent or better results at a faster rate [55]. The features assess the extent and quality of ORFs as well as the number and quality of hits and number of hits in a single frame within BLASTX. The rationale here is that coding RNAs will have better defined ORFs and have better quality hits and more hits in a single reading frame in BLAST analysis. Though methods based on machine learning algorithms are powerful and not heavily dependent on a single feature, they are only identifying coding potentials of RNAs and not functional lncRNAs. The increasing number of functional lncRNAs identified by high-throughput experimental methods [53, 64] will provide a rich data set of functional RNAs, which can be used to train more advanced and powerful machine learning algorithms. However, these methods will likely rely on the idea that lncRNAs will have common characteristics and properties, but there is no a priori reason to assume that all lncRNAs can be grouped into such defined list of distinct lncRNA classes based on their properties. Nevertheless, improvements in machine learning techniques and discovery of novel lncRNA classes and their features may help improve the accuracy with which lncRNAs and their functional subtypes can be predicted.

**FUNCTIONAL ROLES OF lncRNAs AND FUNCTION PREDICTION**

LncRNAs lack sequence conservation even in closely related species [4, 65] and thus were thought to be transcriptional noise [22]. Recent studies show that the lack of conservation does not imply lack of function [65, 66]. Furthermore, functional lncRNAs show patches of conserved regions [66]. Zhou et al postulate that the level of selective pressure on a sequence is directly proportional to the number of interactions it needs to maintain. Thus, if lncRNAs are involved in fewer interactions, the evolutionary constraint to maintain their conservation would be lower. Recent studies have shown that a large number of lncRNAs are conserved in mammals [15] and that they are expressed in a tissue-specific manner [39–43] with characteristic subcellular localization [12, 67–69]. LncRNAs have now been documented to be involved in a diverse set of specific functional roles via a plethora of mechanisms (Figure 2) as detailed below [12, 24, 70, 71] and it has been observed that they exhibit varying levels of conservation both within and across types of mechanisms discussed.

**Regulation of transcription**

LncRNAs play an important role in regulation of transcription via a diverse set of mechanisms [72]. For instance, Naturally Anti-sense Transcripts
(NATs) and long intergenic non-coding RNAs (lincRNAs), which are derived from genomic locations in close proximity to protein coding regions, are known to regulate them. One of the possible mechanisms is via collision of transcriptional machinery [73] (Figure 2C). When a particular genomic region has transcripts on both strands (like in case of NATs), then RNA pol II carrying out transcription on each of these strands collide, leading to repression. In such a mechanism, the extent and pattern of overlap of the NAT with the encoded transcript in the opposite strand dictates the frequency of collision [73]. This mechanism has also been shown to cause transcriptional interference even when transcripts are encoded on the same strand (e.g. yeast’s SER3 gene [74, 75]).

Some lincRNAs regulate the transcription by directly interacting with the DNA. For example, the major promoter of DHFR initiates the transcription of gene coding for dihydrofolate reductase, whereas the minor promoter initiates the transcription of an lincRNA that represses the expression of dihydrofolate reductase. This repression is mediated by the interaction of the lincRNA with the major promoter and the transcription factor IIB, leading to the dissociation of the preinitiator complex at the major promoter [76]. LincRNAs can also control transcription without directly binding to the DNA. In response to DNA damage, several lincRNAs upstream of CCND1 are transcribed and cause the repression of this gene by binding to the RNA-binding protein, TLS (for translocated in liposarcoma) [77]. Together they modulate the Histone Acetyl Transferase activity of transcription cofactors CBP and p300, leading to repression of CCDN1. Additionally, lincRNAs also control the localization of transcription factors [17, 78]. For instance, lincRNA NORN has been shown to bind nuclear transporters like KPNB1, CSE1L and CUL4B and prevent nuclear localization of NFAT [17, 78] (Figure 2D). Similarly, lincRNA linc-p21, which is regulated by p53 and induced in response to DNA damage, represses the expression of p53-regulated genes by binding to and modulating the localization of hnRNP-K, leading to apoptosis [77, 79].

Another example of an lincRNA that co-ordinates transcription is PANDA, which is activated by p53 in response to DNA damage [80]. On activation, PANDA represses the expression of pro-apoptotic genes and enables cell cycle arrest in the G1-S phase, allowing more time for the DNA repair mechanism to act on damaged DNA. It has also

Figure 2: Various mechanisms by which lincRNAs can regulate the expression of genes are shown. (A) lincRNAs can recruit chromatin remodelling complexes to the chromosome. (B) lincRNAs can bind to proteins and act as decoys preventing them from functioning. (C) lincRNAs are also involved in transcriptional interference mechanisms to control the expression of protein coding regions. (D) lincRNAs act as scaffolds and aid the localization of proteins responsible for gene expression (E) lincRNAs can also bind to miRNA, resulting in their titration. This increases the expression of the mRNA, which is the actual target of the miRNA. (A colour version of this figure is available online at: http://bfg.oxfordjournals.org)
been found that many tumour suppressor genes have anti-sense RNA close to them [81]. For instance, an increase in expression of p15AS was found to correspond to a decrease in expression of p15, with p15AS acting via chromatin remodelling in a Dicer-independent manner [81], suggesting the extent to which lncRNAs may be involved in cell cycle control and cancer.

**Role in epigenetics**

The role of lncRNAs in X chromosome inactivation and epigenetic imprinting has been well studied (Figure 2B). X chromosome’s inactivation is initiated in the early stages of development by the expression of Xist [19]. RepA, a lncRNA from Xist locus, after binding to the Polycomb complex (PRC2 -Ezh2) locates to the X chromosome. PRC2 trimethylates lysine 23 on H3 of the X chromosome, leading to chromatin condensation and repression of expression [82]. Simultaneously, Tsix, expressed at the active X chromosome, prevents RepA from binding [83, 84]. Though well studied, there are several aspects of X chromosome silencing that are yet to be explained, like how some of the genes on the inactive X chromosome escape silencing [85, 86] and what causes the localized expression of Tsix and Xist [87, 88].

Imprinting is the phenomenon where for a subset of genes, only one copy is expressed in a parent-of-origin–dependent manner. Kcnq1ot1 and Air are two lncRNAs, both of which are repressed in the maternal chromosome and carry out imprinting only in the paternal chromosome [88]. Kcnq1ot1 acts by recruiting histone methyltransferases like G9a and PRC2 for chromatin remodelling [89]. Kcnq1ot1-induced imprinting is more extensive in the placental cells than in fetal liver cells and documents the lineage-specific imprinting [90]. A recent study has shown that Kcnq1ot1 interacts with Dnmt1 and regulates imprinting of genes close to Kcnq1ot1 by the DNA methylation of somatic Differentially Methylated Regions [90]. Air silences Slc22a3 by recruiting G9a, in a mechanism similar to Kcnq1ot1 [91]. G9a is, however, not necessary for Air-mediated silencing Igfr2, suggesting a different mechanism [91].

LncRNAs are also critical for chromosomal organization and are likely to be responsible for guiding chromosomal factors to critical genomic loci that are important to the maintenance of pluripotency. Twenty-six lincRNAs have been shown to be vital for maintenance of pluripotency [15, 92]. Environmental stimuli can alter the functionality of lncRNAs and maintenance of pluripotency [93].

**Titration of miRNAs and acting as decoys**

lncRNA can bind to miRNA (Figure 2E) and modulate their actions [18]. This interaction between miRNA and lncRNA suggests a very complex organization of the post-transcriptional gene regulatory network, with the latter controlling the availability of functional miRNA [64]. CircularANRIL (cANRIL) expression influences Polycomb group (PcG)-mediated repression of the INK4a/ARF locus to affect atherosclerosis risk [35]. CircRNAs via their miRNA docking capabilities may also function as post-transcriptional regulators of gene expression in specific tissues and during specific developmental phases. Human circRNA CDR1 is an abundant, largely cytoplasmic RNA that harbours 70 target sites of miR-7 and acts as a miR-7 super-sponge (Circular RNA Sponge for miR-7; CiRS-7) [25, 36].

LncRNAs can more generally act as molecular decoys (Figure 2D) and can either bind to regulatory molecules like miRNAs, TFs and splicing factors or to the enhancer and promoter regions and prevent the action of regulatory molecules. For instance, during DNA damage p53 activates PANDA, a lncRNA, which sequesters NF-YA, a nuclear transcription factor NF-YA (Figure 2B) and prevents apoptosis [94]. TERRA, another lncRNA, is believed to act as a tumour suppressor by binding to TERT (telomerase reverse transcriptase), which is responsible for the regulation of telomerase extension [94]. These examples show how lncRNAs can act as negative regulators by repressing the expression of genes.

LncRNAs can also act as positive regulators of gene expression by functioning as miRNA decoys (Figure 2D). PTENP1 functions as a miRNA decoy sequestering miRNAs (miR-17, 21 214 and 19 families) that target the mRNA PTEN, triggering proliferation [96]. This sequestration also results in altered levels of other targets of these miRNAs such as tumour suppressors E2F1, p21 and PDCD4. Similarly, the lncRNA HULC (highly upregulated in liver cancer) forms an autoregulatory loop in liver cancer, which leads to repression of miR-372 and its own upregulation [97].
In addition, lncRNAs act as guides for chromatin remodelling complexes including polycomb complexes [20, 90, 91] (Figure 2A). Khalil et al. [15] demonstrated that a large number of lncRNAs are associated with chromatin modifying complexes. A RIP-CHIP assay was then used to identify lncRNAs associated to polycomb complexes. The RIP-CHIP together with expression data enabled the identification of high confidence set of lncRNAs that interact with Polycomb, coREST and SMCX but not with modified histones H3K4me2 and H3K4me3 [15]. This data set was used to train a SVM classifier that detected 59.4% of polycomb binding lncRNAs in mice [98]. Computational methods like this can provide vital information on how lncRNAs function.

Also, lncRNAs can act as platforms on which different biological components can assemble. For example, TERC plays a major role in binding and catalytic activity of TERT as well as in maintaining the stability of the complex [99]. Mutations in TERC can affect its stability resulting in disease states such as dyskeratosis [100].

**Computational methods to predict the function of lncRNAs**

Experimental studies have made great strides over the past decade identifying several functional lncRNAs, their mechanisms and target genes. However, the diversity of mechanisms by which lncRNAs function makes it difficult to design a single all-encompassing experimental protocol that can identify all lncRNAs and their functional mechanisms. Their tissue-specific and low expression levels also hamper interpretation of the experimental results. In light of these limitations, it is important to develop computational methods that can predict functions of lncRNAs. Few methods have been developed for the functional annotation of lncRNAs [59, 101]. These methods predict functions of lncRNAs based on functional annotations of their associated mRNAs. The premise herein is that transcripts (lncRNAs and mRNAs) that are co-expressed have a high chance of being involved in the same functions [101]. Integrating more biologically relevant information to co-expression data can greatly improve the quality of the predictions. Xingli *et al.* [59] have integrated co-expression data with protein interaction data in a naïve Bayesian fashion to construct a bicouloir network of coding non-coding associations and used a global propagation algorithm to predict functions of lncRNAs.

Experimental studies are providing an ever-increasing data on lncRNAs including their functional domains and modes of interactions. Integration of information from different conceptual platforms will result in development of better machine learning algorithms that predict functions based on features in sequence and secondary structure of lncRNAs.

**ROLE OF lncRNAs IN DISEASE**

lncRNAs play a critical role in several diseases (Table 3) including cancers and neurological disorders [7, 8, 10]. Several lncRNAs are essential for maintaining and attaining the cancer phenotype [8, 150] including repressing the expression of tumour suppressor genes. Overexpression of a set of 5 lncRNAs in a human melanoma cell line results in suppression of PSF, a tumour suppressor and overexpression of GASE6, a proto-oncogene, and development of tumorigenic phenotype [95]. ANRIL recruits polycomb complexes to repress the expression of p15 [102], p16 and p14 [151]. On the other hand, downregulation of GAS5, a tumour suppressor lncRNA, has been observed in rapidly growing leukaemia cells [152] and breast cancer [106]. The role of TERC and TERRA in affecting telomere function has already been discussed [134, 135, 153, 154]. lncRNAs also play an important role in metastasis. For example, MALAT1 promotes cell motility [122], proliferation and invasion [123] and is upregulated in several cancers [105, 123, 124]. Similarly, HOTAIR is upregulated in various cancers and is responsible for metastasis in breast and epithelial cancers [112, 113, 155]. On the other hand, upregulation of treRNA, a translational regulatory lncRNA, stimulates tumour invasion *in vitro* and metastasis *in vivo* [138]. A recent study of 64 archival tumours has demonstrated the differential expression of hundreds of lncRNAs across cancers [9]. It is evident that lncRNA are differentially expressed in cancers and can act as potential markers for identifying cancers.

lncRNAs have also been implicated in several diseases of the nervous system [10]. DiGeorge Syndrome is a neurodevelopmental disorder, caused by the deletion of the 2q11.2 chromosomal region, which encodes for a REST-regulated lncRNA DGCR5 [144]. Similarly, dysregulation of NORN, a lncRNA responsible for the cytoplasmic-nuclear shuttling of the NFAT, is associated with several
features of NORN and Down’s syndrome [78]. These data suggest that the IncRNA play a role in brain development. BACE1-AS, a lncRNA, is involved in the regulation of BACE1, an enzyme responsible for plaque formation in Alzheimer’s disease [146]. Similarly, the expression levels of BC200 correlate with the severity of the Alzheimer’s disease. In advanced cases, BC200 is no longer localized to the synaptic junction but clustered in the cyton [142]. Several lncRNAs have been found to be dysregulated in Huntington’s disease [46]; Qureshi et al. [10] provide a more detailed discussion of other neurological disorders associated with IncRNAs.

The importance of IncRNAs in congenital heart disease and adult cardiac disease is well described [156]. In particular, NATs that are transcribed from genes important in structure and development of heart may alter their regulation [35, 157].

Lastly, mutations in IncRNAs have also been linked to several diseases [35, 158–160]. ANRIL mutations have been linked to cutaneous malignant melanoma, neural system tumours and hereditary predisposition to cancer [103]. Similarly, a large number of SNPs within the ANRIL locus have been associated with type II diabetes [161]. Mutations in IncRNAs may cause subtle changes in the phenotype that are hard to define [162]. However, with increasing in the number of genomic studies investigating the link between SNPs, IncRNAs and diseases, the specific roles of different

<table>
<thead>
<tr>
<th>LncRNA</th>
<th>Disease</th>
<th>Function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANRIL</td>
<td>Leukaemia, prostate cancer</td>
<td>Epigenetic silencing; chromatin remodelling; titration of miRNAs</td>
<td>[81, 102, 103]</td>
</tr>
<tr>
<td>BC200</td>
<td>Multiple cancers</td>
<td>Protein binding, i.e.: SRP, translation inhibitor</td>
<td>[104]</td>
</tr>
<tr>
<td>CDR1as</td>
<td></td>
<td>Titration of miRNAs</td>
<td>[105]</td>
</tr>
<tr>
<td>GASS</td>
<td>Autoimmune disease, leukaemia, breast cancer</td>
<td>Decoy of glucocorticoid receptor/represses several steroid hormones</td>
<td>[106–108]</td>
</tr>
<tr>
<td>H19</td>
<td>Multiple cancers</td>
<td>Transcription regulation</td>
<td>[109–111]</td>
</tr>
<tr>
<td>HOTAIR</td>
<td>Cancer metastasis, breast cancer</td>
<td>Epigenetic silencing of genes by the recruitment of polycomb complexes; chromatin remodelling</td>
<td>[15, 112–115]</td>
</tr>
<tr>
<td>HULC</td>
<td>Liver cancer</td>
<td>MicroRNA decoy</td>
<td>[97, 116, 117]</td>
</tr>
<tr>
<td>Kcnq1ot1</td>
<td>Breast and colon cancer</td>
<td>Epigenetic regulation</td>
<td>[69, 118–120]</td>
</tr>
<tr>
<td>mLSINCT5</td>
<td>Breast and ovarian cancer</td>
<td></td>
<td>[121]</td>
</tr>
<tr>
<td>MALAT1</td>
<td>Cancer metastasis/ prostate, liver, pancreas, breast, lung, uterus, cervix cancers</td>
<td>Regulation transcription, RNA splicing; scaffold</td>
<td>[105, 122–124, 125–127]</td>
</tr>
<tr>
<td>MEG3</td>
<td>Brain cancer</td>
<td>N/A</td>
<td>[128, 129]</td>
</tr>
<tr>
<td>PCAT1</td>
<td>Prostate cancer</td>
<td>Regulation of transcription</td>
<td>[130]</td>
</tr>
<tr>
<td>PTENP1</td>
<td>Prostate cancer</td>
<td>miRNA decoy</td>
<td>[96]</td>
</tr>
<tr>
<td>RNY1</td>
<td>cancer</td>
<td>RNA splicing</td>
<td>[131–133]</td>
</tr>
<tr>
<td>TERC</td>
<td>cancer</td>
<td>Acting as decoys</td>
<td>[134, 135]</td>
</tr>
<tr>
<td>TERRA</td>
<td>Cancer; cell division</td>
<td>Acting as decoys</td>
<td>[136, 137]</td>
</tr>
<tr>
<td>tRNA</td>
<td>Breast cancer metastasis</td>
<td></td>
<td>[138]</td>
</tr>
<tr>
<td>TUGI</td>
<td>Bladder cancer</td>
<td></td>
<td>[139, 140]</td>
</tr>
<tr>
<td>UCAI</td>
<td>multiple carcinoma</td>
<td></td>
<td>[141]</td>
</tr>
<tr>
<td>BC200</td>
<td>Alzheimer’s disease</td>
<td>Protein binding</td>
<td>[142, 143]</td>
</tr>
<tr>
<td>DGCR5</td>
<td>DiGeorge Syndrome</td>
<td></td>
<td>[144]</td>
</tr>
<tr>
<td>DSCR1</td>
<td>Down’s syndrome</td>
<td></td>
<td>[145]</td>
</tr>
<tr>
<td>DYSKIA</td>
<td>Down’s syndrome</td>
<td></td>
<td>[145]</td>
</tr>
<tr>
<td>NORN</td>
<td>Down’s syndrome</td>
<td>Transcription regulation</td>
<td>[78]</td>
</tr>
<tr>
<td>BACE1-AS</td>
<td>Alzheimer’s disease</td>
<td>Increase mRNA stability</td>
<td>[146–148]</td>
</tr>
<tr>
<td>mHTT-AS</td>
<td>Huntington’s disease</td>
<td></td>
<td>[10, 46]</td>
</tr>
<tr>
<td>HARIR</td>
<td>Huntington’s disease</td>
<td></td>
<td>[10]</td>
</tr>
<tr>
<td>ALR</td>
<td>arteriosclerosis</td>
<td>Epigenetic silencing</td>
<td>[35]</td>
</tr>
<tr>
<td>ALCI-as</td>
<td>Congenital heart disease</td>
<td></td>
<td>[149]</td>
</tr>
</tbody>
</table>

**Table 3:** Summary of known IncRNAs commonly studied till date in various diseases, along with their functional roles.
IncRNAs in the causation of disease will become clear [35, 100, 159, 160].

CONCLUSIONS AND FUTURE DIRECTIONS

Recent studies have now convincingly shown that the eukaryotic transcriptome is enormous, and comprises not only a large set of protein-coding messenger RNAs but also large numbers of non-protein coding transcripts that have structural, regulatory or unknown functions. Increasing evidence suggests that a diverse set of functions and phenotypes can be attributed to this class of ncRNA with emerging studies pointing to a complex modular organization of their post-transcriptional regulatory code [64, 163]. For instance, with new kinds of ncRNAs like circular RNAs originating due to the trans-splicing events in the genome, it is possible to speculate that the ncRNA transcriptome is far from complete and is capable of a wider range of functions than is currently believed. It is also unclear if fusion transcripts commonly seen in various cancers have similar abilities to produce circular forms. Although the field of ncRNAs has come a long way with the advent of genomics, our understanding on the dynamic nature of their transcriptomes is just beginning to emerge. However, with the developments in high resolution sequencing technologies extended to the RNA pools, there has been a rapid increase in the number of high-quality genome-wide transcriptome maps in both healthy and disease states [2, 4, 9]. These maps, together with consortium projects (such as the NIH Roadmap Epigenomics consortium), are not only going to revolutionize our current understanding of IncRNAs genomic organization but are also going to deliver improvements in the annotation of their functions by using integrated computational approaches, identification of novel IncRNA genes, elucidation of condition-specific expression patterns and potential regulatory elements that contribute to their behaviour. These new technologies and approaches will prove to be invaluable means for furthering our understanding of the principles governing RNA-based regulatory mechanisms in a genome-wide manner. Given the decreasing cost and unprecedented detail at which these high-throughput technologies can reveal the regulatory elements and expression levels specific to conditions, it is possible to use these approaches in the coming years to interrogate the prevalence of these phenomena in different states and thereby study their relevance to regulation, physiology and disease to link the genotype of IncRNAs with their phenotypes.

Key points

- IncRNA can originate from diverse genomic loci.
- IncRNA are abundant and contribute to the regulation of diverse cellular processes in eukaryotic genomes.
- Computational methods for prediction of IncRNA types and their function are still in their infancy.
- IncRNAs contribute to global, often tissue-specific, regulation by diverse mechanisms.
- Dysregulation of IncRNAs have been observed in a wide range of diseases.

Acknowledgements

S.C.J. acknowledges support from the School of Informatics and Computing at IUPUI in the form of start-up funds. The authors would also like to thank Gabriel Moreno-Hagelseeb, Bobak Kechavarzi, Rajneesh Snivastava, Yasewimi Neelamraju, Tulip Nandu and Deepali Jhamb for critically reading a previous version of the manuscript and providing helpful comments. The authors report no financial or other conflict of interest relevant to the subject of this article.

FUNDING

This work was supported by Indiana University Release Time for Research (RTR) award and Collaborative Research Grant (CRG) to SCJ.

References


