Bioinformatics, 31(14), 2015, 2241–2251
doi: 10.1093/bioinformatics/btv148
Advance Access Publication Date: 15 March 2015
Review

Genome analysis

Computational approaches towards understanding human long non-coding RNA biology

Saakshi Jalali[1,2,†], Shruti Kapoor[1,2,†], Ambily Sivadas[1,†], Deeksha Bhartiya[1,2] and Vinod Scaria[1,2,*]

[1]GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mathura Road, Delhi 110020, India and [2]Academy of Scientific & Innovative Research (AcSIR), 2 Rafi Marg, Anusandhan Bhawan, New Delhi 110001, India

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on October 16, 2014; revised on February 9, 2015; accepted on March 10, 2015

Abstract

Long non-coding RNAs (lncRNAs) form the largest class of non-protein coding genes in the human genome. While a small subset of well-characterized lncRNAs has demonstrated their significant role in diverse biological functions like chromatin modifications, post-transcriptional regulation, imprinting etc., the functional significance of a vast majority of them still remains an enigma. Increasing evidence of the implications of lncRNAs in various diseases including cancer and major developmental processes has further enhanced the need to gain mechanistic insights into the lncRNA functions. Here, we present a comprehensive review of the various computational approaches and tools available for the identification and annotation of long non-coding RNAs. We also discuss a conceptual roadmap to systematically explore the functional properties of the lncRNAs using computational approaches.

Contact: vinods@igib.in

1 Introduction

Non-coding RNAs have been one of the major focuses in the field of functional genomics since the last decade. Non-coding RNAs have been largely classified based on their size into small and long non-coding RNAs. The former class of ncRNAs encompasses well studied candidates including tRNAs, miRNAs, piRNAs and snoRNAs (Birney *et al.*, 2007; Carninci *et al.*, 2005; Kapranov *et al.*, 2007; Okazaki *et al.*, 2002). The latter class of non-coding RNAs encompasses a functionally diverse set of transcripts, which by definition are transcripts greater than 200 nucleotides in length with no potential to encode for functional proteins of more than 30 amino acids (Li *et al.*, 2013; Mercer *et al.*, 2009). A large number of transcripts of this class were initially discovered through earlier studies reporting expressed sequence tags (ESTs) in Human and mouse through the H-invitation (Imanishi *et al.*, 2004) and FANTOM (Carninci *et al.*, 2005) consortia. This repertoire was significantly expanded in the recent years, benefitting from the availability of technologies to annotate transcriptome at single nucleotide resolutions (Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2010). There are over 23 898 transcripts from 13 870 genomic loci as reported by Gencode (Gencode v21) for humans, which encompass classes of transcripts previously annotated as antisense transcripts, intronic transcripts, large intergenic non-coding RNAs (lincRNAs), and processed pseudogenes (Derrien *et al.*, 2012; Fig. 1). Long non-coding RNAs are generally thought to be transcribed by RNA polymerase II, and largely observed to be 5′ capped and 3′ poly-adenylated, though exceptions to this general rule are not uncommon (Dieci *et al.*, 2007; Kiyosawa *et al.*, 2005). Recent evidence suggests that a small subclass of lncRNAs could encode for small peptides and could also be processed to smaller RNAs (Andrews and

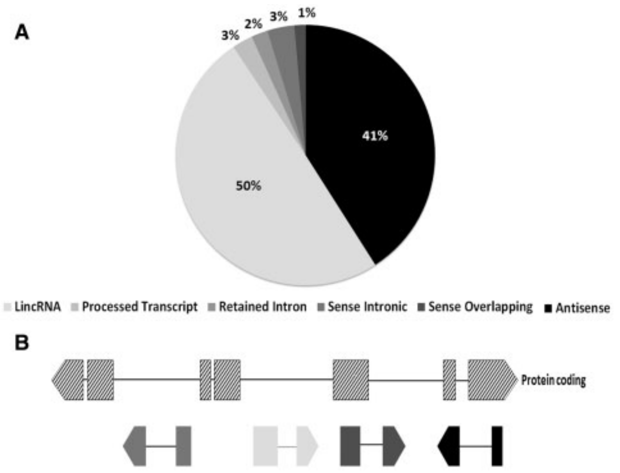

Fig. 1. Distribution of Gencode long non-coding RNAs. (**A**) The pie depicts the number of Gencode lncRNAs distributed in seven subclasses: antisense, lincRNA, processed transcript, retained intron, sense intronic and sense overlapping. (**B**) Depicts the schematic representation of the genomic context of four major lncRNA subclasses: sense intronic, lincRNA, sense overlapping and antisense

Rothnagel, 2014; Bazzini *et al.*, 2014; Fejes-Toth, 2009; Kapranov *et al.*, 2007; Pauli *et al.*, 2014). The functional relevance of these observations has not been explored to great detail. The lncRNA class also includes transcripts classified by some authors as macroRNAs and vlincRNAs. MacroRNAs are long, un-spliced RNAs transcribed from RNA pol II and can readily form secondary structures. MacroRNAs such as Airn or KCNQ1OT1 function mainly in the maintenance of imprinting in normal tissues (Marques and Ponting, 2009). VlincRNAs or very long intergenic RNAs on the other hand, represent a subset of human transcriptome spanning the intergenic genomic space ranging from 50 to 700 kb in size. These transcripts have been identified in various tumors and show cell specific expression. They have also been shown to be associated with pluripotency or the extent of malignancy in some tumours (St *et al.*, 2013). High-throughput techniques have enabled the identification of large number of circular non-coding transcripts commonly known as circRNAs. CircRNAs are non-canonical alternatively spliced RNA structures in which 5′ donor and 3′ acceptor sites are back spliced to form a circular RNA product. Recent studies have discovered large number of circRNAs, majority of which have not yet been functionally characterized (Cocquerelle *et al.*, 1993; Jeck *et al.*, 2013; Salzman *et al.*, 2013; Zhang *et al.*, 2013).Nevertheless, one of the human circRNAs, CDR1as has been shown to harbor 70 binding sites for miR-7, thus potentially acting as microRNA sponge (Bussotti *et al.*, 2013). Of the lncRNA repertoire known till date, only a small subset of lncRNAs appear to be evolutionary conserved (Ponting *et al.*, 2009). LncRNA conservation could potentially be dealt in four dimensions, viz., (i) primary sequence, (ii) secondary or tertiary structure, (iii) function and (iv) conservation in terms of their expression and regulation (Diederichs, 2014). Most of the lncRNAs are not conserved at the level of their primary sequence, though lncRNAs could harbor small stretches of ultra-conserved elements which serve as functional motifs. At the structural level, lncRNAs could have conserved motifs as in the case of SRA lncRNA, where they act as functional domains (Johnsson *et al.*, 2014). Recent analysis from our laboratory show that regions of lncRNAs participating in bio-molecular interactions have a paucity of variations in potential functional domains similar to that of protein coding exons suggesting a strong selective pressure operating at these loci (Bhartiya *et al.*, 2014).

The long non-coding RNA repertoire further encompasses transcripts, which have distinct molecular functions. These functional attributes could be broadly classified as guides, decoys, signals and scaffolds (Wang and Chang, 2011), depending upon the types of molecular interactions with other biomolecules in the cell, namely; DNA, RNA, proteins or small molecules (Braconi *et al.*, 2011; Kirsebom *et al.*, 2006; Okazaki *et al.*, 2002; Tsai *et al.*, 2010). These interactions have been shown to result in a wide spectrum of functional outcomes ranging from chromosomal inactivation as in the case of XIST in X chromosomal inactivation (Brown *et al.*, 1992), epigenetic modifications as in the association with Polycomb repressor proteins (Zhao *et al.*, 2008), RNA interference as in antisense activity (Katayama *et al.*, 2005) or scaffolding of biomolecules as in the case of peri-nuclear bodies (Mohammad *et al.*, 2008). This diversity of functional outcomes is reflected in the myriad ways that lncRNAs have been implicated in disease processes. Long non-coding RNAs are presently implicated in a number of disease processes included cancers, developmental disorders, neurological, metabolic and immunological disorders (Bhartiya *et al.*, 2012; Faghihi *et al.*, 2008; Klattenhoff *et al.*, 2013; Nakamura *et al.*, 2008; Peng *et al.*, 2010; Qureshi and Mehler, 2012)

It would be noteworthy to mention the distinct functional roles and molecular dissection of the functional interactions that have been characterized for a small subset of lncRNAs. The functional roles of a vast majority of the lncRNAs still remain uncharacterized and elusive. The majority of lncRNAs annotated in the recent years have come out of analysis of transcriptome sequencing datasets and the information on these transcripts have been largely limited to the genomic loci and expression patterns in the various conditions studied. This gap in the understanding of the functional roles of lncRNAs has largely been due to the non-availability of tools to characterize specific bio-molecular interactions on genome-scale and of resources, which systematically catalogue these interactions. Recent insights into specific molecular interactions of lncRNAs have been obtained from the integration of genome-scale datasets, both from our laboratory (Jalali *et al.*, 2013) and from others (Agostini *et al.*, 2013; Chu *et al.*, 2012; Lu *et al.*, 2013; Murigneux *et al.*, 2013; Simon, 2013).

Additional evidence on the potential functional attributes of lncRNAs have also been obtained from methodologies, which rely on expression patterns and 'guilt-by-association' methods based on correlation of expression patterns with genes or mutations linked to known phenotypes (Liao *et al.*, 2011a). Computational methodologies integrating genome-scale datasets, expression patterns, motifs and structure annotations provide immense opportunities towards understanding the functional role of lncRNAs.

In the present review, we provide an overview of the computational resources, methodologies and tools available for the identification and functional annotation of human long non-coding RNAs. We also provide a conceptual overview of computational approaches using genome-scale bio-molecular interactions and expression correlation towards understanding the potential function of lncRNAs. We also summarize the current status in identifying functional variations in lncRNAs and discuss the major challenges and opportunities.

2 Computational approaches for LNCRNA annotation and expression analysis

The earliest large-scale genome-wide annotations of long non-coding RNAs have largely emerged from the annotation of ESTs and from orthologous approaches to understand the transcriptome using tiling microarrays. These studies identified a large and previously unknown repertoire of transcripts, many of which did not have any obvious potential to encode for functional proteins (Bertone *et al.*, 2004; Birney *et al.*, 2007; Kapranov *et al.*, 2007). Further in-depth characterization of the transcriptome using deep sequencing from the ENCODE consortium has validated the existence of many of these transcripts (Bernstein *et al.*, 2012). As opposed to conventional microarray-based approach, which is commonly used to profile known transcripts, deep sequencing approaches like RNA-Seq facilitates genome-wide expression profiling thereby unraveling many novel rare transcripts including novel alternatively-spliced isoforms (Cloonan *et al.*, 2008; Marioni *et al.*, 2008; Montgomery *et al.*, 2010; Mortazavi *et al.*, 2008; Mutz *et al.*, 2013; Nagalakshmi *et al.*, 2010; Shendure, 2008). It appears that further depth and focused approaches would uncover an even larger repertoire of transcripts.

A generalized computational pipeline for annotation of lncRNAs includes three steps: (i) alignment of reads to the reference genome, (ii) assembly of transcripts and isoforms and (iii) annotation of the transcripts for coding potential. Additional independent approaches with or without sequence information or transcript annotation, including proteogenomics approaches and RNA secondary structure-based approaches have also been extensively used to annotate long non-coding RNAs and are described briefly in the following sections.

2.1 Alignment of reads and transcript assembly

In a conventional transcriptome analysis pipeline the high-quality reads are aligned to the respective reference genomes using a spliced read aligner like Tophat (Kim *et al.*, 2013) or STAR (Dobin *et al.*, 2013). The aligned reads are further reconstructed into transcripts using reference guided assemblers like Cufflinks (Trapnell *et al.*, 2010) or Scripture (Guttman *et al.*, 2010) which performs an *ab initio* transcriptome assembly, though de novo approaches for transcriptome reconstruction have also been attempted in many organisms. (Grabherr *et al.*, 2011; Li *et al.*, 2009; Zerbino and Birney, 2008). Providing a comprehensive review of the plethora of methods available for spliced alignment and transcript reconstruction exceeds the scope of this review. In addition, an exhaustive evaluation and discussion of more than 25 RNA-Seq aligners (Engstrom *et al.*, 2013) and transcriptome assemblers (Steijger *et al.*, 2013) have been published recently.

2.2 Annotation of lncRNAs

Identification of lncRNAs broadly involves selection of non-protein coding transcripts longer than 200 nucleotides that lack an open reading frame (ORF) of significant length. The ORFs are predicted using tools like GetORF from the EMBOSS suite (Rice *et al.*, 2000). There is no definite ORF length threshold adopted for lncRNAs. More than 95% of the mammalian proteins in public databases such as SwissProt and International Protein Index have an ORF of greater than 100 amino acids (Frith *et al.*, 2006). Some of the landmark lncRNA studies have used ORF length cutoffs ranging from 30 to 100 amino acids. However, it should also be noted that some of the well-characterized lncRNAs such as H19, Xist, Mirg, Gtl2 do have predicted ORFs longer than 100 amino acids (Dinger *et al.*, 2008).

Recent studies have drawn attention to the biological significance of the translated short peptides from short sORFs (Andrews and Rothnagel, 2014) and their potential regulatory role as biologically active peptides. A lot of distinguishing factors based on ORF conservation, nucleotide composition, residue bias and codon usage have been examined towards this goal. Cross-species comparison-based programs like PhyloCSF (Lin *et al.*, 2011) and CRITICA (Badger and Olsen, 1999) have been popular while complementary approaches like CSTMiner (Mignone *et al.*, 2003) and RNAcode (Washietl *et al.*, 2011) use nucleotide substitution and gap patterns as the criteria to assess the coding potential of a sequence. Apart from these, other popular approaches include Support Vector Machine-based classifiers such Coding Potential Calculator (Kong *et al.*, 2007), PORTRAIT (Prediction of Transcriptomic ncRNA by *ab initio* methods; Arrial *et al.*, 2009) which uses multiple distinct sequence features and Coding Potential Assessment Tool (Wang *et al.*, 2013) which uses an alignment-free logistic regression model. Alternately, sequence similarity search programs like BLASTX (Gish and States, 1993) and HMMER3 (Eddy, 2011) help to identify homologous domains in protein data repositories which can also be used to identify and remove transcripts with protein-coding potential. Apart from these computational approaches, experimental techniques like ribosomal profiling have also been employed (Bazzini *et al.*, 2014) to assess the protein coding capacity of lncRNAs relying on the periodicity of ribosome occupancy along the short translated ORFs.

2.3 Proteogenomics-based approaches to evaluate coding potential of transcripts

With the rapidly growing number of newly discovered lncRNAs, several groups have reported that many putative lncRNAs encode for short peptides (Andrews and Rothnagel, 2014). Recent efforts have used innovative approaches to scan transcripts for the presence of putative peptide products using tandem mass spectrometry studies. For example, the recent ENCODE study analysed RNA-Seq data along with the proteomics data and suggested that majority of the annotated lncRNAs (~92%) are not translated (Banfai *et al.*, 2012). Introduction of novel strategies like ribosome profiling (the deep sequencing of ribosome-protected RNA fragments) facilitates genome-wide quantitative monitoring of protein synthesis at high resolution. Ingolia and co-workers developed a machine-learning algorithm to systematically deduce protein products from the riboprofiling data and identified many putative lincRNAs containing regions of high translation, which they called short poly-cistronic ribosome-associated coding RNAs (sprcRNAs; Ingolia *et al.*, 2011). However, examples like H19 mouse lncRNA transcript, which functions as a true lncRNA (Brannan *et al.*, 1990) in spite of harboring highly conserved ORFs with high ribosomal occupancy (Ingolia *et al.*, 2011), demonstrate that ribosomal occupancy alone does not sufficiently indicate the protein coding potential of a transcript. Recently Guttman *et al.* (2013) defined a metric called ribosome release score (RRS), which records the release of the ribosome on encounter with the stop codon to correctly distinguish between coding and non-coding RNA.

2.4 RNA structure-based approaches for lncRNA discovery

Several tools have been developed to aid the classification of transcripts as lncRNAs based on the presence of conserved predicted RNA secondary structures. These include RNAFold (Hofacker, 2003), QRNA (Rivas and Eddy, 2001), EvoFold (Pedersen *et al.*, 2006) and RNAz (Gruber *et al.*, 2010). These approaches have been used extensively in the literature (Humann *et al.*, 2013; Mercer *et al.*, 2010; Ponjavic *et al.*, 2009; Washietl *et al.*, 2005; Woolfe *et al.*, 2005) and have been discussed for their advantages and drawbacks (Gorodkin and Hofacker, 2011). Nevertheless, the use of RNA secondary structure and conservation alone for the identification of lncRNAs could potentially be misleading as conserved RNA secondary structures could also be shared between coding and non-coding transcripts (Dinger *et al.*, 2008). Given these limitations,

Table 1. List of primary, secondary and disease associated databases and resources on long non-coding RNAs

S. No	Database name	Description	Web link	References
Primary databases				
1.	GENCODE	Encyclopædia of genes and gene variants	http://www.gencodegenes.org/	Derrien *et al.* (2012)
2.	LNCipedia	A database for annotated human lncRNA transcript sequences and structures.	http://www.lncipedia.org/	Volders *et al.* (2013)
3.	lncRNAdb	lncRNAdb is a database providing comprehensive annotations of functional long non-coding RNAs (lncRNAs)	http://www.lncrnadb.org/	Amaral *et al.* (2011)
4.	lncRNAMap	Functional Long non-coding RNAs Database	http://lncrnamap.mbc.nctu.edu.tw/php/	Chan *et al.* (2014)
5.	lncRNome	lncRNome is a comprehensive searchable biologically oriented knowledgebase for long non-coding RNAs in Humans.	http://genome.igib.res.in/lncRNome/	Bhartiya *et al.* (2013)
6.	NONCODE 4.0	NONCODE provides an integrative annotation of long non-coding RNAs.	http://noncode.bioinfo.org.cn	Xie *et al.* (2014)
7.	The Functional lncRNA Database	repository of mammalian long non-protein coding transcripts that have been experimentally shown to be both non-coding and functional	http://www.valadkhanlab.org/database	Niazi and Valadkhan (2012)
8.	NRED	A database of long non-coding RNA expression	http://nred.matticklab.com/cgi-bin/ncrnadb.pl	Dinger *et al.* (2009)
Secondary information databases				
9.	ChIPBase	Database for annotating and exploring the expression profiles and the transcriptional regulation of lncRNAs and other ncRNAs.	http://deepbase.sysu.edu.cn/chipbase/	Yang *et al.* (2013)
10.	DIANA-LncBase	Experimentally verified and computationally predicted microRNA targets on long non-coding RNAs.	www.microrna.gr/LncBase	Paraskevopoulou *et al.* (2013)
11.	GeneCards v3	Integrated database of human genes that provides comprehensive, updated, and user-friendly information on all known and predicted human genes.	www.genecards.org	Safran *et al.* (2010)
12.	Linc2GO	A human LincRNA function annotation resource based on ceRNA hypothesis.	http://www.bioinfo.tsinghua.edu.cn/~liuke/Linc2GO/index.html	Liu *et al.* (2013)
13.	lncPro	Prediction of lncRNA–protein interactions	http://202.38.126.151/hmdd/lncpro/	Lu *et al.* (2013)
14.	lncRNABase (starBase v2.0)	A database consisting of RNA–RNA and protein–RNA interaction networks identified from 108 CLIP-Seq datasets generated by 37 independent studies.	http://starbase.sysu.edu.cn/mirLncRNA.php	Li *et al.*, (2014)
15.	miRcode	Transcriptome-wide microRNA target prediction including lncRNAs	http://www.mircode.org/mircode/	Jeggari *et al.* (2012)
16.	lncRNASNP	Provides comprehensive resources of single nucleotide polymorphisms (SNPs) in human/mouse lncRNAs.	http://bioinfo.life.hust.edu.cn/lncRNASNP/	Gong *et al.* (2014)
Disease association database				
17.	LncRNADisease	A database for long non-coding RNA-associated diseases.	http://cmbi.bjmu.edu.cn/lncrnadisease	Chen *et al.* (2013)

recently there is an increased focus on analyzing RNA structure and binding properties for functional annotation of lncRNAs over their detection (Glazko *et al.*, 2012).

3 Databases and resources for long non-coding RNAs

Concomitant with the increasing number of lncRNAs discovered, a number of resources collecting, curating and providing associated functional information of the lncRNAs have been built up in the recent years. In the present collection, described in detail below, we have attempted to provide the salient features of lncRNA resources. A comprehensive listing and description of the resources is summarized in Table 1.

LncRNAdb has been one of the earliest manually curated databases of eukaryotic lncRNAs with information derived from literature and other resources (Dinger *et al.*, 2009; Lein *et al.*, 2007; Sanborn *et al.*, 2011; Su *et al.*, 2004). Apart from lncRNAdb, GENCODE is by and large, one of the most popular lncRNA catalog. The latest version of Gencode (v21) comprises of 15 877 long non-coding RNA genes corresponding to 26 414 transcripts. Another resource, LNCipedia hosts over 21 488 annotated human lncRNA transcripts. The resource provides prediction of secondary structure as well as assessment of the protein coding potential and in addition, incorporates an automated reprocessing pipeline based on data from proteomics experiments to detect ORFs in lncRNAs. NONCODE 4.0 provides information on transcript isoforms and their expression in human tissues while a similar resource lncRNAMap extensively uses RNA-seq datasets available in the public domain to provide insights into the expression levels of lncRNAs in different tissues, cell lines and disease conditions (Chan *et al.*, 2014). The database also incorporates miRNA targets, homologous protein coding genes and endogenous siRNAs (esiRNAs) derived from lncRNAs and their targets.

Two resources lncRNome and the Functional lncRNA database integrate information from other databases to provide insights into

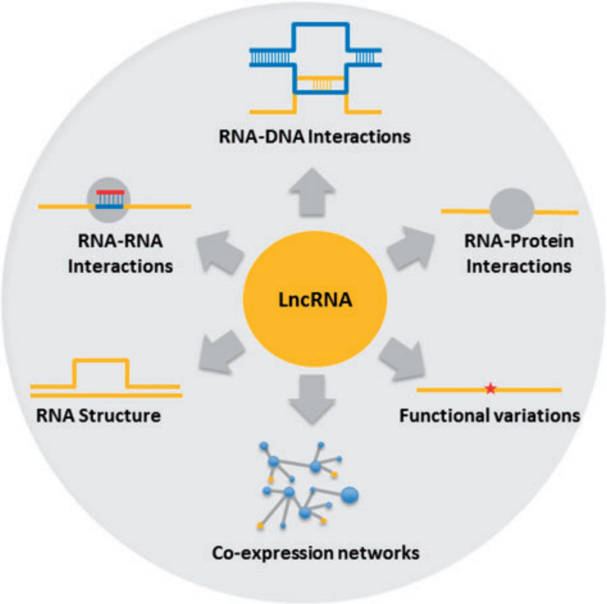

Fig. 2. Schematic representation of current approaches to understand LncRNA functions. A brief summary of the various methodologies used to elucidate the functional characteristics of lncRNAs

the functionality of lncRNAs. lncRNome hosts information of over 18 000 human long non-coding RNA transcripts and is maintained by our group. lncRNome apart from providing information on genomic motifs including sequence and structural motifs also provides information on RNA processing, miRNA binding sites from PAR-CLIP datasets and epigenetic modifications in close proximity with lncRNA promoters (Bhartiya *et al.*, 2013). The database also maps genomic variations in lncRNA sites and features a genome browser. The Functional lncRNA Database hosts a subset of well-curated non-protein coding long non-coding RNAs from literature for Human, Mouse and Rat (Niazi and Valadkhan, 2012), and includes over 180 human lncRNA annotations. Each lncRNA annotation also is accompanied by miRNA precursor information, repeat information and predicted ORFs, apart from the genomic information and sequence.

Apart from the primary lncRNA databases which provide information on lncRNAs, a number of databases also feature functional information of interactions of lncRNAs in specific contexts. Many of these databases also rely on expression information, integration of genome-wide datasets and computational predictions to be able to provide clues into the functional insights of lncRNAs. A subset of the prominent databases is detailed below.

Starbase v2.0 (lncRNABase) integrates public RNA–protein interaction sites generated from high-throughput datasets. The distinct feature of this database is that it provides a comprehensive overview of miRNA-lncRNA interaction maps apart from the expression information. NRED provides microarray and in situ hybridization gene expression information for thousands of transcripts including long ncRNAs in human and mouse (Dinger *et al.*, 2009). Additionally, information for featured ncRNAs evolutionary conservation, secondary structure evidence, genomic context links and antisense relationships are also available.

DIANA-LncBase is a comprehensive resource for computationally predicted and experimentally annotated miRNA-lncRNA interactions (Paraskevopoulou *et al.*, 2013). For the experimental annotation, the database extensively makes use of the CLIP-seq datasets for Argonaute proteins, apart from genome-wide degradome sequencing and other datasets. In addition, the database also provides information on expression, microRNA regulatory modules and their conservation. A similar database lnCeDB catalogs human long non-coding RNAs which can potentially act as competing endogenous RNAs (ceRNA; Das *et al.*, 2014). The database scores the likelihood of lncRNA-mRNA pair for actually being ceRNA based on two methods and integrates predicted mRNA-miRNA interactions from a number of resources. In addition, the Linc2GO database provides functional annotation based on the hypothesis that lncRNAs can function as a ceRNA.

Apart from the databases which largely discuss the expression and function of lncRNAs, LncRNA Disease maintained by the Cui Lab systematically collects diseases associated with lncRNAs, and has created a niche for itself (Chen *et al.*, 2013). The disease associations of lncRNAs have been curated from literature evidence from peer reviewed publications, and in addition, were investigated using bioinformatics methods based on genomic neighborhood.

4 Computational approaches to understand LNCRNA function and disease associations

A large number of lncRNAs have been discovered through computational mapping of the transcriptome, but the functional annotation and intricate molecular mechanisms of gene regulation still remains an enigma. Computational methods offer a new opportunity to understand the potential functional implications of lncRNAs. However, it should be emphasized that follow-up experimental validation would be necessary to examine specific functional correlates unambiguously. The following section provides an overview of the popular computational approaches towards deciphering the functionalities of lncRNAs (Fig. 2). Unlike conservation of the sequence and/or structure, characteristic of members of the small RNA class, like miRNAs and snoRNAs, lncRNAs are largely not conserved. Earlier analyses comparing lncRNAs in humans and other model systems have shown that only a very small subset of the human lncRNAs are evolutionarily conserved, thus severely limiting synteny or orthology-based approaches towards extrapolating functions of lncRNAs from model systems. Unlike protein-coding genes which require strong sequence conservation to preserve their function, lncRNAs may only require short stretches of conserved sequence to maintain structural motifs important for its functional role (Mercer *et al.*, 2009).

4.1 Understanding lncRNA function through expression correlation

The most popular and largely studied approach towards understanding the function of lncRNAs have been based on 'guilt-by-association' from co-expression patterns shared with their protein coding counterparts (Liao *et al.*, 2011a; Necsulea *et al.*, 2014). The basis of the methodology stems from the concept that transcripts sharing common expression patterns should largely share similar biological pathways. A number of different studies have used this approach towards functionally annotating potential lncRNAs. One of the first studies on co-expression analysis of lncRNAs and coding genes at the genome scale (Guttman *et al.*, 2009). The authors ranked the protein coding genes co-expressed with a given lncRNA. The protein coding genes were further analyzed using Gene Set Enrichment Analysis (GSEA) to identify enriched functional terms corresponding to the lncRNAs (Subramanian *et al.*, 2005; Mootha *et al.*, 2003). In another study, a co-expression network between coding and non-coding transcripts was constructed (Liao *et al.*, 2011a).This network takes into account not only the co-expression of lncRNA and protein coding genes but also other lncRNAs, thus being able to identify statistically significant clusters of co-expressed genes, which potentially correspond to functional modules. A couple of computational tools have come up in the recent past, which uses the probe annotation information of Affymetrix microarrays for the analysis of expression of lncRNAs. ncFANs (Liao *et al.*, 2011b) and Noncoder (Gellert *et al.*, 2013) provides a web interface to re-analyze datasets for lncRNA expression. ncFANs also provides functional annotation based on co-expression patterns of the lncRNAs with protein coding genes.

4.2 Understanding lncRNA function as part of interactions

Apart from expression-based approaches, the advent of next-generation sequencing for understanding bio-molecular interactions through a variety of methodologies have opened up new avenues to create genome-wide interaction maps for specific biomolecules. Though many of these methods have largely addressed the problem in the perspective of protein coding genes and their regulation, aspect of lncRNAs still needs to be unraveled. The genome-wide data can be integrated with the lncRNA annotations enabling construction of maps of putative biological interactions and thereby correlating potential functional consequences. The biological function of lncRNAs in the cell could be understood as a function of biological interactions mediated by lncRNAs with other RNA species (RNA–RNA interactions), with proteins (RNA–protein interactions), with DNA (RNA–DNA interactions) and/or processing of lncRNAs to smaller RNAs. In many cases, a combination of one or more of the above-said interactions could be necessary for the functional outcome.

RNA–protein interactions are crucial for the functioning of RNA, either independently or in conjunction with other molecules. One of the well-studied examples is the interaction of lncRNAs with the Polycomb Repressor proteins to modulate epigenetic modification of the targets. A number of computational and experimental approaches have been extensively used to characterize RNA–protein interactions. The earlier computational tools relied on sequence, structure and physicochemical properties or features to predict potential RNA–protein interaction pairs, and the popular algorithms include CatRAPID (Agostini *et al.*, 2013) and RPIseq (Muppirala *et al.*, 2011). A specific approach for lncRNA named lncPro (Lu *et al.*, 2013) has also been developed based on the physicochemical properties of the interaction pairs. Using this resource, the authors could predict the interactions of HOTAIR with PRC2 complex and LSD1 on the basis of their interaction scores, which were in concordance with the experimental findings (Tsai *et al.*, 2010). Experimental methods for understanding RNA–protein interactions have also been extensively used. The most popular approach includes cross-linking immunoprecipitation followed by sequencing (CLIP-seq; Murigneux *et al.*, 2013) and its variants. Other complementary experimental methodologies include protein capture on RNA like RNA affinity in tandem (RAT; Hogg and Collins, 2007), Csy4 Select (Lee *et al.*, 2013) and MS2-TRAP (Yoon *et al.*, 2012). Affinity-based capture approaches including SeqRS (Campbell *et al.*, 2012) and RNAcompete (Ray *et al.*, 2009) have also been used to identify RNA binding proteins interacting with specific RNAs.

Apart from interaction with protein, one of the active areas of interest has been to identify genomic targets of lncRNAs. A number of well-studied lncRNA candidates like Xist, HOTAIR and HOTTIP modulate their function through interactions with genomic DNA (Mercer *et al.*, 2013). A number of experimental approaches to identify RNA–DNA interaction sites have been used in the recent years. Chu *et al.*, introduced a technique of Chromatin Isolation by RNA purification (CHIRP; Chu *et al.*, 2012). The approach used tiling oligonucleotides to capture specific RNAs with bound protein and DNA sequences, the sequence identity of which could be deciphered by sequencing. Simon *et al.*, developed this hybridization-based technique by designing affinity-tagged versions of the complementary oligonucleotides against the RNA called Capture Hybridization Analysis of RNA Targets (CHART; Simon, 2013). This technique could identify interaction of lncRNAs at the chromatin level. It has also been hypothesized that lncRNAs could function through formation of DNA-RNA triplexes where lncRNAs form the third strand. For example, a long non-coding transcript originating from upstream region of the major DHFR promoter represses the transcription of downstream protein coding gene by forming an RNA–DNA triplex structure with the DHFR promoter (Blume *et al.*, 2003; Martianov *et al.*, 2007). Computational approaches like Triplexator (Buske *et al.*, 2012) and R-loop finder (Wongsurawat *et al.*, 2012) offers enormous promises to computationally predict such interactions on a genome-scale.

Besides interactions with DNA and proteins, lncRNAs have also been extensively shown to interact with other RNAs. The most widely studied interactions have been miRNA-lncRNA interactions, where it has been widely proposed that lncRNAs could modulate the action of regulatory microRNAs by acting as a sponge (Ebert *et al.*, 2010; Salmena *et al.*, 2011). Our group (Jalali *et al.*, 2013) has recently re-constructed a genome-wide interactome of lncRNA and miRNAs which might play regulatory roles in gene expression. Argonaute proteins are part of the RISC complex which mediates targeting of microRNAs. PAR-CLIP datasets of argonaute proteins serve as the evidence pointing to microRNA binding. A number of computational tools have also been extensively used to identify conserved and non-conserved microRNA targets in transcripts. It is also noteworthy that some of the lncRNAs serve as templates for the biogenesis of small regulatory RNA. Using small RNA sequencing datasets and comparison of the genomic loci, a genome-wide map of lncRNA processing has also been discussed (Jalali *et al.*, 2012). In addition to this, (Johnson and Guigo, 2014) have proposed a repeat insertion domains of lncRNAs (RIDL) hypothesis where they address that the transposable elements may serve as functional domains for lncRNA interactions with proteins or nucleic acids. LncRNAs harbor a higher frequency for transposable elements (TEs) than protein coding exons. Some lncRNAs like AK046052, a brain specific mouse lncRNA is thought to be derived from TEs. The insertion of TEs in an lncRNA sequence is thought to add functionality by virtue of serving as interacting domains for proteins, RNA and DNA. The outcome could be exemplified in interactions such as Alu/LINE mediated interaction with ribosomal complexes or Alu-mediated mRNA decay by lncRNA (Johnson and Guigo, 2014).

4.3 Understanding lncRNA structure and motifs

A variety of computational approaches including single sequence structure predictors like Mfold (Zuker, 2003), RNAfold (Hofacker, 2003) and Sfold (Ding *et al.*, 2004), sequence alignment-based tools like QRNA (Rivas and Eddy, 2001), EvoFold (Pedersen *et al.*, 2006) and RNAz (Gruber *et al.*, 2010) multiple structural alignment tools like RNAForester (Hochsmann *et al.*, 2004) and MARNA (Siebert and Backofen, 2005) and derivatives of Sankoff's algorithm like Foldalign (Havgaard *et al.*, 2005), Dynalign (Mathews and Turner,

2002) and LocARNA (Will *et al.*, 2007) have been employed for RNA structure predictions of short conserved domains within lncRNAs, though it is still uncertain if all lncRNAs harbor local structured domains. It is also important to note that most of these RNA structure prediction approaches have high false positive rates, around 50%. Minimum free energy folding-based approaches often fail to provide reliable signals that can clearly distinguish non-coding RNA structures from random sequences. While approaches driven by sequence-based alignments are faster than structure-based alignments, the former fails to capture the hidden dimension of conservation at the structure level often missed in weakly conserved primary sequences. Additional algorithmic advances including the use of conservation and additional features to predict pseudoknots have been developed in the recent years (Gruber *et al.*, 2010; Ren *et al.*, 2005; Ruan *et al.*, 2004; Sato *et al.*, 2011). The predictions of structure of lncRNAs were largely limited by the non-availability of accurate methods to predict structures of these long RNAs which are mostly unstructured. Therefore, it is important to integrate computation predictions with other types of data including experimental approaches to ensure precise functional annotation. A detailed discussion on the pros and cons of these approaches has been detailed previously (Gorodkin and Hofacker, 2011). Recent experimental approaches using next generation sequencing technology such as PARS (Kertesz *et al.*, 2010), FragSeq (Underwood *et al.*, 2010), and SHAPE-seq (Lucks *et al.*, 2011) have significantly added to the repertoire of the structural map of lncRNAs. Sequence and structure motifs in RNAs have been an active area of interest. A number of computational tools including RegRNA (Chang *et al.*, 2013) have curated structural and sequence motifs in RNA and offer a user-friendly interface to identify them in sequences. In addition, non-conventional motifs including G-quadruplex motifs and intramolecular RNA triplexes have been actively studied (Carmona and Molina, 2002; Hacht *et al.*, 2014; Nguyen *et al.*, 2014). In an earlier bioinformatics analysis of G-quadruplex motifs in the lncRNome distinct patterns of occurrence have been demonstrated in lncRNAs compared to protein coding genes (Bhartiya *et al.*, 2013). A number of computational tools to predict G-quadruplexes have been reported and offer a unique opportunity to understand their functional attributes (Frees *et al.*, 2014; Kikin *et al.*, 2006; Stegle *et al.*, 2009).

4.2 Computational approaches to understand functional variations in lncRNAs

It is generally observed that a large proportion of the lncRNAs do not show evolutionary conservation unlike well conserved subsets of ncRNAs like microRNAs and snoRNAs. Previous studies have also characterized large variability in the sequence of lncRNAs, which was used to support the argument that lncRNAs do not probably have functional roles. Nevertheless it has also been shown that the well-studied functional lncRNAs such as XIST, HOTAIR show poor conservation across species suggesting conservation is not a prerequisite for functionality (Chodroff *et al.*, 2010; Qu and Adelson, 2012). In a recent analysis, our group (Bhartiya *et al.*, 2014) has shown distinct patterns of variations in lncRNAs and compared the variation densities across potentially functional elements in lncRNAs and protein coding exons. The analysis suggested that the variation densities in functional elements in lncRNAs were comparable to the protein coding exons in mRNAs. The authors also found a number of GWAS associated markers mapped to lncRNA loci. In one example, three isoforms transcribed from gene ENSG00000223891 having binding sites for HUR protein harbor a SNP (rs6017291) which is associated with cognitive test performance. This observation opens up enormous opportunities to map functional elements in lncRNAs and potentially understand the effects of genomic variations in these elements. Additionally the impact of variations on conserved sequence and structural motifs in lncRNAs have not been explored to great detail. Computational tools to map the effect of variations on RNA structure including popular tools like SNPfold (Halvorsen *et al.*, 2010), RNAsnp (Sabarinathan *et al.*, 2013) have been proposed. A starting point in this area has been the analysis which revealed 1900 SNVs which have a potential to alter the RNA structure (Wan *et al.*, 2014).

4.5 Computational framework towards understanding the role of lncRNA in Diseases.

Recent genome-wide association studies (GWAS) have provided immense insights into the potential disease associations of long non-coding RNAs (Bochenek *et al.*, 2013; Jin *et al.*, 2011; Kumar *et al.*, 2013; Wu *et al.*, 2013). A recent analysis of GWAS signals in long non-coding RNAs revealed that a significant proportion of the variations associated with human traits/diseases map to lncRNA loci (Hrdlickova *et al.*, 2014; Kumar *et al.*, 2013). A comprehensive compendium of long non-coding RNAs and traits associated with them have been compiled into a web-based resource lncRNADisease (Chen *et al.*, 2013). Apart from GWAS, a number of expression-based approaches to infer disease associations of lncRNAs have been reported in the recent years. In one of the most recent approaches, Gao and coworkers have used a lncRNA-protein coding gene network based on lncRNAs and protein coding genes implicated in diseases and suggested high predictability of the approach towards finding associations of lncRNAs with diseases (Yang *et al.*, 2014). A similar network-based approach by Liu *et al* used networks built based on expression correlation (Liu *et al.*, 2014) and Sun *et al.*, which used known lncRNA disease associations implemented in the framework called RWRlncD (Sun *et al.*, 2014), both with reliable accuracies.

5 Knowledge gaps, challenges and opportunities

The current understanding of the role of long non-coding RNAs in biological processes is largely limited to a handful of well characterized candidates, while the lion's share of the annotated lncRNAs we know today, do not have any obvious functional annotations. The lack of functional annotations have been mostly contributed by a host of factors including the absence of unified resources to annotate lncRNAs, lack of common biogenesis or mechanism of action for lncRNAs and primarily, lack of evolutionary conservation of lncRNAs, which precludes their study in model systems. The present functional correlations of lncRNAs are largely thus limited to 'guilt by association' studies mainly based on integration of omics datasets to arrive at potential functional correlates of lncRNAs. The challenges aside, the larger role of lncRNAs in mediating disease processes are increasingly being cemented through converging evidence from GWAS studies as well as expression-based studies in disease processes. Computational approaches towards understanding the biogenesis, regulation and function of lncRNAs thus could provide a huge impetus in the field. The incident opportunity provided by the ready availability of omics datasets, thanks to large-scale data on genomics, epigenomics and transcriptomics presently available through consortia including the ENCODE and NIH RoadMap Epigenomics projects (Bernstein *et al.*, 2010) provides the much required baseline to start understanding the regulatory dynamics of lncRNAs and be able to provide insights into the regulation of lncRNAs. The second opportunity has been the availability of sequencing-based approaches to understand biomolecular interactions—be it Protein:RNA, RNA:RNA or RNA:DNA interactions, also described previously in this manuscript. Though the current availability of the datasets for interactions is limited, integration of information from disparate datasets has not been attempted towards understanding unified principles of action of lncRNAs. The third opportunity has been the recent availability of sequencing-based approaches to characterize the secondary structure of RNAs, using chemical or enzymatic approaches followed by sequencing. Insights into the RNA structure or rather the dynamics of RNA structure and its relation with bio-molecular interactions would significantly add to the understanding of lncRNA functionality. The sequence variations in lncRNAs, be it at the genome level or at the RNA level, modulated through RNA editing and its effect on structure and interactions have not been studied to any great extent and is a potential opportunity. The fifth and the most challenging opportunity could be potential new approaches to understand sequence, structure and/or functional orthologs of lncRNAs in models systems to be able to draw parallels between the phenotypes in animal systems modulated through lncRNAs with human diseases or traits. This could potentially be the Holy Grail in the understanding of lncRNA function.

Acknowledgements

Authors acknowledge comments from Dr. S Ramachandran and Sridhar Sivasubbu in preparing the manuscript.

Funding

This work was funded by the Council of Scientific and Industrial Research (CSIR), India through Grant GENCODE-C (BSC0123). SK Acknowledges a Senior Research Fellowship from the Council of Scientific and Industrial Research (CSIR), India.

Conflict of Interest: none declared.

References

Agostini,F. *et al.* (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, **29**, 2928–2930.

Amaral,P.P. *et al.* (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.

Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.

Arrial,R.T. *et al.* (2009) Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC Bioinformatics*, **10**, 239.

Badger,J.H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, **16**, 512–524.

Banfai,B. *et al.* (2012) Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.*, **22**, 1646–1657.

Bazzini,A.A. *et al.* (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.*, **33**, 981–993.

Bernstein,B.E. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, **28**, 1045–1048.

Bernstein,B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Bhartiya,D. *et al.* (2012) Conceptual approaches for lncRNA drug discovery and future strategies. *Expert. Opin. Drug Discov.*, **7**, 503–513.

Bhartiya,D. *et al.* (2013) lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database. (Oxford)*, **2013**, bat034.

Bhartiya,D. *et al.* (2014) Distinct patterns of genetic variations in potential functional elements in long noncoding RNAs. *Hum. Mutat.*, **35**, 192–201.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Blume,S.W. *et al.* (2003) The 5′-untranslated RNA of the human dhfr minor transcript alters transcription pre-initiation complex assembly at the major (core) promoter. *J. Cell Biochem.*, **88**, 165–180.

Bochenek,G. *et al.* (2013) The large non-coding RNA ANRIL, which is associated with atherosclerosis, periodontitis and several forms of cancer, regulates ADIPOR1, VAMP3 and C11ORF10. *Hum. Mol. Genet.*, **22**, 4516–4527.

Braconi,C. *et al.* (2011) microRNA-29 can regulate expression of the long noncoding RNA gene MEG3 in hepatocellular cancer. *Oncogene*, **30**, 4750–4756.

Brannan,C.I. *et al.* (1990) The product of the H19 gene may function as an RNA. *Mol. Cell Biol.*, **10**, 28–36.

Brown,C.J. *et al.* (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, **71**, 527–542.

Buske,F.A. *et al.* (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res.*, **22**, 1372–1381.

Bussotti,G. *et al.* (2013) Detecting and comparing non-coding RNAs in the high-throughput era. *Int. J. Mol. Sci.*, **14**, 15423–15458.

Campbell,Z.T. *et al.* (2012) Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Rep.*, **1**, 570–581.

Carmona,P. and Molina,M. (2002) Binding of oligonucleotides to a viral hairpin forming RNA triplexes with parallel G*G*C triplets. *Nucleic Acids Res.*, **30**, 1333–1337.

Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.

Chan,W.L. *et al.* (2014) lncRNAMap: A map of putative regulatory functions in the long non-coding transcriptome. *Comput. Biol. Chem.*, **50**, 41–49.

Chang,T.H. *et al.* (2013) An enhanced computational platform for investigating the roles of regulatory RNA and for identifying functional RNA motifs. *BMC Bioinformatics*, **14**,S4–S14.

Chen,G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.

Chodroff,R.A. *et al.* (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.*, **11**, R72–R11.

Chu,C. *et al.* (2012) Chromatin isolation by RNA purification (ChIRP). *J. Vis. Exp.*, **3912**, doi:10.3791/3912.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.

Cocquerelle,C. *et al.* (1993) Mis-splicing yields circular RNA molecules. *FASEB J.*, **7**, 155–160.

Das,S. *et al.* (2014) lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. *PLoS One*, **9**, e98965.

Derrien,T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.

Dieci,G. *et al.* (2007) The expanding RNA polymerase III transcriptome. *Trends Genet*, **23**, 614–622.

Diederichs,S. (2014) The four dimensions of noncoding RNA conservation. *Trends Genet.*, **30**, 121–123.

Ding,Y. *et al.* (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.*, **32**, W135–W141.

Dinger,M.E. *et al.* (2009) NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.*, **37**, D122–D126.

Dinger,M.E. *et al.* (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, **4**, e1000176.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Ebert,M.S. *et al.* (2010) Emerging roles for natural microRNA sponges. *Curr Biol*, **20**, R858–R861.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Engstrom,P.G. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.

Faghihi,M.A. *et al.* (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.*, **14**, 723–730.

Fejes-Toth,K. (2009) Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, **457**, 1028–1032.

Frees,S. *et al.* (2014) QGRS-Conserve: a computational method for discovering evolutionarily conserved G-quadruplex motifs. *Hum. Genomics*, **8**, 8.

Frith,M.C. *et al.* (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet.*, **2**, e52.

Gellert,P. *et al.* (2013) Noncoder: a web interface for exon array-based detection of long non-coding RNAs. *Nucleic Acids Res.*, **41**, e20.

Gish,W. and States,D.J. (1993) Identification of protein coding regions by database similarity search. *Nat. Genet.*, **3**, 266–272.

Glazko,G.V. *et al.* (2012) Computational prediction of polycomb-associated long non-coding RNAs. *PLoS One*, **7**, e44878.

Gong,J. *et al.* (2014) lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res.*, D181–D186. doi:10.1093/nar/gku1000.

Gorodkin,J. and Hofacker,I.L. (2011) From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comput. Biol.*, **7**, e1002100.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Gruber,A.R. *et al.* (2010) RNAz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, 69–79. doi: 10.1142/9789814295291_0009.

Guttman,M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**,223–227.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Guttman,M. *et al.* (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, **154**, 240–251.

Hacht,A. *et al.* (2014) Identification and characterization of RNA guanine-quadruplex binding proteins. *Nucleic Acids Res.*, **42**, 6630–6644.

Halvorsen,M. *et al.* (2010) Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet.*, **6**, e1001074. doi: 10.1371/journal.pgen.1001074.

Havgaard,J.H. *et al.* (2005) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.*, **33**, W650–W653.

Hochsmann,M. *et al.* (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM. *Trans. Comput. Biol. Bioinform.*, **1**, 53–62.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Hogg,J.R. and Collins,K. (2007) RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. *RNA*, **13**, 868–880.

Hrdlickova,B. *et al.* (2014) Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochim. Biophys. Acta.*, **1842**, 1910–1922.

Humann,F.C. *et al.* (2013) Sequence and expression characteristics of long noncoding RNAs in honey bee caste development–potential novel regulators for transgressive ovary size. *PLoS One*, **8**, e78915.

Imanishi,T. *et al.* (2004) Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *PLoS. Biol.*, **2**, e162.

Ingolia,N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.

Jalali,S. *et al.* (2013) Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PLoS One*, **8**, e53823.

Jalali,S. *et al.* (2012) Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. *Biol. Direct.*, **7**, 25–27.

Jeck,W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.

Jeggari,A. *et al.* (2012) miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics*, **28**, 2062–2063.

Jin,G. *et al.* (2011) Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis*, **32**, 1655–1659.

Johnson,R. and Guigo,R. (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*, **20**, 959–976.

Johnsson,P. *et al.* (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim. Biophys. Acta.*, **1840**, 1063–1071.

Kapranov,P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.

Katayama,S. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.

Kertesz,M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

Kikin,O. *et al.* (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Kirsebom,L.A. *et al.* (2006) Aminoglycoside interactions with RNAs and nucleases. *Handb. Exp. Pharmacol.*, **173**, 73–96.

Kiyosawa,H. *et al.* (2005) Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. *Genome Res*, **15**, 463–474.

Klattenhoff,C.A. *et al.* (2013) Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell*, **152**, 570–583.

Kong,L. *et al.* (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.

Kumar,V. *et al.* (2013) Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS. Genet.*, **9**, e1003201.

Lee,H.Y. *et al.* (2013) RNA-protein analysis using a conditional CRISPR nuclease. *Proc. Natl. Acad. Sci. USA.*, **110**, 5416–5421.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Li,J.H. *et al.* (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, D92–D97.

Li,R. *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Li,X. *et al.* (2013) Long noncoding RNAs: insights from biological features and functions to diseases. *Med. Res Rev.*, **33**, 517–553.

Liao,Q. *et al.* (2011a) Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.*, **39**, 3864–3878.

Liao,Q. *et al.* (2011b) ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res.*, **39**, W118–W124.

Lin,M.F. *et al.* (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.

Liu,K. *et al.* (2013) Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. *Bioinformatics*, **29**, 2221–2222.

Liu,Z. *et al.* (2014) Microarray profiling and co-expression network analysis of circulating lncRNAs and mRNAs associated with major depressive disorder. *PLoS One*, **9**, e93388.

Lu,Q. *et al.* (2013) Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics*, **14**, 651–614.

Lucks,J.B. *et al.* (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. USA.*, **108**, 11063–11068.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Marques,A.C. and Ponting,C.P. (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.*, **10**, R124–10.

Martianov,I. *et al.* (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*, **445**, 666–670.

Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

Mercer,T.R. *et al.* (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.

Mercer,T.R. *et al.* (2010) Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci.*, **11**, 14.

Mercer,T.R. *et al.* (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol*, **20**, 300–307.

Mignone,F. *et al.* (2003) Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res.*, **31**, 4639–4645.

Mohammad,F. *et al.* (2008) Kcnq1ot1/Lit1 noncoding RNA mediates transcriptional silencing by targeting to the perinucleolar region. *Mol. Cell Biol.*, **28**, 3713–3728.

Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

Mootha,V.K. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267–273.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Muppirala,U.K. *et al* (2011) Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, **12**, 489.

Murigneux,V. *et al.* (2013) Transcriptome-wide identification of RNA binding sites by CLIP-seq. *Methods*, **63**, 32–40.

Mutz,K.O. *et al.* (2013) Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.*, **24**, 22–30.

Nagalakshmi,U. *et al.* (2010) RNA-Seq: a method for comprehensive transcriptome analysis. *Curr. Protoc. Mol. Biol.*, Chapter 4:Unit 4.11.1–13.

Nakamura,Y. *et al.* (2008) The GAS5 (growth arrest-specific transcript 5) gene fuses to BCL6 as a result of t(1;3)(q25;q27) in a patient with B-cell lymphoma. *Cancer Genet. Cytogenet.*, **182**, 144–149.

Necsulea,A. *et al.* (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.

Nguyen,G.H. *et al.* (2014) Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs. *Proc. Natl. Acad. Sci. USA.*, **111**, 9905–9910.

Niazi,F. and Valadkhan,S. (2012) Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA*, **18**, 825–843.

Okazaki,Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.

Paraskevopoulou,M.D. *et al.* (2013) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.*, **41**, D239–D245.

Pauli,A. *et al.* (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science*, **343**, 1248636.

Pedersen,J.S. *et al.* (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.

Peng,X. *et al.* (2010) Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. *MBio.*, **1**, e00206–e00210.

Ponjavic,J. *et al.* (2009) Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet.*, **5**, e1000617.

Ponting,C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Qu,Z. and Adelson,D.L. (2012) Evolutionary conservation and functional roles of ncRNA. *Front Genet.*, **3**, 205.

Qureshi,I.A. and Mehler,M.F. (2012) Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat. Rev. Neurosci.*, **13**, 528–541.

Ray,D. *et al.* (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.*, **27**, 667–670.

Ren,J. *et al.* (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA.*, **11**, 1494-1504.

Rice,P. *et al.* (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Ruan,J. *et al.* (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.

Sabarinathan,R. *et al.* (2013) RNAsnp: efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mutat.*, *546–556*. doi: 10.1002/humu.22273.

Safran,M. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database. (Oxford).*, **2010**, baq020.

Salmena,L. *et al.* (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–358.

Salzman,J. *et al.* (2013) Cell-type specific features of circular RNA expression. *PLoS Genet.*, **9**, e1003777.

Sanborn,J.Z. *et al.* (2011) The UCSC Cancer Genomics Browser: update 2011. *Nucleic Acids Res.*, **39**, D951–D959.

Sato,K. *et al.* (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85–i93.

Shendure,J. (2008) The beginning of the end for microarrays? *Nat. Methods*, **5**, 585–587.

Siebert,S. and Backofen,R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**, 3352–3359.

Simon,M.D. (2013) Capture hybridization analysis of RNA targets (CHART). *Curr. Protoc. Mol. Biol.*, Chapter 21:Unit 21.25.

St,L.G. *et al.* (2013) VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol.*, **14**, R73–14.

Stegle,O. *et al.* (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374–i382.

Steijger,T. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, **101**, 6062–6067.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, **102**, 15545–15550.

Sun,J. *et al.* (2014) Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.*, **10**, 2074–2081.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Tsai,M.C. *et al.* (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.

Underwood,J.G. *et al.* (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.

Volders,P.J. *et al.* (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*, **41**, D246–D251.

Wan,Y. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.

Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell.*, **43**, 904–914.

Wang,L. *et al.* (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.

Washietl,S. *et al.* (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383-–1390.

Washietl,S. *et al.* (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, **17**, 578–594.

Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS. Comput. Biol.*, **3**, e65.

Wongsurawat,T. *et al.* (2012) Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. *Nucleic Acids Res.*, **40**, e16.

Woolfe,A. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.

Wu,H. *et al.* (2013) A genetic polymorphism in lincRNA-uc003opf.1 is associated with susceptibility to esophageal squamous cell carcinoma in Chinese populations. *Carcinogenesis*, **34**, 2908–2917.

Xie,C. *et al*. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res*., **42**, D98–D103.

Yang,J.H. *et al*. (2013) ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res*., **41**, D177–D187.

Yang,X. *et al*. (2014) A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One*, **9**, e87797.

Yoon,J.H. *et al*. (2012) MS2-TRAP (MS2-tagged RNA affinity purification): tagging RNA to identify associated miRNAs. *Methods*, **58**, 81–87.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*., **18**, 821–829.

Zhang,Y. *et al*. (2013) Circular intronic long noncoding RNAs. *Mol. Cell*., **51**, 792–806.

Zhao,J. *et al*. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, **322**, 750–756.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.