Published in final edited form as:

Cancer Lett. 2018 March 28; 417: 161–167. doi:10.1016/j.canlet.2017.12.029.

Statistical Analysis of Non-coding RNA Data

Qianchuan He^{*,†}, Yang Liu[†], and Wei Sun^{*}

Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Abstract

With rapid progress in high-throughput genome technology, the study of noncoding RNA has arisen as a highly popular topic in biomedical research. Noncoding RNA plays fundamental roles in cell proliferation, cell differentiation and epigenetic regulation, and the study of noncoding RNA will yield novel insights into gene regulation and provide new clues for disease treatment. However, due to the large volume and diverse functions of noncoding RNAs, the analysis of these RNAs has proved to be a challenging task. In this review, we review the commonly used computational tools for the identification of noncoding RNAs, and discuss popular statistical tools for their analysis. Due to the large body of noncoding RNA classes, we will focus on the analysis of microRNA and long noncoding RNA, two of the most widely studied classes of noncoding RNAs. Specific examples are provided to show the context of the analysis. This review aims to provide up-to-date information on existing tools and methods for identifying and analyzing noncoding RNA.

Keywords

long noncoding RNA; microRNA; noncoding RNA; statistical analysis; statistical modeling; target prediction

1. Background

Non-coding RNA (ncRNA) refers to RNA that does not encode a protein. Traditionally, coding RNA (i.e., mRNA) has been the major research focus due to its role in the DNA→RNA→Protein dogma, but recent studies suggested that ncRNA comprises a hidden layer of internal signals that are critical for regulating gene expression (Mattick and Makunin, 2006). ncRNAs encompass a huge variety of RNA classes and accomplish a wide range of biological functions, such as regulating gene expression, protecting genome from exogenous DNA, and guiding DNA synthesis (Cech and Steitz, 2014). One class of ncRNA that has been widely studied is micro-RNA (miRNA), which is a short RNA fragment

^{*}Co-corresponding author. qhe@fredhutch.org, wsun@fredhutch.org.

[†]Equally contributed

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest

The authors declare no conflict of interest.

consisting of ~22 nucleotides. The primary function of miRNA is to bind to the 3' UTR of its target-gene's mRNA, and subsequently inhibit the target-gene's expression. Thus, miRNA essentially functions as a silencer for gene expression. The biological roles of miRNA include cell growth, apoptosis, cell differentiation (Pillai, 2005; Zhao, 2007), and carcinogenesis (Shenouda, 2009; Peng et al., 2013), among others. For example, Calin et al. (2002) found that frequent deletion of miR15 and miR16 at 13q14 is strongly associated with chronic lymphocytic leukemia, and later studies revealed that these two miRNA genes actually act as tumor suppressor genes (Calin et al., 2008). More recently, another class of ncRNA, long non-coding RNA (lncRNA) has also drawn considerable research interest. LncRNA, by its name, refers to ncRNA of longer length, usually >200 nucleotides (Guttman et al., 2009). Unlike miRNA which is a relatively homogeneous class of RNAs, lncRNA constitutes a highly heterogenic class of RNAs, such as intergenic lncRNAs, antisense transcripts, and enhancer RNAs (Boon et al., 2016). Consistent with its heterogeneity, lncRNA has a wide array of regulatory roles in gene expression, such as sensory, guiding, scaffolding, and allosteric capacities (Mercer and Mattick, 2013). For example, one role of lncRNA that has been particularly studied is its function in epigenetic regulation, because lncRNA can bind to chromatin-modifying proteins and modulate chromatin states (Gupta et al., 2010).

While existing studies have provided many valuable insights into the biological functions of ncRNA, it is likely that what we have seen is just the tip of the iceberg and many hidden secrets of ncRNAs remain to be elucidated (Bhan et al, 2017). Traditional studies of ncRNA have been confined to a few candidates due to either technology constraint or experimental cost. With new development in genome technology, particularly the next generation sequencing, high-throughput non-coding RNA data have become widely available in the past several years (Ling et al., 2017; Du et al., 2017; Xu et al., 2017). For example, we searched 'high-throughput noncoding RNA' in the PubMed, and it yielded 617 articles between 2012–2017 and only 130 articles between 2007–2012. The massive amount of sequencing data for ncRNA creates tremendous challenges for data analysis, such as identification of ncRNA, identification of the target of ncRNA, and statistical modeling (Sacco et al., 2011). A number of tools and methods have been developed to analyze these data to interrogate the expression, variation, and function of ncRNA at the genomewide scale. Reviewing the scope and utilities of these methods can help researchers to quickly find tools that are appropriate for their data analysis, and identify new areas that require future research development. Besides miRNA and lncRNA, we note that there exist other classes of ncRNA, such as small nuclear RNA (snRNA), small nucleolar RNA (snoRNA) (Matera et al., 2007), and small interfering RNA (siRNA) (McManus and Sharp, 2002). Due to space limitation, our review will be focused on the analysis of miRNA and lncRNA, but the methods reviewed can be potentially used for analyzing other ncRNA as well. We surveyed relevant publications from PubMed and Google Scholar based on their citations and topic areas, and included some representative examples for this review. Our review is by no means to cover all the tools/methods for the statistical analysis of ncRNA, but rather to demonstrate the principles and potential strategies for analyzing ncRNA.

2. Data Processing and Identification of ncRNA

Assume that one starts with a total RNA library prepared from a biological sample. With next generation sequencing, numerous RNA-seq reads will be generated in the form of FASTQ files. If the goal is to study the expression of known ncRNA, then one can map these RNA-seq reads to reference genome with the guidance of gene annotation that include ncRNA as part of the annotation. For example, one popular source of gene annotation is GENCODE (<https://www.encodegenes.org/>) (Derrien et al., 2012) and it includes annotations of both miRNA and lncRNA. One can also obtain ncRNA annotation from specialized ncRNA database such as lncRNADB (<http://www.lncrnadb.org/>) (Amaral et al., 2010). Table 1 lists multiple databases for miRNA and lncRNA. On the other hand, if the goal is to discover new ncRNA, then one needs to map these reads to the genome using tools such as the bowtie (<http://bowtie-bio.sourceforge.net>) (Langmead et al., 2012) or Tophat (Trapnell et al., 2009), and assemble the reads by Cufflinks (Trapnell et al., 2012). Then, transcripts overlapped with known transcripts in the sense direction can be classified as annotated, and these known transcripts include not only mRNA, but also rRNA, tRNA etc. In addition, there are specialized tools to remove certain types of RNAs; for example, the riboPicker (Schmieder et al., 2012) can be used to filter out rRNA sequences. After removing annotated RNA or particular types of RNA, the remaining RNAs are potential candidates for ncRNA. It may also be necessary to further filter out incompletely processed RNAs or genomic DNA contamination based on transcript abundance and recurrence across independent biological samples (Iyer et al., 2015). Once new ncRNAs are identified, biological experiments can be conducted to validate the findings. Fig. 1 illustrates the discovery and validation of ncRNA.

Computational tools are available to identify miRNAs. For example, Yousef et al. (2006) developed the BayesMiRNAfind program based on Naïve Bayes classifier, and used training data from a variety of species (such as human, mouse, drosophila) to build the prediction model. They showed that combining information from multiple species can increase the sensitivity to detect miRNA. Ding et al. (2010) published a method, MiRenSVM, for prediction of miRNA precursors. This method employs an ensemble SVM classifier, which seeks to combine multiple SVM models to gain robustness. Their method also includes multi-loop features in its classifiers and thus is particularly suitable for detecting miRNA genes with multi-loop secondary structures. Wu et al. (2011) invented MiRPara for prediction of most probably miRNA coding regions. The MiRPara employs ~25 parameters in its Support Vector Machine (SVM) algorithm, and achieves the accuracy up to 80% when compared to empirically verified mature miRNA. This tool is particularly useful for identifying miRNA in more divergent sequences or with less conserved structures.

As to miRNA-target identification, a number of tools have been developed for *in silico* prediction. For example, miRDB is an online database for miRNA target prediction and functional annotations (Wong and Wang, 2015). At its publication time, miRDB included 2.1 million predicted gene targets regulated by 6709 miRNAs. The prediction was based on a support vector machine framework developed by Wang and El Naqa (2007). Another tool, TargetScan, predicts miRNA-target sites by searching for the presence of conserved sites that match the seed region of each miRNA (Agarwal et al., 2015). This method employs

several statistical techniques, such as bootstrap, stepwise selection and multiple linear regression, to build the prediction model. ComiR is a tool for predicting whether a given mRNA is targeted by a set of miRNAs (Coronnello and Benos, 2013). The main feature of ComiR is that it uses miRNA expression to improve the prediction. This tool uses a SVM model to combine the prediction scores of four different prediction algorithms, and returns the probability of being a functional target of a set of miRNAs for each gene. Table 2 summarizes the computational tools for the analysis of miRNA. Different methods may have different underlying assumptions and sources of information, and it is often not clear which tool should be preferred in practice, as there is no systematic comparison of these methods regarding their empirical performance. One possible (though conservative) strategy is to apply several tools to identify multiple sets of candidates, and then use the intersection as the final targets. Alternatively, one can assign scores for the miRNA-target prediction from each tool and combine these scores to obtain a final score (Tokar et al., 2017).

The *in silico* prediction of lncRNA is far more challenging than miRNA because the former is of longer length, more variety, and less understood functionality. Despite these challenges, various software tools have been developed to assist in the identification of lncRNA. For example, PhyloCSF (Lin et al., 2011) is a tool that utilizes phylogenetic information in multiple species to distinguish lncRNA from protein-coding RNAs. This tool is based on the theoretical framework of statistical phylogenetic model comparison, and hence requires that at least one of the compared genomes have coding gene annotations of good quality. Other tools based on SVM have also been developed, such as the Coding-Non-Coding Index (CNCI) (Sun et al., 2013), iSeeRNA (Sun et al., 2013), and lncRScan-SVM (Sun et al., 2015). The CNCI employs the radial basis kernel, which can accommodate non-linear effects of predictors. A main advantage of CNCI is that it does not depend on information of known genome annotation or sequence conservation. The iSeeRNA includes a feature-selection step into its SVM algorithm and can reduce the number of noise predictors. The RNAcon (Panwar et al., 2014) explored a variety of algorithms, such as NaiveByes, libSVM and MultilayerPerceptron, and chose the Random Forest as their underlying algorithm because it performs better in their experiments. Table 3 lists some of the characteristics of these computational tools. The practical performance of these methods will depend on the underlying data-generating mechanisms. Since the data-generating mechanism is rarely known in practice, a commonly used practice is to apply different methods and compare their results to identify consensus votes. Parallel to the task of predicting lncRNA, how to predict the targets for lncRNA is also a highly challenging problem. Some tools have been developed, such as the lncRNATargets (Hu and Sun, 2016) which is a web-based platform for predicting lncRNA targets based on thermodynamics. We expect that more tools of this kind will become available in the near future. Once the ncRNA candidates have been identified, one should obtain the read-counts for each candidate ncRNA, and then these counts will be used in the subsequent data analysis.

3. Statistical Modeling/Analysis of ncRNA

The first step for data analysis is usually the data normalization which corrects bias due to variation of library sizes, gene lengths, or experimental conditions. A number of normalization tools have been developed, such as the quantile normalization (Bolstad et al.,

2003; Hu and He, 2007) and the Reads Per Kilobase per Million mapped reads (RPKM) (Mortazavi et al., 2008). The DESeq (Anders and Huber, 2010) also has an embedded function for data normalization. The quantile normalization aims to match distribution of gene counts across different runs, while the RPKM corrects for differences in library sizes and gene length. The normalization function in DESeq corrects for bias introduced by experimental conditions by assuming that most RNAs are not differentially expressed and should have similar read counts across samples. An empirical comparison of different normalization methods can be found in Lin et al. (2016).

Once the data have been normalized, the next step is often to compare the expression of ncRNA for different groups or under different conditions (such as disease versus non-disease, treatment versus placebo, multiple developmental stages). From a statistical point of view, this is to test the difference of the means for different groups. For more than two groups, analysis of variance (ANOVA) can be used to evaluate the global hypothesis: for a given ncRNA, is there any group that has its mean different from other groups? When the research question is to compare two groups, one may choose popular tools, such as the edgeR (Robinson et al., 2010) and the DESeq, to accomplish the task. Both methods adopt the negative binomial distribution (NEB) for read-counts. The NEB can effectively tackle the over-dispersion issue (i.e., higher variance than expected for a Poisson distribution), which is commonly seen in RNA seq data. Both methods have been widely used in real data analysis, and an empirical study suggests that each method has its own advantages under different situations (Zhang et al., 2014). If one does not wish to assume the NEG distribution for read-counts, Fisher's exact test can be used. However, if there is indeed over-dispersion, the results of Fisher's test can lead to inflation of type I error. Sometimes, if the data can be transformed (such as log transformation) to normal or nearly normal, then the 2-sample T test can be applied as well.

Next, we discuss some statistical analysis that is beyond the 2-sample comparison of ncRNAs. Note that the purpose of this review is to show various statistical techniques that can be used for analyzing ncRNA, including but not limited to the methods that are *only* designed for analyzing ncRNA. The analysis methods described in the follows are mostly based on regression modeling, and allow one to associate ncRNA with biological or health features. These methods in general can be applied to a variety of ncRNAs, though in our examples they were primarily used for a specific type of ncRNA. We will discuss the unique challenges for analyzing certain types of ncRNA data when appropriate. Case studies are provided to show the context of the analysis. Table 4 highlights some of the methods for statistical modeling.

eQTL (expression quantitative trait loci) analysis

eQTL has been widely used in the analysis of mRNA, which aims to identify single nucleotide polymorphisms (SNPs) that regulate gene expression. Similar concept can be applied to the expression of miRNA and lncRNA as well. For example, Huan et al. (2015) utilized the linear mixed model to study miRNA-eQTL, and identified thousands of SNPs that are associated with miRNAs expressions. They found that ~50% of these SNPs are located 300–500kb upstream of their associated intergenic miRNAs, suggesting that miRNA

expression can be affected by distal regulatory elements. They further showed that many of these miRNAs display differential expression with regard to the studied trait (HDL cholesterol), suggesting that some miRNAs may be a mediator (of SNPs) to influence complex traits. That is, SNP $\rightarrow$ miRNA $\rightarrow$ traits. These findings may help to identify new opportunities for drug treatment or diagnosis of metabolic diseases. The eQTL analysis we have discussed so far associates SNPs with total expression of miRNA, without distinguishing the two alleles. Using sequencing data, one can collect allele-specific expression and combine total read count and allele-specific read count for eQTL mapping (Sun 2012). This will allow more refined analysis which tends to yield more accurate results for eQTL analysis. For lncRNA, the total read counts of a given lncRNA tend to be low in many samples. In such a situation, it is important to use appropriate count distribution (e.g., negative binomial distribution) for eQTL mapping rather than using a normal distribution approximation for log-transformed count data.

Network analysis to improve identification of miRNA targets

Based on current GENCODE annotation, human genome has 3,837 miRNAs and 19,881 protein coding genes. Each of these miRNAs may regulate the expression of one or more mRNAs. The potential targets of miRNA can be identified by aligning the miRNA sequence with mRNA sequences. However, each miRNA typically aligns with its targets in a short region of 6–8 base pairs (Pasquinelli, 2012), and alignment of such short sequences can lead to many false discoveries of miRNA targets. To reduce false discovery rate, the expression of mRNAs can be used to help identify miRNA targets. Since miRNAs directly influence the expression of their target genes, graphic models can be used to study the relationship between the expression of miRNAs and their potential targets, which yields the miRNA-mRNA co-expression networks. Such graphic models integrate both miRNA and mRNA information and provide direct visualization of the miRNA-mRNA networks. A sophisticated statistical method has been developed to use sequence/structure information as prior knowledge to construct miRNA-mRNA co-expression networks (Stingo et al., 2010). The authors used the proposed approach to identify miRNAs as well as their target genes that are related to neural tube defect disease.

mRNA-lncRNA correlation

Previous studies have demonstrated correlation between intergenic lncRNA and protein-coding genes (Derrien et al., 2012). Spearman correlation (r_s) can be used to quantify the pairwise correlation between the two types of RNAs. Derrien et al. (2012) found that some lncRNAs have extremely strong positive correlation with the expression of antisense coding genes. They observed that while ~3% of lncRNAs have highly positive correlation with neighboring mRNAs within 20kb, only ~0.4% of lncRNAs have such high correlations with neighboring mRNA that is 80–100kb away. This suggests that if an lncRNA affects the expression of nearby protein-coding genes, its effect may be concentrated on overlapping mRNAs or mRNAs in very close proximity. They further found that lncRNAs are generally lower expressed than protein-coding genes, and display more tissue-specific expression patterns. There are several well-known examples of the *cis*-regulatory roles of lncRNA, such as X chromosome inactivation by lncRNA Xist. However, many lncRNAs may also have *trans*-regulatory roles, by binding on regulatory proteins to modify their activity or by acting

as 'decoys' to bind protein complexes and prevent them from binding their targets (Guttman and Rinn, 2012).

Enrichment analysis of the mRNA targets of lncRNA

Zhang et al. (2017) categorized lncRNAs into four types of functional groups: up- or downstream of mRNA genes, antisense transcripts of mRNAs, pre-miRNA, and the lncRNA family. They applied RNAplex to lncRNA from virus-infected cells, and found ~300 lncRNAs to be annotated as antisense transcripts of mRNAs and ~1600 lncRNAs to be adjacent to protein coding genes. Next, they performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of these neighbor and antisense mRNAs. They found that the target genes are enriched in virus infection pathways and immune response ways, suggesting that lncRNAs may act as regulatory elements in the immune system. Hence, analyzing the mRNA targets of lncRNA at the pathway way level can yield valuable insights into the role of lncRNA in gene regulation. Overall, their study provided information about lncRNAs in the immune system and offered new clues for developing antiviral therapies.

lncRNA-mRNA co-expression networks

Since lncRNA and mRNA are closely related to each other, it is interesting to study how a set of lncRNA and another set of mRNA are interacting with each other at the pathway level. Such an analysis can be performed by co-expression network analysis. Wang et al. (2017) used the Cytoscape software (Kohl et al., 2011) to construct the lncRNAs-mRNAs co-expression networks, and studied such networks in patients with Moyamoya disease. They analyzed 3,649 lncRNAs and 2,880 mRNA probes, and showed that the integrated analysis of lncRNA-mRNA co-expression networks were linked to inflammatory response, toll-like signaling pathway, cytokine-cytokine receptor interaction and MAPK signaling pathway. For example, they found that 32 lncRNAs interacted with 11 mRNAs in the inflammatory response pathway, which shed lights on the complex regulation system of this pathway. Their findings led to better understanding of the pathogenesis of Moyamoya disease, and provided new candidates for potential drug development.

Identifying lncRNA associated with survival time or cancer subtypes

For population-based studies, it is possible to study lncRNA with respect to clinical outcomes, such as overall survival time, recurrence free time, or cancer subtypes. Kaplan Meier curve can be used to compare the survival rate of two groups, say patients with high expression (or low expression) of a given lncRNA. If demographic variables (such as ethnicity, age, gender) need to be considered in the analysis, then one can choose Cox proportional hazard regression model to evaluate the association between survival time (recurrence free survival time) and lncRNAs. For example, Du et al. (2013) used the Cox regression model to study the survival time of glioblastoma with respect to lncRNA, and they identified approximately 100 lncRNAs whose expression were markedly correlated with overall survival time. They further used the Mann-Whitney U test to compare the lncRNA levels between different cancer subtypes, and identified a number of lncRNAs showing significant subtype-specific expression patterns. Their analysis provided interesting

candidates for studying the molecular basis of glioblastoma, and has the potential to improve the prediction of clinical outcome.

Build lncRNA signature panel for clinical-outcome prognosis

lncRNA can be used to build prediction model for disease prognosis. The major challenge of building such a prediction model lies in the high dimensionality of the data, i.e., extremely large number of lncRNAs. To tackle such a problem, regularized regression techniques can be potentially employed. Regularized regression differs from traditional regression model in that the former includes a penalty term in the model which seeks to balance the prediction error and the number of predictors in the model. A number of different penalties have been developed, such as the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001), the MCP (Zhang, 2010), etc. If the outcome of interest is survival time, then regularized Cox regression can be conducted; an R package ('glmnet') is available for such a regression model. If the outcome of interest is binary (such as long survival versus short survival), then regularized logistic regression can be applied. For example, Tian et al. (2017) applied the LASSO to lncRNAs data for gastric cancer and identified a 12-lncRNA signature panel. There are many other machine learning tools that can be potentially used for building prediction models, such as random forest, neural network, deep learning, etc.

4. Discussion

Research in noncoding RNA provides exciting opportunities for new discoveries. With rapid progress in genome technologies, especially the next generation sequencing, we are likely to see an explosion of ncRNA data in the coming years. The analysis of existing ncRNA data has already yielded many novel insights into the functions of ncRNA – the genetic material once believed to be the dark matter of genome. Emerging data from high-throughput technologies will only make the field more exciting.

Since ncRNA data are sequencing data, many existing approaches for analyzing mRNA seq can be also applied to ncRNA. On the other hand, the analysis of ncRNA has its unique features, such as identification of ncRNA, identifying the target(s) of ncRNA, and studying the interaction between ncRNA and its target(s). These features entail considerable statistical uncertainties, which are not seen in the analysis of mRNA seq data. For this reason, analyzing ncRNA is inherently more challenging than analyzing mRNA data. While current methods have demonstrated good performance in target identification for ncRNA, the specificity and sensitivity of these methods will likely continue to improve, due to better understanding of the biology and enhanced statistical algorithms for prediction (Zhang et al., 2017).

Currently, many ncRNA studies are focused on a single time point or a single tissue. However, ncRNA is likely to vary across different tissues and fluctuate with time. Exploring the spatial/temporal expression patterns of ncRNA can shed light on the fundamental roles of ncRNA in development and differentiation (Forouzmand et al., 2017; Tang et al., 2017). How to account for the correlations among spatial/temporal points remains to be investigated. Another interesting topic is to integrate ncRNA with other types of genomic data, such as DNA methylation data and mRNA, to predict clinical outcomes. A recent study

showed that adding ncRNA to statistical models can improve prediction accuracy (Xu et al. 2015). Future research on utilizing ncRNA for clinical use is merited.

Acknowledgments

This research is partly supported by Fred Hutchinson Cancer Research Center Institutional Fund, National Institutes of Health [NCI-P30CA015704, and R01 GM105785].

References

1. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *elife*. 2015; 4:e05005.
2. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNADB: a reference database for long noncoding RNAs. *Nucleic acids research*. 2010; 39(suppl_1):D146–D151. [PubMed: 21112873]
3. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology*. 2010; 11(10):R106. [PubMed: 20979621]
4. Bhan A, Soleimani M, Mandal SS. Long noncoding RNA and cancer: a new paradigm. *Cancer research*. 2017; 77(15):3965–3981. [PubMed: 28701486]
5. Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B, Raghava GPS. lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database*. 2013; 2013
6. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19(2):185–193. [PubMed: 12538238]
7. Boon RA, Jaé N, Holdt L, Dimmeler S. Long Noncoding RNAs. *Journal of the American College of Cardiology*. 2016; 67(10):1214–1226. [PubMed: 26965544]
8. Calin GA, Cimmino A, Fabbri M, Ferracin M, Wojcik SE, Shimizu M, Alder H. MiR-15a and miR-16-1 cluster functions in human leukemia. *Proceedings of the National Academy of Sciences*. 2008; 105(13):5166–5171.
9. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Rassenti L. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*. 2002; 99(24):15524–15529.
10. Cech TR, Steitz JA. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*. 2014; 157(1):77–94. [PubMed: 24679528]
11. Coronello C, Benos PV. ComiR: combinatorial microRNA target prediction tool. *Nucleic acids research*. 2013; 41(W1):W159–W164. [PubMed: 23703208]
12. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guigo R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*. 2012; 22(9):1775–1789. [PubMed: 22955988]
13. Ding J, Zhou S, Guan J. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC bioinformatics*. 2010; 11(11):S11.
14. Du Y, Xia W, Zhang J, Wan D, Yang Z, Li X. Comprehensive analysis of long noncoding RNA–mRNA co-expression patterns in thyroid cancer. *Molecular BioSystems*. 2017; 13(10):2107–2115. [PubMed: 28817151]
15. Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, Liu XS. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature structural & molecular biology*. 2013; 20(7):908–913.
16. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*. 2001; 96(456):1348–1360.
17. Forouzmand E, Owens ND, Blitz IL, Paraiso KD, Khokha MK, Gilchrist MJ, Cho KW. Developmentally regulated long non-coding RNAs in *Xenopus tropicalis*. *Developmental biology*. 2017; 426(2):401–408. [PubMed: 27418388]

18. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research*. 2005; 33(suppl_1):D121–D124. [PubMed: 15608160]
19. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Wang Y. Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*. 2010; 464(7291):1071. [PubMed: 20393566]
20. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Cabili MN. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009; 458(7235):223–227. [PubMed: 19182780]
21. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature*. 2012; 482(7385):339–346. [PubMed: 22337053]
22. Hu J, He X. Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles. *Biometrics*. 2007; 63(1):50–59. [PubMed: 17447929]
23. Hu R, Sun X. lncRNATargets: A platform for lncRNA target prediction based on nucleic acid thermodynamics. *Journal of bioinformatics and computational biology*. 2016; 14(04):1650016. [PubMed: 27306075]
24. Huan T, Rong J, Liu C, Zhang X, Tanriverdi K, Joehanes R, Munson PJ. Genome-wide identification of microRNA expression quantitative trait loci. *Nature communications*. 2015; 6:6601.
25. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Poliakov A. The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*. 2015; 47(3):199–208. [PubMed: 25599403]
26. Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. *Data mining in proteomics: from standards to applications*. 2011; 291
27. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012; 9(4):357–359. [PubMed: 22388286]
28. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011; 27(13):i275–i282. [PubMed: 21685081]
29. Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YLS, Sultana H, Harbison ST. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC genomics*. 2016; 17(1):28. [PubMed: 26732976]
30. Ling Y, Xu L, Zhu L, Sui M, Zheng Q, Li W, Zhang X. Identification and analysis of differentially expressed long non-coding RNAs between multiparous and uniparous goat (*Capra hircus*) ovaries. *PloS one*. 2017; 12(9):e0183163. [PubMed: 28934224]
31. McManus MT, Sharp PA. Gene silencing in mammals by small interfering RNAs. *Nature reviews genetics*. 2002; 3(10):737–747.
32. Matera AG, Terns RM, Terns MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature reviews Molecular cell biology*. 2007; 8(3):209–220. [PubMed: 17318225]
33. Mattick JS, Makunin IV. Non-coding RNA. *Human molecular genetics*. 2006; 15(suppl_1):R17–R29. [PubMed: 16651366]
34. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology*. 2013; 20(3):300–307.
35. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*. 2008; 5(7):621–628. [PubMed: 18516045]
36. Park C, Yu N, Choi I, Kim W, Lee S. lncRNator: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics*. 2014; 30(17):2480–2485. [PubMed: 24813212]
37. Panwar B, Arora A, Raghava GP. Prediction and classification of ncRNAs using structural information. *BMC genomics*. 2014; 15(1):127. [PubMed: 24521294]
38. Pasquinelli AE. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*. 2012; 13(4):271–282.

39. Peng Y, Dai Y, Hitchcock C, Yang X, Kassis ES, Liu L, Kim T. Insulin growth factor signaling is regulated by microRNA-486, an underexpressed microRNA in lung cancer. *Proceedings of the National Academy of Sciences*. 2013; 110(37):15043–15048.
40. Pillai RS. MicroRNA function: multiple mechanisms for a tiny RNA? *Rna*. 2005; 11(12):1753–1761. [PubMed: 16314451]
41. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26(1):139–140. [PubMed: 19910308]
42. Sacco LD, Baldassarre A, Masotti A. Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. *International journal of molecular sciences*. 2011; 13(1):97–114. [PubMed: 22312241]
43. Schmieder R, Lim YW, Edwards R. Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics*. 2011; 28(3):433–435. [PubMed: 22155869]
44. Shenouda SK, Alahari SK. MicroRNA function in cancer: oncogene or a tumor suppressor? *Cancer and Metastasis Reviews*. 2009; 28(3–4):369. [PubMed: 20012925]
45. Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. A Bayesian graphical modeling approach to microRNA regulatory network inference. *The annals of applied statistics*. 2010; 4(4):2024. [PubMed: 23946863]
46. Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics*. 2013; 14(2):S7.
47. Sun L, Liu H, Zhang L, Meng J. lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine. *PloS one*. 2015; 10(10):e0139654. [PubMed: 26437338]
48. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Zhao Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research*. 2013; 41(17):e166–e166. [PubMed: 23892401]
49. Sun W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics*. 2012; 68(1):1–11. [PubMed: 21838806]
50. Tang Z, Wu Y, Yang Y, Yang YCT, Wang Z, Yuan J, Zhang Y. Comprehensive analysis of long non-coding RNAs highlights their spatio-temporal expression patterns and evolutionary conservation in *Sus scrofa*. *Scientific Reports*. 2017; 7
51. Tian X, Zhu X, Yan T, Yu C, Shen C, Hong J, Fang JY. Differentially Expressed lncRNAs in Gastric Cancer Patients: A Potential Biomarker for Gastric Cancer Prognosis. *Journal of Cancer*. 2017; 8(13):2575. [PubMed: 28900495]
52. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996:267–288.
53. Tokar T, Pastrello C, Rossos AE, Abovsky M, Hauschild AC, Tsay M, Jurisica I. mirDIP 4.1—integrative database of human microRNA target predictions. *Nucleic Acids Research*. 2017
54. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25(9):1105–1111. [PubMed: 19289445]
55. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*. 2012; 7(3):562–578. [PubMed: 22383036]
56. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, Mestdagh P. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic acids research*. 2012; 41(D1):D246–D251. [PubMed: 23042674]
57. Wang W, Gao F, Zhao Z, Wang H, Zhang L, Zhang D, Zhao J. Integrated Analysis of lncRNA-mRNA Co-Expression Profiles in Patients with Moyamoya Disease. *Scientific Reports*. 2017; 7
58. Wang X, El Naqa IM. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*. 2007; 24(3):325–332. [PubMed: 18048393]
59. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic acids research*. 2014; 43(D1):D146–D152. [PubMed: 25378301]
60. Wu Y, Wei B, Liu H, Li T, Rayner S. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics*. 2011; 12(1):107. [PubMed: 21504621]

61. Xu L, Fengji L, Changning L, Liangcai Z, Yinghui L, Yu L, Jianghui X. Comparison of the prognostic utility of the diverse molecular data among lncRNA, DNA Methylation, microRNA, and mRNA across five human cancers. *PloS one*. 2015; 10(11):e0142433. [PubMed: 26606135]
62. Xu S, Kong D, Chen Q, Ping Y, Pang D. Oncogenic long noncoding RNA landscape in breast cancer. *Molecular cancer*. 2017; 16(1):129. [PubMed: 28738804]
63. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*. 2006; 22(11):1325–1334. [PubMed: 16543277]
64. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*. 2010; 38(2):894–942.
65. Zhang J, Sun P, Gan L, Bai W, Wang Z, Li D, Ma X. Genome-wide analysis of long noncoding RNA profiling in PRRSV-infected PAM cells by RNA sequencing. *Scientific Reports*. 2017; 7
66. Zhang Y, Huang H, Zhang D, Qiu J, Yang J, Wang K, Yang J. A Review on Recent Computational Methods for Predicting Noncoding RNAs. *BioMed research international*. 2017; 2017
67. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Zhao QY. A comparative study of techniques for differential expression analysis on RNA-Seq data. *PloS one*. 2014; 9(8):e103207. [PubMed: 25119138]
68. Zhao Y, Srivastava D. A developmental view of microRNA function. *Trends in biochemical sciences*. 2007; 32(4):189–197. [PubMed: 17350266]
69. Zhao Y, Li H, Fang S, Kang Y, Hao Y, Li Z, Chen R. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic acids research*. 2016; 44(D1):D203–D208. [PubMed: 26586799]

Figure 1.
Flowchart of ncRNA discovery and validation

Table 1**Databases for miRNA and lncRNA**

	Contents	Remarks	Web access	Reference
GENCODE	Extensive gene annotation including 1,881 miRNAs and 15,778 lncRNAs in human genome	High quality reference gene annotation and experimental validation for human and mouse genomes.	https://www.encodegenes.org/	Derrien et al., 2012
LNCipedia	An integrated database of 146,742 human annotated lncRNA transcripts	In addition to basic transcript information and structure, secondary structure information, protein coding potential and microRNA binding sites are also provided.	https://lncipedia.org/	Volders et al., 2012
lncRNome	Hosts information on over 17,000 long noncoding RNAs in Human	Provide chromosomal locations, description on the biological functions and disease associations of long noncoding RNAs.	http://genome.igib.res.in/lncRNome	Bhartiya et al., 2013
lncRNadb	A reference database for functional lncRNAs	Contains gene expression data derived from the Illumina Body Atlas, structural information, subcellular localization, conservation, etc.	http://www.lncrnadb.org/	Amaral et al., 2010
lncRNator	Contains lncRNAs from multiple organisms	Encompass expression profile, binding protein, evolution scores, coding potential.	http://lncrnator.ewha.ac.kr	Park et al., 2014
Rfam	ncRNA	Based on secondary structure, and a covariance model.	http://rfam.xfam.org/	Griffiths-Jones et al., 2005
NONECODE	LncRNAs for 16 species, including 167,150 human lncRNAs	Aims to present the most complete collection and annotation of non-coding RNAs, especially lncRNAs.	http://www.noncode.org/	Zhao et al., 2016

Table 2

Tools for identifying miRNA and its targets

	Underlying Algorithm	Function	Reference
BayesMiRNAfind	Based on Naïve Bayes classifier, and used training data from multiple species to build prediction model	For identifying miRNA	Yousef et al., 2006
MiRenSVM	Based on an ensemble SVM classifier	For identifying miRNA	Ding et al., 2010
MiRPara	Based on SVM	For identifying miRNA	Wu et al., 2011
miRDB	Based on an SVM framework	For identifying the target of miRNA	Wong and Wang, 2015
TargetScan	Bootstrap, stepwise selection and multiple linear regression	For identifying the target of miRNA	Agarwal et al., 2015
ComiR	Integrate the score of TargetScan with other scores by SVM	For identifying the target of miRNA	Coronnello and Benos, 2013

Table 3

Analysis tools for predicting lnc-RNA

	Underlying rationale/algorithm	Remarks	Software access	Reference
PhyloCSF	Based on conservation for evaluating the coding potential of a transcript	Require cross-species nucleotide sequence alignment.	https://github.com/mlin/PhyloCSF/wiki	Lin et al., 2011
CNCI	SVM with radial basis kernel	For moderate-size datasets.	http://www.bioinfo.org/software/cnci	Sun et al., 2013
iSeeRNA	Feature selection + SVM with radial basis kernel	Moderate to large size datasets, also provides users with a program for training a new classification model based on custom dataset.	http://www.myogenesisdb.org/iSeeRNA	Sun et al., 2013
RNAcon	SVM and Random forest	Require single/multiple FASTA RNA sequences.	http://crdd.osdd.net/raghava/rnacon	Panwar et al., 2014
lncRScan-SVM	SVM framework	Integrate information extracted from gene structure, sequence composition and conservation.	http://sourceforge.net/projects/lncscansvm/?source=directory	Sun et al., 2015

Table 4

Methods and tools for statistical modeling

Statistical methods/tools	
Differential analysis	DEseq, edgeR
Regression analysis	Cox regression, linear regression, linear mixed model
Signature analysis	LASSO, Neural network, Deep learning
Enrichment analysis	Gene set enrichment analysis, KEGG
Network analysis	Bayesian graphic model, Gaussian graphic model