doi:10.20944/preprints202006.0116.v1

Computational approaches to functionally annotate long noncoding RNA (lncRNA)

Yashpal Ramakrishnaiah[1,2], Levin Kuhlmann[1], and Sonika Tyagi[2,*]

[1]Faculty of Information Technology, Monash University, Clayton VIC 3800 Australia
[2]School of Biological Science, Monash University, Clayton VIC 3800 Australia
[*]Corresponding author: Sonika Tyagi, sonika.tyagi@monash.edu

June 2020

Abstract

Long noncoding RNA (lncRNA) are implicated in various genetic diseases and cancer, attributed to their critical role in gene regulation. RNA sequencing is used to capture their transcripts from certain cell types or conditions. For some studies, lncRNA interactions with other biomolecules have also been captured, which can give clues to their mechanisms of action. Complementary *in silico* methods have been proposed to predict non-coding nature of transcripts and to analyze available RNA interaction data. Here we provide a critical review of such methods and identify associated challenges. Broadly, these can be categorized as reference-based and reference-free or *ab initio*, with the former category of methods requiring a comprehensive annotated reference. The *ab initio* methods can make use of machine learning classifiers that are trained on features extracted from sequences, making them suitable to predict novel transcripts, especially in non-model species. Machine learning approaches such as Logistic Regression, Support Vector Machines, Random Forest, and Deep Learning are commonly used. Initial approaches relied on basic sequential features to train the model, whereas the use of secondary structural features appears to be a promising approach for functional annotation. However, adding secondary features will result in model complexities, thus demanding an algorithm that can handle it and furthermore, considerably increasing the utilization of computation resources. Computational strategies combining identification and functional annotation which can be easily customized are currently lacking. These can be of immense value to accelerate research in this class of RNAs

Key Words— noncoding RNA, lncRNA, Epigenomics, Gene regulation, Machine Learning, Bioinformatics

1 Introduction

1.1 lncRNA

Noncoding RNA (ncRNA) will not usually directly code for protein and were considered as *junk* regions in the DNA with little to no functional significance. Recently, they are starting to garner attention because of the realization that they play a vital role in genome regulation. Many of them are long noncoding RNA (lncRNA), which form 80% of the noncoding transcriptome [1, 2] and generally act by interacting with other biomolecules during transcriptional, post-transcriptional, translational or post-translational phases in various mechanisms to affect gene activity [3, 4, 5]. These were initially thought to be merely transcriptional noise or high throughput sequencing artefacts [6, 7]. A subset of lncRNA, known as macroRNA, were also reported to act as precursors to other small or long ncRNA [8, 9]. There are also a few examples of lncRNA containing open reading frame (ORF) and having dual action of gene regulation and synthesising protein [10]. Many lncRNA show cross-species conservation [11, 12, 13], and their unstable nature and high turnover rate in nucleus indicate their role in rapid response to external stimuli [14]. A large number of human lncRNA bind to the polycomb group protein repressive complex (PRC) and chromatin modifying complexes strongly suggesting their role in epigenetic regulation [4].

The use of lncRNA as one of the main regulatory players is seen as a cost-effective model. This is because compared to proteins, RNAs are energetically less expensive to produce in the cell and can act locally without having to be transported between nucleus and cytoplasm. More importantly, lncRNA are highly cell and tissue specific; and a specific epigenetic regulation in response to a stimulus can be achieved without changing the transcriptional machinery [15]. Interestingly, an increase in the number of protein coding genes from invertebrates to mammals has been smaller in comparison to drastic increase in the complexity. A large number of annotated lncRNA in mammals, with their diverse roles, may help us understand the complexity of mammalian systems.

The total number of lncRNAs in the human genome is still not accurately known, as novel lncRNAs are identified constantly and the reference databases are still getting updated. This number also varies wildly across these databases because of the differences in the definition of what constitutes lncRNA and what doesn't. Some of the conservatives ones like GENCODE v34 [16] lists 17,960 lncRNA whereas, as high as 270,000 transcripts are reported by LncBook [17]. Other database estimates lie somewhere in between, for example, FANTOM CAT project reported 27919 lncRNA [18], of which 19,175 were shown to be functionally implicated [18]. Although the information on lncRNA is most comprehensive for humans, they have been identified in other animals and plants. The GENCODE database includes annotations for humans and mouse, other databases like ALDB (Domestic-Animal Long Noncoding RNA Database) contain annotations for domestic animals like cow, pig and chicken [19]. Databases exclusively for plant genomes are also in progress (e.g. PlncDB [20]). From a large number of lncRNA loci, only a few hundred have been functionally characterised. More lncRNAs are likely to be discovered and a majority of the novel lncRNAs predicted by various methods are yet to be functionally annotated.

LncRNAs are involved in aging, neurological diseases and various types of cancer, as evidenced by multi-omics studies [21]. This makes them ideal candidates to study molecular pathways of disease and development. Both the biogenesis and computational prediction of short ncRNA (such as miRNA) are now well established. However, biogenic pathways and systematic prediction of sequence and function of lncRNA are yet to be fully understood. This is attributed to their varying sequence lengths, genomic origin and functions; and the new lncRNAs are still emerging. Some of the known challenges in designing lncRNA annotation pipelines are a) incomplete cataloging of lncRNA features, b) the diversity of experimental methods used to generate transcript information, c) low expression d) cell specificity, and e) dynamics of lncRNA function in response to epigenetic signaling. Latest advances in Big Data processing, Machine Learning and Distributed Computing is assisting in overcoming these challenges but the progress in the field is still limited, and more efficient strategies to predict and functionally characterise lncRNA are required.

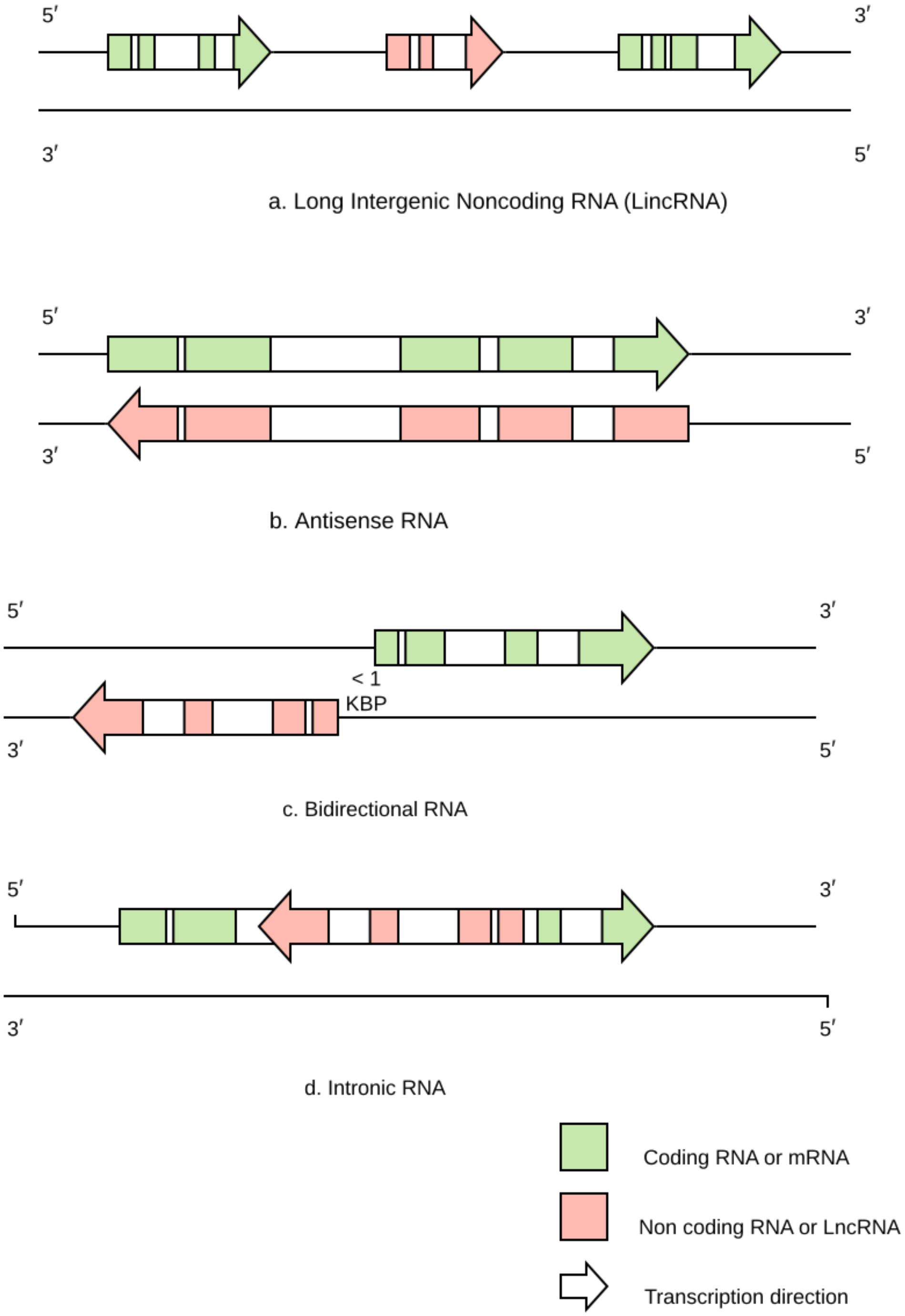
1.2 Classification

LncRNAs can be antisense, intergenic, interleaved, or overlapping with protein-coding genes [22]. This classification is based on their biogenesis loci as shown in Figure 1. If they originated from a region of the genome which lies in between two coding genes they are called as Long Intergenic Noncoding RNA or LincRNA in short. Complementary RNA strands to mRNAs are called antisense RNA. Antisense transcripts can occur with varying degrees of overlap, from none (divergent), to partial (terminal) and complete (nested). Likewise, there are other categories based on the location of the transcription such as intronic RNA, which as the name suggests, lie in the intronic region of another protein coding gene and bidirectional RNA, which are transcribed within 1 kb of a protein coding gene in antisense direction. Other classifications also exist based on their function, conservation, role in regulation or chromatin modeling, and have been previously reviewed [23, 24].

1.3 Characteristics

The lncRNAs are usually arbitrarily defined based on length over 200 bp and having no or low coding potential. The transcripts, however may contain 5' cap and poly-A tail, and composed of multiple exons. They are short lived within the cell and the level of expression varies across different cell types and sub types. LncRNAs are more numerous but have lower abundance than

Figure 1: Classifications of LncRNA based on their site of origin



mRNA within a cell. They are more tissue specific along with high developmental state specificity and cell subtype specificity [25].

LncRNAs vary considerably in length from 200 bp to around 2000 kbp and therefore, are expected to fold into a variety of functional secondary (2D) structures through intramolecular base pairing. Therefore, it is very challenging to identify them conclusively and several efforts are currently underway. Generally conservation of these sequences are found to be moderate compared to their protein coding counterparts, however, the 2D structures are well conserved and so are the expression levels [26]. This suggests that the sequence in itself is less important than the secondary structure. The secondary structures of the lncRNA are thought to be modular, providing interaction sites for other DNA, RNA and proteins. With these interactions they epigenetically regulate the cellular biology, thereby forming a layer of genomic programming on top of the coding genes.

1.4 Role

LncRNAs play many roles and are majorly involved in regulation of proximal or overlapping protein coding genes, also known as *cis*-regulation. However in some cases they can also regulate genes in *trans*, which are further away on the same or on different chromosomes. They are found in different compartments of a cell like nucleus, nucleolus, cytoplasm, and even in the mitochondria, which mainly correlates to its mode of action. LncRNA do not function alone or in a single manner, they interact with other genes and proteins through complex pathways. More comprehensive functional reviews have been published previously [27, 28]. Some of the well known functions of lncRNA are listed in Table 1.

Table 1: Roles of lncRNA.

Role	Description	Examples	References
Recruiters or guide	lncRNA can recruit chromatin remodelling protein factor complexes to the loci, which in turn modify chromatin state. This can happen in *cis*, where the lncRNA is transcribed or in *trans*, the opposite (more distant) side of its transciption site.	. Xist is a ncRNA which remains attached to the transcription site regulating X-chromosome inactivation or XCI. . HOTAIR is a lincRNA in the mammalian HOXC locus that recruits PRC2 (a histamine H3 lysine 27 methylase which is responsible for silencing the developmental gene complex and cancer progression) to the HOXD locus on a different chromosome.	[29, 30, 31]
Tethers	lncRNAs bind with RNA Binding Proteins (RBP), an important class of proteins to form RNA:RBP complexes which play a crucial role in transcriptional or post-translational regulatory functions. They also bind with chromatin via RNA:DNA interactions to form tags facilitating transcription in the *cis*. This is possible due to the fact that lncRNAs are allele specific and also specific to their transcription locus. Further they are well suited to specify the loci for the transcription factors due to their uniqueness.	Protein YY1 contains binding motifs for RNA as well as DNA, it tethers Xist, a lncRNA onto the X chromosome.	[32, 33]

Role	Description	Examples	References
Scaffold	lncRNA can act as a molecular scaffold on which diverse factors involved in the same biological process assemble. For example, facilitation of dynamic formation of heterochromatin or an enhancer-promoter interaction	Heterochromatin is regulated by creation of lncRNAs such as Xist and Tsix, which are involved in X Chromosome Inactivation or XCI, Kcnq1ot1 and Air which are implicated in imprinting by providing scaffolding structures for its assembly.	[34]
Decoys	They can repress gene activity by binding to transcription factor complexes targeting DNA or RNA thereby directly affecting the transcription process. In addition, they affect the post-transcription processes by acting as a sponge or decoy for microRNAs which are involved in processes like mRNA cleavage, direct translational repression and/or mRNA destabilization [35]	LncRNA AK015322 promotes proliferation of spermatogonial stem cell C18-4 by acting as a decoy for microRNA-19b-3p.	[36, 35]
Coregulators	lncRNA can co-regulate its nearby protein coding gene. This association is mostly positive resulting in enhanced expression and sometime negative, repressing the adjacent protein coding gene.	. Igf2as/PEG8 is very highly correlated with Igf2, which is in line with the observation where it is observed to be overexpressed in Wilms' Tumors together with Igf2. . Similarly there are many such coregulatory lncRNA-mRNA pairs like GM42937-ATP1a1, Wt1os-Wt1, GM3235-Fto.	[37]

Role	Description	Examples	References
Pol II inhibitors	In some cases lncRNAs are believed to suppress the ability of RNA polymerase II to perform transcription by binding to its core and forming preinitiation complexes at the promoter regions.	B2 RNA binds to Pol II resulting in lower levels of transcription.	[38]
mRNA processing	. lncRNAs are involved in various stages of mRNA processing like alternative splicing. A class of lncRNA called NAT (natural antisense transcripts) is one of the major contributors in alternative splicing which it performs by masking portions of the splicing regions because of their complementary nature. . They can also lead mRNA to degradation pathways. . LncRNAs can interact with RNA methyltransferases or demethylases and thus regulate mRNA expression post-transcriptionally.	. The FGFR2 gene in humans is alternatively spliced with respect to Exon IIIb-IIIc, is regulated by asFGFR2 lncRNA which is also a NAT. . The lncRNA 1/2-sbsRNA binds the Alu element of the 3'-UTR region on the target gene in the Staufen 1 (STAU1)-mediated mRNA decay (SMD) pathway. . lncRNA mediated reversible m6A methylation modification of mRNA have been reported in animals.	[39]

Role	Description	Examples	References
Stability	Post-transcriptional stability of the transcripts once in cytoplasm is maintained by their corresponding NATs by recruiting stabilizing factors, thus preventing the effect of destabilizing factors on those transcripts.	. Noncoding RNA activated by DNA damage (NORAD) controls the ability of RBMX to form complexes which subdue the instability in the genome. . lncRNA OCC-1 reduces the stability of the RNA binding protein HuR which stabilises many mRNA resulting in the inhibition of corresponding protein coding transcripts.	[40, 41]
Translation	The NATs are found to affect the translation process thereby regulating gene expression by promoting or hindering translation of mRNA transcripts to protein. They do so by providing RNA binding motifs which interact with translation initiation complexes.	lncRNA GAS5 Interacts with the Eukaryotic Translation Initiation Factor 4E to moderate c-Myc synthesis via translation.	[42]
Extracellular vesicle packaging	They can modulate gene expression in the environment when secreted as extracellular vesicles	Extracellular vesicle packaged HIF-1 alpha-stabilizing lncRNA from tumour associated macrophages regulates aerobic glycolysis of breast cancer cells.	[43, 44]

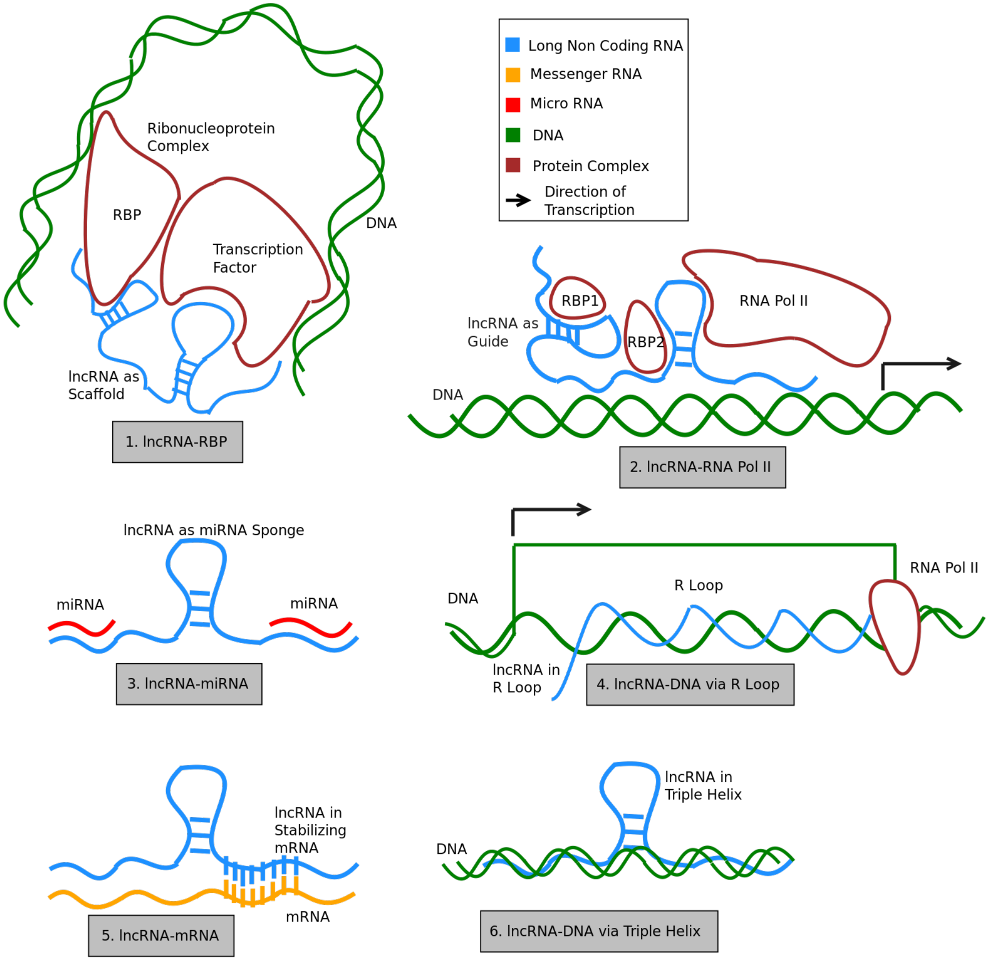
Role	Description	Examples	References
Precursor of small RNA	lncRNA can be precursors of miRNAs through intracellular shearing, and RNAs can be processed to produce specific miRNAs regulating the expression of target genes.	. lncRNA H19 produces a precursor of miRNA675 that suppresses translation of insulin growth factor receptor (Igf1r). . lncRNAs MIR100HG hosts genes of miR-100 and miR-125b known to mediate cancer cell resistance . lncRNA MuLnc1 in plants is cleaved by mul-miR3954 producing secondary siRNA	[45, 46]
Encoding peptides	Some lncRNA have been observed with potential to code for micropeptides which may be functional	Two known examples are from skeletal muscles: . LINC00948 in humans (and AK009351 in mice) encodes a conserved micropeptide (46aa) called myoregulin (MLN) . LncRNA LINC00961 encodes a conserved polypeptide (90aa), named 'small regulatory polypeptide of amino acid response' (SPAR)	[47, 48]

The role of lncRNAs are not limited to the ones mentioned in Table 1. Novel lncRNAs and their roles in cell biology are still emerging due to limited knowledge about this class of RNAs and their functional mechanisms.

1.5 Mechanisms

The secondary and tertiary structures of lncRNA also play a significant role in the mechanism of their actions by creating binding sites for other bio-molecules like DNA/RNA/Proteins to interact with. LncRNAs function by employing multiple mechanisms and actions, unlike their protein coding counterparts. Some of these mechanisms are listed below and illustrated in Figure 2.

Figure 2: Illustration of mechanisms of Long Non Coding RNA. Archetype 1: LncRNA can act as scaffolds bringing RBPs and TF bound at different loci in DNA secondary structure together at the promoter region to start transcription. Archetype 2: LncRNA can guide proteins like chromatin remodelling factor complexes to the loci playing a crucial role in epigenetic regulation. Archetype 3: LncRNA can interact with miRNA which are involved in the post translational process by acting as a sponge or decoy. Archetype 4: LncRNAs can form a three stranded nucleic acid structure called R Loop which is a common mechanism regulating gene expression by various mechanisms. Archetype 5: NATs are a type of LncRNA which maintain stability of its corresponding coding transcript post transcription. Archetype 6: LncRNA forms a triple helix of triplex structure with DNA, this structure interacts with proteins which can stabilize or unwind it.



doi:10.20944/preprints202006.0116.v1

1.5.1 RNA-Protein Interaction

Some lncRNAs interact with proteins called as RNA Binding Proteins (RBP) forming complexes which have significant roles in cellular pathways. Xist is a lncRNA whose functions are well researched, as it is primarily involved in X-chromosome inactivation and assists in differentiation of early pluripotent cells. This transcript interacts with 81 RBPs in a modular and developmentally controlled manner to coordinate chromatin spreading and silencing [49]. LncRNA is also known to interact with transcription factors via Transcription Factor Binding Site (TFBS) and in some cases share the same TFBS with their protein coding counterparts, indicating a direct role in regulating the transcription of coding genes [50]. Computational tools like GraphProt [51] and more recently, Heterogeneous Network Model based method [52] are available to model the binding preference of the RBPs.

1.5.2 RNA-RNA Interaction

LncRNAs interact with other RNAs like mRNAs and/or small RNA such as micro RNAs (miRNAs). The miRNA function by directly interacting with mRNAs regulating gene expression[53]. When these small RNAs are combined with lncRNAs given the tissue specificity of these lncRNAs result in tissue differentiation and cancer development. One such interaction is the formation of MLMI or mRNA-lncRNA-miRNA regulatory network in the case of hepatocellular carcinoma, where miRNAs and lncRNAs are found to be differentially expressed [54]. Another category of interactions includes lncRNA regulating mRNAs by alternate splicing, editing and stabilizing via direct base pairing [55]. RIblast [56] and IntaRNA [57] are *in silico* tools which can be used to predict RNA–RNA interaction.

1.5.3 RNA-DNA Interaction

LncRNAs can interact with chromatin through RNA:DNA interactions. Several mechanisms have been proposed for this interaction [58]. Two of the well known modes are listed below:

By forming RNA loops or R loops: An R loop is a three stranded nucleic acid structure involving an RNA:DNA hybrid and a non-template DNA Figure 2. R-loops are a common occurrence in the genome and they are involved in many regulatory pathways. For example, 1) TERRA lncRNA induces DNA damage response elements by forming an R loop in the telomere region, 2) R-loop formed by the transcription of VIM-AS1 lncRNA regulates the expression of the closest protein-coding gene, transcribed in antisense orientation with respect to the lncRNA, 3) FLC locus in *Arabidopsis thaliana* is repressed by lncRNA called COLDAIR by the same mechanism

[59]. R-loops are also believed to cause instability in the genome. To counter this there are factors preventing the formation of these R loops, failure of which might lead to replication stress, genome instability, chromatin alterations or gene silencing, which generally lead to cancer and other genetic diseases [60]. There are computational tools like QmRLFS-finder [61] to predict and analyse R loop forming sequences, and dedicated databases like R-loopDB [62] for R-loop Forming Sequences (RLFS) and R-loops.

By forming triple helix: RNA-DNA triplex consists of DNA double helix in combination with RNA as the third strand which forms a stable triplex (Figure 2). Orientation of the third strand is important for function [63] and based on its orientation the triplex may be parallel or antiparallel. Some examples of triplex formation are: PRC2 recruitment by the lncRNA Fendrr; Methyltransferase recruitment on rRNA promoter; LncRBA khps1 recruits histone acetyltransferase (p300/CNP) at sense SPHK1 gene and activates transcription. HOTAIR MEG3 and PARTICLE are other examples of lncRNA acting on prompters locally or distantly in epigenetic response. Some proteins like Argonaute may stabilize the triplex structure while some helicases are believed to unwind or remodel it. These interactions point at functional interaction of these proteins with the triple helix [64]. Computational methods Triplexator [65], and LongTarget [65, 66] can be used to study triple helix structures.

2 Data resources

There are many data resources available online containing information about lncRNA sequences, RNA-interactome and functional annotations. Some contain experimental output data, others contain manually curated data obtained from literature and there are even few which provide data obtained by in silico prediction. Table 2 lists some of the well known lncRNA repositories.

Table 2: Datasources containing lncRNA interactome and annotations.

Source name	Type of Data	Latest release[Last modified Date]	URL
Gencode [67]	Comprehensive database consisting of all the experimentally validated annotations of the genes in the human and mouse genome.	Human: 34 [29.04.20] Mouse: M25 [29.04.20]	`https://www.gencodegenes.org`

Source name	Type of Data	Latest release[Last modified Date]	URL
FANTOM-CAT [18]	FANTOM5 CAGE data performed on multiple transcript collections to generate a comprehensive atlas of 27,919 human lncRNA genes with high-confidence 5' ends and expression profiles across 1,829 samples from the major human primary cell types and tissues.	[09.08.2017]	`https://fantom.gsc.riken.jp/cat`
FANTOM-6 [68]	Reference set of genome-wide profiles to establish the basal state of the transcriptome and epigenome in each cell type. Evaluated molecular phenotype of the perturbation by CAGE profiling followed by bioinformatics analysis. Functional characterization of a selected subset of lncRNAs.	[27.04.2020]	`https://fantom.gsc.riken.jp/6`
RNAcentral [69]	Integrated from multiple expert databases (41 in release 15), a comprehensive set of up-to-date lncRNAs currently covering 56 species and different RNA types. LncRNA relevant databases included are LNCipedia, LncBook, NONCODE, lncRNAdb and LncBase.	Release 15 [21.05.2020]	`https://rnacentral.org`
lncBook [17]	A curated knowledgebase of human lncRNAs that features a comprehensive collection of human lncRNAs and systematic curation of lncRNAs by multi-omics data integration, functional annotation and disease association.	Version 1.0 [08.29.2019]	`https://bigd.big.ac.cn/lncbook/index`
lncRNA Disease 2.0 [70]	A database for collection of experimental supported lncRNA-disease associations	Version 2.0	`http://www.rnanut.net/lncrnadisease`
RAID [71]	Manually curated comprehensive database for RNA-associated interactome integrated from 18 different data sources and literature across 60 species.	Version 2.0	`www.rna-society.org/raid/`

doi:10.20944/preprints202006.0116.v1

Source name	Type of Data	Latest release[Last modified Date]	URL
RISE [72]	A database consisting of RNA- RNA interactions obtained from experimental sequencing data and also includes data from other data sources and literature for human, mouse and yeast transcriptome.	Version 1.0 [Aug-2017]	`http://rise.life.tsinghua.edu.cn/`
LncRRI search [73]	A RIblast based prediction tool for lncRNA-lncRNA and lncRNA-mRNA interactions, also contains tissue-specific expression and subcellular localisation data for the lncRNAs.	Version: 1.00 [06.02.2019]	`http://rtools.cbrc.jp/LncRRIsearch`

3 LncRNA transcripts and interactome identification using sequencing

The experimental approaches to identify and annotate these lncRNAs are expensive, time consuming and often are restricted to specific experimental set up. cDNA library preparation followed by sequencing has been one of the traditional approaches for detecting lncRNAs. Earlier, microarray (tiling arrays) based identification was used to detect alternative splicing and discover polymorphisms [74]. High throughput sequencing (HTS) technology using short reads (50-300 bp) provided a revolutionary means for systematic discovery of transcriptional units and has been effective in picking up disease associated with lncRNAs. The wet lab assays to capture RNA-interactome have previously been reviewed in depth [75, 76]. Some of the most widely used sequencing techniques are discussed below.

3.1 Sequencing of RNA transcripts

SAGE or Serial Analysis of Gene Expression is one of the early methods using HTS and made it possible to detect novel transcripts [77]. cDNA is synthesized from the 3' end of mRNAs after cleaving using restriction enzymes. SAGE tags are concatenated, cloned and sequenced using Sanger sequencing [75]. Another similar approach is CAGE or Cap Analysis of Gene Expression where 5' end of mRNA sequences are extracted and cDNA is synthesized from them. The resulting cDNA is then sequenced using HTS techniques to obtain the sequences in the promoter regions of

doi:10.20944/preprints202006.0116.v1

the mRNA which can be used to determine the gene from which they originated [78]. Ultra deep whole RNA sequencing (RNA-seq) is one of the most popular methods to profile both the coding and noncoding transcriptome. The sequencing reads are then aligned against the reference genome or assembled together *in silico* to obtain the transcripts [79]. However, the expression levels of non coding transcripts is much lower in comparison to the coding transcripts and more targeted approaches have been implemented to capture lncRNA expression profiles [80]. Another challenge is to recover full length transcripts from the short read high throughput sequencing, which are currently the more widely used sequencing approaches. Upcoming long read sequencing techniques such as from Oxford Nanopore [81] and Pacific Biosciences (PacBio) [82] can help recover longer isoforms. However, these techniques are currently limited to polyA containing transcripts only. In many cases the sample quantity is insufficient to make use of any of the earlier methods for sequencing. Single-cell techniques Smart-seq [83], DP-seq [84, 83] and Quart-seq [85] can be used in such cases. They can also be used in cases where cell to cell variation in gene expression needs to be qualified to study epigenetic modifications between individual cells [75]. LncRNA can serve as a precursor to other smaller RNA which are processed upon degradation. The technique PARE-Seq [86], GMUCT [87], and degradome-seq [88] have all been used to map transcripts which are in the process of being degraded. Another technique TIF-seq (transcript isoform sequencing) has been developed to detect both the 5' and 3' ends of transcripts [89]. Global run-on sequencing (GRO-seq) methods were developed [90] to detect nascent transcription. In order to assess correlation of half lives of lncRNA with their function RNA decay rates can be measured using 5'-bromo-uridine immunoprecipitation chase–deep sequencing analysis (bric-seq) [91].

3.2 Sequencing of RNA-Protein Interactions

Several approaches exist to study the interactions between protein complexes and RNAs they are bound to. Some of the commonly used ones are CLIP-Seq or crosslinking immunoprecipitation sequencing [92], photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation sequencing (PAR-CLIP-Seq) [93], crosslinking immunoprecipitation sequencing (CLIP-Seq) [94] and RNA immunoprecipitation sequencing (RIP-Seq) [95].

3.3 Sequencing of RNA-Chromatin Interactions

LncRNA interacts with chromatin directly or indirectly via proteins. Chromatin isolation by RNA purification (ChIRP) can be used to identify association between a ncRNA and chromatin [96]. In this method cross-linked Protein-DNA-ncRNA complexes are obtained, which can then be separated into different components for further investigations. ncRNA can be quantified using qPCR and proteins are detected with western blots. Nucleic acid is subjected to sequencing to observe genomic regions associated with lncRNA. RNA antisense purification (RAP) [97] is another technology similar to ChIRP designed with longer ncRNA probes to reduce technical noise. Capture hybridization analysis of RNA targets (CHART) [98] protocol is conceptually similar to ChIRP and RAP with improvement in probe designs that are focused around ncRNA target region as opposed to whole length based probes used previously. Some of the newer techniques developed to capture RNA:DNA interactions are MARGI (Mapping RNA Genome Interactions) [98, 99], iMARGI (*in situ* mapping of RNA–genome interactome) technique [100], GRID-seq (*in situ* global RNA interactions with DNA by deep sequencing), and RADICL-seq or RNA And DNA Interacting Complexes Ligated and Sequenced [101].

3.4 Sequencing of RNA-RNA Interactions

RAP-RNA [102] is a modified RAP protocol to capture RNA:RNA interactions. Direct RNA:RNA hybridization can also be captured by using cross-linking, ligation and sequencing of hybrids (CLASH) [103]. CLASH used UV cross-linking as opposed to chemical cross linking used in RAP-RNA to improve spurious protein-protein cross linking of the capture.

4 LncRNA identification using computational approaches

In silico techniques to identify and annotate lncRNA can be reference-based, which requires a reference genome from which a model is trained by extracting the features; LncRScan-SVM, COME and lncScore fall in this category. Their predictions are restricted to known reference transcripts. Reference-free or *ab initio* methods learn the annotation from the features in the input dataset and hence don't require a reference set. There are many tools for identification of non coding RNAs belonging to this category: CPC, CPC2, CNCI, lncRNA-MFDL, lncScore, LncADeep, DeepLNC, LncRNAnet, COME, CPAT, lncRScan-SVM, longdist, PLEK, FEElnc, lncRNA-ID and LncFinder. Details of all these tools can be found in Table 3.

Table 3: Computational techniques for LncRNA Identification and functional annotation.

Tool	ML model	Feature set	Cons	Pros	Re-trainable
CPC [104]	SVM	. ORF: log-odds score, ORF coverage, ORF integrity. . BLASTX: HSPs (High-scoring Segment Pairs), hit score, frame score.	. Over reliance of ORF might lead to bad predictions for partial length transcripts. . Cannot accurately discriminate transcripts falling entirely within UTR regions	Less feature set makes it run faster with less compute capacity.	No
CPAT [105]	LR	ORF size, ORF coverage, Fickett TESTCODE statistic and hexamer usage bias	Difficult to scale up by adding multiple features.	. More robust . Markedly faster . More convenient to use	Yes
CNCI [106]	SVM	Length and S-score of the most-like coding domain sequence (MLCDS), length-percentage, score-distance and codon-bias	. Not recommended for partial length transcripts. . Requires well assembled high quality transcripts.	. Suitable for non model species. . Can also used to study the evolution of sequential features.	No

Tool	ML model	Feature set	Cons	Pros	Re-trainable
PLEK [107]	SVM	The k-mer frequencies of the transcript sequence.	Classifier built only using human training data.	. Can handle high levels of indel sequencing errors. . Do not require a reference genome.	Yes
lncR Scan SVM [108]	SVM	Transcript length, Stop codon standard deviation, CDS score, exon count, exon length, conservation	. Does not consider structural features. . Biased towards predicting intergenic or lincRNAs. . Prediction of Novel lncRNAs is not possible.	Features derived from gene structure, transcript sequence, potential codon sequence and conservation	Yes
lnc RNA ID [109]	RF	ORF length, ORF coverage, two Kozak motif-related features, ribosome coverage on three regions: transcript, ORF and 3' UTR, ribosome release score, alignment score, alignment length with respect to profile HMM, and the transcript	. Structural features are not accounted for. . No multithreading.	. Uses profile hidden Markov model (profile HMM), hence it is faster. . Model can be re-trained for different species.	Yes

Tool	ML model	Feature set	Cons	Pros	Re-trainable
lnc RNA MFDL [110]	ANN	Multiple features of the ORF, k-mer, the secondary structure and the most-like coding domain sequence (MLCDS)	Only about $\frac{1}{3}$ rd of data is used to train the model.	. Used deep learning architecture instead of shallow learning models. . Strong and robust predictor.	No
lnc Score [111]	LR	. Exon: Hexamer Score (HS), HS-Distance, GC-content. . MCSS: Length, Coding Score, Coding Score Percentage. . ORF: Length and Coverage, Fickett Score, HS, HS-Distance.	Does not make use of ORF features.	. Performs well for partial length transcripts too. . It is multi-threaded hence runs faster.	No
Deep LNC [112]	ANN	Information content stored in k-mer pattern has been used as a sole feature.	Limited feature set.	. Can handle non linear relations well. . Exhibits high performance of accuracy and prediction rate.	No

Tool	ML model	Feature set	Cons	Pros	Re-trainable
CPC2 [113]	SVM	Fickett TESTCODE score, ORF length, ORF integrity and isoelectric point (pI)	Limited feature set.	. Faster than its predecessor CPC . It's species independent.	No
COME [114]	RF	. Sequence-derived: GC content, DNA and protein sequence conservation, RNA 2D conservation. . Expression: Expression abundance from poly(A)+, poly(A)- and small RNA sequencing. . Histone: H3K36me3 and H3K4me3.	. Limited only to 5 species. . Low detection rate for genic lncRNA. . Can not distinguish novel isoforms and poorly assembled known transcripts.	. Can be used for both known and novel transcripts. . Provides supporting evidence for its annotations.	Yes
Long dist [115]	SVM	336 frequencies of nucleotide patterns; and 4 features derived from ORFs.	Only nucleotide pattern frequencies AND ORF lengths considered.	Usage of Principal Component Analysis or PCA reduces the number of features used in the model.	Yes

Tool	ML model	Feature set	Cons	Pros	Re-trainable
FEE lnc [115]	RF	The k-mer frequencies of the transcript sequence.	. Weakly expressed mRNAs might be misclassified. . Week performance in the case of high error rate assemblies.	Feature selection is general enough to capture all lncRNAs classes.	Yes
LncA Deep [116]	DL	2 (ORF length and coverage) + 36 (The EDP of ORF) + 1 (Mean hexamer score) + 4 (UTR length and coverage) + 8 (Fickett nucleotide features) + 3 (HMMER index)	. Features set do not consist of structural features. . Model is pre built and can not be extended. . Need to be trained on more data in order to make more generalized predictions.	. Handles both full length and partial length transcripts. . Can also predict the function of the transcripts.	No
Lnc RNA net [117]	ANN	ORF Indicator	. Limited feature set. . Minor drop in ability to detect long length non coding transcripts.	Shows good performance for short length transcripts.	No

Tool	ML model	Feature set	Cons	Pros	Re-trainable
Lnc Finder [118]	LR, SVM, RF, ELM and ANN	. Sequence Intrinsic Composition: Logarithm-distance of hexamer on ORF, Length of the longest ORF, Coverage of the longest ORF. . Multi-Scale Structural Information: Minimum free energy, UP frequency of paired–unpaired sequence, Logarithm-distance of acguD sequence, Logarithm-distance of acgu-ACGU sequence. . Physicochemical Property: Signal at 1/3 position, SNR, Quantile statistics (Q1, Q2, min and max).	. Doesn't consider interaction, annotation and evolution related features. . Performance varies across species.	. Integrates multiple classifiers. . Easy to use as it is released as an r package. . Algorithm is developed based on the optimal feature combination and the most appropriate classifiers.	Yes

Computational methods to study lncRNA reviewed here make use of Logistic Regression (LR), Support Vector Machines (SM), Random Forest (RF), and Artificial Neural Network (ANN) along with their Deep Learning (DL) variations. Usage of LR for this purpose makes the resulting model simple and generalized as observed in CPAT [105], lncScore [111]. However, it is very important to be careful while selecting the features as using some of the features which may not have a clear association with the prediction or features with a high correlation between them might result in a poorly performing model [119]. If the noise is minimal, SVM can be a good choice and it will provide good performance and accuracy. lncRScan-SVM [108], CPC [104], CPC2 [113], CNC I [106], PLEK [107] and longdist [120] make use of this algorithm. These methods may not perform well for this particular problem where the relationship between predictors and the target variable is complex (not linearly separable) and the number of datapoints is huge. RF which is used in COME [114], lncRNA-ID [109] and FEElnc [115] is well suited for higher-dimensional data with large numbers of data points. It is also easy to parallelize the prediction process and since each process operates on a subset of the data it will run faster. But, the model takes up a lot of memory for large complex training datasets like the one which has to be used in this case. ANN including many of its variations is emerging as the preferred choice recently because of their inherent ability to learn extremely nonlinear and complex relationships by themselves, to train on large numbers of observations, and to scale up the computing process. The models obtained from RF and ANN are difficult to interpret and explain the rationale behind the prediction. For example, a deep learning model is nothing but the weights and biases of the input, hidden, and output nodes. Therefore, even a model that is very good at predicting coding potential may not help us much to understand the underlying biological processes with just those seemingly random numbers. Although there are some methods like Garson's algorithm, Lek's profile, partial dependence plot and local interpretable model-agnostic explanations (LIME) to address this issue to some extent, they can only be used to determine the relationship between the predictors and target variables [121] but this will not be sufficient to explain the complete biological process comprehensively given its nature of complexity. Tools like lncRNA-MFDL [110], LncADeep [116], DeepLNC [112], LncFinder [118] employ ANNs.

These models can be trained on a number of lncRNA features which can be primarily categorised as sequential, structural or conservational. All models use sequential features, additionally, tools like lncRNA-MFDL and LncFinder also use structural features, and COME and lncRScan-SVM utilize conservation-based features to annotate lncRNA.

Sequential features can be extracted by direct parsing of the nucleotide sequence of the transcript which can be obtained in the sequence files (in FASTA format). These consist of features like transcript length, ORF related scores, Exon related scores, GC content, CDS score, Fickett Score, and k-mer pattern information. They are the simplest ones to extract but may have high variability between samples, which needs to be accounted for.

The lncRNA mechanisms of action involve structural interactions in the majority of its functions, thus its secondary structure is crucial in determining its categorization [26, 122]. Moreover, conservation of secondary structure is higher compared to the conservation of the sequence itself, thus making them more reliable. However, incorporating secondary features demand more time and resources. Previous attempts at extracting these features by using algorithms like Stochastic Context-Free Grammar (SCFG) [53] or minimum free energy (MFE) [123] based algorithm have been available.

Every tool performs well only over the subset of use cases, such as, for a specific species or specific type of transcripts, which is fundamentally dictated by its choice of features, data sources, and classifiers. But there is a general lack of approaches that can be used as a tool flexible enough to be extended by training on newer data and/or by adding new features.

5 Conclusion and Future Directions

We understand that lncRNAs make up a large chunk of the human transcriptome, and present diverse molecular structures [Figure 1], roles [Table 1], and mechanisms of action [Figure 2]. As described in section 3, a large selection of experimental protocols exist to capture numerous aspects of lncRNA biogenesis and functions. We also have a number of emerging reference datasets based on automated sequence annotation, experimental verification, disease association, and interaction studies [Table 2]. These add up to a great resource to develop bioinformatic protocols to analyse this vast amount of accessible noncoding transcript data. The knowledge derived on the function of these RNAs and their potential applications in studying various disease and development pathways is of significance.

In addition to these developments, we also identify challenges and gaps in our comprehension of this important new class of RNAs. Inconsistency in the annotation between different databases, incomplete or overlapping annotation of new transcripts and inherent technical noise of the sequencing platforms to generate this data add to intricacy. Owing to varying sequence length, dynamic response to a stimulus, cell specificity, transcript instability, and multitudinal interaction mechanisms, a broad range of biological assays are available to grasp all these different aspects of a lncRNA activities. This further complicates designing computational protocols to answer biological questions from such heterogeneous data and metadata. An ensemble of computational modeling approaches may be required to address unconventional questions from the transcriptomic data. Currently, there are a number of methods available to elucidate if a transcript is coding or noncoding by using computational coding potential criteria involving the measurement of several sequential features as discussed above [Table 3]. Noncoding RNA sequences generally operate by folding into secondary structures, which are important for their function. Some of the methods also exploit these structural features in the annotation process by measuring minimum free energy and looking at the conservation of these features. All these dimensions have a major contribution in determining the functional potential of these transcripts accurately. Another important feature is to understand the mechanism of these RNAs through the prediction of their contact sites on the target biomolecules. None of the existing methods take advantage of interaction motif-based features to train their models. There is a need to include and comprehensively evaluate these features while building the models.

A scarce use of structural features in various pipelines is possibly due to our lack of understanding of how these RNA fold into secondary and tertiary structures and in turn dictate specific interactions with proteins or nucleic acid. A majority of the secondary structure methods used minimum free energy algorithms to find the optimum folding stage of an RNA. Longer RNA sequences can hypothetically fold into a large number of possible structural conformations and this complexity can deviate from its minimum free energy status. Newer approaches making use of convolutional neural networks and dynamic programming to solve the complex structures are becoming available [124]. These improvements in structure predictions can be leveraged to build future lncRNA annotation pipelines.

Sometimes the same transcript might act as coding RNA in some circumstances and non-coding in some others [125, 126]. Therefore, the co-factors affecting the transcription needs to be studied further to identify the conditions leading to gene expression. A comprehensive integrative and meta-analysis approach is useful in understanding the complex lncRNA mechanisms of action at multiple regulatory levels. It is very important to recognize that all these computational predictions have to be verified experimentally, however they to a large extent, they can reduce the search space to conduct experiments.

Finally, since most of the computational methods are predicting lncRNAs with reasonable accuracy, many novel lncRNAs are being predicted. However, there is a gap between the detection of these noncoding transcripts and understanding their mechanisms. Hence, more efforts are to be made to annotate their functionality which is very useful in fully understanding the underlying cell biology.

Author's contributions

Conceptualisation, S. T Formal analysis, S. T, Y. R, L. K; Funding Acquisition, S. T; Investigation, S. T, Y. R, L. K; Resources, S. T; Supervision, S. T, L. K Visualisation Y. R; Writing – review & editing, S. T, L. K, Y. R.

Acknowledgments

S. T acknowledges funding from the Faculty Initiative Fund and Australian Women Research Success Grant at Monash University. S.T. also acknowledges the AISRF-EMCR Fellowship by the Australian Academy of Science. We thank Tyrone Chen for critically reading the manuscript and providing with valuable insights on improvements. We thanks A/Prof Akshat Tanksale for proofreading the manuscript.

doi:10.20944/preprints202006.0116.v1

References

[1] Hui Jia et al. "Genome-wide computational identification and manual annotation of human long noncoding RNA genes". en. In: *RNA* 16.8 (Aug. 2010), pp. 1478–1487.

[2] John S Mattick and Igor V Makunin. "Non-coding RNA". In: *Hum. Mol. Genet.* 15.suppl_1 (Apr. 2006), R17–R29.

[3] Mitchell Guttman et al. "lincRNAs act in the circuitry controlling pluripotency and differentiation". en. In: *Nature* 477.7364 (Aug. 2011), pp. 295–300.

[4] Ahmad M Khalil et al. "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.28 (July 2009), pp. 11667–11672.

[5] U A Ørom et al. "Long noncoding RNAs as enhancers of gene expression". en. In: *Cold Spring Harb. Symp. Quant. Biol.* 75 (2010), pp. 325–331.

[6] Kevin Struhl. "Transcriptional noise and the fidelity of initiation by RNA polymerase II". en. In: *Nat. Struct. Mol. Biol.* 14.2 (Feb. 2007), pp. 103–105.

[7] Harm van Bakel et al. "Most "dark matter" transcripts are associated with known genes". en. In: *PLoS Biol.* 8.5 (May 2010), e1000371.

[8] Masaaki Furuno et al. "Clusters of internally primed transcripts reveal novel long noncoding RNAs". en. In: *PLoS Genet.* 2.4 (Apr. 2006), e37.

[9] Philipp Kapranov et al. "RNA maps reveal new RNA classes and a possible function for pervasive transcription". en. In: *Science* 316.5830 (June 2007), pp. 1484–1488.

[10] Johnny T Y Kung, David Colognori, and Jeannie T Lee. "Long noncoding RNAs: past, present, and future". en. In: *Genetics* 193.3 (Mar. 2013), pp. 651–669.

[11] Michael B Clark et al. "Genome-wide analysis of long noncoding RNA stability". en. In: *Genome Res.* 22.5 (May 2012), pp. 885–898.

[12] Ana C Marques and Chris P Ponting. "Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness". en. In: *Genome Biol.* 10.11 (Nov. 2009), R124.

[13] Igor Ulitsky et al. "Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution". en. In: *Cell* 147.7 (Dec. 2011), pp. 1537–1550.

[14] Tim R Mercer, Marcel E Dinger, and John S Mattick. "Long non-coding RNAs: insights into functions". en. In: *Nat. Rev. Genet.* 10.3 (Mar. 2009), pp. 155–159.

[15] Junichi Iwakiri, Goro Terai, and Michiaki Hamada. "Computational prediction of lncRNA-mRNA interactionsby integrating tissue specificity in human transcriptome". en. In: *Biol. Direct* 12.1 (June 2017), p. 15.

[16] Jennifer Harrow et al. “GENCODE: the reference human genome annotation for The ENCODE Project”. en. In: *Genome Res.* 22.9 (Sept. 2012), pp. 1760–1774.

[17] Lina Ma et al. “LncBook: a curated knowledgebase of human long non-coding RNAs”. en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D128–D134.

[18] Chung-Chau Hon et al. “An atlas of human long non-coding RNAs with accurate 5’ ends”. en. In: *Nature* 543.7644 (Mar. 2017), pp. 199–204.

[19] Aimin Li et al. “ALDB: a domestic-animal long noncoding RNA database”. en. In: *PLoS One* 10.4 (Apr. 2015), e0124003.

[20] Jingjing Jin et al. “PLncDB: plant long non-coding RNA database”. en. In: *Bioinformatics* 29.8 (Apr. 2013), pp. 1068–1071.

[21] Juliane C R Fernandes et al. “Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease”. en. In: *Noncoding RNA* 5.1 (Feb. 2019).

[22] John S Mattick. “The genetic signatures of noncoding RNAs”. en. In: *PLoS Genet.* 5.4 (Apr. 2009), e1000459.

[23] Francesco P Marchese, Ivan Raimondi, and Maite Huarte. “The multidimensional mechanisms of long noncoding RNA function”. en. In: *Genome Biol.* 18.1 (Oct. 2017), p. 206.

[24] Georges St Laurent, Claes Wahlestedt, and Philipp Kapranov. “The Landscape of long noncoding RNA classification”. en. In: *Trends Genet.* 31.5 (May 2015), pp. 239–251.

[25] Dinar Yunusov et al. “HIPSTR and thousands of lncRNAs are heterogeneously expressed in human embryos, primordial germ cells and stable cell lines”. en. In: *Sci. Rep.* 6 (Sept. 2016), p. 32753.

[26] Per Johnsson et al. “Evolutionary conservation of long non-coding RNAs; sequence, structure, function”. en. In: *Biochim. Biophys. Acta* 1840.3 (Mar. 2014), pp. 1063–1071.

[27] Kevin C Wang and Howard Y Chang. “Molecular mechanisms of long noncoding RNAs”. en. In: *Mol. Cell* 43.6 (Sept. 2011), pp. 904–914.

[28] Xiaopei Zhang et al. “Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels”. en. In: *Int. J. Mol. Sci.* 20.22 (Nov. 2019).

[29] Emily Bernstein and C David Allis. “RNA meets chromatin”. en. In: *Genes Dev.* 19.14 (July 2005), pp. 1635–1655.

[30] Rajnish A Gupta et al. “Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis”. en. In: *Nature* 464.7291 (Apr. 2010), pp. 1071–1076.

[31] Jeannie T Lee. “Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome”. en. In: *Genes Dev.* 23.16 (Aug. 2009), pp. 1831–1842.

[32] Yesu Jeon and Jeannie T Lee. “YY1 tethers Xist RNA to the inactive X nucleation center”. en. In: *Cell* 146.1 (July 2011), pp. 119–133.

[33] Jeannie T Lee. “Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control”. en. In: *Nat. Rev. Mol. Cell Biol.* 12.12 (Nov. 2011), pp. 815–826.

[34] Hugh P Cam, Ee Sin Chen, and Shiv I S Grewal. “Transcriptional scaffolds for heterochromatin assembly”. en. In: *Cell* 136.4 (Feb. 2009), pp. 610–614.

[35] Maria D Paraskevopoulou and Artemis G Hatzigeorgiou. *Analyzing MiRNA–LncRNA Interactions.* 2016.

[36] Ke Hu, Jing Zhang, and Meng Liang. “LncRNA AK015322 promotes proliferation of spermatogonial stem cell C18-4 by acting as a decoy for microRNA-19b-3p”. en. In: *In Vitro Cell. Dev. Biol. Anim.* 53.3 (Mar. 2017), pp. 277–284.

[37] Wang Fan et al. “Coregulatory long non-coding RNA and protein-coding genes in serum starved cells”. In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1862.1 (Jan. 2019), pp. 84–95.

[38] Celso A Espinoza et al. “B2 RNA binds directly to RNA polymerase II to repress transcript synthesis”. en. In: *Nat. Struct. Mol. Biol.* 11.9 (Sept. 2004), pp. 822–829.

[39] Inma Gonzalez et al. “A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature”. en. In: *Nat. Struct. Mol. Biol.* 22.5 (May 2015), pp. 370–376.

[40] Rong-Zhang He, Di-Xian Luo, and Yin-Yuan Mo. “Emerging roles of lncRNAs in the post-transcriptional regulation in cancer”. en. In: *Genes Dis* 6.1 (Mar. 2019), pp. 6–15.

[41] Mathias Munschauer et al. “The NORAD lncRNA assembles a topoisomerase complex critical for genome stability”. en. In: *Nature* 561.7721 (Sept. 2018), pp. 132–136.

[42] Guangzhen Hu, Zhenkun Lou, and Mamta Gupta. “The long non-coding RNA GAS5 cooperates with the eukaryotic translation initiation factor 4E to regulate c-Myc translation”. en. In: *PLoS One* 9.9 (Sept. 2014), e107016.

[43] Fei Chen et al. “Extracellular vesicle-packaged HIF-1α-stabilizing lncRNA from tumour-associated macrophages regulates aerobic glycolysis of breast cancer cells”. en. In: *Nat. Cell Biol.* 21.4 (Apr. 2019), pp. 498–510.

[44] Peter J Quesenberry et al. “Role of extracellular RNA-carrying vesicles in cell differentiation and reprogramming”. en. In: *Stem Cell Res. Ther.* 6 (Sept. 2015), p. 153.

[45] Xuezhong Cai and Bryan R Cullen. “The imprinted H19 noncoding RNA is a primary microRNA precursor”. en. In: *RNA* 13.3 (Mar. 2007), pp. 313–316.

[46] Ying-Ping Gai et al. “A Novel LncRNA, , Associated With Environmental Stress in Mulberry ()”. en. In: *Front. Plant Sci.* 9 (May 2018), p. 669.

[47] Douglas M Anderson et al. "A micropeptide encoded by a putative long noncoding RNA regulates muscle performance". en. In: *Cell* 160.4 (Feb. 2015), pp. 595–606.

[48] Akinobu Matsumoto et al. "mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide". en. In: *Nature* 541.7636 (Jan. 2017), pp. 228–232.

[49] Ci Chu et al. "Systematic discovery of Xist RNA binding proteins". en. In: *Cell* 161.2 (Apr. 2015), pp. 404–416.

[50] Zhuomin Wu et al. "Regulation of lncRNA expression". en. In: *Cell. Mol. Biol. Lett.* 19.4 (Dec. 2014), pp. 561–575.

[51] Daniel Maticzka et al. "GraphProt: modeling binding preferences of RNA-binding proteins". en. In: *Genome Biol.* 15.1 (Jan. 2014), R17.

[52] Yuan-Ke Zhou et al. "Predicting lncRNA–Protein Interactions With miRNAs as Mediators in a Heterogeneous Network Model". In: *Frontiers in Genetics* 10 (2020), p. 1341. ISSN: 1664-8021. DOI: 10.3389/fgene.2019.01341. URL: https://www.frontiersin.org/article/10.3389/fgene.2019.01341.

[53] Sonika Tyagi et al. "CID-miRNA: A web server for prediction of novel miRNA precursors in human genome". In: *Biochemical and Biophysical Research Communications* 372.4 (2008), pp. 831–834. ISSN: 0006-291X. DOI: https://doi.org/10.1016/j.bbrc.2008.05.134. URL: http://www.sciencedirect.com/science/article/pii/S0006291X08010553.

[54] Xia Tang et al. "Transcriptomic Analysis of mRNA-lncRNA-miRNA Interactions in Hepatocellular Carcinoma". en. In: *Sci. Rep.* 9.1 (Nov. 2019), pp. 1–12.

[55] Marek Kazimierczyk et al. "Human Long Noncoding RNA Interactome: Detection, Characterization and Function". en. In: *Int. J. Mol. Sci.* 21.3 (Feb. 2020).

[56] Tsukasa Fukunaga and Michiaki Hamada. "RIblast: an ultrafast RNA-RNA interaction prediction system based on a seed-and-extension approach". en. In: *Bioinformatics* 33.17 (Sept. 2017), pp. 2666–2674.

[57] Martin Mann, Patrick R Wright, and Rolf Backofen. "IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions". en. In: *Nucleic Acids Res.* 45.W1 (July 2017), W435–W439.

[58] Sofia Quinodoz and Mitchell Guttman. *Long noncoding RNAs: an emerging link between gene regulation and nuclear organization.* 2014.

[59] Dong-Hwan Kim, Yanpeng Xi, and Sibum Sung. "Modular function of long noncoding RNA, COLDAIR, in the vernalization response". In: *PLOS Genetics* 13.7 (July 2017). Ed. by Richard M. Amasino, e1006939. DOI: 10.1371/journal.pgen.1006939. URL: https://doi.org/10.1371/journal.pgen.1006939.

[60] José M Santos-Pereira and Andrés Aguilera. "R loops: new modulators of genome dynamics and function". en. In: *Nat. Rev. Genet.* 16.10 (Sept. 2015), pp. 583–597.

[61] Piroon Jenjaroenpun et al. "QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences". en. In: *Nucleic Acids Res.* 43.W1 (July 2015), W527–34.

[62] Piroon Jenjaroenpun et al. "R-loopDB: a database for R-loop forming sequences (RLFS) and R-loops". en. In: *Nucleic Acids Res.* 45.D1 (Jan. 2017), pp. D119–D127.

[63] Fabian A Buske, John S Mattick, and Timothy L Bailey. "Potential in vivo roles of nucleic acid triple-helices". en. In: *RNA Biol.* 8.3 (May 2011), pp. 427–439.

[64] Yue Li, Junetha Syed, and Hiroshi Sugiyama. "RNA-DNA Triplex Formation by Long Non-coding RNAs". en. In: *Cell Chem Biol* 23.11 (Nov. 2016), pp. 1325–1333.

[65] Fabian A Buske et al. "Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci". en. In: *Bioinformatics* 29.15 (Aug. 2013), pp. 1895–1897.

[66] Sha He et al. "LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis". en. In: *Bioinformatics* 31.2 (Jan. 2015), pp. 178–186.

[67] Jennifer Harrow et al. "GENCODE: producing a reference annotation for ENCODE". en. In: *Genome Biol.* 7 Suppl 1 (Aug. 2006), S4.1–9.

[68] Jordan Ramilowski et al. "Functional Annotation of Human Long Non-Coding RNAs via Molecular Phenotyping". July 2019.

[69] The RNAcentral Consortium. "RNAcentral: a hub of information for non-coding RNA sequences". en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D221–D229.

[70] Zhenyu Bao et al. "LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases". en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D1034–D1037.

[71] Ying Yi et al. "RAID v2.0: an updated resource of RNA-associated interactions across organisms". en. In: *Nucleic Acids Res.* 45.D1 (Jan. 2017), pp. D115–D118.

[72] Jing Gong et al. "RISE: a database of RNA interactome from sequencing experiments". en. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D194–D201.

[73] Tsukasa Fukunaga et al. "LncRRIsearch: A Web Server for lncRNA-RNA Interaction Prediction Integrated With Tissue-Specific Expression and Subcellular Localization Data". en. In: *Front. Genet.* 10 (May 2019), p. 462.

[74] John L Rinn et al. "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs". en. In: *Cell* 129.7 (June 2007), pp. 1311–1323.

[75] Kashi Kaori et al. "Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1859.1 (Jan. 2016), pp. 3–15.

[76] Masaki Kato and Piero Carninci. “Genome-Wide Technologies to Study RNA-Chromatin Interactions”. en. In: *Noncoding RNA* 6.2 (May 2020).

[77] V E Velculescu et al. “Serial analysis of gene expression”. en. In: *Science* 270.5235 (Oct. 1995), pp. 484–487.

[78] Hazuki Takahashi et al. “5’ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing”. en. In: *Nat. Protoc.* 7.3 (Feb. 2012), pp. 542–561.

[79] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. en. In: *Nat. Rev. Genet.* 10.1 (Jan. 2009), pp. 57–63.

[80] Tim R Mercer et al. “Targeted sequencing for gene discovery and quantification using RNA CaptureSeq”. en. In: *Nat. Protoc.* 9.5 (Apr. 2014), pp. 989–1009.

[81] Michael Eisenstein. “Oxford Nanopore announcement sets sequencing sector abuzz”. en. In: *Nat. Biotechnol.* 30.4 (Apr. 2012), pp. 295–296.

[82] John Eid et al. “Real-time DNA sequencing from single polymerase molecules”. en. In: *Science* 323.5910 (Jan. 2009), pp. 133–138.

[83] Jillian J Goetz and Jeffrey M Trimarchi. “Transcriptome sequencing of single cells with Smart-Seq”. en. In: *Nat. Biotechnol.* 30.8 (Aug. 2012), pp. 763–765.

[84] Vipul Bhargava et al. “Quantitative transcriptomics using designed primer-based amplification”. en. In: *Sci. Rep.* 3 (2013), p. 1740.

[85] Yohei Sasagawa et al. “Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity”. en. In: *Genome Biol.* 14.4 (Apr. 2013), R31.

[86] Marcelo A German et al. “Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends”. en. In: *Nat. Biotechnol.* 26.8 (Aug. 2008), pp. 941–946.

[87] Brian D Gregory et al. “A link between RNA metabolism and silencing affecting Arabidopsis development”. en. In: *Dev. Cell* 14.6 (June 2008), pp. 854–866.

[88] Charles Addo-Quaye et al. “Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome”. en. In: *Curr. Biol.* 18.10 (May 2008), pp. 758–762.

[89] Vicent Pelechano, Wu Wei, and Lars M Steinmetz. “Extensive transcriptional heterogeneity revealed by isoform profiling”. en. In: *Nature* 497.7447 (May 2013), pp. 127–131.

[90] Irene M Min et al. “Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells”. en. In: *Genes Dev.* 25.7 (Apr. 2011), pp. 742–754.

[91] Naoto Imamachi et al. “BRIC-seq: a genome-wide approach for determining RNA stability in mammalian cells”. en. In: *Methods* 67.1 (May 2014), pp. 55–63.

[92] Robert B Darnell. "HITS-CLIP: panoramic views of protein-RNA regulation in living cells". en. In: *Wiley Interdiscip. Rev. RNA* 1.2 (Sept. 2010), pp. 266–286.

[93] Markus Hafner et al. "PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins". en. In: *J. Vis. Exp.* 41 (July 2010).

[94] Jernej Ule et al. "CLIP identifies Nova-regulated RNA networks in the brain". en. In: *Science* 302.5648 (Nov. 2003), pp. 1212–1215.

[95] Jing Zhao et al. "Genome-wide identification of polycomb-associated RNAs by RIP-seq". en. In: *Mol. Cell* 40.6 (Dec. 2010), pp. 939–953.

[96] Ci Chu, Jeffrey Quinn, and Howard Y Chang. "Chromatin isolation by RNA purification (ChIRP)". en. In: *J. Vis. Exp.* 61 (Mar. 2012).

[97] Jesse M Engreitz et al. "The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome". en. In: *Science* 341.6147 (Aug. 2013), p. 1237973.

[98] Matthew D Simon et al. "The genomic binding sites of a noncoding RNA". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.51 (Dec. 2011), pp. 20497–20502.

[99] Bharat Sridhar et al. "Systematic Mapping of RNA-Chromatin Interactions In Vivo". en. In: *Curr. Biol.* 27.4 (Feb. 2017), pp. 610–612.

[100] Weixin Wu et al. "Mapping RNA-chromatin interactions by sequencing with iMARGI". en. In: *Nat. Protoc.* 14.11 (Nov. 2019), pp. 3243–3272.

[101] Alessandro Bonetti et al. "RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions". en. In: *Nat. Commun.* 11.1 (Feb. 2020), p. 1018.

[102] Jesse M Engreitz et al. "RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites". en. In: *Cell* 159.1 (Sept. 2014), pp. 188–199.

[103] Grzegorz Kudla et al. "Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.24 (June 2011), pp. 10010–10015.

[104] Lei Kong et al. "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine". In: *Nucleic Acids Res.* 35.suppl_2 (July 2007), W345–W349.

[105] Liguo Wang et al. "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model". In: *Nucleic Acids Res.* 41.6 (Apr. 2013), e74–e74.

[106] Liang Sun et al. "Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts". In: *Nucleic Acids Res.* 41.17 (Sept. 2013), e166–e166.

[107] Aimin Li, Junying Zhang, and Zhongyin Zhou. "PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k- mer scheme". en. In: *BMC Bioinformatics* 15.1 (Sept. 2014), pp. 1–10.

[108] Lei Sun et al. "lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine". en. In: *PLoS One* 10.10 (Oct. 2015), e0139654.

[109] Rujira Achawanantakun et al. "LncRNA-ID: Long non-coding RNA IDentification using balanced random forests". In: *Bioinformatics* 31.24 (Dec. 2015), pp. 3897–3905.

[110] Xiao-Nan Fan and Shao-Wu Zhang. "lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning". en. In: *Mol. Biosyst.* 11.3 (Feb. 2015), pp. 892–897.

[111] Jian Zhao, Xiaofeng Song, and Kai Wang. "lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts". en. In: *Sci. Rep.* 6 (Oct. 2016), p. 34838.

[112] Rashmi Tripathi et al. *DeepLNC, a long non-coding RNA prediction tool using deep neural network*. 2016.

[113] Yu-Jian Kang et al. "CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features". In: *Nucleic Acids Res.* 45.W1 (July 2017), W12–W16.

[114] Long Hu et al. "COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features". In: *Nucleic Acids Res.* 45.1 (Jan. 2017), e2–e2.

[115] Valentin Wucher et al. "FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome". In: *Nucleic Acids Res.* 45.8 (May 2017), e57–e57.

[116] Cheng Yang et al. "LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning". In: *Bioinformatics* 34.22 (Nov. 2018), pp. 3825–3834.

[117] Junghwan Baek et al. "LncRNAnet: long non-coding RNA identification using deep learning". In: *Bioinformatics* 34.22 (Nov. 2018), pp. 3889–3897.

[118] Siyu Han et al. "LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property". en. In: *Brief. Bioinform.* 20.6 (Nov. 2019), pp. 2009–2027.

[119] Priya Ranganathan, C S Pramesh, and Rakesh Aggarwal. "Common pitfalls in statistical analysis: Logistic regression". en. In: *Perspect. Clin. Res.* 8.3 (July 2017), pp. 148–151.

[120] Hugo W Schneider et al. "A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts". en. In: *BMC Genomics* 18.1 (Oct. 2017), pp. 1–14.

[121] Zhongheng Zhang et al. "Opening the black box of neural networks: methods for interpreting neural network models in clinical applications". en. In: *Ann Transl Med* 6.11 (June 2018), p. 216.

[122] Tim R Mercer and John S Mattick. "Structure and function of long noncoding RNAs in epigenetic regulation". en. In: *Nat. Struct. Mol. Biol.* 20.3 (Mar. 2013), pp. 300–307.

[123] David H Mathews and Douglas H Turner. "Prediction of RNA secondary structure by free energy minimization". In: *Current Opinion in Structural Biology* 16.3 (2006). Nucleic acids/Sequences and topology, pp. 270–278. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2006.05.010. URL: http://www.sciencedirect.com/science/article/pii/S0959440X06000819.

[124] Hao Zhang et al. "A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming". en. In: *Front. Genet.* 10 (May 2019), p. 467.

[125] Damien Ulveling, Claire Francastel, and Florent Hubé. "Identification of potentially new bifunctional RNA based on genome-wide data-mining of alternative splicing events". en. In: *Biochimie* 93.11 (Nov. 2011), pp. 2024–2027.

[126] Damien Ulveling, Claire Francastel, and Florent Hubé. "When one is better than two: RNA with dual functions". en. In: *Biochimie* 93.4 (Apr. 2011), pp. 633–644.