Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.jzus.zju.edu.cn; www.springerlink.com
E-mail: jzus@zju.edu.cn

Review:

Genomic data mining for functional annotation of human long noncoding RNAs*

Brian L. GUDENAS, Jun WANG, Shu-zhen KUANG, An-qi WEI, Steven B. COGILL, Liang-jiang WANG†‡

Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina 29634, USA

†E-mail: liangjw@clemson.edu

Received Mar. 29, 2019; Revision accepted Apr. 15, 2019; Crosschecked Apr. 28, 2019

Abstract: Life may have begun in an RNA world, which is supported by increasing evidence of the vital role that RNAs perform in biological systems. In the human genome, most genes actually do not encode proteins; they are noncoding RNA genes. The largest class of noncoding genes is known as long noncoding RNAs (lncRNAs), which are transcripts greater in length than 200 nucleotides, but with no protein-coding capacity. While some lncRNAs have been demonstrated to be key regulators of gene expression and 3D genome organization, most lncRNAs are still uncharacterized. We thus propose several data mining and machine learning approaches for the functional annotation of human lncRNAs by leveraging the vast amount of data from genetic and genomic studies. Recent results from our studies and those of other groups indicate that genomic data mining can give insights into lncRNA functions and provide valuable information for experimental studies of candidate lncRNAs associated with human disease.

Key words: Long noncoding RNA; Functional annotation; Genomic data mining; Machine learning
https://doi.org/10.1631/jzus.B1900162 **CLC number:** Q522

1 Introduction

The human genome project was a monumental undertaking to sequence the entire set of human chromosomes, which many scientists believed would unlock the secrets of our genome. However, after the completion of human genome sequencing, it was discovered that humans were somewhere between chickens and grapes in terms of the number of protein-encoding genes (Pertea and Salzberg, 2010). Approximately 22000 human genes were discovered, and this outcome was a surprise to the scientific community who had estimated about 100000 genes in the human genome (Pertea and Salzberg, 2010). This relatively small number of genes corresponds to only a few percent of the total human genome, while the rest of the noncoding genome does not encode proteins. The noncoding DNA was referred to as "junk DNA" due to its lack of protein-coding capacity and the presence of pseudogenes, transposons, and repetitive regions.

The advent of high-throughput technologies allowed the genome-wide detection of noncoding RNAs, which showed the pervasiveness of transcription in the genome. From these genome-wide analyses, it is now known that the majority of the genome is actively transcribed (Hangauer et al., 2013). Why would natural selection favor the transcription, which costs energy, of "junk DNA" with no biological purpose? This question rests on the assumption that "junk DNA" has no function, which is now known to be invalid. The term "junk DNA" is obsolete after genomic analyses discovered that over 80% of the human genome possesses biochemical functions (ENCODE Project Consortium, 2012). Genomics research has

‡ Corresponding author
* Project supported by the Self Regional Healthcare Foundation, USA
ORCID: Liang-jiang WANG, https://orcid.org/0000-0002-6316-7962

not only shed light on the dark matter of the genome, but also championed for a redefinition of the term "gene," owing to the vast amount of evidence for functional noncoding RNAs. Genes by definition are no longer always required to encode proteins, thus creating two major classes of genes: those which encode proteins are protein-coding genes while those that do not are noncoding RNAs. This redefinition is of tremendous importance as both noncoding and protein-coding genes are functionally intertwined within the gene network of the genome.

High-throughput RNA-sequencing has been used to discover tens of thousands of long noncoding RNAs (lncRNAs), which are greater than 200 nucleotides in length and do not encode proteins. The nucleotide length threshold of 200 nucleotides is largely arbitrary but does serve an essential purpose in separating these transcripts from the well-known small noncoding RNAs, such as transfer RNAs (tRNAs), microRNAs (miRNAs), and small nucleolar RNAs (snoRNAs). A meta-analysis of 7256 human RNA-sequencing profiles identified 58 648 lncRNA genes, suggesting that 68% of the human transcriptome may be lncRNAs (Iyer et al., 2015). In total, nearly 99 000 human genes were identified, which is close to the estimate of 100 000 prior to the human genome project, but only about 22 000 are protein-coding genes (Iyer et al., 2015). Therefore, it has become evident that lncRNAs are abundant within the human genome with active biological functions.

LncRNAs represent the largest class of noncoding genes with several sub-classes based on their genomic positions relative to protein-coding genes. In order of decreasing prevalence, the major lncRNA sub-classes are long intergenic noncoding RNAs (lincRNAs), antisense lncRNAs (AS-lncRNAs), sense lncRNAs, and bidirectional lncRNAs (Derrien et al., 2012). These various lncRNAs generally share common features, including being predominantly spliced, expressed at low levels, tissue-specific, and having exonic regions with low levels of interspecies sequence conservation (Derrien et al., 2012). Interestingly, lncRNA promoters are conserved at a similar level relative to protein-coding genes, suggesting that lncRNAs are positively selected and thus are functionally important (Derrien et al., 2012). LncRNAs are commonly transcribed by RNA polymerase II and generally modified post-transcriptionally as messenger RNAs (mRNAs), including 5' capping, polyadenylation, and splicing (Quinn and Chang, 2016). While the biogenesis of lncRNAs may be very similar to mRNAs in most cases, a key difference has been discovered recently; knockouts of the ribonuclease Dicer, responsible for generating miRNAs, resulted in the decreased expression levels of hundreds of lncRNAs, but not mRNAs (Zheng et al., 2014). Another intriguing difference between mRNAs and lncRNAs is that some lncRNA transcripts possess higher-order structures, such as 3' secondary cloverleaf structures similar to tRNAs. These 3' secondary structures are cleaved by ribonuclease P to form the mature lncRNA with a 3' triple helix structure which is predicted to increase transcript stability and facilitate nuclear retention (Quinn and Chang, 2016).

2 Functional mechanisms of lncRNAs

Although many lncRNAs are expressed by the human genome, only a few have been functionally characterized. The known functions of lncRNAs are generally within four major mechanistic themes, which are to act as a signal, decoy, guide, or scaffold (Wu et al., 2013). These different mechanisms can act to regulate other genes at the transcriptional, post-transcriptional, translational, or epigenetic levels. As shown in Fig. 1, lncRNAs with signaling functions act as a molecular marker to indicate specific biological conditions, and then induce a response such as histone modifications. An example is the lncRNA, X-inactive specific transcript (*XIST*), which initiates X chromosome inactivation in females for dosage compensation. By coating the X chromosome to be inactivated, *XIST* transcripts signal successive epigenetic modifications such as DNA methylation, histone methylation, and histone ubiquitination (Morris, 2016). Decoy lncRNAs function through the sequence-based competitive binding of molecules. This binding is commonly observed in lncRNAs acting as miRNA sponges to reduce miRNA efficacy (Geisler and Coller, 2013). Guide lncRNAs bind proteins such as transcription factors, thereby recruiting these proteins to specific genomic loci (Werner and Ruthenburg, 2015). Many lncRNAs can function as guides by tethering to chromatin and facilitating the binding of protein complexes such as

polycomb repressive complex 2 (PRC2) and RNA polymerase II (Werner and Ruthenburg, 2015). The last major functional theme of lncRNAs is to act as scaffolds, which mediate the physical interactions between proteins and noncoding RNAs, forming ribonucleoprotein complexes. For example, the lncRNA *HOTAIR* directly facilitates the binding of E3 ubiquitin ligases with multiple substrates for ubiquitination, thereby acting as a scaffold for protein ubiquitination (Yoon et al., 2013). It is important to note that one lncRNA is not confined to a single functional mechanism and can exhibit multiple functions simultaneously.

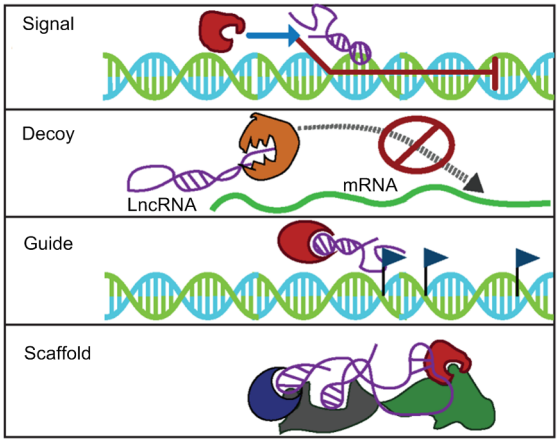

Fig. 1 Mechanistic themes of lncRNA functions
LncRNAs are shown in purple acting as a signal, decoy, guide, or scaffold. Signal lncRNAs act in response to a stimulus to induce gene regulation, such as repression, in a spatiotemporal manner. Decoy lncRNAs act as competitive inhibitors, such as miRNA sponges, thereby preventing the degradation of the targeted mRNA. Guide lncRNAs bind complexes such as chromatin-modifying enzymes and facilitate the targeting of specific genomic loci either in *cis* or *trans*. Scaffold lncRNAs act as a molecular glue to facilitate the interaction of multiple proteins into a ribonucleoprotein complex. Modified from Wang and Chang (2011)

Like proteins, lncRNA functionality is dependent on proper subcellular localization. While most lncRNAs are shown to be enriched in the nucleus, some also localize and function in the cytoplasm (Cabili et al., 2015; Chen, 2016; Carlevaro-Fita and Johnson, 2019). As shown in Fig. 2, lncRNAs play important roles in nuclear chromatin organization, epigenetic modification, transcriptional regulation, and RNA splicing (Sun et al., 2018). For instance, *MALAT1* and *NEAT1* are enriched predominantly in nuclear speckles and paraspeckles, respectively, and are involved in nuclear architecture organization and RNA splicing (Clemson et al., 2009; Tripathi et al., 2010). In the cytoplasm, lncRNAs can regulate gene expression at post-transcriptional levels through modulating translational efficiency, acting as miRNA sponges, affecting RNA stability, and facilitating the subcellular transport of ribonucleoprotein complexes (Rashid et al., 2016). However, the factors that govern lncRNA subcellular localization are mostly unknown. A previous study identified a nuclear retention motif in the lncRNA BMP/OP-responsive gene (*BORG*) through a mutational screen (Zhang et al., 2014). Interestingly, the number of copies of this motif present in lncRNAs correlated with the nuclear to cytoplasmic transcript ratio, whereas mutations of the motif resulted in the loss of nuclear retention.

3 Genomic data mining and machine learning for functional annotation of lncRNAs

The functional characterization of lncRNAs using experimental approaches, such as gene knockouts, is not straightforward and can be highly time-consuming (Cao HF et al., 2018). The current methodology for functional genomics is designed primarily for protein-coding genes. The laborious process of characterizing lncRNA functions can be facilitated by genomic data mining, which utilizes genomic data to extract hidden knowledge regarding a specific biological question. Biological knowledge is gained by using data mining algorithms which identify patterns and relationships within the data. Genomic data mining typically consists of three major steps, including dataset acquisition, data integration, and the application of data mining algorithms. Dataset acquisition involves querying the available databases for biological data which are relevant to the hypothesis at hand. Data integration is the aggregation of diverse or heterogeneous datasets so that generalizable knowledge can be extracted. Lastly, data mining algorithms are applied to the integrated data for knowledge discovery and addressing biological questions. While still a relatively new discipline, genomic data mining is of great importance and will continue to grow in demand proportionally to the vast amount of genomic data being generated.

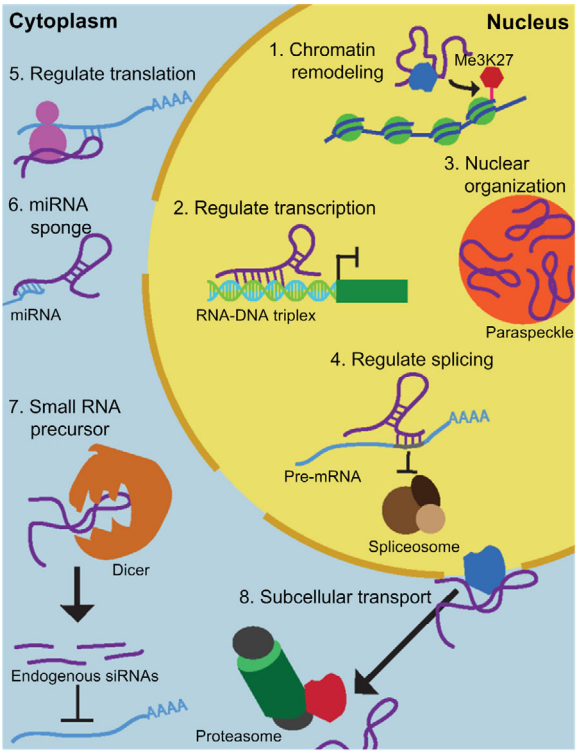

Fig. 2 LncRNA functions in the cellular context

1: LncRNAs can bind chromatin-modifying enzymes and facilitate histone modifications such as the trimethylation of histone 3 at lysine 27 (me3K27) to induce gene silencing. 2: LncRNAs can form an RNA-DNA triplex which blocks accessibility to gene promoter regions. 3: LncRNAs facilitate the organization of nuclear structures such as paraspeckles. 4: LncRNAs can bind pre-mRNAs to affect alternative splicing. 5: LncRNAs can interact with mRNAs and ribosomes to regulate translation. 6: LncRNAs are capable of sequestering miRNAs as a miRNA sponge, thereby preventing the degradation of targeted mRNAs. 7: LncRNAs may be processed by Dicer to produce endogenous small interfering RNAs (siRNAs). 8: LncRNAs can also function in the subcellular localization of proteins to complexes such as the proteasome. Modified from Rashid et al. (2016)

As a major subfield of data mining, machine learning algorithms create models by learning from data on their own without explicit instructions. Machine learning requires a task, a means of scoring the performance of the algorithm on the task, and experience upon which to learn. The experience itself is data, and the nature of the data determines the type of machine learning. If the data instances have labels (known values of the response variable), the task is said to be supervised learning; if the data are unlabeled, then it is unsupervised learning. Because supervised learning can extract generalizable knowledge from labeled data, the resulting model can be used to predict the response variable for new data instances with unknown labels. In contrast, unsupervised learning does not require labeled data, but may identify hidden structures within the unlabeled data for clustering data instances into representative groups.

With the growing size and complexity of genomic data, machine learning algorithms are needed to discover knowledge in a timely and efficient manner. Although both supervised and unsupervised learning algorithms are widely used for genomic data mining, supervised learning is often preferred due to its capability to learn novel complex patterns and make biologically relevant predictions. Two of the most popular supervised learning algorithms are the support vector machine (SVM) and random forest (RF). Both algorithms are easy to implement with a small number of parameters, but can achieve high accuracy for both linear and non-linear problems. SVM and RF have been used for a diverse array of biological problems ranging from the classification of lncRNAs to the prediction of lncRNA–protein interactions and autism spectrum disorder (ASD)-associated lncRNAs (Muppirala et al., 2011; Cogill and Wang, 2016; Pian et al., 2016).

While conventional machine learning algorithms such as SVM and RF can achieve high prediction accuracy for most biological problems, recently a new set of advanced algorithms, known as deep learning algorithms, have shown a better performance for complex problems (Ching et al., 2018). Deep neural networks with multiple layers of artificial neurons are commonly used for deep learning of complex patterns. The number of neuronal layers gives the depth of a deep learning model, with each layer transforming its inputs to derive new and more sophisticated features. The automatic learning of feature representation is regarded as the main advantage of deep learning over conventional machine learning which restricts inputs to human-engineered features. Potential issues with deep learning include the requirement of large datasets to learn generalizable knowledge due to the copious number of parameters that need to be learned. However, in the age of big data, this is becoming less of a concern, which is why deep learning will likely become the next frontier of machine learning in computational genomics.

We propose six major approaches of genomic data mining for the functional annotation of human lncRNAs (Fig. 3). These approaches, as discussed in detail in the following sections, apply machine learning and data mining techniques to the ever-increasing amount of publicly available data from human genetic and genomic studies. The findings are not only useful for lncRNA functional annotation, but they also provide valuable information for experimental studies of candidate lncRNAs.

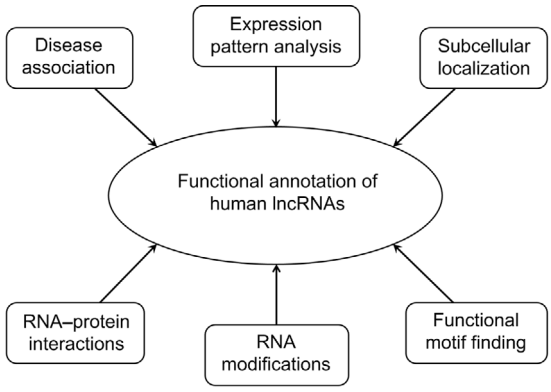

Fig. 3 Genomic data mining approaches proposed for the functional annotation of human lncRNAs

4 Disease association of lncRNAs

LncRNAs have been implicated in many human diseases. This involvement is not surprising as they are versatile regulators of gene expression with known roles in tissue development. Mutations in lncRNAs can alter their functional efficacy, thereby causing aberrant downstream consequences. Interestingly, more than 90% of disease-associated single nucleotide polymorphisms (SNPs) are found within noncoding regions of the human genome (Maurano et al., 2012; Ricaño-Ponce and Wijmenga, 2013). These SNPs may alter the expression levels of lncRNAs (Kumar et al., 2013). In addition, copy number variants (CNVs) have also been shown to change the expression levels of lncRNAs associated with cancer (Xu et al., 2017).

Many lncRNAs are specifically expressed in neuronal tissues, and thus may be associated with brain disorders (Derrien et al., 2012). In mammals, lncRNAs have important roles in neural differentiation and synaptic plasticity (Wu et al., 2013; Clark and Blackshaw, 2014). Therefore, it is not surprising that lncRNAs are implicated in neurodegenerative, psychiatric, and neurodevelopmental disorders. In previous studies, lncRNAs have been shown to be associated with two of the most predominant neurological disorders, intellectual disability (ID) and ASD (van de Vondervoort et al., 2013; Ziats and Rennert, 2013; Cajigas et al., 2015; Wang Y et al., 2015). ID and ASD are clinically and genetically heterogeneous complex disorders, affecting up to 3% and 1% of the human population, respectively (Srivastava and Schwartz, 2014). ID is characterized by diminished intellectual capacity and adaptive reasoning, whereas ASD is recognized by impaired social communications and restrictive or repetitive behavior. Both disorders originate in early childhood, and involve a large number of genes, many of which are associated with the synaptic transmission pathway (Verpelli et al., 2013; de Rubeis et al., 2014). However, in most cases of ID or ASD, the specific genetic factors of the disorders are still unable to be determined (O'Roak et al., 2012; Kiser et al., 2015). Until recently, only protein-coding genes were studied for their involvement in ID and ASD. It is thus likely that many of these genetic factors may reside in lncRNAs.

We have developed an SVM model for the expression-based prediction of ASD risk genes (Cogill and Wang, 2016). The SVM model, trained using brain developmental gene expression profiles of known ASD risk genes (protein-coding), demonstrated the ability to classify and prioritize ASD candidate genes accurately. This model was then used to predict ASD-associated candidate lncRNAs based on their developmental expression patterns. Of brain-expressed lncRNAs, 63 were predicted as ASD-associated candidates with high confidence, and the lncRNAs previously related to brain development and neurodevelopmental disorders were also prioritized highly in the candidate gene list (Cogill and Wang, 2016). Our study proposed a novel approach for knowledge transfer from known ASD risk protein-coding genes to lncRNAs, which should facilitate experimental investigation into these candidate lncRNAs. Moreover, the general machine learning strategy may also be applied to other diseases such as ID and cancer. The disease association of lncRNAs has also been studied using gene co-expression network analysis, as discussed in the next section.

5 Expression pattern analysis

A well-known property of lncRNAs is their tissue and developmental specificity. Thus, the expression pattern of a lncRNA can be used to help understand its biological function. Expression pattern analysis is especially useful for lncRNAs because they do not encode proteins. With the RNA transcript as the functional unit, the biological function of a lncRNA may be investigated by examining the expression differences between various groups of samples, such as diseased versus control tissues or fetal brains versus adult brains. The two most common expression-based approaches are differential expression analysis and co-expression network analysis, both of which have been used to investigate lncRNAs (Liao et al., 2011; Necsulea et al., 2014; Chaudhary et al., 2017; Gudenas et al., 2017; Cogill et al., 2018).

Differential expression analysis identifies genes that show statistically significant differences in expression levels between two or more conditions and is commonly used to find genes associated with a disease, tissue type, or experimental treatment. Gene co-expression network analysis is a clustering method which enables the inference of a gene's biological function based on the strength of connections to other genes with a known function. As shown in Fig. 4, this method is used to cluster genes by their expression profiles into gene groups, known as gene modules. These gene modules can then be functionally annotated through gene set enrichment analysis, which uses a statistical test to check if the overall functional enrichment is different from what would be expected by chance. Thus, gene co-expression network analysis leverages the biological knowledge of known genes to gain insight into the uncharacterized genes through a guilt-by-association heuristic.

We have used differential expression analysis and co-expression network analysis for identification and functional annotation of candidate lncRNAs associated with ASD (Gudenas et al., 2017; Cogill et al., 2018). Various genomic datasets were integrated and then used to identify and prioritize a list of high-confidence candidate lncRNAs that were differentially expressed in the ASD brain, co-expressed with known ASD risk genes during neurodevelopment, and co-located with ASD-associated CNVs (Gudenas et al., 2017). Gene co-expression network analysis also identified two distinct groups of lncRNA modules showing elevated prenatal and postnatal expression patterns, respectively (Cogill et al., 2018). The functional analysis of these modules revealed that the prenatal modules were enriched with transcriptional regulators, while the postnatal modules were associated with synapse formation. Thus, our findings provide insight into the genetic etiology of ASD and the important functions of lncRNAs during early brain development. In addition, gene co-expression network analysis was also used to identify candidate lncRNAs associated with ID (Gudenas and Wang, 2015) and cancer (Cogill and Wang, 2014).

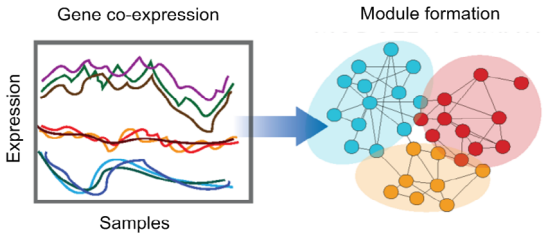

Fig. 4 Schematic diagram of gene co-expression network analysis

Genes are clustered by their correlation of co-expression with other genes, resulting in gene modules shown in different colors. Within each gene module, nodes represent genes while edges represent correlations. The length of an edge is inversely proportional to the correlation of co-expression

6 Subcellular localization

The subcellular localization of a lncRNA can reveal insight into its biological function (Chen, 2016; Carlevaro-Fita and Johnson, 2019). Gene regulation by lncRNAs at the transcriptional, post-transcriptional, or epigenetic level is performed within the nucleus, whereas translational control, binding miRNAs and producing endogenous siRNAs are some lncRNA functions exclusive to the cytoplasm (Rashid et al., 2016; Sun et al., 2018). Therefore, predicting the subcellular localization of lncRNAs can provide useful information about their biological functions.

We have recently developed a deep neural network model, called DeepLncRNA, to predict the subcellular localization of a lncRNA from its transcript sequence (Gudenas and Wang, 2018). The model was constructed using a comprehensive dataset of

nuclear and cytosolic lncRNAs compiled through large-scale analysis of RNA-seq data from the ENCODE project (ENCODE Project Consortium, 2012). DeepLncRNA achieved superior performance when compared with conventional machine learning algorithms such as SVM and RF. The high accuracy of DeepLncRNA suggests that lncRNA transcripts may contain sequence motifs essential for subcellular localization. DeepLncRNA also compares favorably with two other models, lncLocator and iLoc-LncRNA (Su et al., 2018), for predicting lncRNA subcellular localization. LncLocator uses a stacked autoencoder to derive high-level sequence features for an ensemble of SVM and RF models to predict five subcellular localizations (Cao Z et al., 2018), whereas iLoc-LncRNA utilizes pseudo K-tuple nucleotide composition (PseKNC) features to train a multi-class SVM model (Su et al., 2018). However, both iLoc-LncRNA and lncLocator were constructed using a relatively small dataset of lncRNAs (<1000) from various organisms. In contrast, DeepLncRNA has been constructed using a large number of human lncRNAs (>8000), and thus may be particularly suitable for the functional annotation of human lncRNAs.

7 Functional motif discovery

LncRNAs function in various biological processes, and the functional versatilities may rely on their abilities to form different structures and diverse molecular interactions with DNA, RNA, and proteins (Guttman and Rinn, 2012; Zampetaki et al., 2018). However, lncRNA structure prediction is a research area still in its infancy, mainly due to the scarcity of experimentally validated lncRNA structures. Since the primary sequence ultimately dictates the structure, the determinants of lncRNA functionality should be present within the lncRNA transcript sequence. While generally not well conserved at the sequence level, lncRNAs sharing the same function often show similarities in a combination of sequence motifs and structural elements (Achar and Sætrom, 2015). Therefore, finding the motifs present in lncRNA transcripts can provide useful information for functional annotation.

Several well-known RNA motifs are present in lncRNAs and may be critical for their functions. For instance, G-rich lncRNAs can contain G-quadruplexes (G4s), in which four guanines are organized in a planar arrangement to form stacks of G-quartets (Cammas and Millevoi, 2017). G4s often affect cellular activities through interaction with G4-binding proteins and other recruited protein regulators (Brázda et al., 2014). Besides their known regulatory role in RNA metabolism, G4s may also be involved in transcription, recombination, and telomere homeostasis (Cammas and Millevoi, 2017). Another example of functional motifs is the kissing complex, a form of RNA pseudoknot, in which base pairs are formed between the unpaired nucleotides of two hairpin loops. RNAs with the kissing complex can bind to the KH2 domain of Fragile-X mental retardation protein (FMRP), and thus may be the targets for FMRP-mediated translational regulation (Darnell et al., 2005). In addition, sequence motifs can be associated with specific lncRNA higher-order structures such as the AUGC tetraloop motif (Li et al., 2016). However, it is likely that most of the functional motifs in lncRNAs remain to be discovered and characterized. The high functional diversity but low sequence conservation of lncRNAs can make it challenging to analyze functional motifs. The recent development of deep learning techniques, such as convolutional neural networks, should greatly facilitate lncRNA motif discovery and functional annotation.

8 RNA modifications

Post-transcriptionally modified nucleotides such as N^6-methyladenosine (m^6A), 5-methylcytosine (m^5C), and pseudouridine (Ψ) are known to be critical for proper RNA function. In human ribosomal RNAs (rRNAs), there are 91 pseudouridines and 10 methylated nucleotides. These chemical modifications can influence molecular interactions and conformations of rRNAs (Wang and He, 2014). tRNAs also undergo extensive post-transcriptional modifications to ensure their proper structure and function (Jackman and Alfonzo, 2013). More importantly, m^6A methylation is probably the most prevalent modification in mammalian RNAs, and it is a dynamic process mediated by methyltransferases and demethylases (Wang and He, 2014; Song and Yi, 2017). In mRNAs, m^6A modification sites are distributed mainly near the stop codon, but also in other regions (Ke et al., 2015; Linder et al., 2015). A consensus motif, DRACH

(D=A/G/U, R=A/G, H=A/C/U), has been identified for RNA m^6A modification, but not all adenosines within this motif are methylated. This dynamically regulated modification is involved in many aspects of mRNA metabolism, such as alternative splicing, degradation, and translation (Wang X et al., 2014, 2015; Liu et al., 2015).

Relatively less is known about the nucleotide modifications in lncRNAs. It has been shown that the lncRNA *XIST* has at least 78 m^6A residues which are essential for *XIST* function in X chromosome inactivation (Patil et al., 2016). Several other lncRNAs, including *MALAT1*, *TUG1*, and *NEAT1*, also contain multiple m^6A sites, but their roles are still unclear (Wang and He, 2014). Further research is needed to elucidate the patterns and roles of lncRNA nucleotide modifications; genomic data mining methods can facilitate this endeavor. Although several machine learning models have been developed to predict m^6A sites from mRNA sequences (Zhou et al., 2016; Zhang and Hamada, 2018; Zou et al., 2019), it remains to be determined whether these models can also be used for lncRNA functional annotation.

9 RNA–protein interactions

Many functions of lncRNAs are mediated by their interactions with RNA-binding proteins (Ferrè et al., 2016). For instance, some lncRNAs regulate chromatin status by interacting with chromatin modifiers such as PRC2 (Davidovich and Cech, 2015; Jin et al., 2018). LncRNAs can be involved in organizing 3D genome architecture by interacting with proteins such as the CCCTC-binding factor (CTCF) (Sun et al., 2013; Kung et al., 2015). RNA–protein interactions are also essential for the lncRNA *NEAT1* to organize nuclear paraspeckles (Clemson et al., 2009). Moreover, lncRNAs may interact with transcription factors to modulate their regulatory activities on gene expression (Ponting et al., 2009; Wang and Chang, 2011). The importance of lncRNA–protein interactions is further evidenced by their involvement in human diseases, including cancer (Huarte et al., 2010; Yang et al., 2018) and neurological disorders (Guo et al., 2018; Li et al., 2019).

Several machine learning models have been developed to predict lncRNA–protein interactions from their sequences. The model RPISeq was shown to predict lncRNA–protein interactions with an accuracy of 80% from lncRNA and protein sequence pairs using SVM and RF algorithms (Muppirala et al., 2011). LncPro used a structure-based approach to predict lncRNA–protein interactions by first deriving structure-based features from the lncRNA and protein sequences, and achieved a prediction accuracy similar to RPISeq (Lu et al., 2013). Moreover, with the addition of a stacked denoising autoencoder, a type of deep neural network, the IPminer model was able to achieve an accuracy of 89% for predicting lncRNA–protein interactions from their sequences (Pan et al., 2016). While the above models achieved high accuracy on some specific datasets, it has not yet been demonstrated whether they can be used to predict novel lncRNA–protein interactions. Further studies are needed to thoroughly evaluate these models for functional annotation of lncRNAs.

10 Concluding remarks

In this review, we have discussed several approaches of genomic data mining for the functional annotation of human lncRNAs. The human genome encodes a large number of noncoding RNAs, mostly lncRNAs, which are involved in gene regulation and 3D genome organization necessary for cellular function and development. However, most lncRNAs are still functionally uncharacterized as experimental approaches remain difficult and costly. With the rapid accumulation of genomic data, machine learning and data mining algorithms have been utilized to develop novel and integrative approaches for the functional analyses of lncRNAs and the resulting knowledge can provide biological insights into their regulatory roles in development and disease. Since lncRNAs do not encode proteins, these genomic data mining approaches are based on their expression patterns and transcript sequences. In future, as experimentally determined 3D structures of lncRNAs accumulate, structure-based strategies will likely fill in some of the missing pieces for deciphering lncRNA functionality. The genomic and functional annotation of lncRNAs is a very challenging task, which has only just begun, and this endeavor will certainly enhance our understanding of human biology and disease.

Contributors

Brian L. GUDENAS prepared the first draft of the manuscript. Jun WANG, Shu-zhen KUANG, An-qi WEI, and Steven B. COGILL contributed to manuscript writing and revisions. Liang-jiang WANG supervised the project and edited the manuscript. All authors read and approved the final manuscript.

Compliance with ethics guidelines

Brian L. GUDENAS, Jun WANG, Shu-zhen KUANG, An-qi WEI, Steven B. COGILL, and Liang-jiang WANG declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Achar A, Sætrom P, 2015. RNA motif discovery: a computational overview. *Biol Direct*, 10:61. https://doi.org/10.1186/s13062-015-0090-5

Brázda V, Hároníková L, Liao JCC, et al., 2014. DNA and RNA quadruplex-binding proteins. *Int J Mol Sci*, 15(10): 17493-17517. https://doi.org/10.3390/ijms151017493

Cabili MN, Dunagin MC, McClanahan PD, et al., 2015. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol*, 16:20. https://doi.org/10.1186/s13059-015-0586-4

Cajigas I, Leib DE, Cochrane J, et al., 2015. *Evf2* lncRNA/BRG1/DLX1 interactions reveal RNA-dependent inhibition of chromatin remodeling. *Development*, 142(15): 2641-2652. https://doi.org/10.1242/dev.126318

Cammas A, Millevoi S, 2017. RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res*, 45(4):1584-1595. https://doi.org/10.1093/nar/gkw1280

Cao HF, Wahlestedt C, Kapranov P, 2018. Strategies to annotate and characterize long noncoding RNAs: advantages and pitfalls. *Trends Genet*, 34(9):704-721. https://doi.org/10.1016/j.tig.2018.06.002

Cao Z, Pan XY, Yang Y, et al., 2018. The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics*, 34(13):2185-2194. https://doi.org/10.1093/bioinformatics/bty085

Carlevaro-Fita J, Johnson R, 2019. Global positioning system: understanding long noncoding RNAs through subcellular localization. *Mol Cell*, 73(5):869-883. https://doi.org/10.1016/j.molcel.2019.02.008

Chaudhary R, Gryder B, Woods WS, et al., 2017. Prosurvival long noncoding RNA *PINCR* regulates a subset of p53 targets in human colorectal cancer cells by binding to Matrin 3. *eLife*, 6:e23244. https://doi.org/10.7554/eLife.23244

Chen LL, 2016. Linking long noncoding RNA localization and function. *Trends Biochem Sci*, 41(9):761-772. https://doi.org/10.1016/j.tibs.2016.07.003

Ching T, Himmelstein DS, Beaulieu-Jones BK, et al., 2018. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*, 15(141):20170387. https://doi.org/10.1098/rsif.2017.0387

Clark BS, Blackshaw S, 2014. Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease. *Front Genet*, 5:164. https://doi.org/10.3389/fgene.2014.00164

Clemson CM, Hutchinson JN, Sara SA, et al., 2009. An architectural role for a nuclear noncoding RNA: *NEAT1* RNA is essential for the structure of paraspeckles. *Mol Cell*, 33(6):717-726. https://doi.org/10.1016/j.molcel.2009.01.026

Cogill SB, Wang LJ, 2014. Co-expression network analysis of human lncRNAs and cancer genes. *Cancer Inform*, 13(Suppl 5):49-59. https://doi.org/10.4137/CIN.S14070

Cogill SB, Wang LJ, 2016. Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. *Bioinformatics*, 32(23):3611-3618. https://doi.org/10.1093/bioinformatics/btw498

Cogill SB, Srivastava AK, Yang MQ, et al., 2018. Co-expression of long non-coding RNAs and autism risk genes in the developing human brain. *BMC Syst Biol*, 12(Suppl 7):91. https://doi.org/10.1186/s12918-018-0639-x

Darnell JC, Fraser CE, Mostovetsky O, et al., 2005. Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes Dev*, 19(8):903-918. https://doi.org/10.1101/gad.1276805

Davidovich C, Cech TR, 2015. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA*, 21(12):2007-2022. https://doi.org/10.1261/rna.053918.115

de Rubeis S, He X, Goldberg AP, et al., 2014. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209-215. https://doi.org/10.1038/nature13772

Derrien T, Johnson R, Bussotti G, et al., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22(9):1775-1789. https://doi.org/10.1101/gr.132159.111

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57-74. https://doi.org/10.1038/nature11247

Ferrè F, Colantoni A, Helmer-Citterich M, 2016. Revealing protein–lncRNA interaction. *Brief Bioinform*, 17(1):106-116. https://doi.org/10.1093/bib/bbv031

Geisler S, Coller J, 2013. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol*, 14(11):699-712. https://doi.org/10.1038/nrm3679

Gudenas BL, Wang LJ, 2015. Gene coexpression networks in human brain developmental transcriptomes implicate the association of long noncoding RNAs with intellectual disability. *Bioinform Biol Insights*, 9(Suppl 1):21-27. https://doi.org/10.4137/BBI.S29435

Gudenas BL, Wang LJ, 2018. Prediction of lncRNA subcellular localization with deep learning from sequence features. *Sci Rep*, 8(1):16385. https://doi.org/10.1038/s41598-018-34708-w

Gudenas BL, Srivastava AK, Wang LJ, 2017. Integrative genomic analyses for identification and prioritization of long non-coding RNAs associated with autism. *PLoS ONE*, 12(5):e0178532. https://doi.org/10.1371/journal.pone.0178532

Guo Y, Chen X, Xing RX, et al., 2018. Interplay between FMRP and *lncRNA TUG1* regulates axonal development through mediating SnoN-Ccd1 pathway. *Hum Mol Genet*, 27(3):475-485. https://doi.org/10.1093/hmg/ddx417

Guttman M, Rinn JL, 2012. Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385):339-346. https://doi.org/10.1038/nature10887

Hangauer MJ, Vaughn IW, McManus MT, 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*, 9(6):e1003569. https://doi.org/10.1371/journal.pgen.1003569

Huarte M, Guttman M, Feldser D, et al., 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3):409-419. https://doi.org/10.1016/j.cell.2010.06.040

Iyer MK, Niknafs YS, Malik R, et al., 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*, 47(3):199-208. https://doi.org/10.1038/ng.3192

Jackman JE, Alfonzo JD, 2013. Transfer RNA modifications: nature's combinatorial chemistry playground. *Wiley Interdiscip Rev RNA*, 4(1):35-48. https://doi.org/10.1002/wrna.1144

Jin JJ, Lv W, Xia P, et al., 2018. Long noncoding RNA *SYISL* regulates myogenesis by interacting with polycomb repressive complex 2. *Proc Natl Acad Sci USA*, 115(42): E9802-E9811. https://doi.org/10.1073/pnas.1801471115

Ke SD, Alemu EA, Mertens C, et al., 2015. A majority of m^6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev*, 29(19):2037-2053. https://doi.org/10.1101/gad.269415.115

Kiser DP, Rivero O, Lesch KP, 2015. Annual research review: the (epi)genetics of neurodevelopmental disorders in the era of whole-genome sequencing—unveiling the dark matter. *J Child Psychol Psychiatry*, 56(3):278-295. https://doi.org/10.1111/jcpp.12392

Kumar V, Westra HJ, Karjalainen J, et al., 2013. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet*, 9(1):e1003201. https://doi.org/10.1371/journal.pgen.1003201

Kung JT, Kesner B, An JY, et al., 2015. Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol Cell*, 57(2):361-375. https://doi.org/10.1016/j.molcel.2014.12.006

Li L, Zhuang YL, Zhao XS, et al., 2019. Long non-coding RNA in neuronal development and neurological disorders. *Front Genet*, 9:744. https://doi.org/10.3389/fgene.2018.00744

Li R, Zhu HL, Luo YB, 2016. Understanding the functions of long non-coding RNAs through their higher-order structures. *Int J Mol Sci*, 17(5):E702. https://doi.org/10.3390/ijms17050702

Liao Q, Liu CN, Yuan XY, et al., 2011. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res*, 39(9): 3864-3878. https://doi.org/10.1093/nar/gkq1348

Linder B, Grozhik AV, Olarerin-George AO, et al., 2015. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods*, 12(8):767-772. https://doi.org/10.1038/nmeth.3453

Liu N, Dai Q, Zheng GQ, et al., 2015. N^6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, 518(7540):560-564. https://doi.org/10.1038/nature14234

Lu QS, Ren SJ, Lu M, et al., 2013. Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics*, 14:651. https://doi.org/10.1186/1471-2164-14-651

Maurano MT, Humbert R, Rynes E, et al., 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190-1195. https://doi.org/10.1126/science.1222794

Morris KV, 2016. Long Non-coding RNAs in Human Disease. Springer International Publishing, Cham, Germany. https://doi.org/10.1007/978-3-319-23907-1

Muppirala UK, Honavar VG, Dobbs D, 2011. Predicting RNA–protein interactions using only sequence information. *BMC Bioinformatics*, 12:489. https://doi.org/10.1186/1471-2105-12-489

Necsulea A, Soumillon M, Warnefors M, et al., 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635-640. https://doi.org/10.1038/nature12943

O'Roak BJ, Vives L, Girirajan S, et al., 2012. Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature*, 485(7397):246-250. https://doi.org/10.1038/nature10989

Pan XY, Fan YX, Yan JC, et al., 2016. IPMiner: hidden

ncRNA–protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics*, 17:582. https://doi.org/10.1186/s12864-016-2931-8

Patil DP, Chen CK, Pickering BF, et al., 2016. m^{6}A RNA methylation promotes *XIST*-mediated transcriptional repression. *Nature*, 537(7620):369-373. https://doi.org/10.1038/nature19342

Pertea M, Salzberg SL, 2010. Between a chicken and a grape: estimating the number of human genes. *Genome Biol*, 11(5):206. https://doi.org/10.1186/gb-2010-11-5-206

Pian C, Zhang GL, Chen Z, et al., 2016. LncRNApred: classification of long non-coding RNAs and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. *PLoS ONE*, 11(5):e0154567. https://doi.org/10.1371/journal.pone.0154567

Ponting CP, Oliver PL, Reik W, 2009. Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629-641. https://doi.org/10.1016/j.cell.2009.02.006

Quinn JJ, Chang HY, 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet*, 17(1):47-62. https://doi.org/10.1038/nrg.2015.10

Rashid F, Shah A, Shan G, 2016. Long non-coding RNAs in the cytoplasm. *Genomics Proteomics Bioinformatics*, 14(2):73-80. https://doi.org/10.1016/j.gpb.2016.03.005

Ricaño-Ponce I, Wijmenga C, 2013. Mapping of immune-mediated disease genes. *Annu Rev Genomics Hum Genet*, 14:325-353. https://doi.org/10.1146/annurev-genom-091212-153450

Song JH, Yi CQ, 2017. Chemical modifications to RNA: a new layer of gene expression regulation. *ACS Chem Biol*, 12(2):316-325. https://doi.org/10.1021/acschembio.6b00960

Srivastava AK, Schwartz CE, 2014. Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms. *Neurosci Biobehav Rev*, 46:161-174. https://doi.org/10.1016/j.neubiorev.2014.02.015

Su ZD, Huang Y, Zhang ZY, et al., 2018. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics*, 34(24):4196-4204. https://doi.org/10.1093/bioinformatics/bty508

Sun QY, Hao QY, Prasanth KV, 2018. Nuclear long noncoding RNAs: key regulators of gene expression. *Trends Genet*, 34(2):142-157. https://doi.org/10.1016/j.tig.2017.11.005

Sun S, del Rosario BC, Szanto A, et al., 2013. Jpx RNA activates *Xist* by evicting CTCF. *Cell*, 153(7):1537-1551. https://doi.org/10.1016/j.cell.2013.05.028

Tripathi V, Ellis JD, Shen Z, et al., 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39(6):925-938. https://doi.org/10.1016/j.molcel.2010.08.011

van de Vondervoort IIGM, Gordebeke PM, Khoshab N, et al., 2013. Long non-coding RNAs in neurodevelopmental disorders. *Front Mol Neurosci*, 6:53. https://doi.org/10.3389/fnmol.2013.00053

Verpelli C, Montani C, Vicidomini C, et al., 2013. Mutations of the synapse genes and intellectual disability syndromes. *Eur J Pharmacol*, 719(1-3):112-116. https://doi.org/10.1016/j.ejphar.2013.07.023

Wang KC, Chang HY, 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell*, 43(6):904-914. https://doi.org/10.1016/j.molcel.2011.08.018

Wang X, He C, 2014. Dynamic RNA modifications in post-transcriptional regulation. *Mol Cell*, 56(1):5-12. https://doi.org/10.1016/j.molcel.2014.09.001

Wang X, Lu ZK, Gomez A, et al., 2014. N^6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117-120. https://doi.org/10.1038/nature12730

Wang X, Zhao BS, Roundtree IA, et al., 2015. N^6-methyladenosine modulates messenger RNA translation efficiency. *Cell*, 161(6):1388-1399. https://doi.org/10.1016/j.cell.2015.05.014

Wang Y, Zhao X, Ju W, et al., 2015. Genome-wide differential expression of synaptic long noncoding RNAs in autism spectrum disorder. *Transl Psychiatry*, 5(10):e660. https://doi.org/10.1038/tp.2015.144

Werner MS, Ruthenburg AJ, 2015. Nuclear fractionation reveals thousands of chromatin-tethered noncoding RNAs adjacent to active genes. *Cell Rep*, 12(7):1089-1098. https://doi.org/10.1016/j.celrep.2015.07.033

Wu P, Zuo XL, Deng HL, et al., 2013. Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res Bull*, 97:69-80. https://doi.org/10.1016/j.brainresbull.2013.06.001

Xu X, Xu YC, Shi CQ, et al., 2017. A genome-wide comprehensively analyses of long noncoding RNA profiling and metastasis associated lncRNAs in renal cell carcinoma. *Oncotarget*, 8(50):87773-87781. https://doi.org/10.18632/oncotarget.21206

Yang LT, Tang YY, Xiong F, et al., 2018. LncRNAs regulate cancer metastasis via binding to functional proteins. *Oncotarget*, 9(1):1426-1443. https://doi.org/10.18632/oncotarget.22840

Yoon JH, Abdelmohsen K, Kim J, et al., 2013. Scaffold function of long non-coding RNA *HOTAIR* in protein ubiquitination. *Nat Commun*, 4:2939. https://doi.org/10.1038/ncomms3939

Zampetaki A, Albrecht A, Steinhofel K, 2018. Long-noncoding RNA structure and function: is there a link? *Front Physiol*, 9:1201. https://doi.org/10.3389/fphys.2018.01201

Zhang YQ, Hamada M, 2018. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinformatics*, 19(Suppl 19):524. https://doi.org/10.1186/s12859-018-2516-4

Zhang ZH, Jhaveri DJ, Marshall VM, et al., 2014. A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS ONE*, 9(8):e103207. https://doi.org/10.1371/journal.pone.0103207

Zheng GXY, Do BT, Webster DE, et al., 2014. Dicer-microRNA-Myc circuit promotes transcription of hundreds of long noncoding RNAs. *Nat Struct Mol Biol*, 21(7):585-590. https://doi.org/10.1038/nsmb.2842

Zhou Y, Zeng P, Li YH, et al., 2016. SRAMP: prediction of mammalian N^6-methyladenosine (m^6A) sites based on sequence-derived features. *Nucleic Acids Res*, 44(10):e91. https://doi.org/10.1093/nar/gkw104

Ziats MN, Rennert OM, 2013. Aberrant expression of long noncoding RNAs in autistic brain. *J Mol Neurosci*, 49(3):589-593. https://doi.org/10.1007/s12031-012-9880-8

Zou Q, Xing PW, Wei LY, et al., 2019. Gene2vec: gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. *RNA*, 25(2):205-218. https://doi.org/10.1261/rna.069112.118

中文概要

题　目：利用基因组数据挖掘对人类长非编码 RNA 进行功能注释

概　要：越来越多证据表明 RNA 在生物系统中扮演着重要的角色，而这些发现支持了生命起源于 RNA 的假设。在人类基因组中，大部分的基因并不编码蛋白质，被称为非编码 RNA 基因。长非编码 RNA（lncRNA）是其中最大的一类，其转录本长度大于 200 个核苷酸。虽然一些 lncRNA 已被证明是调控基因表达和 3D 基因组结构的重要元件，但是大部分 lncRNA 还未被研究和注释。本课题组利用大量基因组数据，提出一些基于数据挖掘和机器学习的方法，对人类 lncRNA 进行功能注释。我们与其他同领域课题组的近期研究结果表明，基因组数据挖掘可帮助加深对 lncRNA 功能的理解，并为与疾病相关 lncRNA 的实验研究提供重要信息。

关键词：长非编码RNA（lncRNA）；功能注释；基因组数据挖掘；机器学习