Long Non-coding RNAs: Mechanisms, Experimental, and Computational Approaches in Identification, Characterization, and Their Biomarker Potential in Cancer

Anshika Chowdhary, Venkata Satagopam* and Reinhard Schneider

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Long non-coding RNAs are diverse class of non-coding RNA molecules >200 base pairs of length having various functions like gene regulation, dosage compensation, epigenetic regulation. Dysregulation and genomic variations of several lncRNAs have been implicated in several diseases. Their tissue and developmental specific expression are contributing factors for them to be viable indicators of physiological states of the cells. Here we present an comprehensive review the molecular mechanisms and functions, state of the art experimental and computational pipelines and challenges involved in the identification and functional annotation of lncRNAs and their prospects as biomarkers. We also illustrate the application of co-expression networks on the TCGA-LIHC dataset for putative functional predictions of lncRNAs having a therapeutic potential in Hepatocellular carcinoma (HCC).

Keywords: lncRNAs, cancer, biomarker, mechanisms, methods

INTRODUCTION

Advancement in Next Generation Sequencing (NGS) technologies and genome wide analysis of gene expression have revealed at least 80% of the human genome is active (Palazzo and Lee, 2015). However, only up to 1.5% of the genome is translated to protein which implicate RNAs to have more diverse roles than an intermediate component as templates in the genetic flow of information from gene to protein. They are categorized into mRNAs which are translated into proteins and non-coding RNAs (ncRNAs) which have little or no coding potential but are involved in transcriptional regulatory mechanisms.

The evolutionary development of an organism is associated with the increase in complexity of the regulatory potential of these ncRNAs which constitute the majority of the transcriptome. Non-coding RNAs are further categorised as short ncRNAs which include microRNAs (miRNAs), small RNA (sRNA), piwi-interacting RNAs (piRNAs), siRNAs, and long non-coding RNAs (lncRNAs) consisting long intergenic non-coding RNAs (lincRNAs), circular RNAs (circRNAs), and competitive endogenous RNAs (CeRNAs) (Hombach and Kretz, 2016). These RNAs have known to have functions involved in cellular functions like mRNA translation, alternative splicing events, RNA editing and also regulatory mechanisms like RNA silencing involving miRNA and mRNA interference via siRNA (Mattick and Makunin, 2006). LncRNAs have emerged as a latest class of RNA molecules which are more diverse than short ncRNAs having complex gene regulatory
functions in the cells. In this article we present and review the various biological characteristics and mechanisms of IncRNAs in transcriptional regulation and the latest development in experimental and computational methods for their identification, annotation and putative function prediction.

There are more than 30,000 IncRNAs in humans available in the GENCODE (Harrow et al., 2012), and more and more new IncRNAs are being discovered overtime. Long non-coding RNAs are typically longer than 200 nucleotides of length and sometimes have similar features to that of protein-coding genes, such as a 5' cap, exons and poly A tail and are spliced post-transcriptionally, but don’t possess functional open reading frames and cannot be translated to functional proteins. (Fang and Fullwood, 2016). Their varied molecular properties enable them to function in various methods regulating gene expression at various stages of cellular development (Hanahan and Weinberg, 2000).

IncRNAs are also not stable in comparison to mRNAs, localized mainly across the nucleus and cytoplasm and also not conserved across species, transcribed mostly by RNA polymerase II and exhibit tissue specific expression. However, high conservation patterns have been observed in the exonic regions and promoters regions of the IncRNA. Recently, it has been discovered that some IncRNAs can in fact translate to small peptide chains which could have significant biological functions (Hubé and Francastel, 2018; Li and Liu, 2019).

One way to classify IncRNAs is based on the genomic locations from where they are transcribed relative to protein coding genomic regions: (1) IncRNAs: long intergenic non-coding RNAs which are transcribed from the intergenic regions between the protein coding genes; (2) Sense IncRNAs: transcribed from the sense strand of the protein coding genes and may overlap with a part or the entire sequence of a protein coding gene; (3) Antisense IncRNAs: transcribed from the antisense strand of the protein coding genes which may overlap of exons, only from the intronic region and overlapping the entire gene in the antisense strand. (Ma et al., 2013); (4) Intronic IncRNAs: transcribed from the intronic regions between the exomes of a gene. (5) Bidirectional IncRNAs: transcribed from both sense and antisense directions of TSS (Hanahan and Weinberg, 2000; He et al., 2014, 2017).

FUNCTIONS AND MECHANISMS OF LONG NON-CODING RNAs

The elucidation of the mechanisms of long non-coding RNAs is mostly based on empirical evidence of the subcellular localization, developmental stage of the cell and tissue specific expression. The function of IncRNAs can be stratified into four types of molecular mechanisms described and illustrated (Figure 1; Zanella, 2021) below.

Signals

Transcriptional regulation aided by IncRNA where they function as signals are brought by various factors like developmental stages, organismal stress, re-programming of cells and state of the cell at a particular space and time in response to the environment and their expression could be a phenotypical indicator of these states (Wang and Chang, 2011). A prominent example is the chromatin regulation for dosage compensation in females in X-chromosome inactivation (XCI) (Engreitz et al., 2013; Wasko et al., 2019). The mechanism includes expression of XIST IncRNA from one of the X chromosome which coats itself leading to its silencing, which is also aided by the accumulation of the IncRNA Jpx. The antisense transcript of XIST, TSIX represses the activity of XIST in the other chromosome rendering it to be active (Starmer and Magnuson, 2009; Wang and Chang, 2011; Carmona et al., 2018). Another example of epigenetic re-programming that takes place in plants mediated by IncRNAs is to switch between vegetative to reproductive state. In Arabidopsis thaliana with the decrease in temperature for an extended period of time during winter COOLAIR is expressed and accumulated in large amounts which represses the expression of the FLOWERING LOCUS C (FLC). This is gene mediated by the PRC2 complex which when expressed normally in winter stops flowering in the plant. So, gradually upon the approach of spring and warmer temperatures COOLAIR enables vernalization of plants (Swiezewski et al., 2009; Tian et al., 2010; Heo and Sung, 2011; Wang and Chang, 2011).

Guides

As guides IncRNAs bind to proteins and direct them to specific sites, also leading to expression or silencing of the target genomic regions. This essentially involves recruitment of chromatin modifying enzymes which alter the chromatin state with the formation of complex structures with RNA-RNA, RNA-DNA, RNA-DNA-effector proteins. For instance, XIST transcription has also known to be induced by recruiting the Polycomb Repressive Complex 2 (PRC2) by RepA RNA. Additionally XIST also interacts with a matrix protein hnRNP U for its accumulation at the chromosome (Wang and Chang, 2011). Some other examples of IncRNAs acting as signals and guides include COLDAIR, HOTTIP, HOTAIR, ROR and some PRC2-bound RNAs (Rinn et al., 2007; Loewer et al., 2010; Wang et al., 2011; Kim et al., 2017).

Decoys

IncRNAs can also regulate transcription by acting as endogenous target mimics (eTMs) where they bind to intermediary regulatory proteins, RNA, DNA molecules and sequester them away from their respective target site. These otherwise known as competitive endogenous RNA (ceRNA) act as sponges generating a "sponge effect" by base pairing with target molecules which include transcription factors, miRNAs, chromatin modifiers (Wang and Chang, 2011) among others at their active sites and render them to be unavailable for interaction for their target molecules. An example of such activity is that of the IncRNA transcribed at the minor promoter of the DHFR gene which pairs and forms a complex with the DNA at the promoter region of the same gene. The complex inhibits formation of the preinitiation complex and also interacts with transcription factor II B (TFII B) which was also further confirmed by siRNA knockdown of the IncRNA (Martianov et al., 2007; Wang and Chang, 2011). MALAT1 (Tripathi et al., 2010), TERRA (Redon et al., 2010),
Gas5 (Kino et al., 2010) are also examples that exhibit the ‘sponge’/sequestering mechanism. ceRNA mechanism has been extensively studied with several computational algorithms and repositories also being developed in order to identify and store potential and experimentally verified targets of lncRNA (listed in Table 1). However, verification of their mechanism have to be contended with transcriptional levels of miRNA and lncRNA to be sufficient enough for them to function as competitive endogenous RNAs (Denzler et al., 2014, 2016; Zhang et al., 2019).

Scaffolds
LncRNAs serve as structural supports where other effector proteins and DNA/RNA molecules bind to form a functional complex and are then directed to appropriate localization of the complex for its function. Gene repression by HOTAIR forming a complex with the polycomb complex PRC2 for methylation at H3K27 (Rinn et al., 2007; Wang and Chang, 2011) and also forming a complex with LSD1, CoREST and REST (Wang and Chang, 2011) exhibits this mechanism. TERC also assembles the telomerase complex and mediates reverse transcriptase activity by binding with telomere targeting proteins (Balas and Johnson, 2018). The lncRNAs ANRIL (Yap et al., 2010; Kotake et al., 2011), SRP(Signal Recognition component), LINP1(LncRNA In Nonhomologous End Joining Pathway 1) (Sakthianandeswaran et al., 2016) are also found to have similar mechanisms.

IDENTIFICATION AND ANNOTATION
Experimental Approaches
Widely used experimental approaches to identify and annotate lncRNAs include Microarray, RNAseq, SAGE, CAGE among others with customized adaptations to identify and annotate lncRNAs based on their molecular characteristics as described in the following sections and listed in Table 2.

Adaptations in Microarray Technology
Probesets in conventional microarray platforms do not have lncRNAs annotations and not suitable for identifying and measuring lncRNA levels. Some of the mRNAs from these previous microarrays that have been correctly identified as lncRNAs have been re-annotated and their expression levels have been re-analyzed accordingly (Michelhaugh et al., 2011; Ma et al., 2012). ArrayStar Human LncRNA microarrays (V4.0) has been designed to profile both lncRNA and mRNA on the same array with 40,173 lncRNAs with 7,506 gold standard lncRNAs, 20,730 mRNAs among 60,903 distinct probes (Shi and Shang, 2016). As the expression of lncRNAs indicates the relative physiological state of a cell, differential expression between samples at different conditions can provide us information to understand the regulatory lncRNAs at these conditions. (Zhang et al., 2017) identified novel circulating lncRNAs: TINCR, CCAT2, AOC4P, BANCR, and LINC00857 which are differentially expressed in gastric cancer patients and be detected from the plasma of patients and hence function as biomarkers. Similarly, it was found that the lncRNA ENST00000551152 was upregulated and the lncRNA TCO.NS_00001368 was downregulated in cervical cancer cell lines (Huang et al., 2018) in a study by Huang et al using Agilent DNA microarray. Whole-genome tiling arrays are used for the sequenced regions which are not annotated for lncRNA isolation and identification. (Lund et al., 2014) used this in their experimental design where they used tiled probes from chr8: 127,640,000–129,120,000 at locus 8q24 to analyze prostate tissue from prostate cancer patients.
**TABLE 1** | Databases and computational pipelines predicting IncRNAs functioning as ceRNAs.

| Databases/Computational pipeline | Description | References |
|----------------------------------|-------------|------------|
| DIANA-LncBase v3                 | Database dedicated to cataloging miRNA and IncRNA interactions, includes ceRNABase | Karagkouni et al., 2020 |
| StarBase v2.0                    | RNA-RNA and protein-RNA interactions from CLIP-Seq experiments predicting ceRNA function | Li et al., 2014 |
| spongeScan                       | predicts miRNA target sites in IncRNAs | Furió-Tarí et al., 2016 |
| inCeDB                           | stores lncRNAs acting as ceRNAs with targets from StarBase and TargetScan | Grimson et al., 2007 |
| LncCeRBase                       | IncRNA-miRNA-mRNA interactions collected from literature | Das et al., 2014 |
| Linc2GO                          | predicts lncRNAs functions using miRNA and mRNA interactions based ceRNA hypothesis | Plan et al., 2018 |

**TABLE 2** | Experimental approaches in IncRNA profiling.

| Experimental approaches | Features |
|-------------------------|----------|
| RNA-seq                 | Identifies on novel IncRNA transcripts |
| Microarray              | Reannotations of existing microarrays |
| Tiling arrays           | Ability to profile transcriptome for specific regions/whole in the genome. |
| SAGE                    | Accurate quantification and novel transcript identification |
| CAGE                    | Identification of transcription start points |
| PARE, degradome-seq    | Used in RNA degradome analysis |
| GRO-seq                 | Measures nascent RNA regulating gene transcription |
| RIP, CLIP               | LncRNA-protein interaction identification |
| TIF-seq                 | LncRNA structure prediction |
| Selective 2'-hydroxyl acylation by primer extension (SHAPE) | |
| PARS                    | LncRNA structure prediction in vitro |
| FragSeq                 | Transcript structure prediction from RNA fragments |
| nextPARS                | Adaptation of PARS to Illumina technology |

**RNA-Seq Technologies**

RNA-seq is the most prevalent technique used to identify and annotate novel long non-coding transcripts that are less abundant including the isoforms of IncRNAs. RNA-seq offers a broad spectrum of transcript identification with novel transcripts detection and *de novo* assembly as probes are not required in order to hybridize and capture transcripts from samples. Modifications in the RNA-seq pipeline facilitate identification of specific type of IncRNAs, for instance strand-specific RNA-seq allows labeling of *origin of strand* information on the transcripts which allows sense/antisense IncRNA segregation and identification (Mills et al., 2013; Liu et al., 2019).

Wang et al. identified 2895 novel IncRNA in endometrial tissue of pigs; of which 301 were differentially expressed and functionally annotated to be involved in several biological pathways including immune system process and other cellular process of which TCONS_01729386 and TCONS_01325501 have a major functions in embryo pre-implantation (Liu et al., 2017). Functional attributes of IncRNA are validated with qRT-PCR experimental pipelines in which siRNA, GAPmers are designed to knockdown the IncRNA and the resulting change in gene expression is analyzed to identify its effector genes/molecules. However, in order for *in vitro* studies to correlate with *in vivo* studies several contributing factors involved in the knockdown of IncRNA and its effect on resulting varying gene expression need to be considered. Features of the IncRNA to consider while design of the knockdown strategy is the sub-cellular localization of the IncRNA, along with the developmental stage of the cells. Lennox et al. were able to decipher that nuclear IncRNAs were knocked down at higher levels using antisense strands and cytoplasmic IncRNAs were better knocked down using RNAi (Lennox and Behlke, 2016). In a recent study by Nicola Amod et al. a MALAT1-targeting 16mer LNA gapmeR g#5 showed significant anti-tumor activity in humanized murine model. Inference from transcriptome analysis showed proteasome expression was repressed by g#5 and was instead enriched increased *in vivo* in MALAT1 murine model patients (Amadio et al., 2018). RNA CaptureSeq (Mercer et al., 2011), another derivative of RNA-seq involves tiling arrays prepared for specific target regions of the genome. cDNAs against these regions are hybridized and sequenced. This method supports the identification of novel unannotated IncRNAs along with high fold coverage.

**SAGE, CAGE**

Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995) and Cap Analysis of Gene Expression (CAGE) are based on short sequences tags which are complementary to a given RNA of interest (Kashi et al., 2016). In SAGE these cDNA tags are biotinylated, captured on streptavidin beads (Wang and Chekanova, 2019). They are further ligated and later PCR amplified followed by concatenation and sequencing by mapping to reference genes. This method like RNA-seq facilitates discovery to novel transcripts and enables accurate measurement of expression levels of IncRNAs but has a drawback of small cDNA sequences mapping to multiple genes in the reference genome. Gibb et al. analyzed 272 SAGE libraries normal(26) and cancer(19) tissues from human which elucidated the tissue specific and aberrant expression IncRNAs in cancer tissues implicating them in disease development (Gibb et al., 2011). In a study by Jia et al. (2018) SAGE datasets of OPL(Oral premalignant lesions) from GEO were analyzed to identify 10 differentially expressed IncRNAs among with the IncRNA NEAT1 was the highly expressed in OPL. NEAT1 has been
also implicated in lung cancer metastasis and hepatocellular carcinoma (Dong et al., 2018).

Cap analysis gene expression (CAGE), was a development upon SAGE to overcome its drawbacks where cDNA tags can be generated from the 5’ end of the RNA of interest. The cap structure of the transcripts are biotinylated in the CAP-trapper method followed by cDNA tag generation, cleaving by restriction enzymes, PCR, ligation and cloning of tags and mapping to reference genome (Shiraki et al., 2003). CAGE allows the expression analyzes at promoter regions but is restricted only to capped RNAs. CAGE method has better throughput with the use of sequence tags and it is also cheap in comparison to cDNA library (Shiraki et al., 2003). Hon et al. (2017) collated 27,919 human lncRNAs from 1,829 datasets from CAGE and other methods in the FANTOM5 project. HelioscopeCAGE (Kanamori-Katayama et al., 2011) nanoCAGE (Poulain et al., 2017) CAGEscan (Bertin et al., 2017), DeepCAGE (Valen et al., 2009) are also protocols based on the CAGE technology for profiling the mammalian transcriptome.

Other Approaches

Parallel analysis of RNA-ends (PARE) (German et al., 2008), genome-wide mapping of uncapped transcripts (GMUCT) (Gregory et al., 2008), degradome-seq are among other techniques developed to map transcripts that are not stable and get degraded i.e., they act as templates for other non-coding RNAs like miRNA. RNA-seq measures transcripts at equilibrium conditions where as on the other hand Gro-seq (Global run-on sequencing) is able to sequence nascent RNA. This has revealed genome wide view of the transcripts by measuring half life of transcripts at various time points. RNA-seq and GRO-seq analyzes have revealed that divergent transcription occurs at the promoter regions of protein-coding genes (Kashi et al., 2016). 5’-bromo-uridine immunoprecipitation chase—deep sequencing analysis (bric-seq) method involves labeling of transcripts with 5’-bromo-uridine (BrU) which are isolated at sequential time intervals and recovered by immunopurification followed by RT-qPCR (Tani et al., 2012; Kashi et al., 2016). TIF-seq, an approach developed by Pelechano et al. (2013), jointly sequences both 5’ and 3’ ends of RNA molecules enabling characterization isoform heterogeneity of RNA molecules.

Other than perturbation by silencing of lncRNAs by RNA interference as mentioned in above section, functional characterization of lncRNA also involves methods like RNA centric purification methods when the RNA is pulled down exogenously based on in vitro affinity capture methods or endogenously under native or ultraviolet (UV) cross-linking conditions (Cipriano and Ballarino, 2018). On the other hand protein centric purification involves immunoprecipitation of lncRNAs and their target proteins with specific antibodies. RNA immunoprecipitation (RIP) is used to functionally characterize the lncRNA by purifying RNAs associated with target proteins. Cross-linking immunoprecipitation (CLIP), combination of CLIP with high-throughput sequencing (HITS-CLIP or CLIP-seq) and Photo Activatable Ribonucleotide-enhanced (PAR-CLIP) (Spitzer et al., 2014) been developed to analyze interactions of RNA binding proteins but these methods carry disadvantages like loss of cDNAs and de-cross-linking along with being expensive (Barra and Leucci, 2017). Chromatin isolation by RNA purification (ChIRP) has been used to identify lncRNAs and their interactions with chromatin during gene regulation (Chu et al., 2011; Kashi et al., 2016). Furthermore, techniques have been developed to probe the RNA structures, such as Selective 2’-hydroxyl acylation by primer extension (SHAPE) [67], parallel analysis of RNA structure (PARS) (Kertesz et al., 2010) and FragSeq (Underwood et al., 2010) which can provide an extensive evidence on mode of action and interactions with other regulatory molecules (Guo et al., 2016). More recently, Saus et al. described nextPARS an adaptation to PARS technique on the Illumina’s sequencing technology where parallel execution of highly specific enzymatic digestion of single an double stranded genomic regions make the “capable of tagging both the bases in single and double-stranded conformation at a genome-wide scale” making it cost effective with better throughput (Saus et al., 2018). CRISPRlnc, containing manually curated and validated 2184 CRISPR/Cas9 sgRNAs for 335 lncRNAs from different species, (Chen et al., 2019) was developed by Chen et al. which would further help design CRISPR/Cas9 experiments to investigate lncRNAs functions.

Computational Approaches

Novel Computational tools and pipelines are quintessential in combination with novel experimental techniques to identify putative transcripts as lncRNAs and further elucidate their functional roles involving interactions with other DNA, RNA and proteins. Computational pipelines to process NGS data are modified for the annotation of putative lncRNAs from novel transcripts. For the genome wide identification of lncRNA transcripts from data sets generated by the most widely RNA-seq techniques for novel lncRNA identification typically involves the following steps: alignment of reads from the experiment to the target regions in reference genome. This is followed by transcripts assembly and isoform identification and scoring the transcripts for protein coding potential (Coding Potential Calculator) (Jalali et al., 2015) and also include attributes like presence of open reading frames, poly-A tails and exonic regions and strand information into consideration. Standard programs like HISAT2, (Trapnell et al., 2009), STAR (Dobin and Gingeras, 2015) are used for mapping and StringTie (Ghosh and Chan, 2016), Scripture (Schoenbeck, 2016) for assembly. After transcripts of length >200 bp are filtered out, other types of transcripts such as tRNA, rRNA, snoRNA, miRNA, siRNA etc are searched against data repositories like UniProt, PDB and filter RNA transcripts which have similar homologous domains and can be translated to proteins (Gish and States, 1993; Eddy, 2011; Jalali et al., 2015). On comparing the performance of various alignment methods (Zheng et al., 2019) Kallisto or Salmon in combination with full transcriptome annotation performed best
for lncRNA detection on both un-stranded and stranded RNA-Seq datasets.

ORF is also among the features which help categorization of novel transcripts as lncRNAs; for example ORF length predicted by EMBOSS tools (Itaya et al., 2013) (getORF). ORFs of length greater than 100 codons categorised as mRNAs are filtered out as coding transcripts but it is not a definite threshold with certain exceptions like XIST, H19 among others which having ORFs longer than 100 amino acids (Dinger et al., 2008; Jalali et al., 2015).

Another approach is use of machine learning based tools developed on SVM, logistic regression models use sequence features to compute the protein coding potential which predict the transcript to be a lncRNA/mRNA. ORF, conservation of the exonic regions of the transcript, nucleotide composition, sequence motif and codon usage are inclusive feature vectors from the transcript sequences to train the models. In order to compute transcript’s coding potential two methods have been developed CPC (Coding Potential Calculator) (Altschul et al., 1997; Kong et al., 2007; Ma et al., 2012) based on SVM models with sequence features and the comparative genomics features and ii) A later faster version CPC2 that can be for novel transcripts of organism which have improper genome assembly and poorly annotated (Kang et al., 2017). CONC (for coding and non-coding) (Liu et al., 2006) also trains SVM models based on a comprehensive set of RNA features like the peptide length and composition, secondary structure, compositional entropy among others to classify transcripts as lncRNAs and mRNAs. Lu et al. have further integrated quantitative properties like a GC content, conservation patterns, level of expression which is lower of lncRNAs in comparison to mRNAs to predict lncRNAs in *C. elegans* in their machine learning model (Lu et al., 2011; Ma et al., 2012). The pipeline employed by Sun et al. IncRScan-SVM (Sun et al., 2015), which after a standard processing of RNASeq transcripts identifies transcripts as lncRNAs by a SVM model trained on GTF positive and negative samples. iSeeRNA is also a similar tool that identifies putative lncRNAs by using SVM based classifier (Sun et al., 2013). CONE, a coding potential calculator, developed by Hu et al. (2017) integrated multiple features from both sequences and experiments like poly(A) enrichment, methylation taken from RNA-seq data sets had more accuracy over transcripts of different lengths. In the CONE method, an index for the whole genome splitting it into bins of 100-nucleotide(nt) on which the feature vectors were generated and subsequently a balanced random forest (BRF) was trained.

Attempts to functionally characterize novel lncRNAs by computational methods have been challenging. In the case of protein-coding genes a putative function is assigned to transcripts based on their similarity with already characterized proteins (de Hoon et al., 2015); as they have highly conserved regions across species which is not the same with lncRNAs. Their tissue specificity and low abundance along with varied mechanisms involved with various other biological molecules further add to the complexity of modeling their functionality *in-silico.*

### Co-expression Evidence Analysis and Network Inference

Data analysis of microarrays and tiling experiments include identification of differential expressed transcripts followed by network analysis based on co-expression patterns. To infer the putative function of a lncRNA ‘guilt by association’ algorithm has been developed based on the co-expression patterns of lncRNA and protein coding genes (PCGs) which suggest their functional relatedness and regulatory relationships. The tissue and condition specific expression, subcellular localization are distinctive attributes of lncRNA expression which are combined with differential expression to infer putative functions and target proteins interactions of the lncRNA and their role in disease development (Li et al., 2016; Gao et al., 2019a). The correlation scores between expression profiles of lncRNAs and PCGs at a given condition/tissue/time series are calculated which represents a network by a transformed correlation-adjacency matrix. From these networks, clusters of co-expressed lncRNAs and mRNAs are identified. The functional regulation of lncRNAs are annotated based on the functional enrichment of the PCGs in the clusters with which it is co-expressed.

Co-ncRNA is one such tool/database developed by Wu et al. (2016) where they were able to analyze lncRNA-mRNA co-expression patterns, consistent with previous established related lncRNA-mRNAs like HOTAIR, BCRA2, MMP9 and MMP11 and also novel lncRNA RP11-118E18 validated by TANRIC. Such network based clustering approaches have also been further extended to include other non-coding RNAs and regulatory proteins like miRNAs to predict more specific mechanisms like cis-regulatory relationships where whole transcriptomic data is analyzed (Singal et al., 2016).

Several studies have been done to understand the pathogenesis of complex diseases from available data of lncRNA and their interacting proteins (Sumathipala et al., 2019). The approaches consist of Machine learning (ML) based models trained over expression profiles to extract patterns from which lncRNA functionality and disease associations are predicted, random walk based models on networks representing the similar expression patterns or a combination of both. (Chen and Yan, 2013) included disease information into identify lncRNA disease associations from lncRNA expression levels by developing a semi-supervised learning model Laplacian Regularized Least Squares for LncRNA Disease Association (LRLSLDA).

Chen et al. further developed novel lncRNA functional similarity calculation models (LNCSIM) by associating the semantic similarity between lncRNA and disease groups (Chen and Yan, 2013; Chen, 2015a,b). Guo et al. (2019) developed LDASR to identify lncRNA-disease associations where Guassian profile similarities and neural network for dimensional reduction and finally rotating forests were used to predict disease associations (Guo et al., 2019). DishmRF also uses random forest models trained over lncRNA-disease associated protein coding genes in order to score the association of lncRNA for a particular disease (Pan et al., 2019). Liao et al. developed a method called GrwLDA which is based on global network
random walk model in order to predict IncRNA and their associated diseases (Gu et al., 2017). Xuan et al. (2019) also recently proposed a tool graph convolutional network and convolutional neural network (GCN/LDA) to explore network and come up with IncRNA-disease candidate pairs. Bipartite Network inference (LPBNI), a computational pipeline developed by Ge et al. (2016) used two-step propagation in the bipartite network to rank target proteins for IncRNAs; BPLLDA developed to predict IncRNA-disease links from a network of heterogenous IncRNAs and associated diseases based on the node interaction paths (Xiao et al., 2018). TPGLDA also had been developed to predict IncRNA-disease from IncRNA-disease-gene tripartite graph constructed base on was developed by Ding et al. (2018) where they could predict IncRNAs like GAS5, UCA1, implicated in lung, hepatocellular, ovarian cancer (Ding et al., 2018). The above mentioned tools are all based on network propagation and inference. Recently, a similar network diffusion algorithm called LION was developed to infer key candidate IncRNAs (Sumathipala et al., 2019) by Sumathipala et al. with better prediction results for cardiovascular diseases and cancer. Another recent approach IDHI-MIRW by Integrating Diverse Heterogeneous Information(IDHI) with positive pointwise Mutual Information and Random Walk(MIRW) was also proposed by Fan et al. (2019) which integrates IncRNA-miRNA/protein and expression profiles along with disease ontology information.

Conservation and Structure Prediction

Although the conservation scores of IncRNA molecules are lower than mRNA, when used within awareness of biological context including information about potential interactions with other RNA, DNA, proteins, can decipher evidences to categorise novel transcripts to IncRNAs. Algorithms like BLAST (Altschul et al., 1990), ClustalW (Thompson et al., 2002), MAFFT (Katoh et al., 2009), ConSurf (Glaser et al., 2003), MUSCLE (Edgar, 2004) among others perform multiple sequence alignment. Furthermore tools like RNAz 2.0 (Gruber et al., 2010), Evofold (Pedersen et al., 2006) can predict conserved RNA structures from multiple sequence alignment. RNAstructure (Reuter and Mathews, 2010), GTFold (Svensson et al., 2012), CentroidFold (Sato et al., 2009), RNAfold (Denman, 1993), Mfold (Zuker, 2003), CentroidHmfold-LAST (Hamada et al., 2011), and Seqfold (Ouyang et al., 2013), FARNA (Alam et al., 2017), iFoldRNA (Sharma et al., 2008) are among the tools to predict RNA secondary and tertiary structures, respectively from primary sequence. The RNA–RNA interaction prediction methods mainly employ alignment algorithms, comparative (homology) methods and in silico energy calculations (Um and Gardner, 2017). Minimum Free Energy based methods are based on computation of the minimum free energy of the RNA-RNA molecules taking the inter- and/or intra molecular base-pairing into account. On the other hand, as perceivable, alignment and homology based methods include algorithms using tools for multiple sequence alignment and seed match-extension.

IntaRNA (Mann et al., 2017), RNAhybrid (Krüger and Rehmsmeier, 2006), Pairfold (Andronescu et al., 2003), RNApLex (Tafer and Hofacker, 2008), RSeach (Wenzel et al., 2012), RIBlast (Fukunaga and Hamada, 2017), Bindigo (Hodas and Aalberts, 2004), and GUUGle (Gerlach and Giegerich, 2006) are some examples of tools used to predict RNA-RNA interactions. These are also integrated in pipelines to predict IncRNA-RNA interactions in humans. For instance, (Terai et al., 2016), developed a pipeline using RACCESS (Kiryu et al., 2011) to extract accessible regions from RNA molecules followed by masking tandem repeats using TanTan (Frith, 2011) and finding seed match using LAST and then calculate the interaction energy between two RNA molecules using IntaRNA and finally predict the joint secondary structure (RactIP) (Kato et al., 2010) to predict IncRNA-mRNA interactions (Szczęśniak and Makalowska, 2016) proposed a similarity based method to predict RNA-RNA interactions using LAST (Kiobel et al., 2011), miRanda (Betel et al., 2010) tools in some pipelines. Similarly, RNA-protein interactions are also be predicted from sequence based methods which use physiochemical properties of amino/nucleic acids in tools like IncPRO (Lu et al., 2013) and catRAPID (Bellucci et al., 2011). Along with these sequence features, secondary structures of RNA are inl tools like RPI-PreD (Suresh et al., 2015). PARIS (Lu et al., 2016), SPLASH (Aw et al., 2016), LIGR-seq (Sharma et al., 2016), and MARIO (Nguyen et al., 2016) to identify RNA-RNA interactions based on proximity ligation in vivo (Fukunaga and Hamada, 2017).

LncRNA DATABASES

The publicly available datasets from RNA-seq and microarray experiments have led to rapid increase of annotated lncRNAs with dedicated databases for IncRNA and their molecular and disease associations. Many pipelines and tools have been benchmarked from the data available from these knowledge bases. NONCODEv5, the largest database for noncoding RNAs (majorly IncRNAs) contains 548,640 lncRNA transcripts from several model organisms (Fang et al., 2018), of which 96,308 IncRNA genes are from humans. The data has been curated from published literature and annotated with information from public resources like RefSeq, Ensembl, GenBank, IncRNAdb, Incpedia. The FANTOM (Functional ANnoTation Of the Mammalian genome) consortium led by RIKEN has systematically investigated and annotated about 27,919 human IncRNA genes across 1829 samples in the FANTOM database (FANTOM5) (Abougesaisa et al., 2017). Some of the databases provide experimentally validated and/or computationally predicted interactions of lncRNAs with other RNA and proteins. Analysis of data from RNA-seq and microarray experiments on disease cell lines have also helped in discovery of the roles lncRNA in disease mechanisms which have been recorded in disease-association databases. For instance LNCipedia provides lncRNA from humans with experimental and putative annotations along with miRNA-lncRNAs associations (Volders et al., 2013). Similarly, IncRNAdb is repository for functionally annotated IncRNAs along with TF-lncRNA associations. LncRNome, a lncRNA database for human compiled form GENCODE has lncRNAs with annotations of their biomolecular interactions and disease associations. LncATLAS provides information on lncRNA localization in cells from RNA-sequencing data, from GENCODE (Mas-Ponte et al., 2017), Lnc2CAncer has 1,488 entries of lncRNAs from mRNA.
experimentally supported validations which are associated with cancer (Ning et al., 2016). Table 3 contains a list of databases and their references.

**CASE STUDY: CO-EXPRESS**

**NETWORK ANALYSIS IDENTIFYING PRO-INFLAMMATORY LncRNAs IMPLICATED IN HCC**

Cancer is caused by continuous accumulation of unfavourable genetic alterations that cause deregulation of genetic networks and cellular pathways ultimately (Huarte, 2015) leading to unceasing growth of cells and tissue. The mechanisms of these dysregulations are complex, involving altered gene expressions and molecular interactions which are yet to be discovered comprehensively; thus leading to the necessity to analyze the anomalies at all omics levels. In fact, LncRNAs are diversely associated in most of the hallmarks of cancer. Many of the studies on cancer associated LncRNAs have mainly analyzed expression profile variations of LncRNA in cancer vs. healthy tissue and its effects on deregulated pathways and identification their regulatory targets. Also, approaches to identify RNA folding and stable complexes to evaluate LncRNA functions have depicted that genetic alterations like SNPs can also majorly impact the RNA structure and eventually their function with changes in active/binding sites of LncRNAs (Wan et al., 2014; Schmitt and Chang, 2016). Chronic inflammation has known be a vital in

| Table 3 | Overview of databases of LncRNAs. |
|-------------------------------------|-----------------------------------|
| **Database**                        | **Description**                   | **References** |
| NONCODE v5                          | Knowledge base for ncRNAs         | Fang et al., 2018 |
| LNCipedia                           | lncRNA with secondary structure prediction, protein coding potential and microRNA binding sites | Volders et al., 2013 |
| IncRNAIdb v2.0                      | Manually curated IncRNAs from literature | Quek et al., 2015 |
| LncATLAS                            | IncRNA annotated with subcellular localisation | Mas-Ponte et al., 2017 |
| IncRNAAdisease 2.0                  | Experimentally supported IncRNA disease association and molecular targets | Bao et al., 2019 |
| LncRBase                            | IncRNA with information about their subtypes and interactions | Chakraborty et al., 2014 |
| IncRNAome                           | IncRNA with interactions with other RNAs | Bhartya et al., 2013 |
| GreeNC v1.1.2                       | Database for plant IncRNAs        | Paytuvi Gallart et al., 2016 |
| Lnc2Cancer v2.0                     | Manually curated database with experimentally supported IncRNA-cancer associations | Gao et al., 2019b |
| EVlncRNAs                           | Manually curated database validated with low-throughput experiments | Zhou et al., 2018 |
| ChiPBase v2.0                       | IncRNAs and other ncRNA from ChiP seq data | Zhou et al., 2017 |
| DIANA-LncBase v3                    | Database dedicated to cataloging miRNA and IncRNA interactions | Karagkouni et al., 2020 |
| LNCediting                          | Information of IncRNA editing, its impact and interactions with miRNAs | Gong et al., 2017 |
| TCLA                                | Cancer LncRNome Atlas: IncRNAs predicted from TCGA datasets | Yan et al., 2015 |
| MNDR v2.0                           | Experimental and predicted ncRNA-disease associations | Cui et al., 2018 |
| IncRNAASN2P2                       | IncRNA variants and their disease associations | Miao et al., 2018 |
| Lnc2Meth                            | Manually curated database of regulatory relationships between long non-coding RNAs and DNA methylation associated with human disease | Zhi et al., 2018 |
| DES-ncRNA                           | Database of human miRNA and IncRNA from literature | Sahi et al., 2017 |
| LincSNP2.0                          | disease associated SNPs with lncRNAs | Ning et al., 2017 |
| LncVar                              | IncRNAs with associated genetic variations | Chen et al., 2017 |
| deepBase v2.0                       | ncRNA database from deep sequencing data | Zheng et al., 2016 |
| C-It-ioci                           | Tissue specific transcriptome data (protein coding genes and ncRNA) | Weirick et al., 2015 |
| LncRNA2Target v2.0                  | IncRNA and IncRNA-to-target genes after IncRNA knockdown and over expression | Cheng et al., 2019 |
| LncTarD                             | Manually curated database of IncRNAs and target regulations | Zhao et al., 2020 |
| CRlncRNA                            | Cancer related lncRNAs along with associations and interactions | Wang et al., 2018 |
| IncRNA4KB                           | Cancer related lncRNAs along with associations and interactions | Sefufudin et al., 2020 |
| Cancer LncRNA                       | IncRNAs from GENCODE involved in cancer | Carlevaro-Fita et al., 2020 |
| Census (CLC)                        | lncRNAs from GENCODE involved in cancer | Carlevaro-Fita et al., 2020 |

| Table 4 | Details of datasets used in the case study. |
|---------|-------------------------------------------|
| **Dataset** | **Project** | **Number of samples** | **Number of modules** | **Data source** |
| RNA-Seq expression data HCC         | TCGA-LIHC | 372 | 27 | Tomczak et al., 2015 |
| RNA-Seq expression data NAT         | TCGA-LIHC | 50 | 76 | Tomczak et al., 2015 |
| RNA-Seq expression from liver       | GTEx | 208 | 43 | Lonsdale et al., 2013 |
cancer progression in case of Hepatocellular carcinoma (HCC). Some of the pathways known to be chronically upregulated causing hepatoma cell proliferation include JAK/STAT signalling, NF-Kappa B signalling, PI3K/AKT/mTOR pathway, WNT pathway, and MAPK pathway (Chen et al., 2018; Yang et al., 2019). In order to investigate the application of co-expression network based on the “guilt by association” principle analysis of RNA-seq data, we applied the Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008) on the following datasets: The RNA-Seq dataset from The Cancer Genome Atlas (Tomczak et al., 2015) Liver Hepatocellular Carcinoma (TCGA-LIHC) project and the GTEx dataset (Lonsdale et al., 2013) (Table 4) samples to identify the pathways dysregulated in HCC with regards to chronic inflammation in HCC progression. The steps in the pipeline are illustrated in Figure 2. The datasets were collected and analyzed using the TCGAbiolinks, WGCNA packages in R.

WGCNA analysis consists of the following steps: correlations across the normalized expression values of the samples are computed and raised to a soft threshold power based on the scale free topology criterion generating an adjacency matrix representing the co-expression network. This is followed by hierarchical clustering is used to identify clusters of co-expressed lncRNAs and protein coding genes among the network, each of which is labeled with a color/number. Co-expression Network using WGCNA was generated across all the 3 datasets and modules obtained in each case were enriched for functional process by cluster profiler. The modules which were identified for pathways dysregulated in case of HCC were selected and the lncRNAs which were highly connected, i.e., being significant for each module were identified for having biomarker prognostic potential.

For HCC, NAT and GTEx profiles 27, 76 and 43 modules were identified, respectively from the hierarchical clustering with the cut height being selected 0.99, 0.98, 0.98 (Figure 3), respectively. These includes all the PCGs and lncRNAs transcripts. Each module was labeled with a color allocated by the WGCNA function and were enriched for KEGG pathways with threshold \( p < 0.05 \). The red, yellow in TCGA-HCC dataset and turquoise, green modules in TCGA-NAT dataset were enriched for the pathways involved in inflammation including JAK-STAT signalling pathway, cytokine-cytokine receptor interaction, NF-kappa B signaling pathway, T cell receptor signaling pathway among others contributing in inflammatory response. The network properties of all the networks were calculated based on which the transcripts in these modules were sorted according to their connectivity. The top highly connected lncRNAs(top 10) putatively having important regulatory mechanisms in these modules were selected for having biomarker potential in regards to chronic inflammation both in the tumour and its surrounding micro environment proceeding to NAT. The common lncRNAs among the both phenotypes across other TCGA datasets (HCC and NAT) but not the GTEx dataset. This could be attributed to the batch effects of the RNA-Seq experiments across the GTEx and TCGA projects.

For HCC, NAT and GTEx profiles 27, 76 and 43 modules were identified, respectively from the hierarchical clustering with the cut height being selected 0.99, 0.98, 0.98 (Figure 3), respectively. These includes all the PCGs and lncRNAs transcripts. Each module was labeled with a color allocated by the WGCNA function and were enriched for KEGG pathways with threshold \( p < 0.05 \). The red, yellow in TCGA-HCC dataset and turquoise, green modules in TCGA-NAT dataset were enriched for the pathways involved in inflammation including JAK-STAT signalling pathway, cytokine-cytokine receptor interaction, NF-kappa B signaling pathway, T cell receptor signaling pathway among others contributing in inflammatory response. The network properties of all the networks were calculated based on which the transcripts in these modules were sorted according to their connectivity. The top highly connected lncRNAs(top 10) putatively having important regulatory mechanisms in these modules were selected for having biomarker potential in regards to chronic inflammation both in the tumour and its surrounding micro environment proceeding to NAT. The common lncRNAs among the both phenotypes across these modules were PCED1B-ASI, TRG-ASI, MIR155HG, MIAT, LINC00996. MIAT has been known to be implicated in several cancers such as breast cancer, gastrointestinal cancer and NSCLC and also its silencing has known to inhibit cell proliferation and tumorogenesis in HCC (Zhao et al., 2019). In a recent study by Peng et al. (2020) it has been postulated that MIAT regulates the expression of JAK2 among other genes and has an important role in controlling the tumour microenvironment in HCC.

LINC00996 has also been known to have regulatory mechanism in the JAK-STAT signalling pathway in colorectal cancer in a study by Ge et al. (2018). These pathways are dysregulated in the case of HCC as seen in the clusters from the TCGA datasets (HCC and NAT) but not the GTEx dataset. This provides us with corroboration pointing that NAT is subjected to an inflammatory environment prompted by the malignant tissue. This is similar to micro tumour environment with higher proliferation rate than a healthy hepatocyte. Identification of these modules and LncRNAs provides extended empirical evidence of lncRNA regulation in inflammation and pertaining to cancer progression. This analysis provides support to the “guilt by association” hypothesis of co-expression of lncRNAs with the genes involving in similar functions. However, few of the lncRNAs like MEG3, MALAT1, H19, UCA1 which have been studied for their implications in HCC didn’t show an expression in the GTEx greater the variance threshold and could not be characterized in the co-expression networks while comparing to the TCGA datasets. This could be attributed to the batch effects of the RNA-Seq experiments across the GTEx and TCGA projects.
which can be addressed and corrected while pre-processing the raw reads together from all the datasets. The understanding of such complex networks in which dysregulation of IncRNAs occurs impacting cancer progression and metastasis, which also being tissue specific can set IncRNAs to become excellent biomarkers in cancer therapy (Schmitt and Chang, 2016).

CONCLUDING REMARKS

The recent discovery of the IncRNAs in the non-coding genome has led to a paradigm shift in the understanding of the mechanism of information flow from the genetic code and the genotype-phenotype map. But, as discussed, the mechanisms in which IncRNAs functions are very complex involving interactions with various molecules from other 'omic' levels. Advancements in RNA technologies have helped to elucidate some of the diverse mechanisms of IncRNAs but the regulatory potential of the majority of these noncoding genes have yet to be discovered. Differential co-expression of IncRNAs, RNA secondary structure and sequence analysis and prediction, ML based approaches in computational pipelines have aided in the identification and characterization of IncRNAs from RNA-seq experiments. This has to be supported by experimental validations and clarifications on cis-trans regulatory processes. Genome wide transcriptome profiling has identified several IncRNAs which have significant roles in diseases like cancer exhibiting cell- and/or tissue/tumor-specific expression and hence can be excellent candidate targets for therapy. It has been demonstrated that silencing of certain disease associated IncRNAs exhibited tumor suppression. In summary, a comprehensive knowledge of IncRNAs shall provide researchers insights into genotype-phenotype distinction and genetic disorders leading to more effective therapeutic strategies for diseases and with emergence of new experimental designs and computational pipelines we can advance our understanding of the transcriptome.

AUTHOR CONTRIBUTIONS

VS and RS: original idea for the manuscript, contributed to design and conceptualization of the study, and supervision. AC: literature, data analysis, and writing initial draft. VS: writing, review, and editing. RS: project administration and funding acquisition. All authors also critically reviewed, wrote, and approved the final version.

FUNDING

This work was supported by an IRP grant of the University of Luxembourg to Iris Behrmann and Reinhard Schneider (RS) (IL6longliv).

REFERENCES

Abuagassisa, I., Noguchi, S., Carninci, P., and Kasukawa, T. (2017). The FANTOM5 Computation Ecosystem: Genomic Information Hub for Promoters and Active Enhancers. Methods Mol. Biol. 1611, 199–217. doi: 10.1007/978-1-4939-7015-5_15

Alam, T., Uludag, M., Essack, M., Salhi, A., Ashoor, H., Hanks, J. B., et al. (2017). FARNA: knowledgebase of inferred functions of non-coding RNA transcripts. Nucleic Acids Res. 45, 2838–2848. doi: 10.1093/nar/gkw973

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Amodio, N., Stamat, M. A., Juli, G., Morelli, E., Fulciniti, M., Manzoni, M., et al. (2018). Drugging the IncRNA MALAT1 via LNA gapmeR ASO inhibits gene expression of proteasome subunits and triggers anti-multiple myeloma activity. Leukemia. 32, 1948–1957. doi: 10.1038/s41375-018-0067-3
Guo, C., Liao, B., Li, X., Cai, L., Li, Z., Li, K., et al. (2017). Global network random walk for predicting potential human lncRNA-disease associations. Sci. Rep. 7, 12442. doi: 10.1038/s41598-017-12763-z

Guo, X., Gao, L., Wang, Y., Chiu, D. K., Wang, T., and Deng, Y. (2016). Advances in long noncoding RNAs: identification, structure prediction and function annotation. Brief Funct. Genomics 15, 38–46. doi: 10.1093/bfgp/elv022

Guo, Z. H., You, Z. H., Wang, Y. B., Yi, H. C., and Chen, Z. H. (2019). A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest. Sciencing 19, 786–795. doi: 10.3109/1566506X.2019.1608688

Hamada, M., Yamada, K., Sato, K., Frith, M. C., and Asai, K. (2011). CentroidHomfold-LAST: accurate prediction of RNA secondary structure using automatically collected homologous sequences. Nucleic Acids Res. 39, W100–106. doi: 10.1093/nar/gkr290

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. Cell 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE project. Genome Res. 22, 1760–1774. doi: 10.1101/gr.135550.111

He, X., Ou, C., Xiao, Y., Han, Q., Li, H., and Zhou, S. (2017). LncRNAs key players and novel insights into diabetes mellitus. Oncotarget 8, 7125–7134. doi: 10.18632/oncotarget.19921

He, Y., Meng, X. M., Huang, C., Wu, B. M., Zhang, L., Lv, X. W., et al. (2014). Long noncoding RNAs: Novel insights into hepatocellular carcinoma. Cancer Lett. 344, 20–27. doi: 10.1016/j.canlet.2013.10.021

Heo, J. B., and Sung, S. (2011). Vascularization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331, 76–79. doi: 10.1126/science.1197349

Hodas, N. O., and Aalberts, D. P. (2004). Efficient computation of optimal oligo-RNA binding. Nucleic Acids Res. 32, 6636–6642. doi: 10.1093/nar/gkh1008

Hombach, S., and Kretz, M. (2016). Non-coding RNAs: classification, biology and functioning. Adv. Exp. Med. Biol. 937, 3–17. doi: 10.1007/978-3-319-42059-2_1

Hon, C. C., Ramlowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5’ ends. Nature 543, 199–204. doi: 10.1038/nature21374

Hu, L., Xu, Z., Hu, B., and Lu, Z. J. (2017). COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. Nucleic Acids Res. 45, e2. doi: 10.1093/nar/gkw798

Huang, J., Liu, T., Shang, C., Zhao, Y., Wang, W., Liang, Y., et al. (2018). Identification of lncRNAs by microarray analysis reveals the potential role of lncRNAs in cervical cancer pathogenesis. Oncol. Lett. 15, 5584–5592. doi: 10.3892/ol.2018.8037

Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat. Med. 21, 1253–1261. doi: 10.1038/nm.3981

Hubé, F., and Francazelt, C. (2018). Coding and Non-coding RNAs, the Frontier Has Never Been So Blurred. Front. Genet. 9:140. doi: 10.3389/fgene.2018.00140

Iwaya, H., Oshita, K., Arakawa, K., and Tomita, M. (2013). GEMBASSY: an EMBOSS associated software package for comprehensive genome analyses. Source Code Biol. Med. 8:17. doi: 10.1186/1751-0473-8-17

Jalali, S., Kapoor, S., Sivadas, A., Bhatiaya, D., and Scaria, V. (2015). Computational approaches towards understanding human long non-coding RNA biology. Bioinformatics 31, 2241–2251. doi: 10.1093/bioinformatics/btv148

Jia, H., Wang, X., and Sun, Z. (2018). Exploring the molecular pathogenesis and biomarkers of high risk oral premalignant lesions on the basis of long noncoding RNA expression profiling by serial analysis of gene expression. Eur. J. Cancer Prev. 27, 370–378. doi: 10.1097/CJE.0000000000000346

Kamamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., et al. (2011). Unsupervised cap analysis of gene expression on a single-molecule sequencer. Genome Res. 21, 1150–1159. doi: 10.1101/gr.115469.110

Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45, W12–W16. doi: 10.1093/nar/gkx428

Karagkouni, D., Paraskevopoulou, M. D., Tastoglu, S., Skoufos, G., Karavangeli, A., Pierros, V., et al. (2020). DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. Nucleic Acids Res. 48, D101–D110. doi: 10.1093/nar/gkz1036

Kashi, K., Henderson, L., Bonetti, A., and Carrincli, P. (2016). Discovery and functional analysis of lncRNAs: methodologies to investigate an uncharacterized transcriptome. Biochim. Biophys. Acta 1859, 3–15. doi: 10.1016/j.bbagrm.2015.10.010

Kato, Y., Sato, K., Hamada, M., Watanabe, Y., Asai, K., and Akutsu, T. (2010). RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. Bioinformatics 26, 1460–1466. doi: 10.1093/bioinformatics/btq372

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. 537, 59–64. doi: 10.1007/978-1-59745-251-9_3

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., et al. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature 467, 103–107. doi: 10.1038/nature09322

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. 21, 487–493. doi: 10.1101/gr.113985.110

Kim, D. H., Xi, Y., and Sung, S. (2017). Modular function of long noncoding RNA, GOLDAIR in the vernalization response. PLoS Genet. 13:e1006939. doi: 10.1371/journal.pgen.1006939

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., and Chrousos, G. P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci. Signal 3, ra8. doi: 10.1126/scisignal.200568

Kiryu, H., Terai, G., Imamura, O., Yoneyama, H., Suzuki, K., and Asai, K. (2011). A detailed investigation of accessibility around target sites of siRNAs and miRNAs. Bioinformatics 27, 1778–1779. doi: 10.1093/bioinformatics/btr276

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2010). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 38, W345–W349. doi: 10.1093/nar/gkm390

Kotake, Y., Nakagawa, T., Kitagawa, K., Liu, N., Kitagawa, M., et al. (2011). Long non-coding RNA ANRIL is required for the PRC2 recruitment to
Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., et al. (2019). IncRNA miat functions as a ceRNA to upregulate sirt1 by sponging miR-22-3p in HCC: cellular senescence. Aging (Albany NY) 11, 7098–7122. doi: 10.18632/aging.102240

Zheng, H., Brennan, K., Hernandez, M., and Gevaert, O. (2019). Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. Gigascience 8:giz145. doi: 10.1093/gigascience/giz145

Zheng, L. L., Li, J. H., Wu, J., Sun, W. J., Liu, S., Wang, Z. L., et al. (2016). deepBase v2.0: identification, expression, evolution and function of small RNAs, lncRNAs and circular RNAs from deep-sequencing data. Nucleic Acids Res. 44, 196–202. doi: 10.1093/nar/gkv1273

Zhi, H., Li, X., Wang, P., Gao, Y., Gao, B., Zhou, D., et al. (2018). Lnc2Meth: a manually curated database of regulatory relationships between long non-coding RNAs and DNA methylation associated with human disease. Nucleic Acids Res. 46, D133–D138. doi: 10.1093/nar/gkx985

Zhou, B., Zhao, H., Yu, J., Gao, C., Dou, X., Song, F., et al. (2018). EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. Nucleic Acids Res. 46, D100–D105. doi: 10.1093/nar/gkx677

Zhou, K. R., Liu, S., Sun, W. J., Zheng, L. L., Zhou, H., Yang, J. H., et al. (2017). ChiPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChiP-seq data. Nucleic Acids Res. 45, D43–D50. doi: 10.1093/nar/gkw965

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415. doi: 10.1093/nar/gkg595

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Chowdhary, Satagopam and Schneider. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.