Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 317642, 5 pages
http://dx.doi.org/10.1155/2014/317642

Research Article

TF2LncRNA: Identifying Common Transcription Factors for a List of lncRNA Genes from ChIP-Seq Data

Qinghua Jiang,[1] Jixuan Wang,[2] Yadong Wang,[3] Rui Ma,[3] Xiaoliang Wu,[3] and Yu Li[1]

[1] *School of Life Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[2] *School of Software, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*

Correspondence should be addressed to Yadong Wang; ydwang@hit.edu.cn

Received 23 October 2013; Revised 14 January 2014; Accepted 27 January 2014; Published 4 March 2014

Academic Editor: Hesham H. Ali

High-throughput genomic technologies like lncRNA microarray and RNA-Seq often generate a set of lncRNAs of interest, yet little is known about the transcriptional regulation of the set of lncRNA genes. Here, based on ChIP-Seq peak lists of transcription factors (TFs) from ENCODE and annotated human lncRNAs from GENCODE, we developed a web-based interface titled "TF2lncRNA," where TF peaks from each ChIP-Seq experiment are crossed with the genomic coordinates of a set of input lncRNAs, to identify which TFs present a statistically significant number of binding sites (peaks) within the regulatory region of the input lncRNA genes. The input can be a set of coexpressed lncRNA genes or any other cluster of lncRNA genes. Users can thus infer which TFs are likely to be common transcription regulators of the set of lncRNAs. In addition, users can retrieve all lncRNAs potentially regulated by a specific TF in a specific cell line of interest or retrieve all TFs that have one or more binding sites in the regulatory region of a given lncRNA in the specific cell line. TF2LncRNA is an efficient and easy-to-use web-based tool.

1. Introduction

The Encyclopedia of DNA Elements (ENCODE) project has expanded our knowledge of what lies in the dark recesses of the human genome. One of these important findings is that only a small fraction of the human genome encodes proteins; almost 60% is represented in processed transcripts that seem to lack protein-coding capacity [1]. Long noncoding RNAs (lncRNAs) are non-protein-coding transcripts longer than 200 nucleotides in length. Compared to microRNAs or snoRNAs that exhibit strong conservation across diverse species, lncRNAs lack strong conservation, which is often cited as evidence of nonfunctionality [2, 3]. However, many well-described lncRNAs, such as Air and Xist, are poorly conserved [4], and increasing studies have demonstrated that a number of lncRNAs are not transcriptional noise but have important functions, such as regulating gene expression at various molecular levels, including protein, RNA, miRNA, and DNA [5–7].

Thousands of human lncRNAs have been identified [8], and accumulating studies have revealed that a number of lncRNAs play important roles in organismal development and various human diseases [9], such as cancers [10, 11], cardiovascular diseases [12], and neurodegeneration diseases [13]. However, few studies focus on how lncRNA genes themselves are transcriptionally regulated. Yang et al. developed a system by which users can browse transcription factor (TF) binding sites in the regulatory region of a lncRNA [14]. However, given a cell line or tissue of interest, users cannot obtain which transcription factors (TFs) have binding sites in the regulatory region of a specific lncRNA and which lncRNAs are regulated by a specific TF in the specific cell line or tissue.

Furthermore, high-throughput genomic technologies like lncRNA microarray (Arraystar Inc., Rockville, MD, USA) and RNA-Seq often generate a set of lncRNA genes of interest (e.g., coexpressed lncRNA genes). Given a set of lncRNA genes showing similar expression patterns, researchers often wonder how to find out which TFs are responsible for the observed expression pattern of the set of lncRNAs. For these kinds of problems, researchers used to examine whether the regulatory regions of the set of lncRNA genes contain an overrepresented sequence motif by using de novo sequence motif finding tools [15] or descriptors of the binding specificity of TFs, which may provide clues on which TFs could be common transcriptional regulators of the set of lncRNA genes. However, lncRNAs are temporally and spatially expressed and regulated, and motif-based sequence analysis cannot capture the dynamic regulation of lncRNAs by TFs in different cell lines.

Fortunately, chromatin immunoprecipitation followed by sequencing (ChIP-Seq) has enabled detecting transcription factor binding sites (TFBSs) with unprecedented sensitivity. The ENCODE project has completed ChIP-Seq experiments for many human TFs for a number type of cell lines. Enriched peak regions from the ChIP-Seq experiments of TFs can be crossed with the genomic coordinates of lncRNAs, which facilitate the discovery of TF-lncRNA regulatory relationships in a diversity of cell lines and also give us a better opportunity to identify common TFs for a given set of lncRNA genes in a cell line of interest.

Therefore, based on ChIP-Seq peak data from ENCODE and all annotated human lncRNAs from GENCODE, we developed a web-based tool titled "TF2LncRNA," accessible at http://mlg.hit.edu.cn/tf2lncrna, which enables users to identify which TFs present a statistically significant number of peaks within the regulatory regions of a set of input lncRNA genes and thus identify common TFs that are likely to regulate the set of lncRNA genes. In addition, our tool enables researchers to easily browse and retrieve TF-lncRNA regulatory relationships for a specific TF or lncRNA in a specific cell line of interest.

2. Materials and Methods

2.1. Materials

2.1.1. Genomic Annotations of lncRNAs. Genomic annotations of 13,249 human lncRNA genes and 22,531 lncRNA transcripts were downloaded from the GENCODE website (GENCODE version 15 that is identical to the Ensembl release 70) [16].

2.1.2. Genome-Wide Binding Sites of TFs. Peak lists of 425 ChIP-Seq datasets performed on 148 TFs generated from uniform processing pipeline were downloaded from UCSC ENCODE Project Portal [1], where the PeakSeq [17] peak calling method was used to identify peaks (regions of enrichment) by comparing each ChIP-Seq experiment to a corresponding input DNA control experiment. Peak calling was performed independently on each replicate of a ChIP-Seq dataset, and a measure of consistency of peak calling results between replicates, known as the irreproducible discovery rate (IDR), was used to determine an optimal number of reproducible peaks.

2.2. Methods

2.2.1. Associating Peaks of Transcription Factors with lncRNA Genes. A lncRNA gene was defined to be regulated by a TF, if the TF has at least one peak in the regulatory region of the lncRNA gene. Here, the regulatory region of a lncRNA gene is defined as a region that extends 2000 bp upstream and 1000 bp downstream from its transcription start site (denoted as −2 kb/+1 kb). We also considered other regulatory regions, such as −50 kb/+5 kb, −30 kb/+2 kb, −20 kb/+1 kb, and −10 kb/+1 kb.

2.2.2. Finding Common Transcription Factors for a Set of lncRNA Genes Using Hypergeometric Test. The annotated human lncRNAs were downloaded from GENCODE website (version 15, i.e., Ensembl v70), which includes 13,249 annotated lncRNA genes and 22,531 lncRNA transcripts. Hypergeometric test (this method is usually applied to assess gene ontology or pathway enrichment for a list of protein-coding genes) is used to identify common TFs for a set of lncRNA genes or transcripts. For each ChIP-Seq experiment of a TF, a P value (i.e., a probability of obtaining m or more input lncRNA genes targeted by the TF by chance) is calculated by the formula below:

$$P = \frac{\sum_{i=m}^{\min(n,M)} \binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}, \tag{1}$$

where

(1) N is the number of all annotated lncRNA genes (transcripts) in *Homo sapiens*;

(2) M is the number of annotated lncRNA genes (transcripts) that contain at least one ChIP-Seq peak of the TF in their regulatory regions;

(3) n is the number of lncRNA genes (transcripts) that users input;

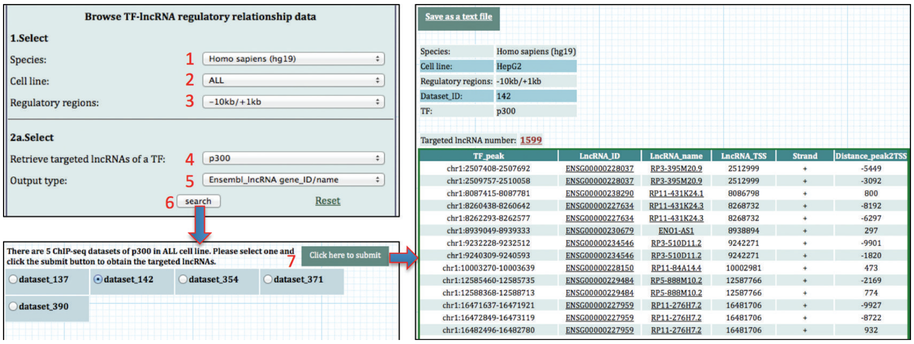
(4) m is the number of lncRNA genes (transcripts) in the inputted lncRNAs that have at least one peak of the TF within their regulatory regions.

All P values were adjusted by the Benjamini-Hochberg procedure for multiple testing corrections. A TF is defined as common TF of a set of lncRNAs if its adjusted P value is equal to or smaller than 0.05.

3. Web Interface

The web interface contains two panels on the left and right hand side, which allow users to input a set of lncRNAs for finding their common TFs or for browsing and retrieving TF-lncRNA regulatory relationships for a specific TF or lncRNA in a specific cell line of interest.

3.1. Browse and Retrieve TF-lncRNA Regulatory Relationships. The right hand panel allows users to browse and retrieve TF-lncRNA regulatory relationships in a specific cell line of interest. Users can select (i) the source organism and the TF, (ii) the cell line in which the ChIP experiment was performed, (iii) the regulatory region of lncRNA genes (e.g., 2000 bp upstream and 1000 bp downstream from its TSS), and (iv) the lncRNA ID/name to be used to display the results. Therefore,



Browse TF-lncRNA regulatory relationship data

1.Select

Species: Homo sapiens (hg19)

Cell line: ALL

Regulatory regions: -10kb/+1kb

2a.Select

Retrieve targeted lncRNAs of a TF: p300

Output type: Ensembl_lncRNA gene_ID/name

search Reset

There are 5 ChIP-seq datasets of p300 in ALL cell line. Please select one and click the submit button to obtain the targeted lncRNAs. Click here to submit

○ **dataset_137** ● **dataset_142** ○ **dataset_354** ○ **dataset_371**

○ **dataset_390**

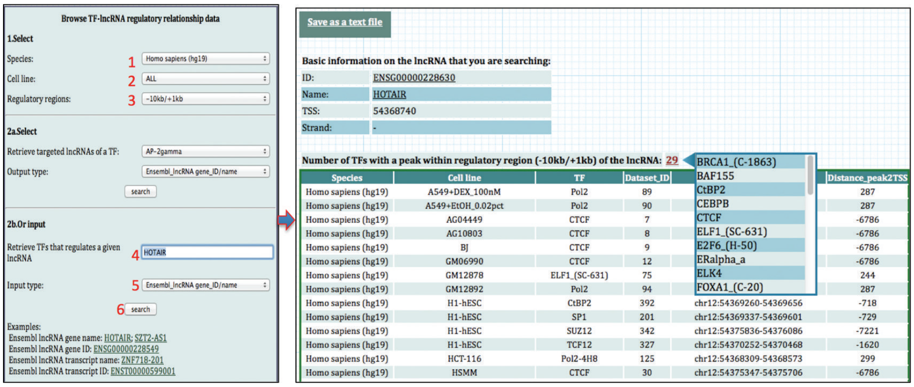
Save as a text file

Species:	Homo sapiens (hg19)
Cell line:	HepG2
Regulatory regions:	-10kb/+1kb
Dataset_ID:	142
TF:	p300

Targeted lncRNA number: 1599

TF_peak	LncRNA_ID	LncRNA_name	LncRNA_TSS	Strand	Distance_peak2TSS
chr1:2507408-2507692	ENSG00000228037	RP3-395M20.9	2512999	+	-5449
chr1:2509757-2510058	ENSG00000228037	RP3-395M20.9	2512999	+	-3092
chr1:8087415-8087781	ENSG00000238290	RP11-431K24.1	8086798	+	800
chr1:8260438-8260642	ENSG00000227634	RP11-431K24.3	8268732	+	-8192
chr1:8262293-8262577	ENSG00000227634	RP11-431K24.3	8268732	+	-6297
chr1:8939049-8939333	ENSG00000230679	ENO1-AS1	8938894	+	297
chr1:9232228-9232512	ENSG00000234546	RP3-510D11.2	9242271	+	-9901
chr1:9240309-9240593	ENSG00000234546	RP3-510D11.2	9242271	+	-1820
chr1:10003270-10003639	ENSG00000228150	RP11-84A14.4	10002981	+	473
chr1:12585460-12585735	ENSG00000229484	RP5-888M10.2	12587766	+	-2169
chr1:12588368-12588713	ENSG00000229484	RP5-888M10.2	12587766	+	774
chr1:16471637-16471921	ENSG00000227959	RP11-276H7.2	16481706	+	-9927
chr1:16472849-16473119	ENSG00000227959	RP11-276H7.2	16481706	+	-8722
chr1:16482496-16482780	ENSG00000227959	RP11-276H7.2	16481706	+	932

Figure 1: Browse and retrieve all lncRNAs potentially targeted by a specific TF in a specific cell line of interest.



Browse TF-lncRNA regulatory relationship data

1.Select

Species: Homo sapiens (hg19)

Cell line: ALL

Regulatory regions: -10kb/+1kb

2a.Select

Retrieve targeted lncRNAs of a TF: AP-2gamma

Output type: Ensembl_lncRNA gene_ID/name

search

2b.Or input

Retrieve TFs that regulates a given lncRNA: HOTAIR

Input type: Ensembl_lncRNA gene_ID/name

search

Examples:
Ensembl lncRNA gene name: HOTAIR; SZT2-AS1
Ensembl lncRNA gene ID: ENSG00000228549
Ensembl lncRNA transcript name: ZNF718-201
Ensembl lncRNA transcript ID: ENST00000599001

Save as a text file

Basic information on the lncRNA that you are searching:

ID:	ENSG00000228630
Name:	HOTAIR
TSS:	54368740
Strand:	-

Number of TFs with a peak within regulatory region (-10kb/+1kb) of the lncRNA: 29

BRCA1_(C-1863)
BAF155
CtBP2
CEBPB
CTCF
ELF1_(SC-631)
E2F6_(H-50)
ERalpha_a
ELK4
FOXA1_(C-20)

Species	Cell line	TF	Dataset_ID		Distance_peak2TSS
Homo sapiens (hg19)	A549+DEX_100nM	Pol2	89		287
Homo sapiens (hg19)	A549+EtOH_0.02pct	Pol2	90		287
Homo sapiens (hg19)	AG04449	CTCF	7		-6786
Homo sapiens (hg19)	AG10803	CTCF	8		-6786
Homo sapiens (hg19)	BJ	CTCF	9		-6786
Homo sapiens (hg19)	GM06990	CTCF	12		-6786
Homo sapiens (hg19)	GM12878	ELF1_(SC-631)	75		244
Homo sapiens (hg19)	GM12892	Pol2	94		287
Homo sapiens (hg19)	H1-hESC	CtBP2	392	chr12:54369260-54369656	-718
Homo sapiens (hg19)	H1-hESC	SP1	201	chr12:54369337-54369601	-729
Homo sapiens (hg19)	H1-hESC	SUZ12	342	chr12:54375836-54376086	-7221
Homo sapiens (hg19)	H1-hESC	TCF12	327	chr12:54370252-54370468	-1620
Homo sapiens (hg19)	HCT-116	Pol2-4H8	125	chr12:54368309-54368573	299
Homo sapiens (hg19)	HSMM	CTCF	30	chr12:54375347-54375706	-6786

Figure 2: Browse and retrieve all TFs that have at least one peak in the regulatory region of specific lncRNA in various cell lines.

given a TF of interest, users can retrieve all lncRNAs whose regulatory regions have at least one peak of the TF in the condition that users select (Figure 1). In addition, given a lncRNA of interest, users can also retrieve all TFs that have at least one peak in the regulatory region of the lncRNA in the condition that users design (Figure 2).

3.2. lncRNA Input. The left hand panel enables users to paste a set of lncRNA genes (the Ensembl lncRNA gene ID/name or lncRNA transcript ID/name) and to find TFs that have a significantly high number of peaks associated with the set of lncRNAs. Users then can select (i) the source organism of the lncRNA genes, (ii) the cell line in which the ChIP experiment was performed, (iii) the regulatory region, relative to the TSS of lncRNAs, and (iv) the input type, that is, what kind of lncRNA ID or name that users input, and, (v) optionally, users can upload a list of lncRNAs to define their own reference sets of lncRNAs. For example, if a lncRNA microarray study revealed x changing lncRNAs with a particular treatment, the reference set would not be all annotated human lncRNAs (default in TF2LncRNA system), but the user would provide a set of lncRNAs detected by the microarray to serve as the reference set or something similar.

3.3. Output. After users input a set of lncRNAs of interest and upload a reference set (optionally), they click on the "Run" button. The system will first examine whether or not the input lncRNA IDs or names are correct and show the information in the message box and then identify common TFs based on the hypergeometric test. A schematic workflow is shown in Figure 3. From left to right the columns of the output result table summarize the following.

Species. From which species the transcription factor is.

Cell Line. In which cell line the ChIP-Seq experiment was performed.

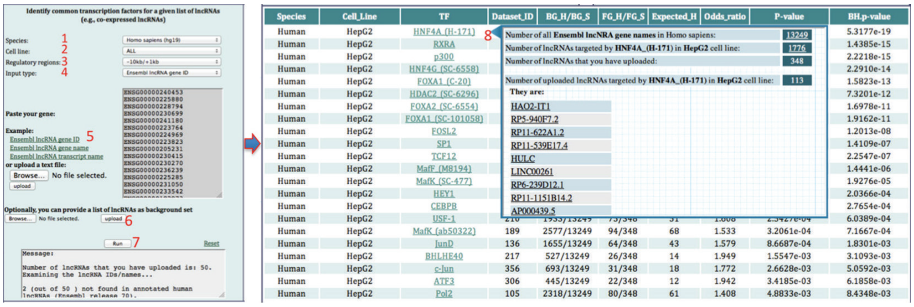
TF. Transcription factor or other DNA-binding protein in the ChIP-Seq experiment.

Dataset_ID. An ID was assigned for each ChIP-Seq experiment.

BG_H/BG_S. Number of targeted lncRNA genes (transcripts) of the TF/number of all annotated lncRNA genes (transcripts) in *Homo sapiens*. If users upload a reference set, the BG_S will be the number of lncRNAs in the reference set.

FG_H/FG_S. Number of input lncRNA genes (transcripts) targeted by the TF/number of uploaded lncRNA genes (transcripts).

Expected_H. Expected number of lncRNA genes (transcripts) targeted by the TF within lncRNAs that users input and Expected_H = FG_S ∗ (BG_H/BG_S).



Identify common transcription factors for a given list of lncRNAs (e.g., co-expressed lncRNAs)

Species: Homo sapiens (hg19)
Cell line: ALL
Regulatory regions: -10kb/+1kb
Input type: Ensembl lncRNA gene ID

Paste your gene:

```
ENSG00000240453
ENSG00000225880
ENSG00000228794
ENSG00000230699
ENSG00000241180
ENSG00000223764
ENSG00000224969
ENSG00000223823
ENSG00000205231
ENSG00000230415
ENSG00000230270
ENSG00000236239
ENSG00000225285
ENSG00000231050
ENSG00000233542
```

Example:
Ensembl lncRNA gene ID
Ensembl lncRNA gene name
Ensembl lncRNA transcript name
or upload a text file:
Browse... No file selected.
upload

Optionally, you can provide a list of lncRNAs as background set
Browse... No file selected. upload

Run Reset

```
Message:

Number of lncRNAs that you have uploaded is: 50.
Examining the lncRNA IDs/names...

2 (out of 50 ) not found in annotated human
lncRNAs (Ensembl release 70).
```

Species	Cell_Line	TF	Dataset_ID	BG_H/BG_S	FG_H/FG_S	Expected_H	Odds_ratio	P-value	BH.p-value
Human	HepG2	HNF4A_(H-171)							5.3177e-19
Human	HepG2	RXRA							1.4385e-15
Human	HepG2	p300							2.2218e-15
Human	HepG2	HNF4G_(SC-6558)							2.2910e-14
Human	HepG2	FOXA1_(C-20)							1.5823e-13
Human	HepG2	HDAC2_(SC-6296)							7.3201e-12
Human	HepG2	FOXA2_(SC-6554)							1.6978e-11
Human	HepG2	FOXA1_(SC-101058)							1.9162e-11
Human	HepG2	FOSL2							1.2013e-08
Human	HepG2	SP1							1.4109e-07
Human	HepG2	TCF12							2.2547e-07
Human	HepG2	MafF_(M8194)							1.4441e-06
Human	HepG2	MafK_(SC-477)							1.9276e-05
Human	HepG2	HEY1							2.0366e-04
Human	HepG2	CEBPB							2.7654e-04
Human	HepG2	USF-1	210	1933/13249	73/348	51	1.608	2.3427e-04	6.0389e-04
Human	HepG2	MafK_(ab50322)	189	2577/13249	94/348	68	1.533	3.2061e-04	7.1667e-04
Human	HepG2	JunD	136	1655/13249	64/348	43	1.579	8.6687e-04	1.8301e-03
Human	HepG2	BHLHE40	217	527/13249	26/348	14	1.949	1.5547e-03	3.1093e-03
Human	HepG2	c-Jun	356	693/13249	31/348	18	1.772	2.6628e-03	5.0592e-03
Human	HepG2	ATF3	306	445/13249	22/348	12	1.942	3.4185e-03	6.1858e-03
Human	HepG2	Pol2	105	2318/13249	80/348	61	1.408	4.8833e-03	8.4348e-03

Number of all **Ensembl lncNRA gene names** in Homo sapiens:	13249
Number of lncRNAs targeted by **HNF4A_(H-171)** in **HepG2** cell line:	1776
Number of lncRNAs that you have uploaded:	348
Number of uploaded lncRNAs targeted by **HNF4A_(H-171)** in **HepG2** cell line:	113

They are:
HAO2-IT1
RP5-940F7.2
RP11-622A1.2
RP11-539E17.4
HULC
LINC00261
RP6-239D12.1
RP11-1151B14.2
AP000439.5

Figure 3: A schematic workflow of finding common TFs for a set of lncRNAs of interest in various cell lines.

Odds Ratio. Ratio of the odds of lncRNAs targeted by a specific TF in your uploaded lncRNAs to the odds of lncRNAs targeted by a specific TF in all human lncRNAs and Odds_ratio = FG_H/(FG_S − FG_H)/BG_H/(BG_S − BG_H).

P Value. Each *P* value is computed by the hypergeometric test.

BH.P Value. All *P* values are corrected by the Benjamini-Hochberg method.

4. Conclusion

We developed a web-based tool titled "TF2LncRNA" that enables researchers to easily find common transcription factors for a set of lncRNAs of interest, such as coexpressed lncRNAs. In addition, users conveniently browse and retrieve TF-lncRNA regulatory relationships for a specific TF or lncRNA gene in a specific cell line of interest. As the GENCODE annotations of lncRNAs will continue evolving and more ChIP-Seq data of TFs will become available, we will continue to maintain and improve TF2lncRNA as more data become available for facilitating the research on the transcriptional regulation of a set of lncRNAs.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Natural Science Foundation of China (61102149), Special Financial Grant from the China Postdoctoral Science Foundation (2012T50357), China Postdoctoral Science Foundation (20110490108), Fundamental Research Funds for the Central Universities (HIT NSRIF. 2010057, HIT BRETIII. 201219) and China 863 program project (2012AA020404).

References

[1] ENCODE Project Consortium, B. E. Bernstein, E. Birney et al., "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[2] J. Brosius, "Waste not, want not—transcript excess in multicellular eukaryotes," *Trends in Genetics*, vol. 21, no. 5, pp. 287–288, 2005.

[3] K. Struhl, "Transcriptional noise and the fidelity of initiation by RNA polymerase II," *Nature Structural and Molecular Biology*, vol. 14, no. 2, pp. 103–105, 2007.

[4] T. B. Nesterova, S. C. Barton, M. A. Surani, and N. Brockdorff, "Loss of Xist imprinting in diploid parthenogenetic preimplantation embryos," *Developmental Biology*, vol. 235, no. 2, pp. 343–350, 2001.

[5] S. Geisler and J. Coller, "RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts," *Nature Reviews Molecular Cell Biology*, vol. 14, pp. 699–712, 2013.

[6] A. Fatica and I. Bozzoni, "Long non-coding RNAs: new players in cell differentiation and development," *Nature Reviews Genetics*, vol. 15, pp. 7–21, 2014.

[7] A. E. Kornienko, P. M. Guenzl, D. P. Barlow, and F. M. Pauler, "Gene regulation by the act of long non-coding RNA transcription," *BMC Biology*, vol. 11, article 59, 2013.

[8] T. Derrien, R. Johnson, G. Bussotti et al., "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression," *Genome Research*, vol. 22, no. 9, pp. 1775–1789, 2012.

[9] G. Chen, Z. Wang, D. Wang et al., "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, pp. D983–D986, 2013.

[10] R. Spizzo, M. I. Almeida, A. Colombatti, and G. A. Calin, "Long non-coding RNAs and cancer: a new frontier of translational research?" *Oncogene*, vol. 31, pp. 4577–4587, 2012.

[11] T. Gutschner and S. Diederichs, "The hallmarks of cancer: a long non-coding RNA point of view," *RNA Biology*, vol. 9, pp. 703–719, 2012.

[12] A. Congrains, K. Kamide, R. Oguro et al., "Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B," *Atherosclerosis*, vol. 220, no. 2, pp. 449–455, 2012.

[13] R. Johnson, "Long non-coding RNAs in Huntington's disease neurodegeneration," *Neurobiology of Disease*, vol. 46, no. 2, pp. 245–254, 2012.

[14] J. H. Yang, J. H. Li, S. Jiang, H. Zhou, and L. H. Qu, "ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data," *Nucleic Acids Research*, vol. 41, pp. D177–D187, 2013.

[15] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," *Pacific Symposium on Biocomputing*, pp. 127–138, 2001.

[16] J. Harrow, A. Frankish, J. M. Gonzalez et al., “GENCODE: the reference human genome annotation for The ENCODE Project,” *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[17] J. Rozowsky, G. Euskirchen, R. K. Auerbach et al., “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls,” *Nature Biotechnology*, vol. 27, no. 1, pp. 66–75, 2009.