University of Tampere
BioMedTech
Computational biology group

Computational functional prediction of novel long noncoding RNA in TCGA Glioblastoma multiforme sample

Adossa, Nigatu Ayele
Master of Science Thesis
Supervisor: Prof. Matti Nykter
Reviewers: Prof. Matti Nykter, Dr. Juha Kesseli
Date: 10.02.2016

ABSTRACT

University of Tampere
Master's Degree programme in Bioinformatics
Adossa, Nigatu Ayele, Computational functional prediction of novel long noncoding RNA in TCGA glioblastoma multiforme sample
Master of Science Thesis, 81 Pages, 1 Appendix
February 2016
Supervisor: Prof. Matti Nykter
Reviewers: Prof. Matti Nykter, Dr. Juha Kesseli
Keywords: RNA, DNA, lincRNA, lncRNA, miRNA, mRNA, novel lincRNA, co-expression analysis, DNA-protein interaction, PWM, lincRNA-protein interaction, lincRNA-miRNA interaction, correlation, mutual information, ARCENE, functional prediction.

According to international human genome sequencing consortium 2004[43], it was known that only less than 2% of the total human genome code for proteins. This ignited quite a surprise in the scientific community. Since then, a lot of researchers are attracted towards the noncoding part of the genome. There are explosion of researches addressing the role of the 98% of the human untranslated regions of the genome. This shows that the transcription is not only limited to the protein coding regions of the genome rather more than 90% of the genome are likely to be transcribed. [43] This will result in the transcription of tens and thousands of the long noncoding RNAs (lncRNAs) with little or no coding potential. However, the molecular mechanism and function of long noncoding RNAs are still an open research topic. Although the functions of limited lncRNAs are identified, there is still a gap in identifying the function of novel lncRNAs.

This project implements different computational methods to predict the function of novel lncRNAs identified from TCGA glioblastoma multiforme samples. The methods used in this functional prediction include both expression and sequence-based analysis approach. In expression-based analysis, the co-expressing genes with lncRNAs are used to predict the possible functional relation. In sequence based analysis, the gene-protein and lncRNA-protein interactions together with miRNA-lncRNA interactions are considered towards the possible functional predictions.

The result from the integrated functional prediction on the novel lncRNAs show that TCGA_gbm-3-153501 novel lncRNA which is co-expressed together with the THBS1 gene with correlation coefficient of more that 0.5 is predicted to function in cell-cell and cell-to-matrix interactions, platelet aggregation, angiogenesis, and tumorigenesis. [202] MSI1, RBM3 and RBM8A are RNA binding proteins (RBPs) that have binding site on both the first top five differentially expressed lncRNAs which are *TCGA_gbm-2-104096501, TCGA_gbm-3-153501, TCGA_gbm-5-63687001 and TCGA_gbm-17-10671251* and IGF2 which is among the top 10 differentially expressed genes. Therefore, these lncRNAs are predicted to have functional role in cell proliferation and maintenance of stem cells in the central nervous system.

PREFACE

I would like to acknowledge the role of the teaching staff from the computational biology group in equipping me with the relevant knowledge in the biological and data science to tackle the future scientific challenges in the field of bioinformatics. Juha kesseli's lectures with blue and white backgrounded slide and his 5/6 exam questions, Kirsi Granberg's special way of teaching with her overwhelming smile and Ville Kytölä's humble assistance during the practical exercise sessions are part of the memories that I attach with this master's degree. I also would like to give special thanks to prof. Matti Nykter in advising me throughout my thesis work.

Tampere, 10.02.2016
Nigatu A. Adossa

Table of Contents

TERMS AND ABBREVIATIONS

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
siRNA	Small interfering
snoRNA	Small nucleolar RNAs
miRNA	microRNA
rRNA	Ribosomal RNA
LncRNA	long noncoding RNA
LincRNA	long intergenic RNA
mRNA	messenger RNA
ENCODE	The Encyclopedia of DNA Elements
FANTOM	Functional Annotation Of Mammalian genome
cDNA	Complementary DNA
LNCipedia	Database for annotated human lncRNA transcript Sequences and structures
ORF	Open reading frame
HOTAIR	HOX transcript antisense RNA
RNA splicing	Modification of the nascent pre-messenger RNA (pre-mRNA) transcript in which introns are removed and exons are joined
Chromatin remodeling	Dynamic modification of chromatin architecture to allow access of condensed genomic DNA
RNA-seq	RNA sequencing

SAGE	A sequencing technology capable of producing large number of short sequence tags identifying both known and unknown transcripts
GENCODE	Part of the pilot phase of the ENCODE project to identify and map all protein-coding genes within from ENCODE project.
MSigDB	Molecular Signatures Database
Onco-lncRNAs	Cancer causing long noncoding RNAs
Apoptosis	A process of programmed cell death
catRAPID	Algorithm to estimate the binding propensity of protein-RNA pairs
MCL	Markov cluster algorithm
MALAT1	Metastasis associated lung adenocarcinoma transcript 1
GBM	Glioblastoma multiforme
Novel transcript	The transcripts that has not been observed before
Promoter	A region in DNA that initiates the transcription of the particular gene
Exon	Part of the gene that becomes the final mature mRNA.
Intron	non coding sections of an premature mRNA transcript
Start codon	The first triplet of nucleotides on mRNA to be translated by rRNA
Stop codon	A nucleotide triplet that ignites the termination of translation in mRNA
PWM	Position weight matrix
PFM	Position frequency matrix
Mutual information	The measure of variable's mutual dependency.
Chromosome	A structure in a cell that contain the DNA molecules. Human cells hold 23 pairs of chromosomes.

DEesq2	Differential expression analysis algorithm for high throughput count data such as RNA-seq
RPKM	Reads per kilobase pairs per million of mapped reads, a normalization method in sequencing count data.
p-value	The measure of statistical significance in the hypothesis testing.
RBP	RNA binding proteins
miRanda	A software package to predict the microRNA target proteins
3' UTR	A section of mRNA that immediately follows the translation termination codon
5' UTR	The region in mRNA that is directly upstream from start codon.
KEGG	Database resource for understanding high-level functions and utilities of the biological system.
RBPDB	RNA binding protein database.
hDPI	Human DNA-protein interaction database

1 INTRODUCTION

Ribonucleic acid (RNA) is one of the three basic biological macromolecules that constitute living cells alongside with deoxyribonucleic acid (DNA) and protein. The flow of genetic information originates at DNA to be transcribed or copied to the RNA and then translated into protein. Proteins play the central role in all cellular endeavors of living things, ranging from playing important role as an enzyme, structural component of the cell and in cell signal transduction.

DNA is a blueprint that holds all information regarding the cell and the RNA is a photocopy of the part of DNA. When the cell is in need of a certain protein, it turns on the portion of DNA or gene that codes for the particular protein. The RNA copies the information as mRNA. The genetic code in mRNA then translated into protein using the cell's protein manufacturing machinery, ribosome, which is composed of ribosomal RNA (rRNAs).

Not all transcripts that copy genetic information from DNA can be translated into protein. Surprisingly, 97% of transcribed transcripts are non-protein coding. In general RNA can be categorized into two: protein-coding RNA and non-protein coding RNA (ncRNA). Protein coding RNA, which is mRNA, has been the central topic of interest for long period of time in the field of molecular biology. Considering the change in the concentration of mRNA at the transcript level and the defect in protein synthesis are widely researched topic. Transcriptional and translational defects were assumed due to the mutation, chromosomal aberration or DNA damages at the genomic level. The non-protein coding RNAs were also considered to be only the building block of transfer RNA (tRNA), which acts as an adapter that bridges the mRNA sequence & the amino acid produced during translation and structural component of ribosomal RNA (rRNA) that acts as a protein manufacturing machinery in a cell.

But in recent years, researches have shown that non-coding RNAs can act as enzymes, carrier for viral genetic information and regulation of different cellular process ranging from cell division, cell differentiation, cell growth, cell aging and apoptosis. There are different types of non-coding RNAs that specialized in particular functions in a cell. For example, microRNA (micrRNA) regulates the translation of mRNA by binding to 3'UTR region of the mRNA or in combination with small interfering RNA (siRNA), it causes the degradation of target mRNAs. Small nucleolar RNAs (snoRNA) act on modification of small nuclear RNA (snRNAs), whose primary function is in the processing of the pre mRNA, mRNAs and rRNA in ribosome biogenesis.

The other non-coding RNA that has attracted the attentions of many researchers recently is a long noncoding RNA (lncRNA), which is different from the other ncRNAs in that it has a longer nucleotide base, about 200 or longer. The whole-transcriptome analysis revealed vast number of lncRNAs.

Significant numbers of lncRNA are involved in various biological functions. Yet most of them are not annotated and the functional and molecular mechanisms have not been known. As the numbers of discovered lncRNA are huge, it is difficult to experiment the functionality of lncRNA in the laboratory for all of them. Hence, the need to implement computational method to predict the functionality of different lncRNAs is vital. The goal of this thesis project is to fill the gap on computational functional prediction for lncRNA by integrating the gene expression and the newly identified lncRNAs expression in search of the co-expressing genes and novel lncRNAs together with the sequence based DNA-protein, lncRNA-protein and miRNA-lncRNA interaction.

2 BACKGROUND

LncRNAs are RNAs that have length of 200 nucleotides or more without protein-coding potential. It can be distinguished from small RNAs, such as microRNA, siRNA, snRNA and snoRNAs, by its length and from mRNA by its coding potential [3]. The genomic origin, subcellular compartmentalization and modifications of RNA reveal its possible functionality [2]. Out of all transcribed nucleotide bases, 79.6% can be mapped back to the nuclear long RNAs while 15.1% are exclusively cytosolic long RNA as it is shown in the figure 1B [2]. 5.3% of the sequences are also exclusively small RNAs as it is shown in figure 1A.

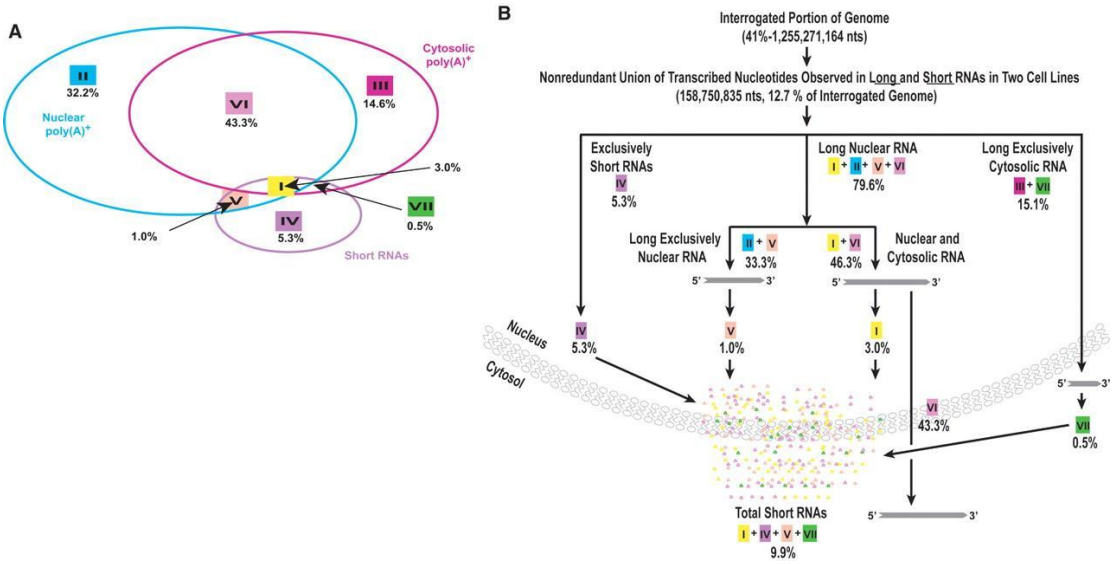


Figure 1. The nuclear and cytosolic distribution of lRNA & sRNA in human genome [2]

Although lncRNAs are such an abundant transcript in the genome, limited numbers of them are annotated. Recently, there has been a project that aims to create a catalog of lncRNAs and information on their tissue specific expression. According to ENCODE database version 3.0, there are 73,370 lncRNA entries from 1,239 organisms. Among these, only less than 200 were functionally annotated. [10] In 2000, the FANTOM [4] consortium discovered set of 34,030 lncRNAs catalog for mouse genome from the cDNA sequencing. [4] In 2010, 5446 lncRNAs catalog for the human genome are manually annotated. [5, 7] As to August 28, 2014, the LNCipedia 3.0 database contains the annotations for 113,513 human lncRNAs. [8, 9]

2.1 Calcification of lncRNA

LncRNAs are believed to play roles in transcription, translation, protein localization, cellular structure integrity, imprinting, cell cycle & apoptosis, stem cell pluripotency, reprogramming heat shock response, cancer progression and development of other diseases. In order to perform functional studies on lncRNAs, it is important to see the different category of lncRNAs. LncRNAs are classified into several categories depending on the genomic location and context, the effects exerted on DNA sequence, functional mechanism and targeting mechanism. [10]

2.1.1 Genomic location and context

Based on the genomic location and context, the lncRNAs can be classified as an intergenic long non-coding RNA (lincRNAs) and intronic long noncoding RNA. LincRNAs are long noncoding RNAs that are transcribed from the non-coding part of the DNA between the two protein coding genes as it is illustrated in figure 2A. Structurally, lincRNAs are similar with the protein coding genes. They are 3' polyadenylated, 5' capped and exhibit the transcriptional activation. [10] But they don't have an open reading frame (ORF) and they don't code for protein. LincRNAs are the largest class of the non-coding RNA molecules in the human genome. In 2011, annotation catalog with 4662 human lincRNAs are created. [6, 7]

Long noncoding RNA transcripts that are transcribed from the intron parts of the protein coding genes are intronic long noncoding RNAs as it is shown in figure 2B. Most of the intronic lncRNAs have the same tissue expression patterns as that of the corresponding protein coding genes. Hence, they may stabilize protein-coding transcripts or regulate their alternative splicing. [11]

The transcription of intronic and intergenic long noncoding RNAs are regulated in different transcription activation mechanism. As a result, they may have different poly (A) modification and exhibit activities in different cellular localization. A small portion of intronic lncRNA's function is studied in contrast with the lincRNAs that function through different mechanisms including cis or trans transcriptional regulation, translational control, splicing regulation, and other post-transcriptional regulation. [10]

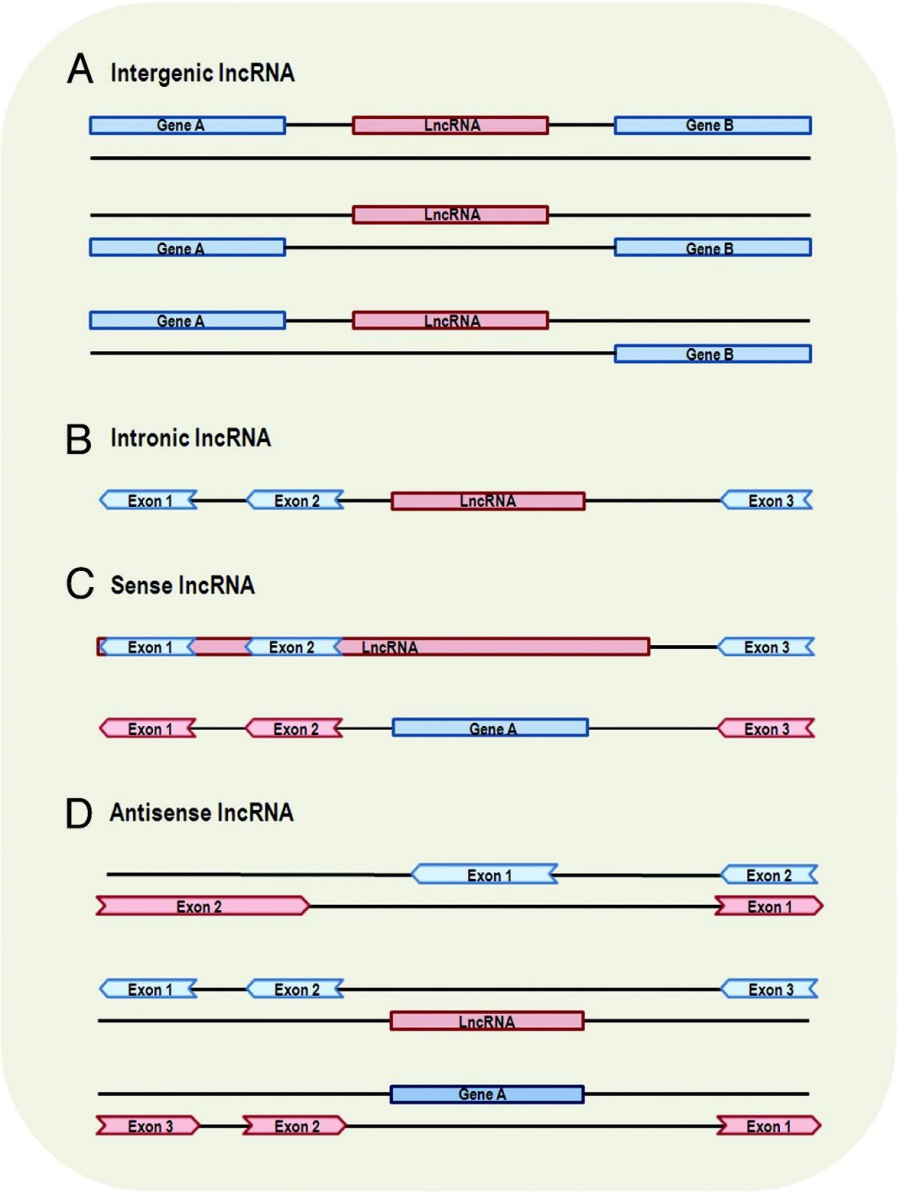


Figure 2. Classifications of lncRNAs [10]

The other category of lncRNAs based on the genomic location and context are sense and antisense lncRNAs. Sense lncRNAs are the ones that are transcribed from the sense strand of the protein-coding gene containing the exons as it is illustrated in figure 2C. In contrast, the antisense lncRNAs are the ones that are transcribed from the antisense strands of the protein-coding gene. The sense strand of the DNA is the strand that has the same sequence as that of transcribed mRNA while the antisense strand is the DNA strand that is used as a template during transcription. [10]

Yet another category of the lncRNA based on the genomic location and context is bidirectional lncRNA. This type of lncRNA sequence is transcribed from the opposite strand of the protein-coding gene whose transcription initiation site is less than 1000 base pair away. [12]

2.1.2 Effects exerted on the DNA sequence

Based on the effects exerted on the DNA sequences, lncRNAs can also be classified into cis and trans lncRNAs. As significant number of lncRNAs are involved in the transcriptional regulation, cis-lncRNAs are involved in regulation of gene expression in the close proximity while the trans-lncRNAs are the ones that involve in remote regulation of gene expression. [10]

Cis-lncRNA regulates the expression of the nearby gene by transcriptional interference or chromatin modification mechanism. The transcriptional interference takes place as the lncRNA binds to the promoter of the target gene to block the PIC (preinitiation complex) formation or by the interaction of the lncRNA with the transcription factor. Such cis-lncRNAs are transcribed from the promoter region of the genes. [10]

Cis-lncRNAs that function through the chromosomal modifications usually recruit the chromatin modification complex such as PRC (polycomb repressive complex) or Rpd3S HDAC (Rpd3 small histone deacetylase complexes). [10] PRC is the most studied chromatin modification complex. X inactive-specific transcript (Xist), the human 19 kb lncRNA, binds to PRC2 to induces the H3K27me3 modification to make a transcriptional silencing in X chromosomes. In figure 3, L1 & L3 are labeled to be cis-lncRNAs. [10]

Trans-lncRNAs act in trans-acting mode to target distant gene loci. For example, HOTAIR (HOX antisense intergenic RNA), which is approximately 2.2 kb pair long, lncRNA that is transcribed from the HOXC (homeobox C cluster) gene locus in chromosome 12, can be transported by the Suz-Twelve protein to regulate the homologous target sites at HOXD (homeobox D cluster) gene locus in chromosome 2. [10] The trans-lncRNAs function independently of sequence complementary to target the gene loci. Trans-lncRNAs binds to the transcription elongation factors or RNA polymerases in addition to the chromatin modification complex to affect transcription. For example, lncRNA, B2 SINE RNA, has been found to bind the polymerase II complex to block its activity during heat shock response. [10]

2.1.3 Cellular molecular mechanism

Based on the cellular molecular mechanism, lncRNAs can also be classified into those that affect the transcriptional regulation, post-transcriptional regulation and other functions. lncRNAs that involve in the transcriptional regulation function in two mechanisms. The first one is transcriptional interference

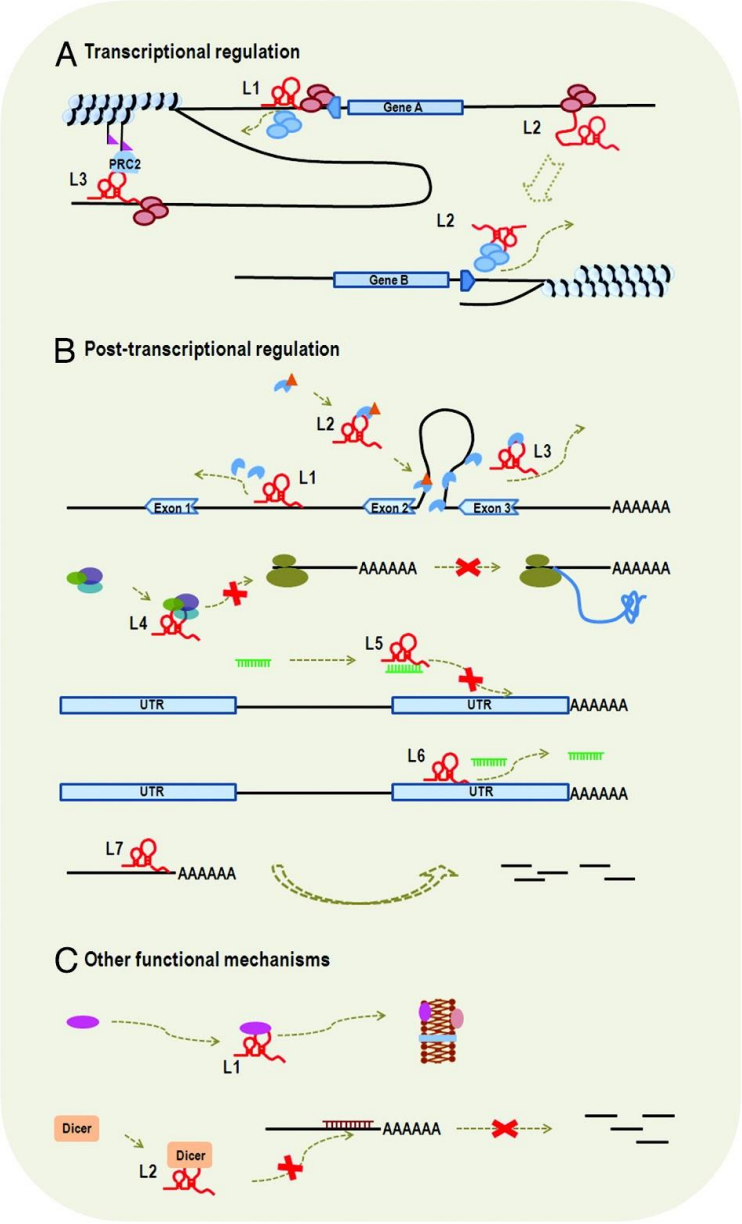


Figure 3. Functional mechanism of lncRNA [10]

mechanism. As it is illustrated in the figure 3A, L1 is a cis-lncRNA that is transcribed from the promoter region of gene A and it binds to the promoter regions of gene A to block the binding of transcription factors. Hence, L1 is acting in transcriptional interference mechanism affecting the transcriptional initiation of gene A. In the same way L2 influences the transcription of gene B by interacting with the transcription factors and RNA polymerase as it is illustrated in figure 3A.

The second transcriptional regulation functional mechanism is chromatin remodeling. As it is shown in the figure 3A, lncRNA L3, functions to modify the chromatin protein by recruiting the chromatin modification complex PRC2.

In post-transcriptional regulation, lncRNAs are actively involved in the splicing regulation and translational control as it is shown in figure 3B. LncRNAs that involve in splicing regulation may influence the splicing of mRNA by binding to the splicing factors or by directly hybridizing with mRNA sequence to block the splicing. For example, MIAT (myocardial infarction associated transcript), which is approximately 9–10 kb long lncRNA, binds to SF1 (splicing factor 1) to inhibit splicing and spliceosomal complex formation. [10] In the same way, Malat1 (metastasis-associated lung adenocarcinoma transcript 1), which is approximately 7 kb long lncRNA, can bind to SR splicing factor [serine–arginine (SR)-rich splicing factor] and regulate its distribution in nuclear speckle domains. [10] Malat1 also influences the alternative splicing of pre-mRNAs. [10]

Those lncRNAs that function in translational control may function either in binding to the translational factors or ribosomes. There are two lncRNAs, BC1 (brain cytoplasmic RNA 1) and BC200 (200 nt brain cytoplasmic RNA), which can bind to eIF4A (eukaryotic translation initiation factor 4A), PABP (poly (A)-binding protein) and other factors, to repress translation initiation by blocking assembly of the required complex. [10] snaR (small NF90-associated RNAs), a cytoplasmic lncRNAs and Gadd7 (growth arrested DNA-damage inducible gene 7) bind to the ribosome to influence the translation of mRNA. [10]

Besides from the splicing and translational control mechanism, there are other post-translational mechanisms that are utilized by the lncRNAs. lncRNAs involved in the process of siRNA (small interfering RNA) mechanism and interact with the miRNA to stabilize the target mRNA as it is shown in figure 3B. Such kind of lncRNAs are called ceRNAs (competing endogenous RNAs). For instance, linc-MD1 (long intergenic ncRNA that is associated with muscle differentiation), which is approximately 0.5 kb pair long lncRNA, acts as sponge/target mimic of miR-133 and miR-135 to regulate the expression of two transcription factors: MAML1 (mastermind-like protein 1) and MEF2C (myocyte-specific enhancer factor 2C), which activate the expression of muscle-specific genes. [10]

LncRNAs may also function as a natural antisense inhibitors to promote degradation of mRNA as it is shown in figure 3B. It has been found that 21A, which is approximately 300 bp long lncRNA, which shows high sequence homology to CENP-F (centromere protein F) intronic portions, can reduce CENP-F expression at both mRNA and protein level through antisense inhibitor.[13]

The other functional mechanism of lncRNAs beside the transcriptional and post-transcriptional regulation, are the protein localization, telomerase replication and RNA interference as it is shown in the figure 3C. [10]

2.1.4 Targeting mechanism

The last criterion to classify lncRNAs is based on their targeting mechanisms. Based on their targeting mechanism, lncRNAs are classified into signal, decoy, guide and scaffold lncRNAs. The signal lncRNA shows the cell type specific expression and its transcription is to deliver response to the stimuli. The decoy lncRNA binds and titrates away the protein target but does not exert any additional function. The guide lncRNA binds to protein to direct the localization of the ribonucleoprotein complex to specific target. The scaffold lncRNAs serve as a platform to bring multiple proteins together to form a ribonucleoprotein complex. [10]

However, one lncRNA might have several targeting mechanism that the mode of action does not solely drawn from a single targeting mechanism. For example, Xist, Air, COLDAIR, HOTTIP, HOTAIR and lincRNA-p21 lncRNAs operate in a dual mode as both signal and decoy lncRNAs. Some lncRNAs such as HOTAIR have more than two archetype or targeting mechanisms: anatomic signal, guiding the chromatin-modifying complexes to the target gene, and as a scaffold for PRC2 and LSD1. [10]

The other way to categorize the lncRNAs are based on the type of interaction they make with their targets such as RNA-RNA pairing, RNA-DNA hybrids, RNA structure mediated interactions, and protein linkers.

2.2 Molecular mechanism of lncRNA

Unlike other small noncoding RNAs, which are highly conserved and function via transcriptional and posttranscriptional gene silencing through specific base pairing with their targets, lncRNAs are poorly conserved and involved in gene regulation by diverse mechanisms that are not fully known yet. But, it is generally believed that lncRNAs may function by interacting with DNA, RNA and protein molecules and has four types of archetype that play roles as signal, decoy, scaffold and guide in the cell.

2.2.1 Signal

One of the molecular mechanisms for lncRNA is acting as a signaling molecule as the transcription of the lncRNA takes place at the specific time and place to integrate the developmental cues, interpret cellular context, or respond to diverse stimuli. This signifies that their expression is in response to specific transcriptional control mechanisms by the diverse cellular stimuli such as cellular stress and temperature. After transcription, lncRNAs might be involved in other regulatory mechanism or they might also be byproduct of the transcription. The signal lncRNAs can be used as a marker for the functionally significant biological event in the cell. [15]

2.2.2 Decoy

The second archetype in the molecular mechanism of the lncRNA is Decoy. After transcription, decoy lncRNAs bind and titrate away protein factors such as transcription factors, chromatin modifier and other regulatory factors from its target without exerting any additional function. Decoy lncRNAs "sponge" the protein factors leading to a broad change in the cellular transcriptome as it is shown in the figure 4 II. [15]

2.2.3 Guide

The third archetype of lncRNA is guide in which the lncRNA binds to the protein and directs the localization of the ribonucleoprotein complex to specific target. As a result the guide lncRNA can guide changes in gene expression either in cis (to its neighboring genes) or in trans (to the distance genes) that cannot be easily predicted by just the lncRNA sequence itself as it is illustrated in the figure 4 III. [15]

2.2.4 Scaffold

Assembly of complex protein complexes can be supported by fourth archetype of lncRNAs, which serve as central platforms upon which relevant molecular components are assembled. The scaffold lncRNAs link factors together to form new functions. Some lncRNAs possesses different domains that bind distinct protein factors that altogether may impact transcriptional activation or repression. Figure 4 IV shows how scaffold works. [15]

2.3 LncRNA identification

The identification process of the lncRNA starts from obtaining all transcripts including ncRNAs and mRNAs. Then the next step is going to be classifying each transcripts based on the coding potential. After distinguishing the mRNA from the ncRNA, then lncRNA are from the other small ncRNAs. This can be achieved by both experimental and computational methods.

Experimentally, the traditional microarray are designed to detect the protein coding transcripts or mRNA. Therefore, there has to be other unbiased experimental methods, such as tilling array, serial analysis of gene expression (SAGE), high-throughput RNA sequencing (RNA-seq), RNA-immunoprecipitation (RNA-IP) and Chromatin Signature-Based Approach, to identify of the lncRNAs from mRNAs.

Computationally, the lncRNAs can be identified in three ways. The first one is using ORF Length Strategy. The second approach is sequence and secondary structure conservation strategy while the third approach is using the machine learning strategy.

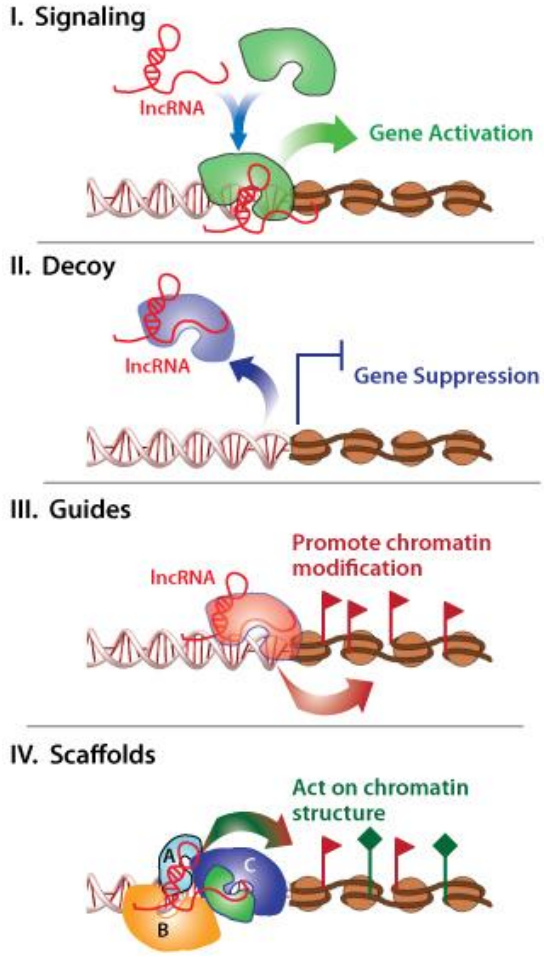


Figure 4. The molecular mechanisms of lncRNAs [15]

2.3.1 Experimental methods

2.3.1.1 Tilling array

Tilling array is subtype of microarray in which cDNA is hybridized to the microarray slides carrying overlapping oligonucleotides that cover either specific chromosomal regions or a complete genome. It allows the analysis of global transcription from specific genomic regions and it was initially used for both identification and expression analysis of lncRNAs. [16]

2.3.1.2 Serial analysis of gene expression (SAGE)

Serial analysis of gene expression (SAGE) is a technology capable of producing large number of short sequence tags identifying both known and unknown transcripts. It is based on the generation of short stretches of unbiased cDNA sequence SAGE tags by restriction enzymes. SAGE tags are then concatenated before cloning and sequencing. [19]

It has been used and proved to be an efficient approach in identification and analysis of lncRNAs. For example, Gibb et al. compiled 272 human serial analysis of gene expression (SAGE) libraries to delineate lncRNA transcription patterns across a broad spectrum of normal human tissues and cancers. [17]

2.3.1.3 High-throughput RNA sequencing (RNA-seq)

As the sequencing technology is advancing rapidly, RNA-seq or whole transcriptome shotgun sequencing experiment is becoming the mainstream technique in novel discovery of transcripts and gene expression analysis. RNA-seq has advantage over the other techniques in that it is sensitive in detecting the less abundant transcripts. In addition to that it is also effective in identifying de novo splicing isoforms and novel ncRNAs. Figure 5 show the workflow of identification of the lncRNAs. [14]

RNA-seq is the most widely used technique in identification of lncRNAs. For example, Li et applied the RNA-seq experiment in the identification of lncRNAs during chicken muscle development. [20] In the other way, Prensner et al also used RNA-seq experiment to study lncRNAs in prostate cancer from 102 prostate cancer tissues and cell lines. Based on the study, it has been concluded that lncRNAs may be used in cancer subtype classification. [21]

2.3.1.4 RNA-immunoprecipitation (RNA-IP)

RNA-IP is a technique that is used to identify lncRNAs, which are interacting with the specific protein molecule. The antibody of the protein is used to isolate the lncRNA-protein complexes, and then deep sequencing the cDNA library results in the sequences of lncRNA that has been interacting with the given protein. [14]

2.3.1.5 Chromatin Signature-Based Approach

Chromatin signature-based approach used the chromatin signatures to study the transcription of the lncRNA genes and others. This chromatin signature includes H3K4me3, the marker of active

promoters, and H3K36me3, the marker of transcribed region. Guttman et al. has used this approach in identifying 1600 large multiexonic lncRNAs regulated by the p53 and NFkB transcription factors. [22]

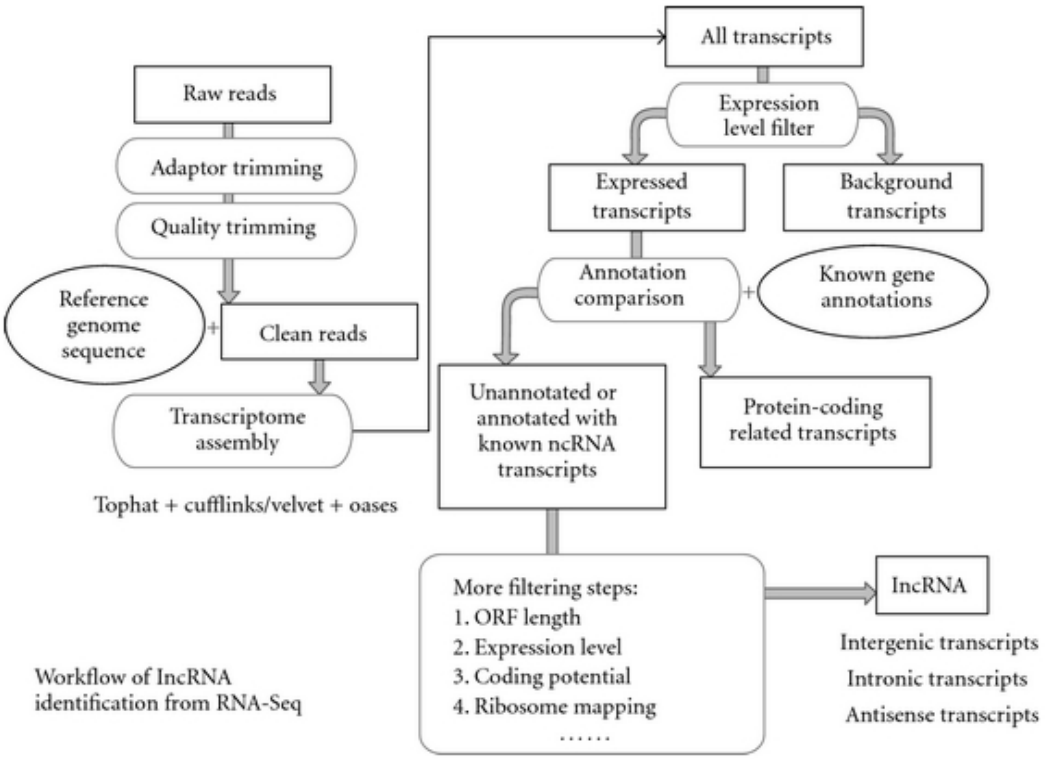


Figure 5. The workflow in the identification of lncRNAs using RNA-seq [14]

2.3.2 Computational Methods

2.3.2.1 ORF Length Strategy

It is known that the start and termination codon for lncRNA are distributed randomly. This implies that the ORF in lncRNAs are hardly any to find, if there exists, it is short in length, not more than 100 codon. In this method, transcripts that lack the open reading frame (ORF) or have short ORF are computationally identified. [14]

However, this method is not efficient way to identify lncRNAs from mRNAs as there are protein-coding transcripts with shorter ORF and noncoding RNAs with ORF of longer ORFs. To overcome this problem, Jia et al. utilized the comparative genomic approach in which they used the cDNA that have no homologous protein more than 30 amino acids across the mammalian genome are regarded

as lncRNAs. [23] As this approach is heavily dependent on the completeness of the database, the lack of the genomic annotation highly influences the result. [14]

2.3.2.2 Sequence and secondary structure conservation strategy

LncRNAs are generally less conserved as compared to the protein coding RNAs. Thus, they are subjected to mutation. The codon substitution frequency (CSF) is one of the criteria to measure mutation. Therefore, it could be used as the means to identify the lncRNAs. For example, Guttman et al. used the CSF score as a means of calculating the coding potential for the RNA sequences and identified the lncRNAs from the coding RNAs. [24] In the other hand, Clamp et al. and Lin et al. combined the CSF with reading frame conservation (RFC) to come up with better way of identifying the lncRNAs from the mRNAs. [25][26] PhyloCSF [28] is a tool that implement two phylogenetic models based on the intrinsic features of the sequence and sequence conservation to distinguish the coding and noncoding RNAs. [33]

There are also other approaches that are engaged in identifying the lncRNAs from the mRNAs using the conservation of the RNA's secondary structure. Tools that are implemented in such a way includes QRNA [29], RNAz[30] and EvoFOLD[31]. But, this approach has a major limitation in that the lncRNAs do not have common secondary structures. [14]

2.3.2.3 Machine learning strategy

There are a number of tools that has been developed to identify the lncRNA from the mRNA utilizing different machine learning techniques. For example, coding-non-coding (CONC) is a tool that has been developed using a support vector machine (SVM). It classifies transcripts according to features they would have if they were protein coding. This features, such as peptide length, amino acid composition, predicted secondary structure content, predicted percentage of exposed residues, compositional entropy, number of homologs from database searches, and alignment entropy, are used to train the SVM model. [34][14]

Another tool that has been developed using the machine learning approach is coding potential calculator (CPC). [35] CPC also uses SVM for modeling and extracting the sequence features and comparative genomic features to calculate the coding potential of the transcripts. [35] In the other hand, Lu et al. has also developed a tool using machine learning method that uses the GC contents, DNA conservation and expression information to identify the coding and noncoding transcripts in C. elegans genomes. [38] Table 1 summarizes the machine learning methods and the features used to train the models.

However, as different machine learning methods shown effectiveness in identifying lncRNAs, there are concerns that there are still exceptional cases. For example, some genes are bifunctional and exhibit both coding and non-coding isoforms such as in the case of steroid receptor RNA activator (SRA). [36] In the other hand, the transcriptability of RNAs might change during the course of evolution as Xist lncRNA evolved from protein coding gene. [37]

Table 1. Machine learning methods for identification of lncRNAs [14]

Method	Features	Algorithm	References
CONC	Peptide length Amino acid composition Hydrophobicity Secondary structure content Percentage of residues exposed to solvent Sequence compositional entropy Number of homologs obtained by PSI-BLAST Alignment entropy	SVM	[34]
CPC	ORF prediction quality Number of homologs obtained by BLASTX Alignment quality Segment distribution	SVM	[35]
Lu et al.	RNA-seq experiments Tilling arrays poly-A + RNA-seq experiments poly-A + tilling arrays GC content DNA conservation Predicted protein sequence conservation Predicted secondary structure free energy Predicted secondary structure conservation	Naïve Bayes Bayes Net Decision Tree Random Forest Logistic Regression SVM	[38]

2.4 Function of lncRNA

Recent studies have shown that lncRNAs have diverse cellular functions. Figure 6. Shows different functions that lncRNA is involved in. lncRNAs play a major role in different epigenetic regulations in a cell by recruitment of chromatin remodeling complex (CRC) to specific loci. In this regard, Pandey et al. [39] showed that Kcnq1ot1, lncRNA found in KCNQ1 loci, is important for the bidirectional silencing of genes in Kcnq1 domain by interacting with chromatin, H3K9- and H3K27-specific histone methyltransferases G9a and PRC2 complex in a lineage-specific manner. In other study Nagano et al. [40], Air, which is another lncRNA, is responsible for allele specific transcriptional silencing of cis-

linked Slc22a3, Slc22a2, and Igf2r genes in mouse placenta by interacting with the Slc22a3 promoter chromatin and the H3K9 histone methyltransferase G9a in placenta.

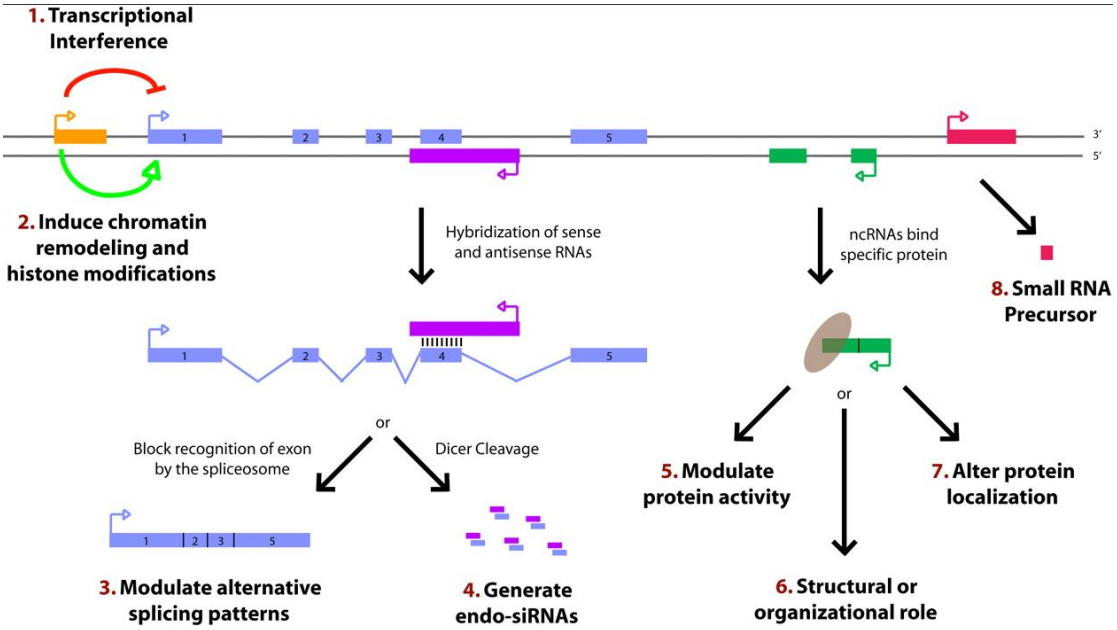


Figure 6. Function of lncRNAs[43]

The second cellular function of lncRNA is regulation of gene expression by interacting with protein partners in the biological process such as protein synthesis, imprinting (Kcnq1ot1, Air), cell cycle control (TERRA), alternative splicing (MALAT1), and chromatin structure regulation (DNMT3b, PANDA). MALAT-1 regulates alternative splicing through its interaction with the serine/arginine-rich (SR) family of nuclear phosphoproteins, which are involved in the splicing machinery. Another lncRNAs such as Gas5 and p21 (lincRNA-p21) are involved in apoptosis and cell cycle control. [41]

Thirdly, lncRNAs plays significant role in enhancer-regulating gene activation (eRNAs) in which they interact with distal genomic regions. The distal genomic region, such as enhancers, transmits transcriptional regulation instructions by interacting with lncRNAs. [42]

Fourthly, lncRNAs has functional role in processing of small RNA molecules. lncRNAs are processed to yield small noncoding RNAs. They can also modulate how other small RNAs are processed. Therefore, lncRNAs can act as an interacting partner or precursor for the small RNAs. For example, microRNA (miRNA) are usually the result of sequential cleavage of lncRNAs. Figure 7A. Illustrates how lncRNAs are cleaved to yield two small ncRNAs. Figure 7B shows that how MALAT1 lncRNA transcript is processed to yield two small ncRNAs. The Piwi-interacting RNA (piRNAs) can also be produced by processing single lncRNA transcript. [43]

Fifthly, lncRNAs serve as structural component. A number of RNA-binding proteins, including paraspeckle protein component 1 (PSPC1, also known as PSP1α) and NONO (also known as

p54/nrb), localize to paraspeckles. paraspeckles is an irregularly shaped compartment of the cell in the nucleus's interchromatin space. This dynamic structure can be altered in response to

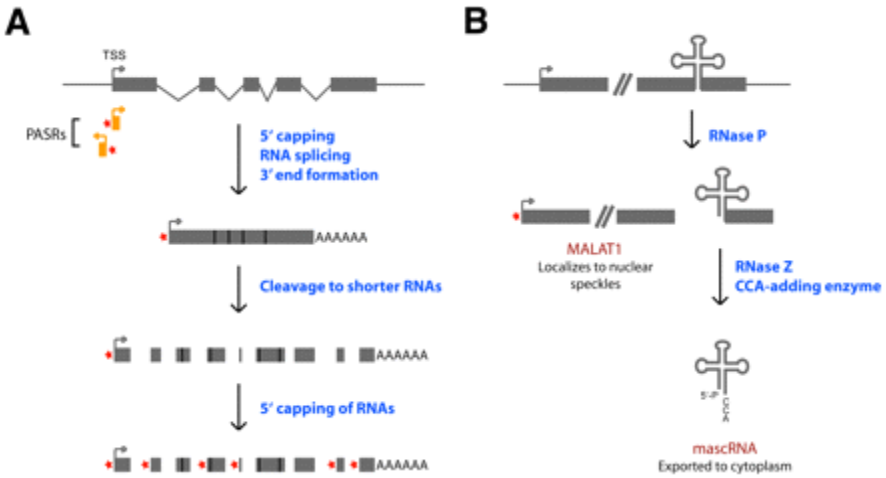


Figure 7. LncRNA processing to yield small ncRNA[43]

cellular metabolic activity. They are transcription dependent. As the exact function of the paraspeckles not known, they are suggested to act as the storage site for nuclear-retained RNAs. The RNase A treatment of the paraspeckles result in the disruption of structural integrity of paraspeckles implying that lncRNAs can be a crucial components of the nuclear structure. [43]

2.5 LncRNA and disease

Studies show that the expression of lncRNAs is tissue specific. The majority of lncRNAs have brain-specific expression patterns implying that the deregulation of lncRNAs have important role in the central nervous system (CNS) pathologies such as neurodevelopmental, neurodegenerative and neuroimmunological disorders, primary brain tumors and psychiatric diseases. The loss-of-function studies on mouse embryonic stem (ES) cell also showed that the knockdown of the lincRNA have significant role in the regulation of gene expression pattern in trans. This indicates that the lncRNAs play important role in the regulation of the diverse cellular processes such as stem cell pluripotency, development, cell growth and apoptosis, and cancer invasion and/or metastasis. [44]

There is an increasing number of evidences that show the mutation and dysregulation of lncRNAs including the alteration of the primary and secondary structure together with the expression level of the lncRNAs and the RNA-binding protein associated with it lead to diverse human diseases ranging from neuronal to cancer diseases.[45]

The dysfunctionality of lncRNAs appears to be the cause for majority of complex human diseases such as leukaemia[143], colon cancer[144], prostate cancer[145], breast cancer[146], hepatocellular carcinoma[143],[147], psoriasis[148], ischaemic heart disease, [149],[150], Alzheimer's disease[151] and spinocerebellar ataxia type 8[152]. The following table summarizes the name of lncRNAs and their function in most frequent cancer types.

Table 2. LncRNAs in frequent cancer types [46]

LncRNA	Function	Reference
Prostate		
CBR3-AS1	Oncogene	[49]
CTBP1-AS	Oncogene	[50]
GAS5	Tumor suppressor	[51],[52]
H19	Putative susceptibility and diagnostic marker	[53],[54]
MALAT1	Putative marker	[55]
PCA3	Diagnostic marker	[56],[57]
PCAT1	Putative marker and oncogene	[58]
PCGEM1	high-risk pridictive marker and oncogene	[59],[60],[61],[62],[63],[64],[65]
PRNCR1	Oncogene	[66],[67]
PTENP1	Oncogene, tumor suppressor	[68],[69]
ucRNAs	Putative oncogene	[70]
XIST	Diagnostic and prognostic marker	[71],[72]
Breast		
ANRIL	Tumor suppression	[73]
BC040587	Prognostic biomarker	[74]
BCAR4	Oncogene	[75]
BCYRN1	Prognostic biomarker	[76]
DSCAM-AS1	Involved in malignant progression	[77]
GAS5	Tumor suppression	[78]

H19	Oncogenic, involved in imprinting	[79]
HOTAIR	Prognostic marker of metastasis	[80]
LincRNA-BC4	Associated with advanced diseases	[81]
LincRNA-BC5	associated with advanced diseases	[81]
LSINCT5	Proliferation	[82]
Loc554202	proliferation and cell migration	[83]
MALAT1	Splicing and metastasis	[79]
MEG3	Suppression of growth	[84]
MIR31HG	Epigenetics	[85]
PINC	Cell survival and progression	[86]
PVT1	Inhibition of apoptosis	[87]
SRA1	Co-activator of estrogen receptor alpha	[87]
XIST	Epigenetics	[89]
ZNFX1-AS1	Tumor suppressor	[90]
Lung		
CCAT2	Invasion	[91]
HOTAIR	Proliferation and invasion	[92]
LincRNA-P21	Mediated gene repression	[93]
LincRNA-MVIH	Proliferation and invasion	[94]
MALAT1	Metastasis	[95]
MEG3	Suppression of growth	[84]
MINA	Oncogene	[96]
TUG1	Suppression of cell proliferation	[97]
Colorectal		

CCAT1	Diagnostic marker and involved in proliferation and invasion	[98],[99]
CRNDE	Diagnostic marker	[100]
HULC	Liver metastasis	[101]
HOTAIR	Metastasis	[102]
KCNQ1OT	Epigenetic modifier	[103]
LincRNA-P21	Suppression of P53 target gene and enhances sensitivity to radiotherapy	[104],[105]
lnc-LET1	Hypoxia mediated metastasis via HDAC3	[106]
MALAT1	Metastasis	[107]
SNGH16	Cell migration and invasion	[108]
Melanoma		
ANRIL	Scaffold of the chromatin modifying complex	[109]
BANCR	Cell migration	[110]
C9orf14	candidate tumor suppressor	[111]
SPRY4-IT1	Modulates apoptosis and invasion	[112]
Bladder		
Linc-UBC1	Negative prognostic factor	[113]
H19	Recurrent marker	[114]
MALAT1	Mediator of TGF-B-induced epithelial mesenchymal transition	[115]
MEG3	Possible tumor suppressor	[116]
SNHG16	Proliferation, invasion and affects sensitivity of chemotherapy	[117]
TUG1	Proliferation	[118]

UCA1	Cell proliferation, migration and chemoresistance	[119] and [120]
Non-Hodgkin lymphoma		
BIC	Up-regulation in B-cell lymphomas	[121]
Kidney (renal cell and renal pelvis) cancer		
GAS5	Tumor suppressor	[122]
H19	Tumor suppressor	[123]
HIF1A-AS1	Up-regulated in cancer	[124]
HIF1A-AS2	UP-regulated in cancer	[124]
KCNQ1OT1	Oncogene	[125]
MALAT1	Tumor progression and poor prognosis	[126]
MEG3	Tumor suppressor	[127]
Thyroid		
AK023948	Down-regulated in thyroid carcinoma	[128]
NAMA	Activated by BRAF mutation,contributes to cell proliferation and activates autophagy	[129]
PTCSC3	Down-regulated in papillary thyroid carcinoma	[130]
Endometrial		
CASC2	Potential tumor suppressor	[131]
HOTAIR	Proliferation and invasion	[132]
MALAT1	Up-regulated	[133]
Leukemia (all types)		
ANRIL	Silencing of CDKN2B	[134]
DLEU1	Tumor suppressor, harbors mir-15-a and mir-16-1	[135]

DLEU2	Tumor suppressor, harbors mir-15-a and mir-16-1	[136]
MEG3	Tumor suppressor methylated in acute myeloid leukemia	[137]
MIR155HG	Direct NF-κB target gene	[138]
TCL6	Involved in leukemogenesis	[139]
WT1	AS,Epigenetic and splicing defects	[140]
Pancreatic		
DAPK	High expression in metastatic disease	[141]
HOTAIR	Predictive of metastasis and disease progression	[102]
MALAT1	Splicing,metastasis	[79]
MAP3K14	High expression in metastatic disease	[141]
PPP3CB	High expression in metastatic disease	[141]
PVT1	Drug sensitivity regulation	[142]

Recent studies show that some lncRNAs are involved in mediating invasion and metastasis in the cancer cells by altering the gene expression pattern epigenetically. Such lncRNAs are identified to be polycomb repressive complexes (PRC) protein dependent in transcriptional control. PRCs are the polycomb group (PcG) proteins that function in multiprotein complex. [47] PcG proteins play major role in repressing the promoter of the genes that are crucial in the cell fate determination and embryonic development. In cancer, the PcG target genes are frequently silenced epigenetically by DNA methylation due to the high level of expression of PcG proteins in the cancer cell. HOTAIR, PCAT-1 and MALAT1 are lncRNAs that are involved in the PRC dependent transcriptional control that their overexpression induces metastasis and invasion in different cancer cells. [47]

2.5.1 HOTAIR (HOX transcript antisense intergenic RNA)

The lncRNA HOTAIR is located at the 12th chromosome HOXC locus (12q13.13) and it regulates the expression of HOXD genes at chromosome 2 by binding to PRC2 in trans. It induces the epigenetic

silencing of several tumor suppressor genes at HOXC loci by metelation. Gupta *et al. [153]* revealed the important role of *HOTAIR* in breast cancer metastasis. Additionally, the high level expression of the HOTAIR in primary breast tumor is a powerful indicator of metastasis and mortality. Conversely, the low level of expression of HOTAIR inhibits cancer invasiveness in cells with high excessive PRC2 activity.[153][47] HOTAIR is involved in one or other way in the following cancer types: Lung, breast, colorectal, esophageal, laryngeal, nasopharyngeal, hepatocellular, gastric, pancreatic carcinoma, non-small cell lung cancer, mesenchymal glioma cancers.

2.5.2 PCAT-1 (prostate cancer-associated ncRNA transcripts-1)

The PCAT-1 lncRNA is located at 8q24 which is considered as a "gene desert" loci, around the well-studied SNP associated with prostate cancer risk, nearly 725 kb upstream of the *c-MYC* oncogene. Recent studies Prensner *et al. [154]* show that the PCAT-1 is associated with metastasis and can promote the cell proliferation in prostate cancer by repressing the PRC2 complex. However, further studies is still needed in order to well explain the regulatory mechanism of PCAT-1. [154]

2.5.3 MALAT1 (metastasis-associated lung adenocarcinoma transcript 1)

MALAT1 is lncRNA later referred to as NEAT2 (nuclear-enriched abundant transcript 2) that act as the prognostic marker of metastasis and patient survival in non-small cell lung cancer. [155][156] As MALAT1 is the most abundant in human cell types and highly conserved across several species, the deeper the study of this lncRNA is expected to reveal the functional mechanism and importance of the lncRNAs. MALAT1 is upregulated in lung, breast, prostate, colon and liver cancers.[46] The molecular mechanism of MALAT1 is identified as the regulator of the alternative splicing for certain genes[158], the genetic regulator of the metastasis-associated genes in lung cancer and activator of gene expression in mediating the assembly of coactivator complexes by binding to the unmethylated polycomb 2 (Pc2) which is a component of the polycomb repressive complex 2(PRC2).[157]

Many of the LncRNAs are recently being associated with p53 signaling pathway. Huarte *et al.[158]* has showed significant number of lncRNAs are a key constituent of the p53-dependent transcriptional pathways. For example, MALAT1, H19 and CCAT2 lncRNAs are involved in the Wnt/β-catenin pathway. [47]

2.5.4 H19

The H19 lncRNA, which is imprinted and maternally expressed at 11p15.5 in the genome, near the insulin like growth factor 2(IGF2) gene, has crucial role in genomic imprinting during the cellular growth and development. H19 is expressed marginally in nearly all normal adult tissues while the aberrant expression is observed in several cancer types including breast, hepatocellular, colon, bladder and esophageal cancers. It is also indicated that H19 plays important roles in gastric cancer.

H19 is involved in the down regulation of RB tumor suppressor gene in colorectal cancer by acting as a precursor for miRNA-675. [46][47]

2.5.5 GAS5 (The Growth Arrest Specific 5)

GAS5 at 1q25 is important player in mammalian apoptosis and cell growth. It binds to the receptors in a cell to inhibit the association of the receptors by blocking their binding domain. As a result several genes including cellular inhibitor of apoptosis 2 gene are suppressed and there will be a reduced cellular metabolism leading to cellular death. In both prostate and breast cancers GAS5 induces apoptosis directly or indirectly. [46] GAS5 has significantly reduced expression level in breast cancer cell as compared to the normal.[78] It also act as a tumor suppressor in renal cell carcinoma.[160] Studies also show that the decreased expression of GAS5 implicates that poor progenesis in cervical cancer. [161]

2.6 LncRNAs in diagnosis and therapy of cancer

As cancer is multifactorial and multistep disease, the need to have molecular malignancy biomarker is immense in cancer patient management. In the past, several biomarkers in several cancer types are discovered and validated, but recently, the lncRNAs are being introduced as a potential biomarker in the diagnostics and prognostics of different cancer types via both oncogenic and tumor-suppressive pathways. For example, the increased level of expression of HOTAIR is associated with metastasis in breast cancer. In addition, the expression of HOTAIR is correlated with metastasis in colorectal and many other malignant neoplasms in liver, stomach, nasopharynx, esophagus, and skin. [47]

In addition to the diagnostic potential of lncRNAs, studies also has shown lncRNAs can be potential therapeutic target. As the inhibition of some lncRNAs, such as MALAT1, are involved in metastasis and poor progenesis in non-small lung cancer, and the inhibition of MALAT1 does not affect the normal cell, there is a big potential of using such lncRNAs as a therapeutic target in metastatic cancers. [155][162] In recent study, Wheeler et al. [163], showed that the abolition of MALAT1 using gene therapy in vivo has a significant potential of curing a metastatic cancers. Other study, Ren S et al. [164], also showed that MALAT1 is involved in the maintenance of prostate tumorigenicity and be a potential therapeutic target for castration-resistant prostate cancer in nude mice by delayed tumor growth and reduced metastasis of the prostate cancer cells.

2.7 lncRNA databases

The accessibility of lncRNAs in public dataset is becoming more realistic in recent years. There are a number of lncRNA databases, with different data coverage and quality, are developed to give services via web interface. The number of lncRNAs stored in the existing databases is different based on the source of transcripts. The lncRNAs stored in the databases might be collected in one of the following

means: literature, computational predictions, or primary data repositories such as informations from GENCODE project. But Functional lncRNA Database and lncRNADisease databases are mairly rely on the manually curated, literature-extracted annotations. Most of the lncRNA databases contain human and mouse lncRNAs while LncRNAdb and Noncode v4.0 databases covers large number of species' lncRNAs ranging from yeast to plants.[165] Table 3 summarizes the lncRNA databases.

NONCODEv4 is one of the largest databases that hosts around 210 831 lncRNAs transcripts for human and mouse. It gives the graphical expression profile of lncRNA genes based on public RNA-seq data and predicts their functionality for both human and mouse.[166] CHIPBase, DIANA-LncBase, lncRNAdb, Noncode v3.0, and lncRNome databases provide cell or tissue specificity of the lncRNAs, but only lncRNAdb and Noncode v3.0 designate the cellular localization of the lncRNAs.[165]

DIANA-LncBase is another lncRNA database with the largest number of experimentally verified, about 5000, and computationally predicted, about 10 million, microRNA targets on the lncRNAs. It provides the detail information on miRNA-lncRNA interaction with external links, graphic plots of transcripts' genomic location, representation of the binding sites, lncRNA tissue expression as well as MREs conservation and prediction scores. [168]

Some lncRNA databases such as DIANA-lncBase, lncRNAdb, Noncode v3.0, and lncRNome describe lncRNAs based on the validated and putative biological functional annotations. Functional lncRNA database exclusively contains the lncRNA-associated diseases that are experimentally validated while DIANA-lncBase, lncRNAdb, lncRNADisease, Noncode v3.0, and lncRNome provide putative or validated associations between lncRNAs and diseases. [165]

Table 3. List of lncRNA databases

Database	Web access link	Description	Reference
lncRNAdb	http://www.lncrnadb.org/	Contain comprehensive list of lncRNAs in eukaryotes, and mRNAs with regulatory roles	[168]
NONCODEv4	http://noncode.org/	Integrative annotation of noncoding RNA(210 831 lncRNAs)	[166]
LNCipedia	http://www.lncipedia.org/	offers 21 488 Annotated human lncRNA transcripts with secondary structure information, protein coding potential, and microRNA binding sites	[169]
The functional LncRNA database	http://www.valadkhanlab.org/database	Contains studied lncRNAs manually culled from the literature along with a parallel	[170]

		database containing all annotated protein-coding human RNAs	
ChIPBase	http://deepbase.sysu.edu.cn/chipbase/	Decods the transcriptional regulation of microRNA and lncRNA genes from ChIP-Seq data	[171]
DIANA-LncBase	www.microrna.gr/LncBase	Experimentally verified and computationally predicted microRNA targets on long non-coding RNAs	[172]
LncRNADisease	http://cmbi.bjmu.edu.cn/lncrnadisease	A database for long-non-coding RNA-associated diseases	[173]
LncRNA2Target	http://www.lncrna2target.org/	a database for differentially expressed genes after lncRNA knockdown or overexpression	[174]
lncRNASNP	http://bioinfo.life.hust.edu.cn/lncRNASNP/	a database of SNPs in lncRNAs and their potential functions	[175]
LncRNAWiki	http://lncrna.big.ac.cn/index.php/Main_Page	Community Curated Database For LncRNA	[176]
lncRNome	http://genome.igib.res.in/lncRNome	A comprehensive knowledgebase of human long noncoding RNAs	[177]
PLncDB	http://chualab.rockefeller.edu/gbrowse2/homepage.html	Plant Long noncoding RNA Database	[178]
StarBase v2.0	http://starbase.sysu.edu.cn/panCancer.php	Decods Pan-Cancer and Interaction Networks of lncRNAs from TCGA 14 cancer types	[179]
ALDB	http://res.xaut.edu.cn/aldb/index.jsp	Comprehensive database with a focus on the domestic-animal lncRNAs	[180]

2.8 lncRNA functional prediction

Unlike the protein coding genes, it is difficult to predict the functionality of lncRNA based on their sequence motifs and secondary structures as lncRNAs are not conserved and does not have conserved sequence motifs. As lncRNAs are involved in regulation of cellular activities by interacting with other molecules, recent studies has concentrated their attention towards exploring the relation between lncRNAs and proteins, protein coding genes and miRNAs. In general, there are three approaches that has been deployed for the computational functional prediction of lncRNAs such as Comparative genomic approach, co-expression with coding genes approach and interaction with miRNAs and proteins approach. [181][14]

3.8.1 Comparative genomic approach

Though lncRNAs are not conserved across different species, Amit et al. [182] identified 78 lncRNAs that are conserved both in human and mouse and found 70 of them are located within or close to 1000nt distance from the coding genes that are also conserved in both human and mouse. This implies that lncRNAs that are in a close proximity with the coding genes might have functional relationship. But this approach lacks efficiency because of poor conservation of lncRNAs and it cannot be applied at the genomic scale. [14]

2.8.2 Co-expression with coding gene approach

This approach is based on the assumption that the lncRNA that is regulating certain biological process might be co-expressed with the protein coding genes that are involved in same biological process. [14] Hao et al. [184] utilized this assumption to predict functionality of lncRNAs in carcinogenesis of esophageal taking samples from four patients with Primary ESCC tumors and adjacent non-neoplastic tissues conducted the differential expression analysis followed by the co-expression analysis using limma & RedeR R packages respectively. Then, the experimental quantitative real time polymerase chain reaction with small interfering RNA-mediated knockdown and apoptosis & invasion assays are applied in vitro to predict the onco-lncRNAs.

Yun Xiao et al. [183] applied a bayesian network model to identify the co-expression relationship between the lncRNAs and protein coding genes. Using 58 prostate cancer samples, Yun Xiao et al. identified and constructed the expression profiles of both lncRNAs and protein coding genes. Then, the bayesian network method is applied to construct the regulatory network based on the relationship between the lncRNA and protein coding genes. Finally, each protein coding genes that are linked to the lncRNAs from the regulatory network is mapped to the protein-protein interaction network and subsequently the functionality of the lncRNAs is predicted.

Liao et al. [185] also applied the coding-noncoding (CNC) gene co-expression network in two ways to predict the function of the lncRNAs: the hub-based and the network-module-based approaches. In hub-based network model, the function of the lncRNA is predicted based on the functional enrichment of its neighboring genes while network-module-based approach utilizes the Markov cluster algorithm (MCL) to identify the co-expressed functional module in the CNC network. [185]

2.8.3 Protein & miRNA Interaction approach

It is speculated that lncRNA is involved in the regulatory network in synergy with miRNA and proteins. Some miRNA and proteins bind to lncRNA to carry out their activity in the cell. Therefore, identifying those miRNAs and proteins that interact with the lncRNA is expected to reveal the possible function of the lncRNAs. With this regard, Jeggari et al. [187] developed an algorithm called "miRcode" that can predict the possible miRNA binding site on lncRNAs based on their seed complementarity and evolutionary conservation. Jeggari et al. used the algorithm to build the genome-wide network of validated miRNA mediated interaction to reveal the previously unknown mediatory role of lncRNA and miRNA.

Bellucci et al. [188] on the other hand came up with a method called "catRAPID" that can correlate the lncRNAs with proteins based on calculating their interaction potential using physicochemical characteristics such as secondary structure, hydrogen bonding, van der Waals, and others.

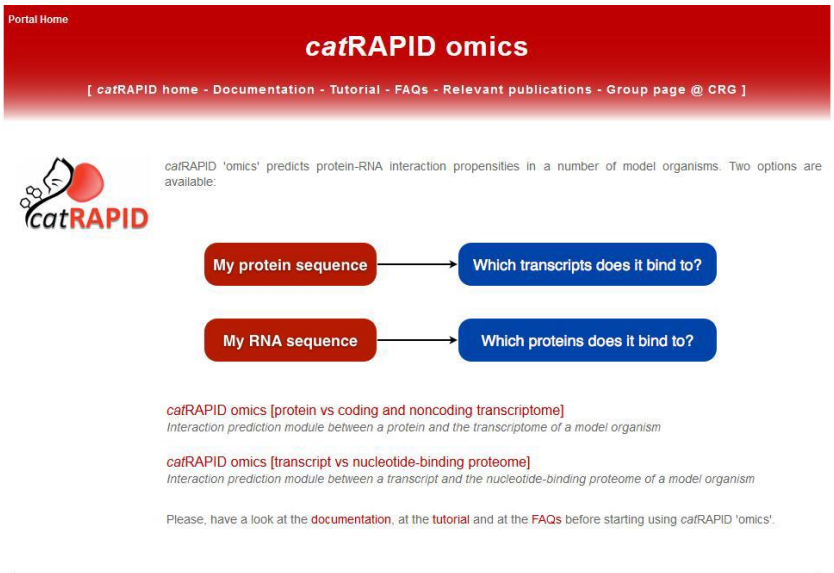


Figure 8 the snapshot of the catRAPID web service for protein-RNA interaction

The RPISeq [200][201] is another web based tool that uses only sequence information to train the support vector machine(SVM) and random forest(RF) machine learning algorithms.

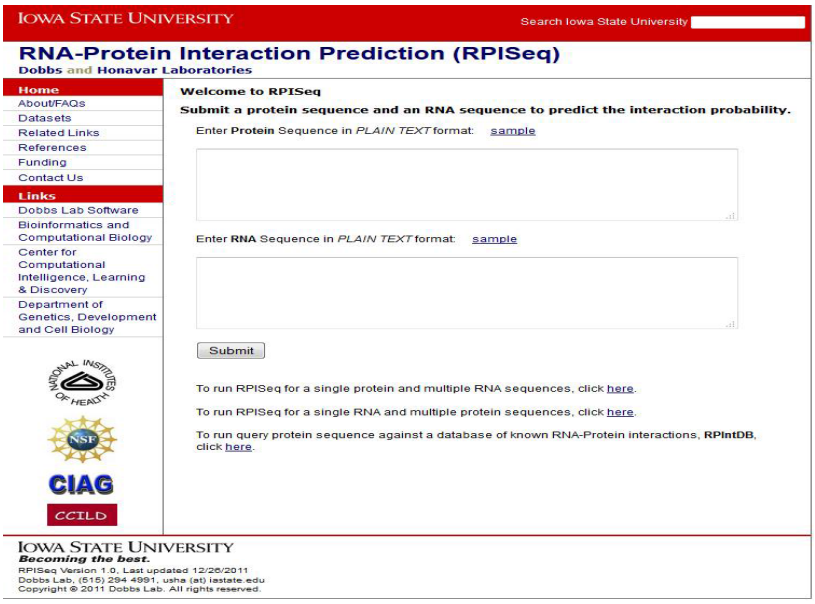


Figure 9. *The snapshot of the RPISeq web service for protein-RNA interaction*

However, applying this interaction-based approach in miRNAs and proteins is only limited to specific lncRNAs as mechanism of interaction for the lncRNAs with other bimolecular elements are still unclear. On the top of that, there are small number of lncRNAs whose functions are well studied making the computational functional prediction difficult in such a way that validation and optimization of computational algorithms are unattainable. [14]

2.9 RNA-seq data analysis

RNA sequencing technology is one of the high-throughput sequencing technique getting popularity in transcriptome profiling using deep sequencing approach. The overall work flow of RNA-seq protocol is shown in the figure below.

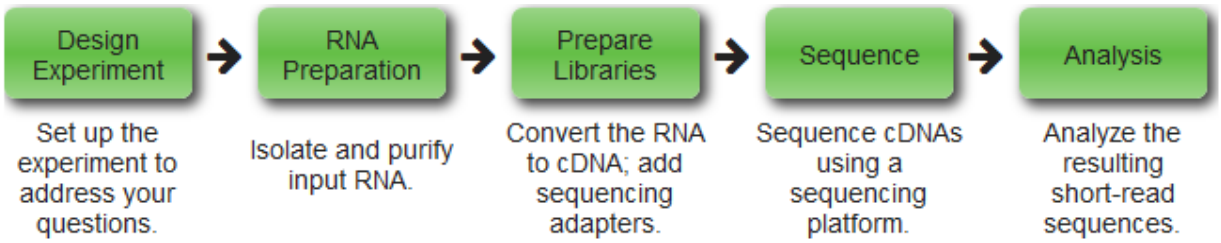


Figure 10. The work flow of RNA-seq sequencing [206]

The first step in RNA-seq work flow is to identify the research questions to be addressed by setting up experimental design such as identifying expressed transcript, identifying boundary exon/intron

junctions and transcription start site (TSS) or poly-A sites. The research question at hand can also be to identify the difference in expression among two or more sample groups. The next step in the work flow is RNA preparation by isolating and purifying the RNA samples. The isolated RNAs are converted to cDNA and then the sequencing adapter is attached to each of the cDNAs as shown in the figure 11. This step is known as a library preparation step and the cDNAs with sequencing adapter are referred as a libraries.

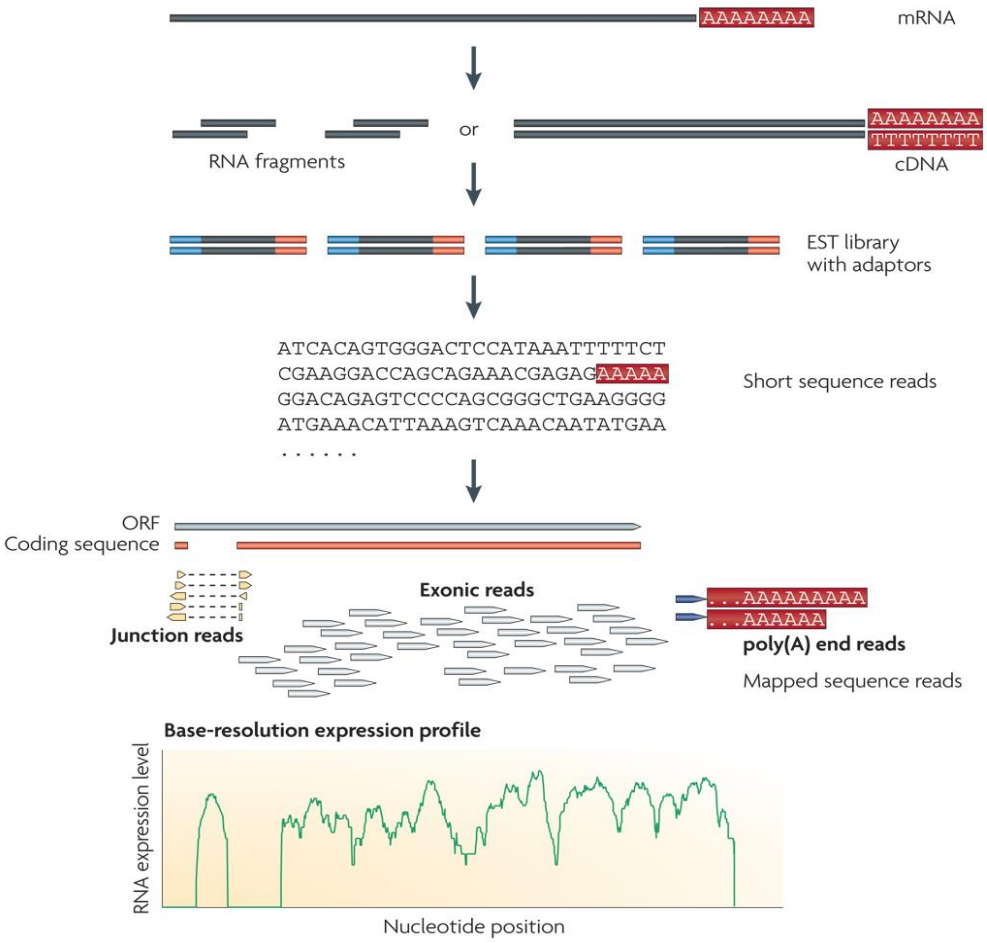


Figure 11. RNA-seq experiment

The next step in the RNA-seq work flow is to fragment the cDNA libraries and sequence it using sequencing platforms such as Illumina or SOLiD sequencer. The reads that are sequenced by the sequencer are of three types: exonic read, junction read and poly-A end reads. Finally, by aligning the reads with reference genome and quantify the expression of the transcript by counting the number of reads mapped to certain region in the genome, one can undergoing through different analysis technique to address the targeted research questions.

2.9.1 Normalization

Before directly proceeding with applying different statistical analysis methods to address the research question, the expression raw count data has to be comparable across features (genes and transcripts) and across the different libraries or samples. It also has to be in human friendly scale or magnitude. The process of preparing the row count data in such a way is known as normalization.

There are a number of techniques that are proposed to perform normalization for RNA-seq data such as total Count (TC), upper quartile (UQ), Median (Med), DESeq2, Trimmed Mean of M values (TMM), Quantile (Q) and Reads per Kilobase per Million mapped reads (RPKM) normalization. [187]

- **Total count (TC)**: Gene counts are divided by the total number of mapped reads (or library size) associated with their lane and multiplied by the mean total count across all the samples of the dataset.
- **Upper Quartile (UQ)**: Very similar in principle to TC, the total counts are replaced by the upper quartile of counts different from 0 in the computation of the normalization factors.
- **Median (Med)**: Also similar to TC, the total counts are replaced by the median counts different from 0 in the computation of the normalization factors.
- **DESeq**: This normalization method is included in the DESeq Bioconductor package (version 1.6.0) and is based on the hypothesis that most genes are not DE.
- **Trimmed Mean of *M*-values (TMM)**: This normalization method is implemented in the edgeR Bioconductor package (version 2.4.0). It is also based on the hypothesis that most genes are not DE.
- **Quantile (Q)**: First proposed in the context of microarray data, this normalization method consists in matching distributions of gene counts across lanes.
- **Reads Per Kilobase per Million mapped reads (RPKM)**: This approach was initially introduced to facilitate comparisons between genes within a sample and combines between- and within-sample normalization.

As different libraries are sequenced with different depth, DEesq2 uses statistical model to model the offsets and it makes sure that the parameters are comparable. In normalizing raw count data, DEseq2 defines virtual reference sample by taking the median of each gene value across sample and computes the size factor as the median of the ratio of each sample to reference sample. Thus, dividing each column of the count table with the corresponding size factor should yield the normalized count value that can be scaled for interpretation. [187]

2.9.2 Deferential expression analysis

The expression or abundance of transcript in a given target sample can be inferred by examining the probability of the randomly drown read from millions of reads in the library uniquely map to the target. This probability distribution can be statistically modeled by discrete distribution model, continuous distribution model or nonparametric distribution model. In differential expression analysis, having the library of two samples A & B, if the probability of randomly drown read from library A mapping to the target transcript is higher than that of the read that is randomly drown

from library B, then the target transcript is said to be differently expressed among the two library samples.

In discrete distribution model, randomly drown reads can be modeled in person's probability distribution. However, in RNA-seq data the variance of the probabilities among the individuals in the library is significantly higher than that of mean. As the person's distribution assumes equal variance and mean, modeling RNA-seq data with person's distribution will lead to "overdispersed" fit. This shortcoming of overdespersion in the person's distribution modeling can be improved by adding additional desperation parameter to adopt another model called negative binomial (NB) distribution model. There are several R packages that implement the negative binomial models analyzing the deferential expression such as edgeR and DESeq.

In the other way, normalizing and transforming RNA-seq read count values can be considered as a continuous distribution variable. If this continuous distribution is approximately normal, then it can be used to infer the differential expression among two groups using the continuous distribution models such as t-test.
The nonparametric model method in analyzing the differential expression is employed if the real data does not conform any specific assumptions. In this approach the rank based test statistics such as Mann-Whitney is calculated for analyzing the differential expression among two sample groups. Taking the log ratio among the top and the lowest quartile of sample expression can also be used in the differential analysis among unknown samples.

2.9.3 Gene list enrichment analysis

By performing gene list enrichment analysis, one is identifying whether those genes list of interest overlap in certain gene lists from the database that are known for certain biological pathways, ontologies and other biological information that are intended to be revealed more than expected by chance.

The team more than expected by chance signifies that the comparison between the two genes list have to have statistical certainty of enrichment. This can be achieved by utilizing different statistical approaches such as:

- ✓ Fisher's exact test
- ✓ Hypergeometric test
- ✓ Chi-squared test
- ✓ Z test
- ✓ Kolmogorov-Smirnov test
- ✓ Permutation test

In this particular project the gene lists of pathway from KEGG, the gene lists of the gene ontologies from GO database, gene lists that share motifs for miRNA and transcription factor proteins are

downloaded from MSigDB database and the Fisher's exact test is used to statistically test the enrichment of the genes list of interest with corresponding gene lists from the database.

Fisher's exact test is a statistical testing method that is applied on contingency table. Contingency table in statistics is a table that contains the multivariate frequency distributions of variables. In the case of enrichment analysis the contingency table looks table 6.

Table 6. The contingency table for gene list enrichment analysis

		The number of all genes in the gene lists biologically related genes from database		
		Yes	No	
The number of top DE genes in gene lists from database	Yes	a	b	a+b
	No	c	d	c+d
		a+c	b+d	N=a+b+c+d

Given the above contingency table for the datasets in the analysis, the null hypothesis for the above contingency table will be the number of overlapped genes from the top differently expressed gene with the one in the database are overlapped by chance. The p-value of significance is calculated from the above contingency table as:

$$P - value = \frac{(a+b)! + (b+d)! + (a+c)! + (c+d)!}{N!.\, a!.\, b!.\, c!.\, d!}$$

Therefore, with Fisher's exact statistical testing it can be determined whether the number of overlapped genes are significant to say the top differently expressed genes are enriched in certain biological gene ontology, pathways and other biological activities.

2.10 Co-expression analysis methods

The co-expression of two expression values from different data type can be analyzed using either Pearson's correlation or mutual information as a measure of association.

2.10.1 Correlation

Correlation is a measure of how well the two data are related. Pearson Product Moment Correlation (PPMC) shows the linear relationship between the two sets of data. The Pearson correlation coefficient(r), which ranges from -1 to 1, is a measure of similarity between the dataset. -1 means the two datasets are highly negatively correlated, +1 means the two datasets are highly positively correlated and 0 means there is no linear correlation between the two datasets. Mathematical expression for the Pearson's correlation coefficient is given by:

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right)}}$$

where:

r is the correlation coefficient
X and Y are datasets and
n is the number of data in the dataset

2.10.2 Mutual information

If the dataset are not linear, applying the Pearson's correlation might not lead to the ultimate solutions. In this case, applying the information theoretic based mutual information approach of a measure of association between the two datasets might be crucial. In information theory, information contained in variable can be quantified as "entropy". Given that random discrete variable X, the entropy is given by:

$$H(X) = -\Sigma p(x) log_b p(x)$$

where:

$H(X)$ is entropy of the discrete variable X,
$p(x)$ is the probability of the single discrete element from variable X

Depending on the use of logarithmic base, the unit of entropy varies. For example, if the base used in the logarithmic calculation is 2, the unit will be "bit". If base used is Euler's number e, the unit will be "nat", and "dit" for the base 10 calculations.

Pairwise mutual information is a measure of shared information between the two random variables. Given two random discrete variables X & Y their mutual information can be given as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

Where;

$I(X,Y)$ is a the pairwise mutual information of random variable X and Y,
$H(X)$ entropy of random variable X,
$H(Y)$ is entropy of random variable Y and
$H(X,Y)$is the joint entropy of the random variable X and Y

Joint entropy, $H\ (X,Y)$ is the entropy of joint probability distributions of random variables X and Y. Mathematically the joint entropy of two random variables X and Y is given by:

$$H(X,Y) = -\sum_{x\varepsilon X}^{n}\sum_{y\varepsilon Y}^{n} p(x,y)\log p(x,y)$$

Where:

$H(X,Y)$ is a joint entropy between two discrete variables X and Y

$p(x,y)$ is a joint probability distribution of dataset x and y

$p(x)$ and $p(y)$are factorized marginal distribution of dataset x and y

There are several algorithms that utilizes the mutual information as a measure of association between two dataset. For example, Algorithm for the Reconstruction of Accurate Cellular Networks (ARCENE), the context likelihood of relatedness (CLR), maximum relevance minimum redundancy network (MRNet) and network deconvolution algorithm use mutual information as a measure of association.

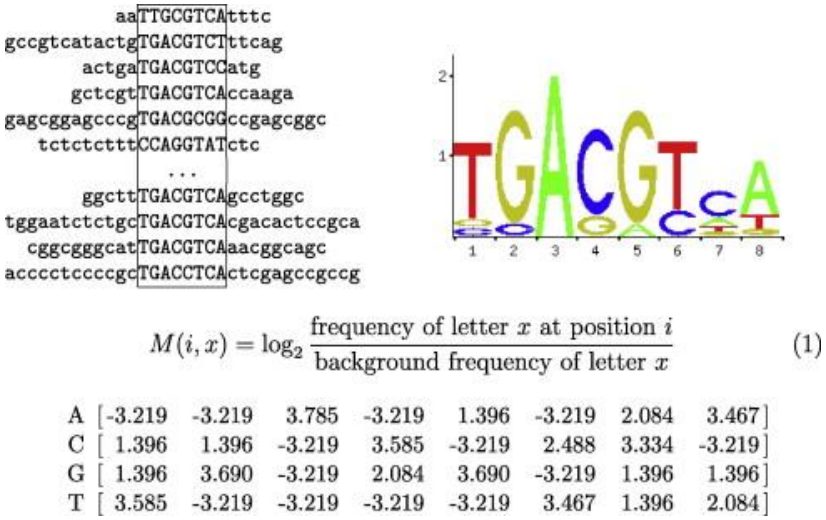
ARCENE algorithm is widely used in inferring the gene regulatory network and it is implemented in finding out the association between the genes, miRNAs and non-mRNA transcripts of this project to identify the association among them. There are three steps involved in association inference based on the ARACNE algorithm. The first one is pairwise mutual information calculation. The second step is based on significance threshold, it builds a graph of significant pairwise mutual information. Finally, it remove indirect connections from the network based on the violation of data processing inequality (DPI) principles.

According to the information theory, the data processing inequality (DPI) principles states that if there are three points A, B, C connected each where B is in the middle, then MI(A, B) >MI(A,C) & MI(B,C)>MI(A,C). Any connection that violates this principle, is indirect connection and removed from the network.

2.11 Sequence based analysis

3.11.1 Interaction Motif scanning

There are several approaches in figuring out the interactions between the DNA or RNA and proteins. These includes combinatorial approaches, using hamming distance, and probabilistic approaches such as position weight matrix (PWM) model and hidden markov model (HMM). PWM is widely used model in computational motif scanning.



$$M(i,x) = \log_2 \frac{\text{frequency of letter } x \text{ at position } i}{\text{background frequency of letter } x} \qquad (1)$$

A	[-3.219	-3.219	3.785	-3.219	1.396	-3.219	2.084	3.467]
C	[1.396	1.396	-3.219	3.585	-3.219	2.488	3.334	-3.219]
G	[1.396	3.690	-3.219	2.084	3.690	-3.219	1.396	1.396]
T	[3.585	-3.219	-3.219	-3.219	-3.219	3.467	1.396	2.084]

Figure 12. *From parallel alignment to PWM [188]*

The implementation of PWM algorithm begins with the parallel-aligned sets of binding sites with the length of m and the background distribution of q as it is shown in figure 12. The parallel-aligned sets of binding sites will be transformed into the position frequency matrix:

N_{ij}=									
	A	9	11	49	51	0	1	1	4
	C	19	3	0	0	0	45	25	16
	G	5	1	2	0	17	0	4	21
	T	18	36	0	0	34	5	21	10

The second step is to add the pseudocount to each of the value in the position frequency matrix. This substitutes the zero count position with comparably smaller number, as it will result in zero probability in calculating the position probability matrix.

The third step in calculating the PWM is to transform the position frequency matrix from step two to the position probability matrix. The position probability matrix is calculated by normalizing the individual frequency value from the position frequency matrix by the sum of column frequency value at particular position and the product of scaling parameter β and the background distribution q. mathematically, the position probability matrix is given by:

$$f_{ij} = \frac{N_{ij}}{\sum_{i=0}^{n} N_{ij} + \Sigma \beta q}$$

The forth and the final step in calculating the PWM is to change the position probability matrix into the weighted position matrix by taking the logarithmic 2 ratio between the signal and the background frequencies as it is illustrated in the figure 12. Once the PWM is constructed, there need to be a score for each motif scan against the certain DNA sequence S with sequence length of L. The PWM score of the sequence S given the PWM is given by:

$$PWM_{score}(S/PWM) = \sum_{i=0}^{n} PWM_{si}$$

Where, S is the DNA sequence to be scanned for binding motif

n is the length of the motif to be scanned

PWM is the Position weight matrix

$PWM_{score(S/PWM)}$ is the position weight matrix score for sequence S given PWM

3 OBJECTIVES

In order to implement different approaches towards the functional prediction of the long non coding RNAs, publicly available human glioblastoma multiforme (GBM) for 169 samples are used from the Cancer Genome Atlas (TCGA) glioblastoma project (Brennan et al. 2013). Based on the result from a novel gene and transcript identification algorithm, Novellette [204], which is RNA-seq analysis pipeline for gene and transcript identification, it has been identified that there exist about 53 novel lincRNA transcripts from 169 human glioblastoma multiforme (GBM) samples. Nothing is known about these novel lincRNA transcripts. Therefore, there is a need to predict the functions of those lincRNA computationally.

The goals of this project are:

- ✓ To analyze the differential gene expression across the 169 human GBM samples
- ✓ To analyze the differential novel transcript expressions across the GBM samples
- ✓ To analyze differential miRNA expressions across GBM samples
- ✓ To analyze the GO and pathway enrichment analysis with highly differently expressed genes
- ✓ To identify the co-expressing novel lincRNAs with the well-studied genes
- ✓ To identify the co-expressing novel lincRNA with known miRNAs
- ✓ To apply motif scanning methods to figure out the interactions between top differently expressed genes and proteins
- ✓ To apply motif scanning methods to figure out the interaction between the novel lincRNAs and proteins
- ✓ To Identify the common proteins binding to both the top differently expressed gene and novel lincRNAs
- ✓ To identify the RNA-RNA interaction between the top differently expressed miRNAs and novel lincRNAs
- ✓ To predict the function of the novel lincRNAs by integrating the expression and interaction based analysis.

4 MATERIALS AND METHODS

The functional prediction of the lncRNAs is carried out in combination with different approaches. The dataset used for this are the gene expression data, the novel transcript expression data and miRNA expression data for the glioblastoma multiforme from the TCGA samples. The primary stage of the analysis is identifying the gene of interest by applying quartile based differential gene expression analysis. Then, the gene list enrichment analysis for Gene ontology and pathway is performed to make sure that the genes of interest are related to glioblastoma.

The first part of the analysis is expression-based approach towards the computational functional prediction of lincRNA. In this approach, the genes that are co-expressed with the lincRNAs are identified to infer the function of the novel lincRNAs. Therefore, the co-expressed lincRNAs are in some ways related to that of the co-expressed genes.

The second part of the analysis is implemented based on the sequence information. This is used to identify the interaction between different proteins and the Gene/DNA. In the same way the interaction between RNA binding proteins and the lincRNAs are investigated to see how the lncRNA influence the gene regulation. Finally the potential miroRNA-lincRNA interaction is taken into consideration using a miRanda software which identifies the miRNA targets on lincRNAs based on the complementary sequence alignment score and minimum fold energy score.

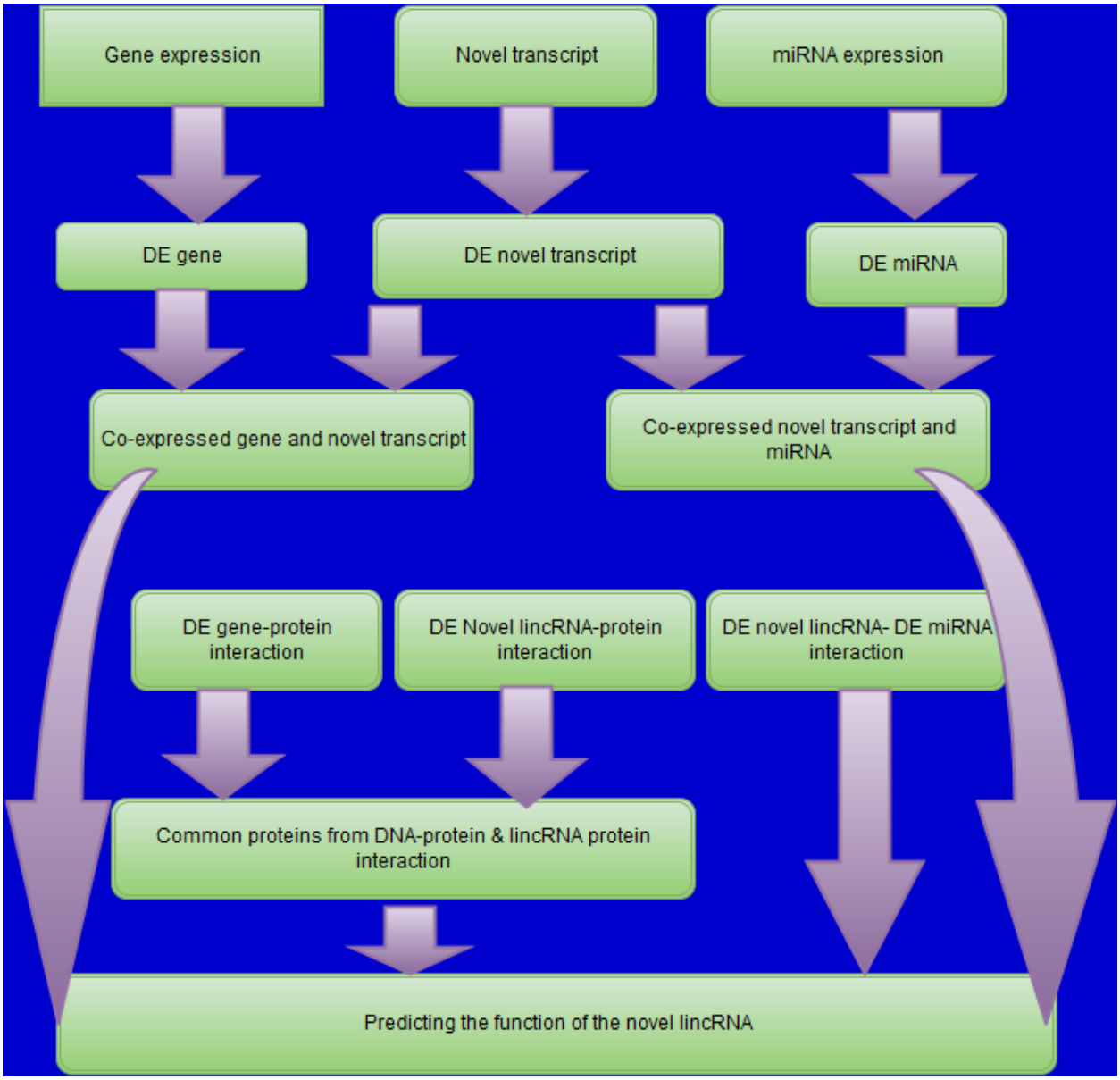


Figure 13. The analysis pipeline towards the functional prediction of novel lincRNAs

Bringing the result from the co-expression analysis, protein-DNA interactions, protein-lincRNA interactions and miRNA-lincRNA interactions together, one can predict the possible functional insight for the given novel lincRNA transcript. Figure 13 illustrates the analysis pipeline for the functional prediction of lincRNAs.

4.1 Expression based analysis

The TCGA gene expression count data for glioblastoma is used in [204] which contain the raw expression level of genes for each of 169 samples. [Appendix 1] 156 of the samples are from the primary solid tumor and the rest of 13 samples are from recursive solid tumor. The other dataset

used in this project also contains the raw count expression data for 243 novel transcripts from same samples. Out of these transcripts only 53 of them are identified as a long noncoding RNA (lncRNA) transcripts [Appendix 1]. The normalized expression data of 534 miRNA transcripts in 137 different glioblastoma multiforme samples which are among the samples we study for gene expression and transcript expression analysis are another dataset used in this project.

4.1.1 Normalization and pre-processing of raw count expression data

As the TCGA RNA-seq data for gene expression and the transcript expression are raw count data there is a need to perform normalization and preprocessing analysis. Therefore, the DESeq2 R package from Bioconductor is used to normalize the dataset. After normalizing the read count from the expression data, the next step is to make the normalized read count to be easier for interpretation. In this regards, we can scale the normalized read counts using a logarithmic transformation. The following figure shows that the expression values plot after DEseq2 normalization and transforming it into a logarithmic scale for the three dataset: gene expression dataset, transcript expression dataset and miRNA dataset.

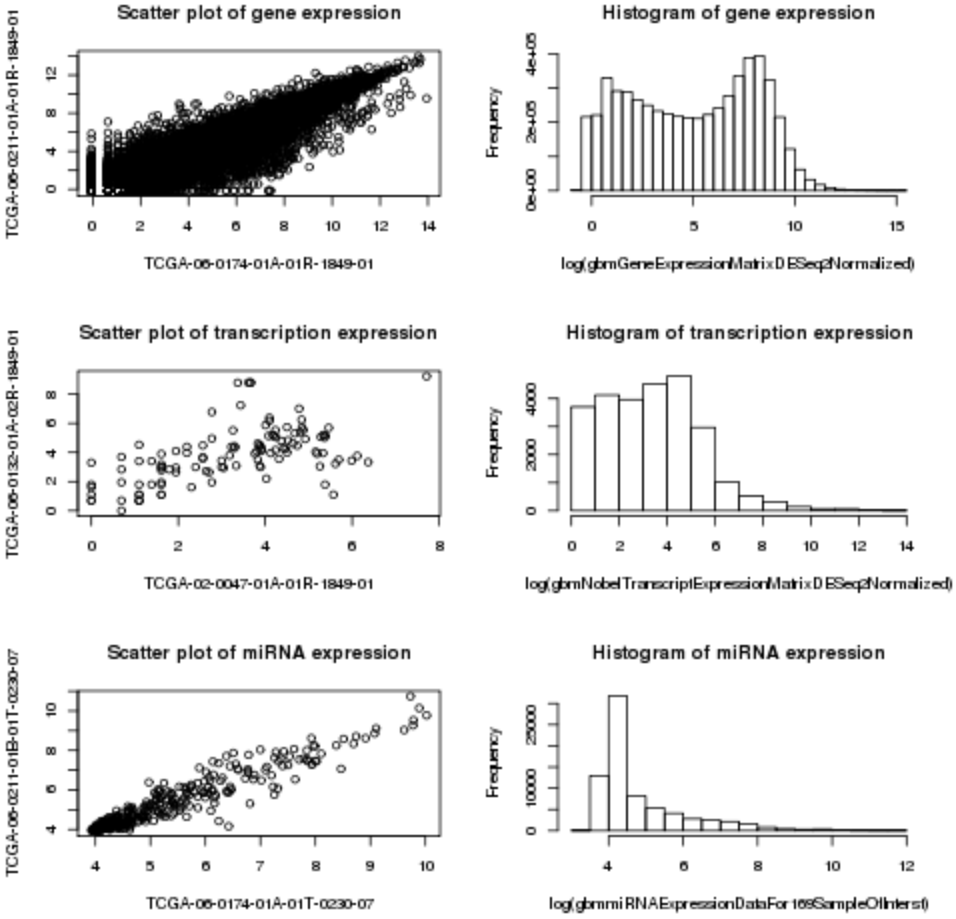


Figure 14. DEseq2 normalized and log transformed read counts for gene expression, novel transcript expression and miRNA expressions

As all the three datasets are of different types, one can expect that the expression levels are not absolutely comparable to each other rather there might be a slight difference in their average expression levels. The following box plot for each of the three dataset shows that the comparability of expression levels across the datasets.

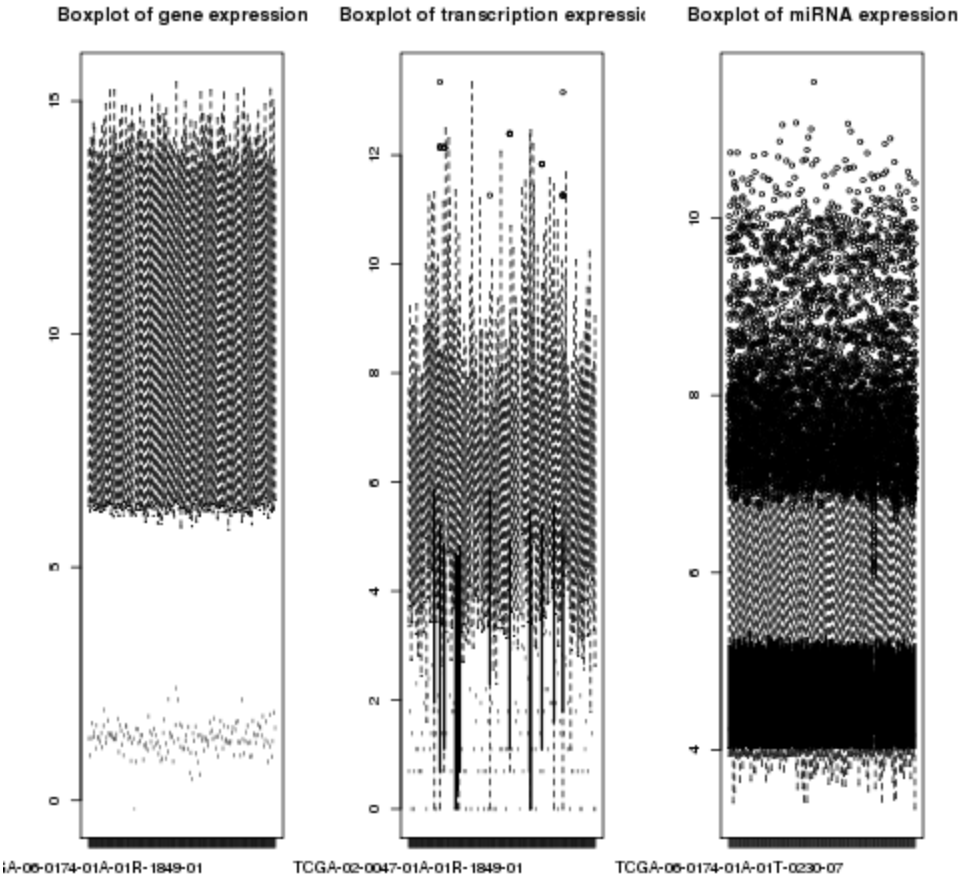


Figure 15. Box plots for the gene expression, transcription expression and miRNA expression

4.1.2 Gene differential expression analysis

The differential analysis of the gene among the given samples is calculated by quartile based nonparametric model method. The quartile based nonparametric method sorts the expression values of the gene among all the samples and compares the group of sample that are on the top 5% of high expression and the top 5% of low expression. Then taking the mean expression among the two sample groups for each gene, the differential expression is calculated as the logarithmic fold change ratio between the two mean values. Applying this approach for differential analysis will give a lot of flexibility to study what is going on the dataset irrespective of having information about the data samples. In this way, different genes might have different sample groups associated with their differential value of logarithmic fold changes.

Based on this analysis result, the first 61 genes with high logarithmic fold changes are selected as the interesting genes in their respective sample groups. The reason behind selecting the first top 61 genes is, as there are only 61 transcript that are differently expressed out of 243 novel transcripts

and the main research question for this project also lies on this dataset, having same number of interesting genes and miRNAs as the number of differently expressed transcript help in the subsequent co-expression analysis. Therefore, the following table shows that the first top 61 genes with their logarithmic fold changes.

Table 3. Top 61 differentially expressed genes with ensembl id, gene symbol, and fold change

Gene Symbol	*Log fold Change*	*Gene Symbol*	*Log fold Change*
"LTF"	11.3642074205527	"KCNQ2"	8.33818388296843
"COL6A3"	11.0190144223067	"ATP1A3"	8.30191789685553
"POSTN"	10.4981829710414	"FGFR3"	8.29225466676582
"PDGFRA"	9.84543985263434	"MXRA5"	8.26158154552933
"F13A1"	9.7374404284162	"OLIG2"	8.23721063576021
"IGF2"	9.55334306300648	"SEZ6"	8.19144975203853
"COL1A1"	9.50027627479324	"OLIG1"	8.18036915760204
"COL3A1"	9.32970966215948	"AVIL"	8.16003353885608
"RN7SK"	9.1158443808604	"RARRES2"	8.15928840573475
"CHI3L1"	8.98782027021159	"MOXD1"	8.14771737273237
"CXCL14"	8.91824581507234	"PCSK5"	8.11828585281088
"BCAS1"	8.90807793060724	"B4GALNT1"	8.11544039041408
"MBP"	8.87567709804349	"NNMT"	8.0896075541968
"NPTX2"	8.82043483049901	"EGFR"	8.07807112804236
"SMOC1"	8.77740068108231	"COL12A1"	8.07769146698688
"FMOD"	8.74501375002601	"HSPB8"	8.06737660966411
"LUM"	8.71006150778892	"COL14A1"	8.04081221144058
"SNAP25"	8.65970863179063	"HLA-DRB5"	7.99699423664328
"COL1A2"	8.57867730526177	"SEZ6L"	7.97690149381729
"COL5A1"	8.5659300445678	"SPOCD1"	7.93288568306023
"LRRN2"	8.51709906203847	"CHI3L2"	7.93024694934674
"COL6A2"	8.47002556846174	"IGFBP3"	7.91537975165864
"MEG3"	8.41778172021755	"BCAN"	7.91222139951645
"NCAN"	8.41754899188062	"PTX3"	7.89781989141177
"ADGRB1"	8.41368640326492	"GABBR2"	7.88484517360006
"THBS1"	8.41004327779113	"EPHB1"	7.8811984482472
"COL7A1"	7.83011425800144	"MFAP2"	7.85222460081362
"ISLR"	7.77533836641366	"ANGPTL4"	7.84611117485458
"PTPRN"	7.76980198170599	"FBXL16"	7.8329238759881

4.1.3 Transcript differential expression analysis

The same method used in the differential analysis of genes is applied to the differential expression analysis of novel transcript samples. Top 61 novel transcripts with the highest fold change are considered to be list of interesting novel transcript in their corresponding sample groups. The following table illustrates the lists of interesting novel transcript.

Table 4. List of top 61 novel transcripts with their logarithmic fold change

Novel transcript Id	log Fold Change Value	Novel transcript Id	log Fold Change Value
"TCGA_gbm-21-32283251"	15.0088341771392	"TCGA_gbm-X-3135001"	10.1719586190771
"TCGA_gbm-17-51805501"	14.9834167340625	"TCGA_gbm-X-87976751"	10.0014081943928
"TCGA_gbm-11-24180751"	14.8088531773541	"TCGA_gbm-1-5787751"	9.37565333409204
"TCGA_gbm-17-8682251"	14.6205643198004	"TCGA_gbm-3-153501"	8.87892673308683
"TCGA_gbm-9-23156751"	12.5927959421181	"TCGA_gbm-1-68830501"	8.75627831981759
"TCGA_gbm-2-123619751"	12.5511991627051	"TCGA_gbm-13-104880751"	8.20013231456476
"TCGA_gbm-3-163945501"	12.5339632555393	"TCGA_gbm-18-64289001"	7.86781411603516
"TCGA_gbm-12-69383001"	12.5212581678299	"TCGA_gbm-5-63687001"	7.28355142317431
"TCGA_gbm-X-151738001"	12.4738008608724	"TCGA_gbm-21-27587751"	7.01648058864122
"TCGA_gbm-Y-7085501"	12.4438824448183	"TCGA_gbm-17-10671251"	6.24253899444822
"TCGA_gbm-X-144413251"	12.367483007346	"TCGA_gbm-9-42103251"	6.15987133677839
"TCGA_gbm-12-38495001"	12.3657414672672	"TCGA_gbm-21-46975501"	5.96213496979665
"TCGA_gbm-16-64693251"	12.2999949885017	"TCGA_gbm-21-48002751"	5.77942968642626
"TCGA_gbm-7-154817501"	12.2790623894994	"TCGA_gbm-19-46930501"	5.44720132985189
"TCGA_gbm-2-153081251"	12.2768513869177	"TCGA_gbm-1-153561751"	5.43596189543763
"TCGA_gbm-4-130546751"	12.2618011071953	"TCGA_gbm-19-46927751"	5.38760426915573
"TCGA_gbm-10-109839251"	12.2594122648908	"TCGA_gbm-17-30454751"	5.38052274879821
"TCGA_gbm-7-54753751"	12.2241524496637	"TCGA_gbm-22-29576001"	5.2871767517316
"TCGA_gbm-8-37946001"	11.6130987942092	"TCGA_gbm-9-42202751"	5.08834657484054
"TCGA_gbm-3-28093501"	11.5324776871188	"TCGA_gbm-13-110076001"	4.91096771489633
"TCGA_gbm-4-109412001"	11.475670109043	"TCGA_gbm-2-112797251"	4.86052992695827
"TCGA_gbm-15-79922001"	11.2615807643097	"TCGA_gbm-9-70597501"	4.85435840680639
"TCGA_gbm-5-18266001"	11.2330204264745	"TCGA_gbm-X-53205501"	4.65105169117893
"TCGA_gbm-1-5525501"	10.9469062744564	"TCGA_gbm-9-70631251"	4.63461597035594
"TCGA_gbm-8-114503501"	10.9155993254444	"TCGA_gbm-17-21730501"	4.57041795292621
"TCGA_gbm-2-104066001"	10.8868397058844	"TCGA_gbm-9-42236501"	4.56074492041121
"TCGA_gbm-X-110058501"	10.7682877252458	"TCGA_gbm-9-68284501"	4.51945638970952
"TCGA_gbm-2-196150751"	10.5642685979355	"TCGA_gbm-13-50529001"	4.41069595209362
"TCGA_gbm-2-104096501"	10.5191447879974	"TCGA_gbm-6-28389001"	3.82044308746537
"TCGA_gbm-X-141374501"	10.462630074789	"TCGA_gbm-8-90598251"	3.06579303170701

```
                                        "TCGA_gbm-1-120693251"      0
```

4.1.4 miRNA differential expression Analysis

In the same method as that of differential expression analysis with genes and transcript, the first top 61 highly expressed miRNAs with the highest logarithmic fold changes are illustrated with the following table.

Table 5. List of top 61 differently expressed miRNAs with their logarithmic fold change

miRNA transcripts	Log fold Change	miRNA transcripts	Log fold Change
hsa-miR-370	5.747241	hsa-miR-181a	3.154147
hsa-miR-219	5.709165	hsa-miR-143	3.132416
hsa-miR-9	5.249204	hsa-let-7e	3.123281
hsa-miR-222	5.143633	hsa-let-7i	3.098401
hsa-miR-9*	5.024168	hsa-miR-130b	3.095383
hsa-miR-338	4.664476	hsa-miR-22	3.079704
hsa-miR-21	4.513117	hsa-miR-145	3.069379
hsa-miR-451	4.460724	hsa-miR-100	3.056662
hsa-miR-26a	4.320483	hsa-miR-320	2.989183
hsa-miR-638	4.219905	hsa-miR-29b	2.976133
hsa-miR-801	4.181721	hsa-miR-487b	2.962234
hsa-miR-34a	4.168021	hsa-miR-15b	2.952667
hsa-miR-210	4.111611	hsa-miR-106a	2.936700
hsa-miR-494	4.041169	hsa-miR-99a	2.926939
hsa-miR-575	3.892887	hsa-miR-26b	2.925090
hsa-miR-223	3.755507	hsa-miR-181c	2.918160
hsa-miR-663	3.728135	hsa-miR-23b	2.893517
hsa-miR-30a-5p	3.694828	hsa-miR-181d	2.828504
hsa-miR-155	3.557142	hsa-miR-374	2.822835
hsa-miR-149	3.539161	hsa-let-7g	2.813461
hsa-miR-17-3p	3.510356	hsa-miR-20b	2.805066
hsa-miR-126	3.424362	hsa-miR-768-5p	2.784178
hsa-miR-195	3.377252	hsa-miR-19b	2.782292
hsa-miR-20a	3.374421	hsa-miR-92b	2.775912
hsa-miR-19a	3.352281	hsa-miR-130a	2.768690

hsa-miR-301	3.334236	hsa-let-7a	2.739558
hsa-miR-574	3.288644	hsa-miR-181b	2.724688
hsa-let-7f	3.249088	hsa-let-7d	2.723302
hsa-miR-142-3p	3.235623	hsa-miR-99b	2.690628
hsa-miR-424	3.229504	hsa-miR-17-5p	2.689587
		hsa-miR-768-3p	2.675927

4.1.5 Gene list enrichment analysis

At this point the gene lists of interesting genes are identified according to their log fold change values. In this case, only the first top 61 genes with higher log fold change values are identified to be interesting genes. From the real biological perspectives, certain set of genes are involved in certain biological processes, biological functions or cellular compartments. This gene lists that are involved in certain biological ontologies are identified and organized in certain databases.

There are different databases that contain the gene sets for different biological pathways, ontologies, genes found in same chromosome or genes sharing same motifs with proteins or miRNAs. For example, Molecular Signatures Database (MSigDB) [203], which contains a collection of annotated gene lists for gene ontology, pathways, genes found in same chromosome and many more that can be used to compare against the gene list of interest.

The Fisher's exact test p-value with threshold value of 0.05 is applied to find the significantly enriched KEGG pathways, gene ontologies, miRNA and transcription factor proteins for the top 61 differently expressed gene lists. Table 2 show the enrichment analysis result for the first top 61 differently expressed gene lists.

Table 7. Significant enrichment results with threshold enrichment p-value of 0.05

GO MF enrichment result	*GO BP Enrichment result*	*GO CC enrichment result*
TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY	RNA_METABOLIC_PROCESS	PROTEINACEOUS_EXTRACELLULAR_MATRIX
METABOTROPIC_GLUTAMATEGABA_B_LIKE_RECEPTOR_ACTIVITY	REGULATION_OF_CELLULAR_PH	EXTRACELLULAR_MATRIX_PART
	SYSTEM_DEVELOPMENT	
INSULIN_LIKE_GROWTH_FACTOR_RECEPTOR_BINDING	ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	EXTRACELLULAR_REGION
	NERVOUS_SYSTEM_DEVELOPMENT	
PROTEIN_TYROSINE_KINASE_ACTIVITY	REGULATION_OF_PROTEIN_AMINO_ACID_PHOSPHORYLATION	

RECEPTOR_ACTIVITY	SKELETAL_DEVELOPMENT	EXTRACELLULAR_REGION_PART
TRANSMEMBRANE_RECEPTOR_ACTIVITY	REGULATION_OF_PROTEIN_MODIFICATION_PROCESS	EXTRACELLULAR_MATRIX
TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_ACTIVITY	REGULATION_OF_PHOSPHORYLATION	COLLAGEN
	MULTICELLULAR_ORGANISMAL_DEVELOPMENT	
	CELL_CELL_SIGNALING	
	PEPTIDE_METABOLIC_PROCESS	
	SYNAPTIC_TRANSMISSION	
	EPIDERMIS_DEVELOPMENT	
	ANATOMICAL_STRUCTURE_DEVELOPMENT	
	POSITIVE_REGULATION_OF_EPITHELIAL_CELL_PROLIFERATION	
	POSITIVE_REGULATION_OF_CELL_MIGRATION	
	POSITIVE_REGULATION_OF_ANGIOGENESIS	
	NEGATIVE_REGULATION_OF_CATALYTIC_ACTIVITY	
	TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_SIGNALING_PATHWAY	
	ORGAN_DEVELOPMENT	
KEGG enrichment result	*miRNA enrichment result*	*TF enrichment result*
KEGG_P53_SIGNALING_PATHWAY	CTACCTC,LET-7A,LET-7B,LET-7C,LET-7D,LET-7E,LET-7F,MIR-98,LET-7G,LET-7I	V$MEF2_01 V$IK1_01
KEGG_FOCAL_ADHESION	CACTGTG,MIR-128A,MIR-128B	V$SRF_Q6
KEGG_ECM_RECEPTOR_INTERACTION	ACTACCT,MIR-196A,MIR-196B	V$NKX25_02
KEGG_TYPE_I_DIABETES_MELLITUS	GTGTGAG,MIR-342	V$NRSF_01
KEGG_GLIOMA	GTGACTT,MIR-224	V$HNF4_DR1_Q3
KEGG_MELANOMA	GAGCCAG,MIR-149	V$AP1_Q2_01
KEGG_BLADDER_CANCER		

V$MEF2_Q6_01

GGGYGTGNY_UNKNOWN
YATTNATC_UNKNOWN
YCATTAA_UNKNOWN
SCGGAAGY_V$ELK1_02
RCGCANGCGY_V$NRF1_Q6
GGGTGGRR_V$PAX4_03TAAW
WATAG_V$RSRFC4_Q2

4.1.6 Co-expression analysis

The co-expression analysis of two expression values of different data type, such as, gene expression and lincRNA expression values, gene expression and miRNA expression values or lincRNA and miRNA expression values, can be analyzed using either Pearson's correlation or one can apply a mutual information based approach to find the similarity between the two datasets.

5.1.6.1 Correlation based co-expression analysis

Implementing Pearson's correlation based measure of association in R on the normalized gene expression data and on the novel transcript data, any associations with greater that 0.5 Pearson's correlation coefficient (r) and less than -0.5 Pearson's correlation coefficient is assumed to be significant for the further analysis steps. Figure 16 shows distributions of the correlation coefficient values for the gene-novel transcripts, gene-miRNA and miRNA-novel transcripts together with the co-expressing pairs of transcripts.

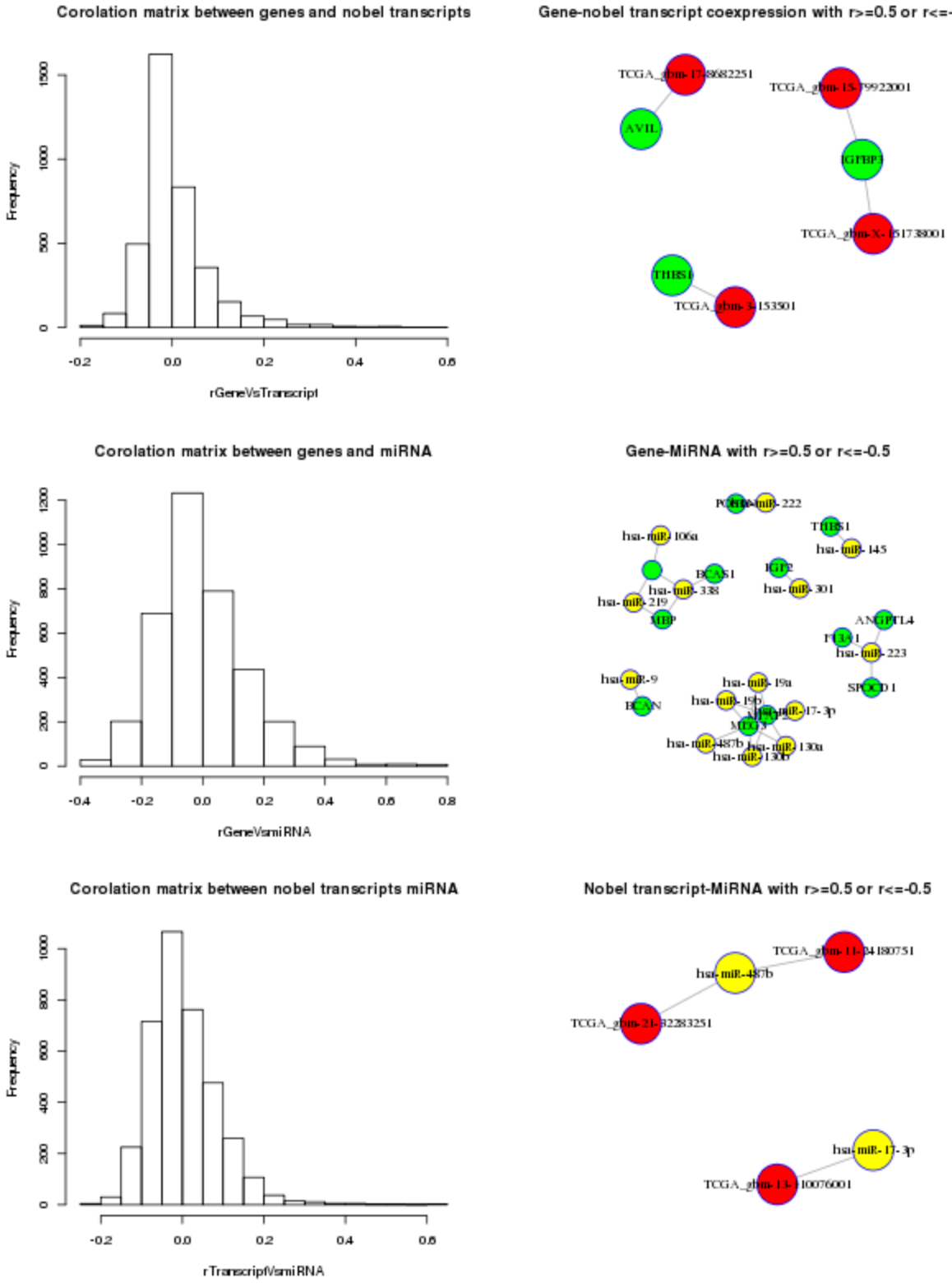


Figure 16. Correlation based a co-expression analysis

From the result shown in the figure 16, out of all genes, there are only three genes that are co-expressed or highly correlated with the threshold correlation coefficient of 0.5, either in a positive or negative way, with the novel transcripts. These genes are AVIL, THBS1 and IGFBP3. Only THBS1 gene is co-expressed with the novel lincRNA TCGA_gbm-3-153501 that is identified by Novellette algorithm.

4.1.6.2 Mutual information based co-expression analysis

Applying ARACNE algorithm in R to identify the co-expression of the genes with novel transcripts and miRNAs of interest, the result of the analysis is summarized in the figure 17 below. As the dataset in which the co-expression analysis is made are of different types, it is obvious that the mutual information threshold is small. For the gene-novel transcript co-expression analysis the mutual information threshold is 0.2. The threshold values of mutual information used for the gene-miRNA and miRNA-novel transcript are 0.4 and 0.26 respectively.

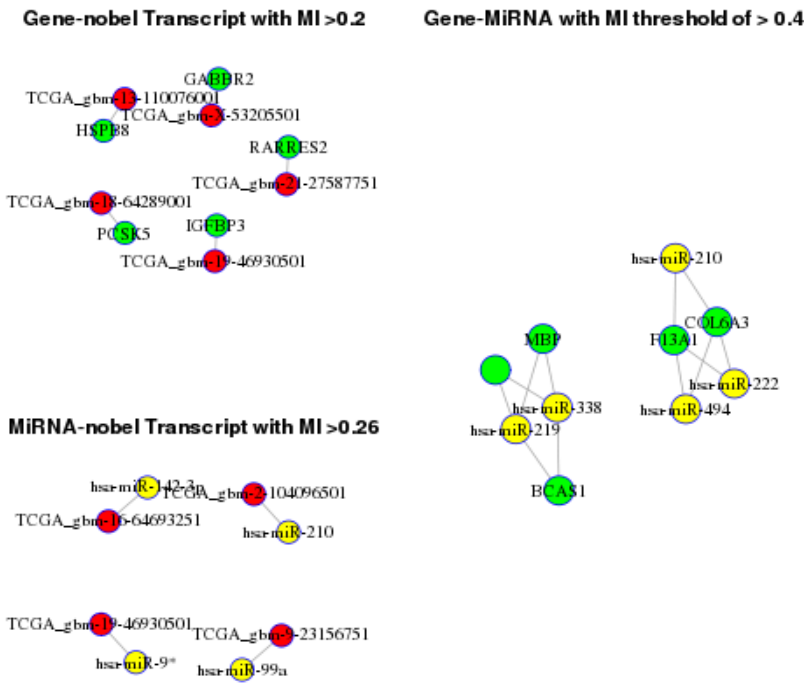


Figure 17. Mutual information based co-expression with ARSENE

Looking at the co-expression between the genes and novel transcripts with the cutoff threshold mutual information of 0.2, only five genes namely, *PCSK5, RARRES2, IGFBP3, GABBR2* and *HSPB8* are co-expressed with one of the novel transcript at hand as it is shown in the figure 17.

4.2 Sequence based analysis

4.2.1 Protein-DNA interaction

Moving from expression-based analysis to sequence based analysis, the interaction between the DNA sequence of the promoter regions of interesting genes and different type of proteins can be examined. In this regards, the DNA of the promoter regions of the interesting genes might interact with different transcription factor proteins, RNA binding proteins or other types of protein that might affect the function of the lincRNA directly or indirectly. Therefore, the investigation of the involvement of different proteins interaction with lincRNA will be the crucial part in revealing the possible functional mechanism of the lincRNA.

In PWM model, the position frequency matrices are constructed by aligning the sets of binding sites from different experiments given the background distribution. There are several databases with PFM or PWM for different proteins that interact with DNA based on the chip-seq experiments or literature mining. Some of the PWM databases are:

- **UNIPROBE** is one of the database with both experimental and literature mined PWMs
- **JASPAR** is the other one with 123 PFM matrices for different organism based on the SELEX method
- **TRANSFAC** is not open source database which has about 848 matrices for different organism.
- **hPDI(Human Protein-DNA Interactome)** with 17,718 preferable DNA binding sequences for 1013 human DNA-binding proteins

As hPDI database has abundant and experimentally validated PFM for human DNA-binding proteins, this database is used to scan specific protein binding motif. In this project, 437 experimentally verified human protein-DNA binding motifs are extracted from hDPI database. [189]. The 437 proteins are of different types:

- 52 mitochondrial protein
- 80 RNA binding protein
- 18 chromatin associated protein
- 5 kinase protein

- 14 Nucleic Acid binding proteins
- 55 other non-transcription factor proteins
- 212 Transcription factor proteins

A gene is a nucleotide sequence of specific region in the DNA. For protein coding genes, the structure of a gene is composed of four specific parts as shown in the figure below.

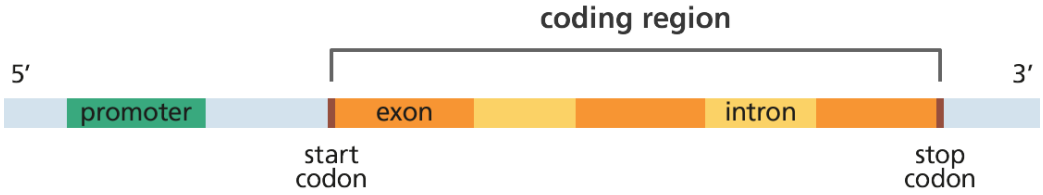


Figure 18. Protein coding gene structure

The promoter region at the 5' part of the gene structure is a place where the transcription factor proteins bind and send a signal for the RNA polymerase to start transcription right from the start codon. The start codon is a three nucleotide base sequence, in most case "AUG", in which the transcription begins. Stop codon at the 3' part of the gene in the other hand is also the three-nucleotide base sequences, in most case "TAG", where the RNA polymerase stops the transcription. The genomic region between the start and stop codon in which transcription takes place is known as coding region or open reading frame (ORF).

Scanning the promoter regions of the interesting genes from the previous analysis pipeline might reveal the proteins that play major role in the transcription of those specific genes. In addition, scanning the novel lincRNA sequences with proteins might unveil the function of the lincRNA in transcriptional interference.

In order to extract the genomic sequences of the highly differentially expressed genes, the BioMart package from bioconductor is used. The sequence of 2000 nucleotides upstream from the transcription start site or start codon are extracted as the promoter sequence. In the same way, for the retrieval of the lincRNA sequences, the BEDtools from samtools together with the bed file with exonic coordinate of the lincRNA's as identified in [204] is used to query it from the HG19 genome assembly, the same genome assembly that is used in [204].

Applying the PWM model in calculating the PWM score for each promoter region of the first top 61 highly expressed genes in both strand and calculating the significant PWM score will identify the proteins that are interacting with the promoter region of the genes. Thus, this protein-DNA interaction, when it is integrated with the result from lincRNA-protein interaction, might provide insight in the possible transcriptional interference functional mechanism of the lincRNA.

The PWM for 437 human protein-DNA interaction motifs are implemented in R using the PFMs from the hPDI database. Chunking the promoter sequences with the length of the binding motif to form a scanning window for the PWM, in all possible chunks or window the PWM score is calculated both in the forward and reverse strand of the promoter sequence.

Once the PWM score is calculated for each of the proteins binding motifs and the promoter regions of the genes of interest, the next step is to figure out which of the protein binding motifs or windows has significant PWM score. In this regards, the significant protein binding motifs can be identified by randomly scanning equal length of sequence as promoter sequence from any intergenic region of DNA with the protein binding motif and compare how the PWM scores of the randomly scanned and that of the promoter regions of the gene differ by calculating the p-value of significance.

The control nucleotide sequence is taken from the chromosome 4 starting from 54985200 to 54987200 nucleotide coordinate. The p-value of significance for the PWM score is calculated using the nonparametric Wilcoxon statistical testing method. The following table summarizes the significant proteins with binding site in both strands of promoter regions of the 5 highly differentially expressed genes. Parts of the table are omitted due to large number of rows.

Table 8. Proteins with the binding site in the top four highly differentially expressed genes

LTF gene		COL6A3 gene		POSTN gene		PDGFRA gene	
protein	p_value	Protein	p_value	Protein	p_value	Protein	p_value
AFF4	1.698688e-05	AGGF1	7.306652e-08	BAT4	6.732246e-10	TAF9	0.006032266
AKR1A1	1.357184e-04	AKR1A1	1.607913e-03	BOLL	2.568877e-05		
BAT4	1.374678e-05	ASCC1	8.955983e-08	C9orf156	2.279876e-07		
C9orf156	2.156080e-08	CAT	5.340410e-06	CDK2AP1	4.713121e-08		
CCDC25	6.816969e-03	CYCS	6.184013e-03	DIABLO	7.039418e-05		
CYCS	8.203411e-04	DIS3	7.269489e-09	FAM127B	7.422401e-03		

4.2.2 Protein-lincRNA interaction

The biological binding principles behind the proteins and the RNA molecules are explained in terms of the electrostatic interactions, hydrogen bonding, secondary & tertiary structures and the van der waal's interactions. The computational prediction of the RNA-protein interactions are to some extent takes the above principles into considerations. The computational approach to predict the RNA-protein interactions are carried out in two ways, either it predicts the RNA binding protein (RBP) motifs on the RNA molecules or it predicts the possible RNA binding motifs on the protein molecules.

In this project, the computational prediction of RNA-protein interaction in search of the RNA binding protein motifs on the novel lincRNAs is carried out based on only sequence information of the lincRNA molecules. If the structural aspect of the RNA-protein interaction is ignored, the methods to implement the RBP binding motifs on the RNAs are similar to that of the DNA-protein interaction.

As the DNA binding proteins there are different databases that hold a CLIP-seq experimental or literature mined RNA-protein interaction data that contains the PFM of different proteins that bind to the RNA molecules. Some of the database that provides the PFM of the interaction includes:

- ✓ **RNAcontext**
- ✓ **MEMERIS**
- ✓ **MatrixREDUCE**
- ✓ **RNAcompete**
- ✓ **RBPDB**

One thing that differentiate the PWM model applied in the protein binding motif scanning on DNA from RNA binding proteins scanning on RNA is that in the case of scanning for RNA binding proteins on RNA there is no reverse strand scanning as the RNA molecule is a single stranded molecule. In addition to that, the PFM contains uracil (U) base pair instead of thymine (T) base pair.

The PFM for 53 experimentally verified RBPs from the RBPDB database and 102 PFMs for RBPs from RNAcompete database is used in the RBP site scanning on the 17 novel lincRNA sequences. The lincRNA sequences are extracted from the HG19 genome assembly based on the genomic coordinate identified by [204]. Out of the top 61 differentially expressed novel transcripts, only 27 of them are found to be the novel lincRNA transcript.

"TCGA_gbm--2--104066001"	"TCGA_gbm--2--104096501"	"TCGA_gbm--3--153501"
"TCGA_gbm--1--68830501"	"TCGA_gbm--13--104880751"	"TCGA_gbm--18--64289001"
"TCGA_gbm--5--63687001"	"TCGA_gbm--21--27587751"	"TCGA_gbm--17--10671251"
"TCGA_gbm--21--46975501"	"TCGA_gbm--21--48002751"	"TCGA_gbm--1--153561751"
"TCGA_gbm--19--46927751"	"TCGA_gbm--17--30454751"	"TCGA_gbm--22--29576001"
"TCGA_gbm--9--42202751"	"TCGA_gbm--13--110076001"	"TCGA_gbm--2--112797251"
"TCGA_gbm--9--70597501"	"TCGA_gbm--X--53205501"	"TCGA_gbm--9--70631251"
"TCGA_gbm--17--21730501"	"TCGA_gbm--9--42236501"	"TCGA_gbm--9--68284501"
"TCGA_gbm--13--50529001"	"TCGA_gbm--8--90598251"	"TCGA_gbm--1--120693251"

Out of the above 27 identified novel lincRNAs only 17 of them are with enough information on the transcription direction, start and end coordinate of their exons. These 17 novel transcripts with enough detail information are:

"TCGA_gbm--2--104096501" "TCGA_gbm--3--153501" "TCGA_gbm--5--63687001"
"TCGA_gbm-21--27587751" "TCGA_gbm--17--10671251" "TCGA_gbm--1--153561751"
"TCGA_gbm--17--30454751" "TCGA_gbm--8--90598251" "TCGA_gbm--22--29576001"
"TCGA_gbm--2--112797251" "TCGA_gbm--9--70597501" "TCGA_gbm--9--70631251"
"TCGA_gbm--17--21730501" "TCGA_gbm--9-42236501" "TCGA_gbm--1--120693251"
"TCGA_gbm--9--68284501" "TCGA_gbm--13--50529001"

The genomic locations of the above novel lincRNAs are given in the following bed file format table.

Table 9. Genomic location of the novel lincRNAs

novel lincRNA name	Chr	Start	End	Exons	Strand
TCGA_gbm-1-120693251	chr1	120693368	120697115	TCGA_gbm-1-120693251Ex1	+
TCGA_gbm-1-153561751	chr1	153557550	153562200	TCGA_gbm-1-153561751Ex1	-
TCGA_gbm-2-104096501	chr2	104066248	104066478	TCGA_gbm-2-104096501Ex1	-
	chr2	104066930	104067073	TCGA_gbm-2-104096501Ex2	-
	chr2	104096754	104097000	TCGA_gbm-2-104096501Ex3	-
TCGA_gbm-2-112797251	chr2	112796951	112798143	TCGA_gbm-2-112797251Ex1	+
TCGA_gbm-3-153501	chr3	153750	154250	TCGA_gbm-3-153501Ex1	-
TCGA_gbm-5-63687001	chr5	63682339	63688666	TCGA_gbm-5-63687001Ex1	+
TCGA_gbm-8-90598251	chr8	90598127	90600261	TCGA_gbm-8-90598251Ex1	-
TCGA_gbm-9-70597501	chr9	70596970	70600171	TCGA_gbm-9-70597501Ex1	-
TCGA_gbm-9-70631251	chr9	70631128	70631944	TCGA_gbm-9-70631251Ex1	+
TCGA_gbm-9-42236501	chr9	42235789	42237190	TCGA_gbm-9-42236501Ex1	-
TCGA_gbm-9-68284501	chr9	68284810	68285084	TCGA_gbm-9-68284501Ex1	+
TCGA_gbm-13-50529001	chr13	50528852	50530202	TCGA_gbm-13-50529001Ex1	-
TCGA_gbm-17-30454751	chr17	30454391	30455334	TCGA_gbm-17-30454751Ex1	-
TCGA_gbm-17-10671251	chr17	10670126	10672237	TCGA_gbm-17-10671251Ex1	-
TCGA_gbm-17-21730501	chr17	21730767	21731530	TCGA_gbm-17-21730501Ex1	+
TCGA_gbm-21-27587751	chr21	27588024	27589704	TCGA_gbm-21-27587751Ex1	+
TCGA_gbm-22-29576001	chr22	29574536	29576616	TCGA_gbm-22-29576001Ex1	+

Once the sequence of each transcript is retrieved from the corresponding hg19 assembly using samtools, it will be converted into the RNA sequence. The Biostrings R package from the bioconductor is used to convert the DNA sequences into the RNA sequences.

In the same ways as the PWM score is calculated for the protein-DNA interactions, the PWM score is calculated for RBPs and the novel lincRNAs. Once having the PWM scores for each of the window in the lincRNA sequence, the need to calculate the significant motif hit is the next step. As it is done for the DNA-protein interactions, in this case the reference or control sequence is taken from any random exonic sequence from the genome with equal sequence length with the lincRNA. After the Control sequence is extracted from the genomic exonic region, it is then converted to RNA sequence using the Bioconductor's Biostrings package in R.

Then, having sequence length of the control sequence the same as that of each lincRNAs, the RBP motif scan is carried out by calculating the PWM score for each windows or motifs on the control sequence. Then the PWM score of both the lincRNA sequence and that of the control sequence are used to find out which RBPs have statistically significant binding site. Based on this approaches, the lincRNAs with the statistically significant binding hit for certain RBPs from the PWM model are fully illustrated in the appendix 1.

In order to show the RBPs that have significant binding site on the lincRNAs, only the significant RBPs that have significant binding site on the top 5 highly differentially expressed novel lincRNAs with the fold changes illustrated in the table below are selected due to large table columns.

Table 10. top 5 DE novel lincRNA

top 5 DE novel lincRNA	Log fold Change
TCGA_gbm-2-104096501	10.519145
TCGA_gbm-3-153501	8.878927
TCGA_gbm-5-63687001	7.283551
TCGA_gbm-17-10671251	6.242539
TCGA_gbm-21-27587751	7.016481

Table 11. Proteins with binding site on novel lincRNAs based on PFM from RBPDB database

TCGA_gbm-2-104096501		TCGA_gbm-3-153501		TCGA_gbm-17-10671251		TCGA_gbm-5-63687001	
RBP	p-value	RBP	p-value	RBP	p-value	RBP	p-value
A2BP1	1.483784e-04	A2BP1	7.994252e-04	A2BP1	2.041465e-09	PABPC1	6.497751e-21
KHDRBS3	4.143959e-12	KHDRBS3	4.128248e-07	KHDRBS3	2.437164e-13	A2BP1	6.583837e-03
QKI	5.165369e-03	QKI	8.723493e-04	QKI	1.350722e-05	KHDRBS3	2.670768e-17
ZFP36	2.752346e-04	SFRS2	6.411125e-05	SFRS2	9.231877e-09	ZFP36	5.998443e-17
SFRS2	1.213953e-08	SFRS1	1.541553e-03	SFRS1	7.744355e-07	SFRS1	4.458376e-80
SFRS1	3.413817e-08	EIF4B	1.251044e-03	EIF4B	8.207584e-04	EIF4B	2.182950e-27
EIF4B	4.044587e-05	IGF2BP1	2.399029e-03	IGF2BP1	2.053731e-03	IGF2BP1	1.939343e-05
IGF2BP1	6.615535e-03	KHDRBS3	1.651886e-06	KHDRBS3	1.260073e-10	KHDRBS3	1.202122e-06
KHDRBS3	4.083160e-11	NONO	3.719077e-11	NONO	2.080393e-16	QKI	2.148875e-08
NONO	2.014912e-22	HNRNPA1	4.037018e-06	HNRNPA1	5.729528e-10	NONO	3.498104e-43
HNRNPA	1.658171e-11	YBX1	9.888623e-16	YBX1	2.500646e-25	HNRNPA1	2.131723e-33
YBX1	1.550585e-20	KHDRBS3	3.539882e-08	KHDRBS3	1.584971e-05	YBX1	2.545392e-58
KHDRBS3	1.672579e-13	SNRPA	4.911790e-03	SNRPA	9.113207e-03	SNRPA	2.002908e-04

RBM4	5.497882e-06	MBNL1	2.548278e-06	MBNL1	1.254569e-07	KHDRBS3	5.545314e-16
SNRPA	9.562542e-03	NOVA2	1.450842e-04	NOVA2	2.555746e-06	SFRS1	1.592015e-06
MBNL1	3.933100e-17	ELAVL2	3.028037e-03	ELAVL2	6.309175e-04	RBM4	4.632046e-75
NOVA2	7.255056e-05	RBMX	8.545470e-08	SFRS2	2.575048e-04	PTBP1	7.694061e-03
ELAVL2	3.393680e-05	YTHDC1	2.082768e-14	SFRS7	1.620752e-05	ELAVL1	9.026422e-04
SFRS2	2.334640e-03	RBMY1A1	1.024184e-04	RBMX	2.361421e-12	SFRS13A	2.145211e-22
SFRS7	1.252532e-06			YTHDC1	7.869544e-16	PABPC1	3.329851e-05
RBMX	8.883769e-09			NCL	6.003961e-03	SFRS1	2.902830e-15
YTHDC1	7.060927e-30			RBMY1A1	2.492161e-08	SFRS7	1.079706e-03
RBMY1A1	4.978744e-08					SFRS2	1.135660e-10
ZRANB2	8.907087e-04					ZFP36	1.946519e-20
						SNRPA	5.603732e-07
						FUS	2.038293e-13
						MBNL1	1.405556e-91
						ELAVL2	1.310357e-06
						SFRS2	2.967565e-20
						SFRS7	3.758398e-43
						SFRS9	6.014165e-07
						RBMX	8.173692e-14
						YTHDC1	3.241395e-151
						NCL	2.742761e-04
						RBMY1A1	6.750634e-14
						RBMY1A1	1.562333e-13
						ZRANB2	7.263636e-55

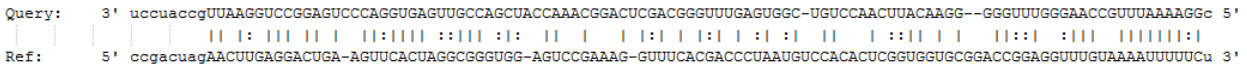
4.2.3 miRNA-lincRNA interaction

RNA-RNA interactions are one of the mechanism by which ncRNAs achieve their diverse functions. One of the well-studied functions of miRNA on protein coding genes is destabilizing and repressing the translation of protein coding transcripts by binding at the 3' UTR regions of the mRNAs. However, recent studies are showing that miRNAs has influence on the function of the lncRNAs and lncRNA also in some way influence the function of the miRNA. The stability of the lncRNAs in some case will be degraded by the interaction with miRNAs. In the other hand, lncRNAs might act as the miRNA decoy and some of them are degraded to produce miRNAs. [190]

In this project, the interaction of novel lincRNAs with the miRNAs are investigated to examine their functional effect on gene expression and in turn reveal the molecular and functional mechanisms of the novel transcripts of interest. The computational approach to predict the interaction of lncRNAs and miRNAs is based on the identification of miRNA binding target site on the lncRNAs. miRanda is an algorithm developed by Memorial Sloan-Kettering Cancer Center, New York and distributed

under GNU Public License to scan the miRNA target on other RNA sequences. It utilizes the dynamic programming alignment together with the thermodynamics approach to find the possible miRNA binding site on the reference RNAs.

In miRanda, the potential miRNA target site is identified in two step strategies. In the first step the algorithm carries out the dynamic programming local alignment between miRNA sequences and the reference. In the first step the algorithm produces the alignment score based on the sequence complementarity, gap desired and gap penalty. The RNA sequence complementaries used in the alignment are A:U, G:C and G:U.



Once the target scanning passed the minimum threshold of alignment score from step one, the algorithm will proceed to the second step in which the thermodynamic stability of RNA duplexes based on these alignments are examined. The folding routine from RNAlib library that is the ViennaRNA package written by Ivo Hofacker is utilized to generate constrained fictional single-stranded RNA composed of the query sequence, a linker and the reference sequence. This structure is then folded using RNAlib and the minimum free energy (DG kcal/mol) is calculated. The final result is composed of the target sites with less than the minimum fold energy threshold value. The command used to scan for the miRNA target site on the lincRNA of interest using the miRanda software is given below:

miranda miRNASeqmiRbaseDB.fasta LincRNASeqHg19.fasta -sc 220 -en -110 -go -9 -ge -4 -out myresult.txt

where:

miRNASeqmiRbaseDB.fasta is the multiple miRNA sequences that are differently expressed in the glioma TCGA samples and

LincRNASeqHg19.fasta is the 17 lincRNAs that are identified by the seppala's work.

-sc 200 is the minimum alignment score threshold and -en -110 is the minimum fold energy threshold

-out myresult.txt option will write the output to the specified text file.

-go -9 is the gap open penalty

-ge -4 is the gap extend penalty

One of the output for the above command that satisfy the threshold value for both the maximum alignment score and minimum folding energy is shown below.

```
Read Sequence:TCGA_gbm-13-50529001 (1365 nt)
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
Performing Scan: hsa-mir-181c vs TCGA_gbm-13-50529001
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

   Forward: Score: 268.000000  Q:2 to 109  R:93 to 211 Align Len (120) (55.83%) (65.00%)

   Query:    3' ucCUACCGUUAAGGUCCGGAG-UCCCAGGUGAG------UUG--CCAGCUAC--CAAACGGACUC-GACGGGUUUG-AGUGGCUGUCCAACUUACAAGGGGGUUUGGGAACCGUUUAAAAGGc 5'
                  ||||| | |||||  || |  || |:| |||      |||  ||  | ||   ||| || |:| :|| :|||||  ||:  | ||| | ||   | |:::|||:::|| |   |||||||
   Ref:      5' guGAUGGGACUUCCAACCCACUUGGUUUCCCUCGGGGAAAACCAGGAGGUUGUAAUUUCCCGGGGUUUGGUCAAACGACAUAAAGAGGGUCAAACCU-CUUUAAAUUUUUCG---AUUUUCCg 3'

   Energy:  -111.150002 kCal/Mol

Scores for this hit:
>hsa-mir-181c   TCGA_gbm-13-50529001    268.00  -111.15 2 109   93 211   120 55.83%  65.00%

Score for this Scan:
    Seq1,        Seq2,                  Tot Score,  Tot Energy, Max Score,  Max Energy, Strand, Len1,   Len2,    Positions
>>hsa-mir-181c  TCGA_gbm-13-50529001   268.00      -111.15     268.00      -111.15      199     110     1365     93
Complete
```

The miRNAs and the novel lincRNAs that has passed the minimum alignment scores threshold and the minimum folding energy across the top 61 differently expressed miRNAs and the 17 novel lincRNAs in miRanda algorithm are illustrated in table 12.

Table 12. miRNA with binding site on novel lincRNAs by miRanda algorithm

miRNA seq	lincRNA seq	Alignmet Score	min Folding Energy	Strand	Len1	Len2	Positions
hsa-mir-638	TCGA_gbm-13-50529001	223.00	-114.34	658	100	1365	1073
hsa-mir-181d	TCGA_gbm-22-29576001	223.00	-122.98	578	137	2095	578
hsa-mir-181d	TCGA_gbm-21-27587751	221.00	-118.22	577	137	1695	748
hsa-mir-181c	TCGA_gbm-13-50529001	268.00	-111.15	199	110	1365	93
hsa-mir-181d	TCGA_gbm-17-10671251	252.00	-136.84	575	137	2126	852, 1294 1547,1139
hsa-mir-210	TCGA_gbm-17-30454751	222.00	-118.02	217	110	943	413
hsa-mir-210	TCGA_gbm-17-21730501	225.00	-117.15	219	110	778	74

Based on the result from miRanda algorithm, *hsa-mir-181d,* miRNA, is predicted to have an interaction with *TCGA_gbm-22-29576001*, *TCGA_gbm-21-27587751* and *TCGA_gbm-17-10671251* novel lincRNAs. hsa-mir-210 also have a binding site on TCGA_gbm-17-21730501 and TCGA_gbm-17-30454751 lincRNAs. hsa-mir-638 is expected to interact with *TCGA_gbm-13-50529001* and hsa-mir-181c is predicted to interact with *TCGA_gbm-13-50529001*.

5 RESULT

The functional prediction of the novel lincRNA transcripts that are identified by [204] is carried out by integrating the different results in the analysis pipeline such as gene list enrichment analysis, co-expression analysis, protein-DNA interaction, protein-lncRNA interaction and miRNA-lincRNA interaction.

The result from gene list enrichment analysis reveals that relation between the first 61 highly differentially expressed genes and the Gene ontologies and pathways associated with it. Based on this result the list of highly differently expressed genes have pathways and gene ontologies related to the glioblastoma.

As one of the mechanism to predict the functions of the lincRNA is examining the functionality of the co-expressed gene along with the lincRNA, the functionality of the gene that is co-expressed with the novel transcript are predicted to be the possible function of the newly identified novel lincRNA. The result from the protein-DNA and protein-lincRNA interaction assert the transcriptional interference, posttranscriptional regulation and the roles of the novel lincRNAs in mRNA splicing. As the interaction of the miRNA with lincRNA has a destabilizing effect on lincRNA and in return it affects the normal function of the miRNA, it is possible to predict that functions associated to the interacting miRNA might infer functional mechanism of the newly identified lincRNA. In addition, lincRNAs and miRNAs might be involved in competition for binding site on mRNA.

As it is illustrated in table 7 of the gene list enrichment analysis result, the first 61 highly differentially expressed gene lists are enriched among the gene ontology molecular functions of transmembrane receptor protein kinase activity, transmembrane receptor protein tyrosine kinase activity, transmembrane receptor activity and insulin like growth factor receptor binding.

The gene ontology biological processes in which the top 61 differently expressed genes involved are organ development, skeletal development, nervous system development, anatomical structure development, enzyme linked receptor protein signaling pathway, RNA metabolic process , regulation of cellular PH, regulation of protein modification process, regulation of phosphorylation, cell-cell signaling, synaptic transmission, epidermis development, positive regulation of the epithelial cell proliferation, positive regulation of cell migration, positive regulation of angiogenesis, negative regulation of catalytic activities and transmembrane receptor protein trypsin kinase signaling pathway.

The KEGG enrichment pathway analysis result that are shown in the table 7 indicates that the top 61 highly differentially expressed genes are involved in KEGG glioma pathway, KEGG ECM receptor interaction pathway, KEGG p53 signaling pathway, KEGG melanoma pathway, KEGG bladder cancer pathway, KEGG focal adhesion pathway and KEGG type 1 diabetes mellitus pathway.

As the co-expression analysis of the novel lincRNAs with the genes give insight in the possible functional mechanism of the newly discovered lincRNA, the correlation based co-expression analysis result with the threshold correlation coefficient of 0.5 from figure 19 shows that the THBS1 gene and TCGA_gbm--3--153501 novel lincRNA are co-expressed.

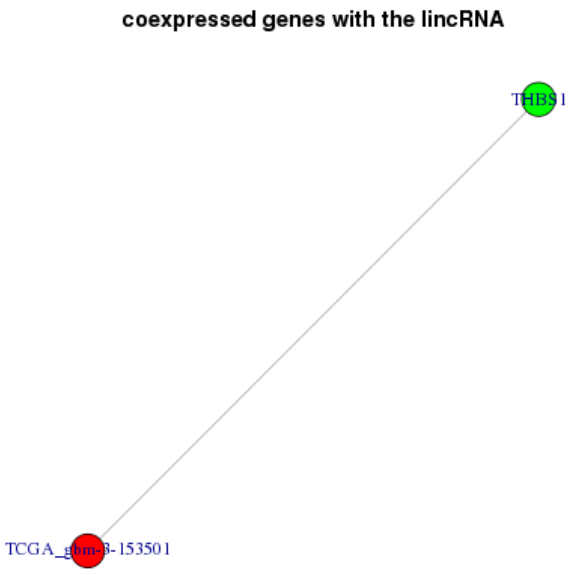


Figure 19. Co-expression network between the novel lincRNAs and genes with cutoff correlation coefficient of 0.5

Looking at the co-expression networks of genes with the miRNA in figure 20 with a threshold correlation coefficient of 0.5, THB1 gene is again co-expressed with the has-miR-145 microRNA. The IGF2 gene is co-expressed with the has-miR-301

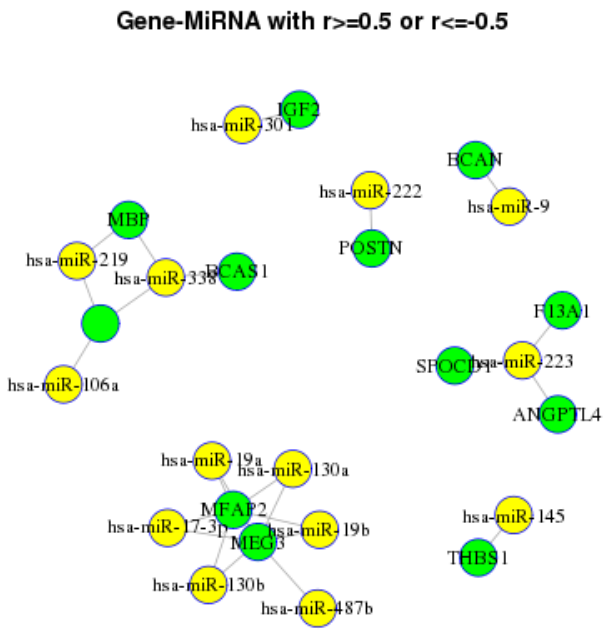


Figure 20. Gene-miRNA network with threshold correlation coefficient of 0.5

Out of these all miRNAs that are co-expressed with genes there are only four of them that are among top ten highest log fold change values. The sub-network of co-expression between the top ten highest log fold change miRNAs and genes are shown in the figure below.

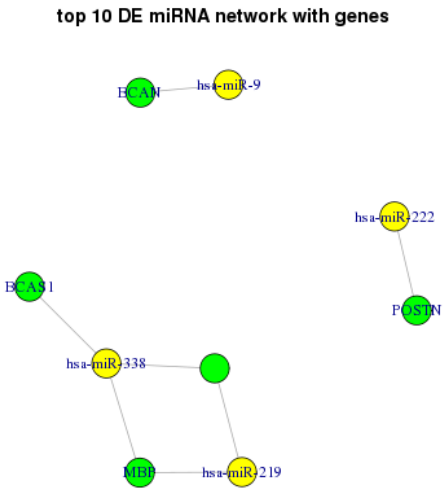


Figure 21. The sub network of the top 10 highest log fold change valued miRNA with genes

As it is show in the figure 21, hsa-miR-222 is co-expressed with POSTN, hsa-miR-219 with MBP and hsa-miR-338 is co-expressed with BCAS1 genes. Looking at the transcript and miRNA co-expression in figure 22, there are two miRNAs that are co-expressed with the transcripts with threshold correlation coefficient of 0.5.

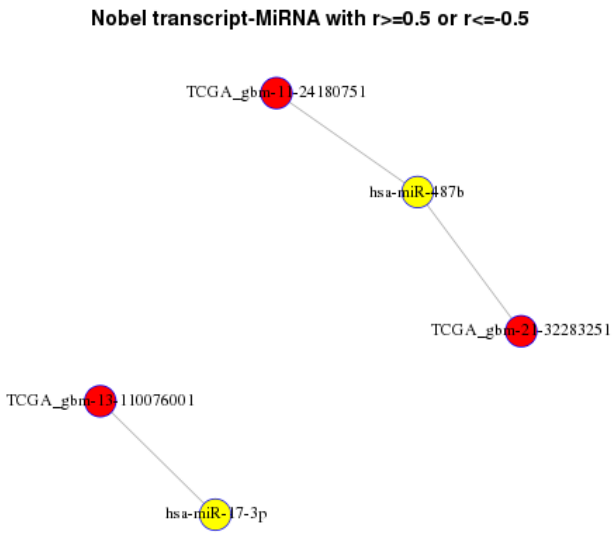


Figure 22. The network of co-expression between the novel transcripts and miRNAs with threshold correlation coefficient of 0.5

As it is shown in the figure 22 none of the novel lincRNA transcripts identified with Novellett algorithm are identified to be co-expressed with any of the miRNAs. The co-expression network between the miRNA and transcripts shows that hsa-miR-17-3p is co-expressed with TCGA_gbm-13-110076001 and hsa-miR-487b is co-expressed with both TCGA_gbm-21-32283251 and TCGA_gbm-11-24180751.

The mutual information based co-expression analysis with the minimum threshold mutual information of 0.2 for gene and novel transcript association is shown in figure 23. Out of the novel transcripts that are co-expressed with the genes, there is only one novel lincRNA transcript which is *TCGA_gbm-21-27587751* co-expressed with *RARRES2* gene as it is illustrated in figure 23.

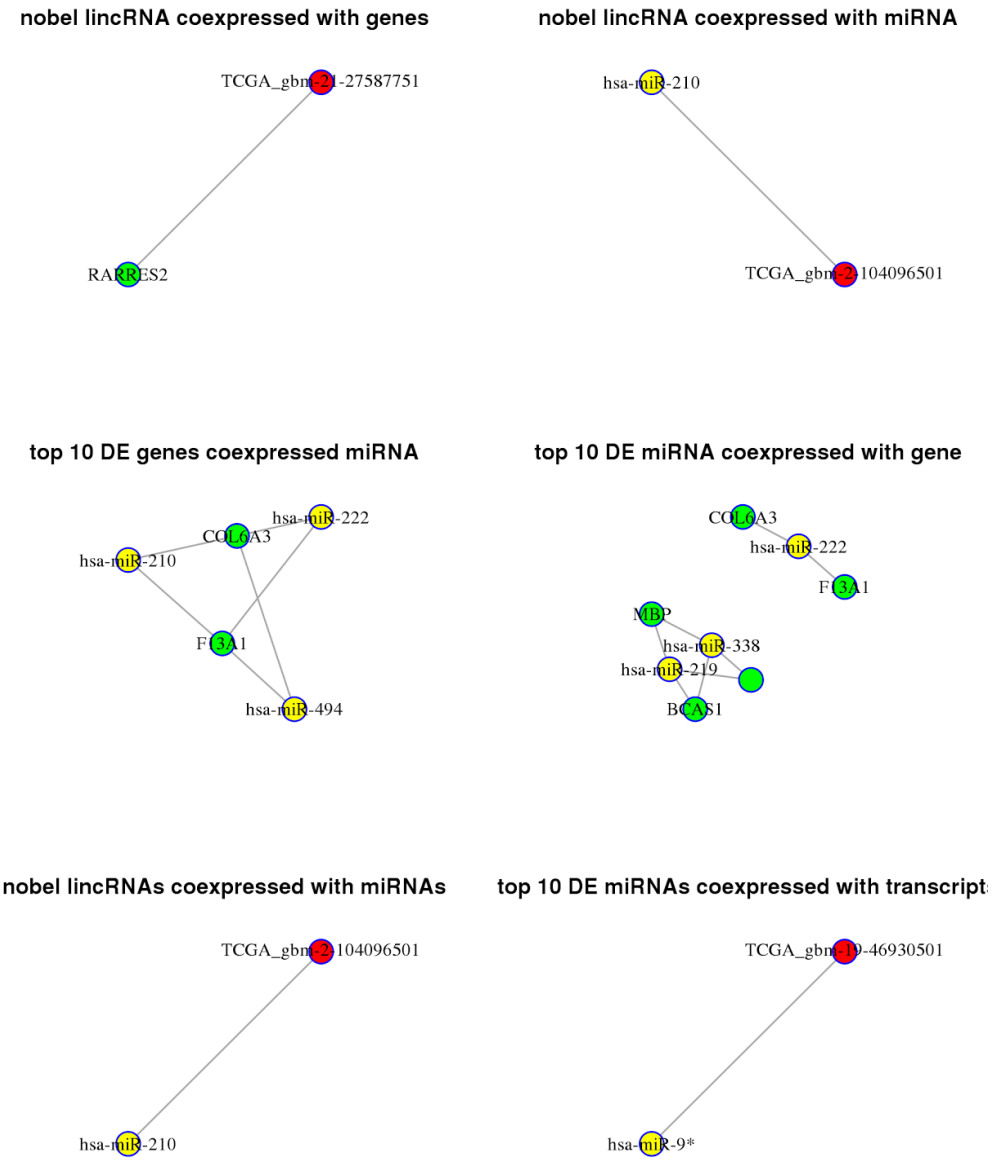


Figure 23. Co-expression of top 10 genes and miRNAs with novel transcripts and novel lincRNAs

Looking at the miRNA and novel transcript co-expression shown in figure 17 with the threshold mutual information of 0.26, there are four miRNAs, *hsa-miR-9*, hsa-miR-210, hsa-miR-142-3p* and *hsa-miR-99a* that are co-expressed with *TCGA_gbm-19-46930501, TCGA_gbm-2-104096501, TCGA_gbm-16-64693251* and *TCGA_gbm-9-23156751* respectively. As it is shown in the figure 23, *TCGA_gbm-2-104096501* is the only novel lincRNA that is co-expressed with hsa-miR-210 miRNA.

Taking cutoff threshold mutual information to 0.4, the co-expression between the gene and miRNA are illustrated in the figure 23. *COL6A3* gene is co-expressed with *hsa-miR-222, hsa-miR-210* and *hsa-miR-494*. In the same way, BCAS1 is co-expressed with *hsa-miR-338* and *hsa-miR-219* miRNAs. MBP gene is expressed with *hsa-miR-219* and *hsa-miR-338* miRNAs while *F13A1* genes is co-expressed with *hsa-miR-210, hsa-miR-222* and *hsa-miR-494* miRNAs.

As it is illustrated in the figure 23, *COL6A3* and *F13A1* genes are among the top 10 differentially expressed genes that are co-expressed with *hsa-miR-222*, *hsa-miR-494* and *hsa-miR-210* while *hsa-miR-222*, hsa-miR-338 and *hsa-miR-219* miRNAs are among top 10 differentially expressed miRNAs that are co-expressed with *COL6A3*, *F13A1, MBP and BCAS1.*

Based on the result from the DNA-protein and protein-lincRNA interaction, it is possible to predict the transcriptional interference of certain proteins that in turn reveals the possible functional mechanisms of the lincRNA. Hence, identifying proteins that have binding site on both the promoter regions of the highly differentially expressed genes and on the newly identified lincRNAs are in some way highlights the possible functional roles of the lincRNAs. Table 7 illustrates proteins that have a binding site on both the novel lincRNA and promoter regions of the highly differentially expressed genes.

Table 7. Proteins that have binding site on both novel lincRNA and promoter regions of top 10 DE expressed genes

novel lincRNA	Top 10 DE Genes	common binding proteins
TCGA_gbm-1-120693251,TCGA_gbm-8-90598251, TCGA_gbm-17-21730501	IGF2	"DAZAP1", "MSI1", "RBM3", "RBM8A", "TIA1"
	F13A1, COL1A1, COL3A1	"TIA1"
TCGA_gbm-1-153561751, TCGA_gbm-5-63687001,TCGA_gbm-22-29576001,	F13A1, COL1A1, COL3A1	"TIA1"
	IGF2	"BRUNOL5", "DAZAP1", "MSI1", "RBM3", "RBM8A", "TIA1"
TCGA_gbm-2-104096501,TCGA_gbm-2-112797251, TCGA_gbm-9-70597501,TCGA_gbm-9-42236501, TCGA_gbm-9-68284501	IGF2	"DAZAP1", "MSI1", "RBM3", "RBM8A"
TCGA_gbm-2-112797251, TCGA_gbm-8-90598251	IGF2	"BRUNOL5", "MSI1", "RBM3", "RBM8A"
TCGA_gbm-3-153501,	IGF2	"MSI1", "RBM3", "RBM8A
TCGA_gbm-9-70631251	IGF2	"DAZAP1", "RBM3", "RBM8A"

TCGA_gbm-13-50529001	IGF2	"DAZAP1", "RBM8A"
TCGA_gbm-1-153561751, TCGA_gbm-2-112797251,TCGA_gbm-9-70631251,TCGA_gbm-9-70597501,TCGA_gbm-21-27587751	IGF2, COL3A1	ACO1(RBPDB)

Insulin-like growth factor receptor (IGF2) gene is one of the genes that are highly differentially expressed as it is shown in the previous analysis steps and it has the RBP binding site in its promoter region. IGF2 protein hormone is preferentially expressed after birth in the liver and it is involved in regulation of cellular proliferation growth, migration, differentiation and survival. Adult IGF2 expression occurs in liver and in epithelial cells lining the surface of the brain. IGF2 is imprinted and is expressed exclusively from the paternal allele except in adult liver and central nervous system, where it is expressed biallelically [193]. This gene has MSI1, RBM3, RBM8A, BRUNOL5 and DAZAP1 proteins that have binding sites in both its promoter region and the identified novel lincRNA.

Factor XIII, a1 subunit (F13A1) is another gene with common protein binding site in its promoter region and novel lincRNAs as it is shown in table 7 that encodes the coagulation factor XIII A subunit. Coagulation factor XIII is an enzyme activated in the blood coagulation cascade. This enzyme acts as a transglutaminase to catalyze the formation of gamma-glutamyl-epsilon-lysine crosslinking between fibrin molecules, thus stabilizing the fibrin clot. Defects in this gene can result in a lifelong bleeding tendency, defective wound healing, and habitual abortion. [194]

Collagen, type I, alpha 1(COL1A1) is a gene among highly differentially expressed genes which encodes to the pro-alpha1 chains of type I collagen protein which is the most abundant protein in the human body and it is a substance that holds the whole body together. It is found in most connective tissues and it is abundant in bone, cornea, dermis and tendon. Defect on this gene results in a particular type of skin tumor called dermatofibrosarcoma protuberans, resulting from unregulated expression of the growth factor. [193] collagen, type III, alpha 1(COL3A1) gene also encodes for the pro-alpha1 chains of type III collagen, a fibrillar collagen that is found in extensible connective tissues such as skin, lung, uterus, intestine and the vascular system, frequently in association with type I collagen. [195]

TCGA_gbm-2-104096501, TCGA_gbm-3-153501, TCGA_gbm-5-63687001 and TCGA_gbm-17-10671251, among the first top 5 differentially expressed novel lincRNAs, and IGF2 gene, among the first top 10 differentially expressed gene, have common protein binding site for MSI1, RBM3 and RBM8A proteins. MSI1 is an RNA binding protein that plays a role in the proliferation and maintenance of stem cells in the central nervous system. It is involved in translational regulation of target mRNA. Therefore, it is predicted that the interaction of the above novel lincRNAs and MSI1 might have translational regulatory effect on IGF2 gene which is expressed in adult liver and central nervous system. [193]

In the other hand RBM3 protein is Cold-inducible mRNA binding protein that enhances global protein synthesis. It is involved in positive regulation of translation by reducing the relative abundance of microRNAs during overexpression. [199] Hence, this protein might be interacting with the above mentioned lincRNAs to regulate the translation of the gene by regulating the relative abundance of the miRNAs.

RNA-binding protein 8A (RBM8A) is a core component of the splicing-dependent multiprotein exon junction complex (EJC) deposited at splice junctions on mRNAs. The EJC marks the position of the exon-exon junction in the mature mRNA for the gene expression machinery and the core components remain bound to spliced mRNAs throughout all stages of mRNA metabolism thereby influencing downstream processes including nuclear mRNA export, subcellular mRNA localization, translation efficiency and nonsense-mediated mRNA decay (NMD).[196] Thus, the interaction of this protein with both IGF2 gene and the novel lincRNAs might have functional significance on splicing and metabolism of IGF2 gene and the newly identified lincRNAs. [196]

Based on the miRNA-lincRNA interaction using the miRanda software, the function of the *TCGA_gbm-22-29576001*, *TCGA_gbm-21-27587751* and *TCGA_gbm-17-10671251* novel lincRNAs are predicted to influence the gene regulation by either competing for binding site on *hsa-mir-181d* with its target genes and/or influencing the normal functioning of hsa-mir-181d. hsa-mir-181d targets and modulates protein expression by inhibiting translation or inducing degradation of target messenger RNAs.[197]

hsa-mir-210, which is linked to hypoxia pathway usually overexpressed in cells affected by cardiac disease and tumours, is known for its up-regulation of angiogenesis and inhibition of cardiomyocyte apoptosis. It has also target on TCGA_gbm-17-21730501 and TCGA_gbm-17-30454751 novel lincRNAs. Therefore, in one way or another the interaction of the TCGA_gbm-17-21730501 and TCGA_gbm-17-30454751 with hsa-mir-210 reveals the functional involvement of novel lincRNAs on the regulation of angiogenesis and cardiomyocyte apoptosis. [198]

6 DISCUSSION

As the identification and discovery of the novel lncRNAs are growing due to the highly efficient and affordable high throughput sequencing technologies, the functional involvement of lncRNAs in various cellular system is the central research topic. Recent studies have shown that the lncRNAs are functionally involved in transcriptional regulation, posttranscriptional regulations, translational regulation and RNA processing.

The result from the co-expression analysis shows that TCGA_gbm--3--153501 novel lincRNA which co-expressed with THBS1 gene is involved in cell-cell and cell-to-matrix interactions. These highlight that the role of such lincRNA might be involved in chromatin remodeling by recruiting the chromatin modification complex. This might also be evidence for the involvement of this lincRNA in the transcriptional regulation. Thus, these effect might be exhibited on platelet aggregation, angiogenesis, and tumorigenesis. [202] Based on the mutual information co-expression analysis, TCGA_gbm-21--27587751, which co-expressed with *RARRES2* gene, is expected to have functional involvement in the initiation of chemotaxis via the ChemR23 G protein-coupled seven-transmembrane domain ligand [205]. Chemotaxis is the movement of cells in response to the chemical stimuli.

From the protein-DNA and protein-lincRNA interaction results, the proteins that bind to both the gene and lincRNAs are identified. The top 5 highly differentially expressed novel lincRNAs and IGF2 gene have common proteins that bind to the promoter region of the IGF2 gene. These proteins are MSI1, RBM3, RBM8A, BRUNOL5 and DAZAP1. The interaction of MSI1 and RBM3 on the promoter regions of IGF2 gene and the novel lincRNA, which is associated with regulation of cellular proliferation, growth, migration, differentiation and survival, might have translational regulation by targeting mRNA or reducing the relative abundance of microRNAs during overexpression. [199]

The interaction of RBM8A protein with the IGF2 gene and the novel lincRNA is predicted to have a functional significance in RNA splicing and overall process of RNA metabolism. As a result, it influences downstream processes including nuclear mRNA export, subcellular mRNA localization, translation efficiency and nonsense-mediated mRNA decay (NMD).[196] The fact that IGF2 gene is expressed in adult liver and in epithelial cells lining surface of the brain indicates that it has the protein interactions have significant functional roles in the glioblastoma tumorigenesis. [193]

Based on the result from the miRNA-lincRNA interactions, *TCGA_gbm-22-29576001*, *TCGA_gbm-21-27587751* and *TCGA_gbm-17-10671251 lincRNAs and hsa-mir-181d interact* in such a way that it results in the transcriptional regulation by competing for the binding site on the miRNA with its target gene. In addition to that it regulates the gene expression by inhibiting translation or inducing degradation of target messenger RNAs. [198] The functional aspect of those interacting novel lincRNAs might be related to the functionalities of the interacting miRNAs or they might influence

the interaction between the 3' UTR of mRNAs and miRNAs. This has a post transcriptional regulatory role.

The efficacy of the computational functional predication of the long noncoding RNAs depends on the algorithmic efficiency of identifying the lncRNAs and computational methods used to predict the function. One of the drawback in this project is that it relayed on the Novellette algorithm [204] for the identification of the novel lncRNAs. The exons, polyA tails and UTR regions of lincRNA's genomic coordinates from [204] are ambiguous. Thus, it has made the exact lincRNA sequence retrieval difficult.

In the protein-DNA and protein-lincRNA interactions analysis, this project only considers the sequence-based approaches ignoring the secondary & tertiary structures and the van der waals interactions between the interacting molecules. This might be the limitation of protein-DNA and protein-lincRNA interactions analysis steps in this project. In addition to that, for protein-DNA interaction analysis, the PFM from hDPI database might not contain all of the possible protein-DNA interaction motifs. In the same way, for protein-lincRNA interaction analysis, there are only 53 RBP motifs used from RBPDB database and 102 RBP motifs from RNAcompete database. This PFM might not be the only RBP motifs. Therefore, it might considered as a limitation.

The future researches in the field of the functional computational prediction of long noncoding RNAs have to consider optimizing the algorithms used in identifications of novel lncRNAs. In addition, together with the gene expression, novel transcripts expression and miRNA expression, it is recommended to add the protein level expression on the analysis pipeline. This will give a chance to investigate the posttranslational effects of different interacting molecules such as the interaction of miRNA with lincRNA and miRNA with mRNA. The computational methods used in protein-DNA and protein-lincRNA interaction analysis should also consider the secondary and tertiary structures together with the van der waal's interactions.

7 CONCLUSIONS

Long noncoding RNAs are RNA molecules without coding potential and longer than 200 nucleotides. The vast majority, about 97%, of the transcribed RNAs in the genome are ncRNAs. Based on the genomic location and context lncRNAs are classified into long intergenic RNA (lncRNA), intronic RNA, sene lncRNA and anti-sense lncRNA. Based on the effects exerted on the DNA sequence, lncRNAs are classified into cis-lncRNA and trans-lncRNA. Based on the cellular molecular mechanism, lncRNAs are grouped in transcriptional regulation, post-transcriptional regulation and other functions. Signal, Decoy, guide and scaffolds are types of lncRNA based on the targeting mechanism.

LncRNAs are identified using both experimental and computational methods. Some of the experimental methods include tilling array, serial analysis of gene expression (SAGE), RNA sequencing (RNA-seq), RNA immunoprecipitation (RNA-IP) and chromatin signature based approaches. The computational approaches to identify the lncRNAs include ORF length stratagy, sequence and secondary structure conservation strategy, and machine learning approaches.

LncRNAs function in different epigenetic regulation, transcriptional regulation of gene expression and in the processing of other small RNAs. They also function as structural component by interacting with other proteins. HOTAIR, PCAT-1 and MALAT1 are some of the lncRNAs that are associated with diseases such as Lung, breast, colorectal and prostate cancers. However, some lncRNAs are being introduced as a potential therapeutic target and biomarker in the diagnostics and prognostics of different cancer types via both oncogenic and tumor-suppressive pathways.

The result from different computational methods are combined to predict the novel lincRNA functions that are identified from TCGA glioblastoma multiforme datasets. In the analysis, the co-expression based method and sequence based methods like protein-DNA, protein-lincRNA and miRNA-lincRNA interactions are considered to predict the functionality of the novel lincRNAs. As a result, the functional prediction of those novel lincRNAs are associated with transcriptional regulation by transcriptional interference, translational regulation by reducing the expression of miRNAs or by competing for target sites on miRNAs with mRNAs. They are also predicted to be involved in the RNA processing and splicing regulation.

In conclusion, further researches has to be made by integrating the finding from this project and protein level expression to assert the posttranscriptional regulation that has been predicted in this project is the valid one. In addition to that, enhancing the computational method that is used to identify the lncRNA by using the modern and efficient machine learning algorithms increase the functional prediction accuracy. The feature research has to also consider the secondary & tertiary structure together with the vander waal's interaction while analyzing the protein-DNA, protein-lincRNA and miRNA-lincRNA interaction.

8 REFERENCES

1. Mattick, John S., Makunin, Igor V. Non-coding RNA 2006
2. Kapranov P, Cheng J, Dike S, et al. (June 2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription". Science 316 (5830): 1484–8.
3. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research* 2012;22(9):1775-1789. doi:10.1101/gr.132159.111.
4. Maeda N, Kasukawa T, Oyama R, et al. Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs. *PLoS Genetics* 2006;2(4):e62. doi:10.1371/journal.pgen.0020062.
5. Jia, Hui et al. "Genome-Wide Computational Identification and Manual Annotation of Human Long Noncoding RNA Genes." *RNA* 16.8 (2010): 1478–1487. *PMC*. Web. 13 Dec. 2014.
6. Cabili, Moran N. et al. "Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses." *Genes & Development* 25.18 (2011): 1915–1927. *PMC*. Web. 14 Dec. 2014.
7. Chen et al., LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013 Jan 1;41(D1):D983-6.
8. Volders, Pieter-Jan et al. "LNCipedia: a Database for Annotated Human lncRNA Transcript Sequences and Structures." *Nucleic Acids Research* 41.Database issue (2013): D246–D251. *PMC*. Web. 26 Dec. 2014.
9. Volders, Pieter-Jan, Verheggen, Kenneth, Menschaert, Gerben, et al. An update on LNCipedia: a database for annotated human lncRNA sequences 2014
10. Ma, Lina, Vladimir B. Bajic, and Zhang Zhang. "On the Classification of Long Non-Coding RNAs." *RNA Biology* 10.6 (2013): 924–933. *PMC*. Web. 29 Dec. 2014.
11. online Systematic LncRNA Classification, Arraystar Inc. accessed on: 25 Dec. 2014. world wide web, URL http://www.arraystar.com/Services/Services_main.asp?ID=307
12. online mcmanuslab. University of California, San Francisco accessed on: 25 Mar. 2015 http://mcmanuslab.ucsf.edu/node/251
13. Pagano A, Castelnuovo M, Tortelli F, Ferrari R, Dieci G, Cancedda R. New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. PLoS Genet. 2007;3:e1. doi: 10.1371/journal.pgen.0030001
14. Handong Ma, Yun Hao, Xinran Dong, et al., "Molecular Mechanisms and Function Prediction of Long Noncoding RNA," The Scientific World Journal, vol. 2012, Article ID 541786, 11 pages, 2012. doi:10.1100/2012/541786.
15. Wang, Kevin C., and Howard Y. Chang. "Molecular Mechanisms of Long Noncoding RNAs." *Molecular cell* 43.6 (2011): 904–914. *PMC*. Web. 30 Mar. 2015.
16. Rinn, J. L. *et al*. The transcriptional activity of human chromosome 22. Genes Dev. 17, 529–540 (2003).
17. E. A. Gibb, E. A. Vucic, K. S. Enfield, G. L. Stewart, K. M. Lonergan, et al., "Human cancer long non-coding RNA transcriptomes," PLoS One, vol. 6, Article ID e25915, 2011.
18. T. L. Lee, A. Xiao, and O. M. Rennert, "Identification of novel long noncoding RNA transcripts in male germ cells," Methods in Molecular Biology, vol. 825, pp. 105–114, 2012
19. Fatica, Alessandro, Bozzoni, Irene "Long non-coding RNAs: new players in cell differentiation and development" Nat Rev Genet, Nature Publishing Group, a division of Macmillan Publishers Limited, 2014.
20. T. Li, S. Wang, R. Wu, X. Zhou, D. Zhu, et al., "Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing," Genomics, vol. 99, pp. 292–298, 2012.
21. J. R. Prensner, M. K. Iyer, O. A. Balbin et al., "Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression," Nature Biotechnology, vol. 29, no. 8, pp. 742–749, 2011.
22. M. Guttman, I. Amit, M. Garber et al., "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals," Nature, vol. 458, no. 7235, pp. 223–227, 2009.
23. H. Jia, M. Osak, G. K. Bogu, L. W. Stanton, R. Johnson, and L. Lipovich, "Genome-wide computational identification and manual annotation of human long noncoding RNA genes," RNA, vol. 16, no. 8, pp. 1478–1487, 2010.
24. M. Guttman, I. Amit, M. Garber et al., "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals," Nature, vol. 458, no. 7235, pp. 223–227, 2009.
25. M. F. Lin, J. W. Carlson, M. A. Crosby et al., "Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes," Genome Research, vol. 17, no. 12, pp. 1823–1836, 2007.

26. M. Clamp, B. Fry, M. Kamal et al., "Distinguishing protein-coding and noncoding genes in the human genome," Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 49, pp. 19428–19433, 2007.
27. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35(Web Server issue):W345–W349. doi: 10.1093/nar/gkm391.
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. doi: 10.1093/bioinformatics/btr209.
29. E. Rivas and S. R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," BMC Bioinformatics, vol. 2, article no. 8, 2001.
30. S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and reliable prediction of noncoding RNAs," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 7, pp. 2454–2459, 2005.
31. J. S. Pedersen, G. Bejerano, A. Siepel et al., "Identification and classification of conserved RNA secondary structures in the human genome," PLoS Computational Biology, vol. 2, no. 4, article no. e33, pp. 251–262, 2006.
32. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. 2013;41(17):e166. doi: 10.1093/nar/gkt646.

33. Li, Aimin, Junying Zhang, and Zhongyin Zhou. "PLEK: A Tool for Predicting Long Non-Coding RNAs and Messenger RNAs Based on an Improved *k*-mer Scheme." *BMC Bioinformatics* 15.1 (2014): 311. *PMC*. Web. 6 Apr. 2015.
34. J.Liu, J. Gough, and B. Rost, "Distinguishing protein-coding from non-coding RNAs through support vector machines," PLoS genetics, vol. 2, no. 4, article no. e29, 2006.
35. L. Kong, Y. Zhang, Z. Q. Ye et al., "CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine," Nucleic Acids Research, vol. 35, pp. W345–W349, 2007.
36. S. Chooniedass-Kothari, E. Emberley, M. K. Hamedani et al., "The steroid receptor RNA activator is the first functional RNA encoding a protein," FEBS Letters, vol. 566, no. 1-3, pp. 43–47, 2004.
37. L. Duret, C. Chureau, S. Samain, J. Weissanbach, and P. Avner, "The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene," Science, vol. 312, no. 5780, pp. 1653–1655, 2006.
38. Z. J. Lu, K. Y. Yip, G. Wang et al., "Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data," Genome Research, vol. 21, no. 5, pp. 276–285, 2011.
39. R. R. Pandey, T. Mondal, F. Mohammad et al., "Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation," Molecular Cell, vol. 32, no. 2, pp. 232–246, 2008.
40. T. Nagano, J. A. Mitchell, L. A. Sanz et al., "The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin," Science, vol. 322, no. 5908, pp. 1717–1720, 2008.
41. O. Wapinski and H. Y. Chang, "Long noncoding RNAs and human disease," Trends in Cell Biology, vol. 21, no. 6, pp. 354–361, 2011.
42. U. A. Ørom and R. Shiekhattar, "Noncoding RNAs and enhancers: complications of a long-distance relationship," Trends in Genetics, vol. 27, pp. 433–439, 2011.
43. J. E. Wilusz, H. Sunwoo, and D. L. Spector, "Long noncoding RNAs: functional surprises from the RNA world," Genes and Development, vol. 23, no. 13, pp. 1494–1504, 2009
44. Guttman, Mitchell et al. "lincRNAs Act in the Circuitry Controlling Pluripotency and Differentiation." *Nature* 477.7364 (2011): 295–300. *PMC*. Web. 13 Apr. 2015.
45. Orly Wapinski, Howard Y. Chang, Long noncoding RNAs and human disease, Trends in Cell Biology, Volume 21, Issue 6, June 2011, Pages 354-361, ISSN 0962-8924, http://dx.doi.org/10.1016/j.tcb.2011.04.001.
46. Mustafa Isin, Nejat Dalay, LncRNAs and neoplasia, Clinica Chimica Acta, Volume 444, 15 April 2015, Pages 280-288, ISSN 0009-8981, http://dx.doi.org/10.1016/j.cca.2015.02.046.
47. Shen, Xiao-han Qi, Peng Du, Xiang, Long non-coding RNAs in cancer invasion and metastasis, Mod Pathol. 2015/01//print, United States & Canadian Academy of Pathology. http://dx.doi.org/10.1038/modpathol.2014.75. 10.1038/modpathol.2014.75
48. Yang, Xiaofei et al. "A Network Based Method for Analysis of lncRNA-Disease Associations and Prediction of lncRNAs Implicated in Diseases." Ed. Paolo Provero. *PLoS ONE* 9.1 (2014): e87797. *PMC*. Web. 16 Apr. 2015.

49. Z. Cui, S. Ren, J. Lu, F. Wang, W. Xu, Y. Sun, *et al.* "The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor" Urol Oncol, 31 (2013), pp. 1117–1123
50. K. Takayama, K. Horie-Inoue, S. Katayama, T. Suzuki, S. Tsutsumi, K. Ikeda, *et al.* Androgen responsive long noncoding RNA CTBP1-AS promotes prostate cancer. EMBO J, 32 (2013), pp. 1665–1680
51. T. Kino, D.E. Hurt, T. Ichijo, N. Nader, G.P. Chrousos Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. Sci Signal (3) (2010), p. ra8
52. E.S. Martens-Uzunova, S.E. Jalava, N.F. Dits, G.J. van Leenders, S. Møller, J. Trapman, *et al.* "Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer" Oncogene, 31 (2012), pp. 978–991
53. V.X. Fu, S.R. Schwarze, M.L. Kenowski, S. Leblanc, J. Svaren, D.F. Jarrard "A loss of insulin-like growth factor-2 imprinting is modulated by CCCTC-binding factor down-regulation at senescence in human epithelial cells" J Biol Chem, 279 (2004), pp. 52218–52226
54. V.X. Fu, J.R. Dobosy, J.A. Desotelle, N. Almassi, J.A. Ewald, R. Srinivasan, *et al.* "Aging and cancer-related loss of insulin-like growth factor 2 imprinting in the mouse and human prostate Cancer" Res, 68 (2008), pp. 6797–6802
55. R. Lin, S. Maeda, C. Liu, M. Karin, T.S. Edgington "A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas" Oncogene, 26 (2007), pp. 851–858
56. M.J. Bussemakers, A. van Bokhoven, G.W. Verhaegh, F.P. Smit, H.F. Karthaus, J.A. Schalken, *et al.* "DD3: a new prostate-specific gene, highly overexpressed in prostate cancer" Cancer Res, 59 (1999), pp. 5975–5979
57. E.D. Crawford, K.O. Rove, E.J. Trabulsi, J. Qian, K.P. Drewnowska, J.C. Kaminetsky, *et al.* "Diagnostic performance of PCA3 to detect prostate cancer in men with increased prostate specific antigen: a prospective study of 1,962 cases" J Urol, 188 (2012), pp. 1726–1731
58. J.R. Prensner, M.K. Iyer, O.A. Balbin, S.M. Dhanasekaran, Q. Cao, J.C. Brenner, *et al.* "Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression" Nat Biotechnol, 29 (2011), pp. 742–749
59. V. Srikantan, Z. Zou, G. Petrovics, L. Xu, M. Augustus, L. Davis, *et al.* "PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer" Proc Natl Acad Sci U S A, 97 (2000), pp. 12216–12221
60. G. Petrovics, W. Zhang, M. Makarem, J.P. Street, R. Connelly, L. Sun, *et al.* "Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients" Oncogene, 23 (2004), pp. 605–611
61. Y. Xue, M. Wang, M. Kang, Q. Wang, B. Wu, H. Chu, *et al.* "Association between lncrna PCGEM1 polymorphisms and prostate cancer risk" Prostate Cancer Prostatic Dis, 16 (2013), pp. 139–144
62. G.O. Ifere, G.A. Ananaba "Prostate cancer gene expression marker 1 (PCGEM1): a patented prostate-specific non-coding gene and regulator of prostate cancer progression" Recent Pat DNA Gene Seq, 3 (2009), pp. 151–163
63. G.O. Ifere, E. Barr, A. Equan, K. Gordon, U.P. Singh, J. Chaudhary, *et al.* "Differential effects of cholesterol and phytosterols on cell proliferation, apoptosis and expression of a prostate specific gene in prostate cancer cell lines" Cancer Detect Prev, 32 (2009), pp. 319–328
64. X. Fu, L. Ravindranath, N. Tran, G. Petrovics, S. Srivastava "Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1" DNA Cell Biol, 25 (2006), pp. 135–141
65. L. Yang, C. Lin, C. Jin, J.C. Yang, B. Tanasa, W. Li, *et al.* "lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs" Nature, 500 (2013), pp. 598–602
66. L. Yang, C. Lin, C. Jin, J.C. Yang, B. Tanasa, W. Li, *et al.* "lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs" Nature, 500 (2013), pp. 598–602
67. S. Chung, H. Nakagawa, M. Uemura, L. Piao, K. Ashikawa, N. Hosono, *et al.* "Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility" Cancer Sci, 102 (2011), pp. 245–252
68. L. Poliseno, L. Salmena, J. Zhang, B. Carver, W.J. Haveman, P.P. Pandolfi "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology" Nature, 465 (2010), pp. 1033–1038
69. P. Johnsson, A. Ackley, L. Vidarsdottir, W.O. Lui, M. Corcoran, D. Grandér, *et al.* "A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells" Nat Struct Mol Biol, 20 (2013), pp. 440–446
70. R.S. Hudson, M. Yi, N. Volfovsky, R.L. Prueitt, D. Esposito, S. Volinia, *et al.* "Transcription signatures encoded by ultraconserved genomic regions in human prostate cancer" Mol Cancer, 12 (2013), p. 13
71. T. Laner, W.A. Schulz, R. Engers, M. Muller, A.R. Florl "Hypomethylation of the XIST gene promoter in prostate cancer" Oncol Res, 15 (2005), pp. 257–264

72. M.A. Song, J.H. Park, K.S. Jeong, D.S. Park, M.S. Kang, S. Lee "Quantification of CpG methylation at the 50-region of XIST by pyrosequencing from human serum" Electrophoresis, 28 (2007), pp. 2379–2384
73. E. Pasmant, A. Sabbagh, M. Vidaud, I. Bièche "ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS" FASEB J, 25 (2011), pp. 444–448
74. Y. Chi, S. Huang, L. Yuan, M. Liu, N. Huang, S. Zhou, *et al.* "Role of BC040587 as a predictor of poor outcome in breast cancer" Cancer Cell Int, 14 (2014), p. 123
75. Z. Xing, A. Lin, C. Li, K. Liang, S. Wang, Y. Liu, *et al.* "lncRNA directs cooperative epigenetic regulation downstream of chemokine signals" Cell, 159 (2014), pp. 1110–1125
76. A. Iacoangeli, Y. Lin, E.J. Morley, I.A. Muslimov, R. Bianchi, J. Reilly, *et al.* "BC200 RNA in invasive and preinvasive breast cancer" Carcinogenesis, 25 (2004), pp. 2125–2133
77. D. Liu, P.S. Rudland, D.R. Sibson, R. Barraclough "Identification of mRNAs differentially-expressed between benign and malignant breast tumour cells" Br J Cancer, 87 (2002), pp. 423–431
78. M. Mourtada-Maarabouni, M.R. Pickard, V.L. Hedge, F. Farzaneh, G.T. Williams "GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer" Oncogene, 28 (2009), pp. 195–208
79. P. Qi, X. Du "The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine" Mod Pathol, 26 (2013), pp. 155–165
80. K.P. Sørensen, M. Thomassen, Q. Tan, M. Bak, S. Cold, M. Burton, *et al.* "Long non-coding RNA HOTAIR is an independent prognostic marker of metastasis in estrogen receptor-positive primary breast cancer" Breast Cancer Res Treat, 142 (2013), pp. 529–536
81. X. Ding, L. Zhu, T. Ji, X. Zhang, F. Wang, S. Gan, *et al.* "Long intergenic non-coding RNAs (LincRNAs) identified by RNA-seq in breast cancer" PLoS One, 9 (2014), p. e103270
82. Y. Shi, J. Lu, J. Zhou, X. Tan, Y. He, J. Ding, *et al.* "Long non-coding RNA Loc554202 regulates proliferation and migration in breast cancer cells" Biochem Biophys Res Commun, 446 (2014), pp. 448–453
83. Y. Shi, J. Lu, J. Zhou, X. Tan, Y. He, J. Ding, *et al.* "Long non-coding RNA Loc554202 regulates proliferation and migration in breast cancer cells" Biochem Biophys Res Commun, 446 (2014), pp. 448–453
84. X. Zhang, Y. Zhou, K.R. Mehta, D.C. Danila, S. Scolavino, S.R. Johnson, *et al.* "A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells" J Clin Endocrinol Metab, 88 (2003), pp. 5119–5126
85. K. Augoff, B. McCue, E.F. Plow, K. Sossey-Alaoui "miR-31 and its host gene lncRNA LOC554202 are regulated by promoter hypermethylation in triple-negative breast cancer" Mol Cancer, 11 (2012), p. 5
86. L. Lipovich, R. Johnson, C.Y. Lin "MacroRNA underdogs in a microRNA world: evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA" Biochim Biophys Acta, 1799 (2010), pp. 597–615
87. Y. Guan, W.L. Kuo, J.L. Stilwell, H. Takano, A.V. Lapuk, J. Fridlyand, *et al.*" Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer" Clin Cancer Res, 13 (2007), pp. 5745–5755
88. N. Borth, J. Massier, C. Franke, K. Sachse, H.P. Saluz, F. Hänel "Chlamydial protease CT441 interacts with SRAP1 co-activator of estrogen receptor alpha and partially alleviates its co-activation activity" J Steroid Biochem Mol Biol, 119 (2010), pp. 89–95
89. A. Vincent-Salomon, C. Ganem-Elbaz, E. Manié, V. Raynal, X. Sastre-Garau, D. Stoppa-Lyonnet, *et al.* "X inactive-specific transcript RNA coating and genetic instability of the X chromosome in BRCA1 breast tumors" Cancer Res, 67 (2007), pp. 5134–5140
90. M.E. Askarian-Amiri, J. Crawford, J.D. French, C.E. Smart, M.A. Smith, M.B. Clark, K. Ru, T.R. Mercer, E.R. Thompson, S.R. Lakhani, A.C. Vargas, I.G. Campbell, M.A. Brown, M.E. Dinger, J.S. Mattick "SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer" RNA, 17 (5) (2011 May), pp. 878–891
91. M. Qiu, Y. Xu, X. Yang, J. Wang, J. Hu, L. Xu, *et al.* "CCAT2 is a lung adenocarcinoma-specific long non-coding RNA and promotes invasion of non-small cell lung cancer" Tumour Biol, 35 (2014), pp. 5375–5380
92. H. Ono, N. Motoi, H. Nagano, E. Miyauchi, M. Ushijima, M. Matsuura, *et al.* "Long noncoding RNA HOTAIR is relevant to cellular proliferation, invasiveness, and clinical relapse in small-cell lung cancer" Cancer Med, 3 (2014), pp. 632–642
93. M. Huarte, M. Guttman, D. Feldser, M. Garber, M.J. Koziol, D. Kenzelmann-Broz, *et al.* "A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response" Cell, 142 (2010), pp. 409–419
94. F.Q. Nie, Q. Zhu, T.P. Xu, Y.F. Zou, M. Xie, M. Sun, *et al.* "Long non-coding RNA MVIH indicates a poor prognosis for non-small cell lung cancer and promotes cell proliferation and invasion" Tumour Biol, 35 (2014), pp. 7587–7594

95. P. Ji, S. Diederichs, W. Wang, S. Böing, R. Metzger, P.M. Schneider, *et al.* "MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer" Oncogene, 22 (2003), pp. 8031–8041
96. K. Komiya, N. Sueoka-Aragane, A. Sato, T. Hisatomi, T. Sakuragi, M. Mitsuoka, *et al.* "*Mina53*, a novel c-Myc target gene, is frequently expressed in lung cancers and exerts oncogenic property in NIH/3T3 cell"J Cancer Res Clin Oncol, 136 (3) (2010), pp. 465–473
97. E.B. Zhang, D.D. Yin, M. Sun, R. Kong, X.H. Liu, L.H. You, *et al.* "P53-regulated long non-coding RNA TUG1 affects cell proliferation in human non-small cell lung cancer, partly through epigenetically regulating HOXB7 expression"Cell Death Dis, 5 (2014), p. e1243
98. A. Nissan, A. Stojadinovic, S. Mitrani-Rosenbaum, D. Halle, R. Grinbaum, M. Roistacher, *et al.* "Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues" Int J Cancer, 130 (2012), pp. 1598–1606
99. X. He, X. Tan, X. Wang, H. Jin, L. Liu, L. Ma, *et al.* "c-Myc-activated long noncoding RNA CCAT1 promotes colon cancer cell proliferation and invasion" Tumour Biol, 35 (2014), pp. 12181–12188
100. L.D. Graham, S.K. Pedersen, G.S. Brown, T. Ho, Z. Kassir, A.T. Moynihan, *et al.* "Colorectal neoplasia differentially expressed (CRNDE), a novel gene with elevated expression in colorectal adenomas and adenocarcinomas" Genes Cancer, 2 (2011), pp. 829–840
101. I.J. Matouk, I. Abbasi, A. Hochberg, E. Galun, H. Dweik, M. Akkawi "Highly upregulated in liver cancer noncoding RNA is overexpressed in hepatic colorectal metastasis" Eur J Gastroenterol Hepatol, 21 (2009), pp. 688–692
102. B. Hrdlickova, R.C. de Almeida, Z. Borek, S. Withoff "Genetic variation in the non-coding genome: involvement of micro-RNAs and long non-coding RNAs in disease" Biochim Biophys Acta, 2014 (1842), pp. 1910–1922
103. S. Nakano, K. Murakami, M. Meguro, H. Soejima, K. Higashimoto, T. Urano, *et al.* "Expression profile of LIT1/KCNQ1OT1 and epigenetic status at the KvDMR1 in colorectal cancers" Cancer Sci, 97 (2006), pp. 1147–1154
104. H. Zhai, A. Fesler, K. Schee, O. Fodstad, K. Flatmark, J. Ju "Clinical significance of long intergenic noncoding RNA-p21 in colorectal cancer" Clin Colorectal Cancer, 12 (2013), pp. 261–266
105. G. Wang, Z. Li, Q. Zhao, Y. Zhu, C. Zhao, X. Li, *et al.* "LincRNA-p21 enhances the sensitivity of radiotherapy for human colorectal cancer by targeting the Wnt/β-catenin signaling pathway" Oncol Rep, 31 (2014), pp. 1839–1845
106. F. Yang, X.S. Huo, S.X. Yuan, L. Zhang, W.P. Zhou, F. Wang, *et al.* "Repression of the long noncoding RNA-LET by histone deacetylase 3 contributes to hypoxia-mediated metastasis" Mol Cell, 49 (2013), pp. 1083–1096
107. C. Xu, M. Yang, J. Tian, X. Wang, Z. Li "MALAT-1: a long non-coding RNA and its important 3′ end functional motif in colorectal cancer metastasis" Int J Oncol, 39 (2011), pp. 169–175
108. P. Qi, M.D. Xu, S.J. Ni, X.H. Shen, P. Wei, D. Huang, *et al.* "Down-regulation of ncRAN, a long non-coding RNA, contributes to colorectal cancer cell migration and invasion and predicts poor overall survival for colorectal cancer patients" Mol Carcinog (Feb 12 2014) http://dx.doi.org/10.1002/mc-22137 [Epub ahead of print]
109. K.L. Yap, S. Li, A.M. Muñoz-Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, *et al.* "Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a" Mol Cell, 38 (2010), pp. 662–674
110. S. Hombach, M. Kretz "The non-coding skin: exploring the roles of long non-coding RNAs in epidermal homeostasis and disease" Bioessays, 35 (2013), pp. 1093–1100
111. J. Valls, X. Solé, P. Hernández, C. Cerrato, I. Madrigal, R. de Cid, *et al.* "Molecular characterization of a t(9;12)(p21;q13) balanced chromosome translocation in combination with integrative genomics analysis identifies C9orf14 as a candidate tumor-suppressor" Genes Chromosomes Cancer, 46 (2007), pp. 155–162
112. D. Khaitan, M.E. Dinger, J. Mazar, J. Crawford, M.A. Smith, J.S. Mattick, *et al.* "The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion" Cancer Res, 71 (2011), pp. 3852–3862
113. W. He, Q. Cai, F. Sun, G. Zhong, P. Wang, H. Liu, *et al.* "linc-UBC1 physically associates with polycomb repressive complex 2 (PRC2) and acts as a negative prognostic factor for lymph node metastasis and survival in bladder cancer" Biochim Biophys Acta, 1832 (2013), pp. 1528–1537
114. I. Ariel, M. Sughayer, Y. Fellig, G. Pizov, S. Ayesh, D. Podeh, B.A. Libdeh, C. Levy, T. Birman, M.L. Tykocinski, N. de Groot, A. Hochberg "The imprinted H19 gene is a marker of early recurrence in human bladder carcinoma" Mol Pathol, 53 (2000), pp. 320–323
115. Y. Fan, B. Shen, M. Tan, X. Mu, Y. Qin, F. Zhang, Y. Liu "TGF-β-induced upregulation of malat1 promotes bladder cancer metastasis by associating with suz12" Clin Cancer Res, 20 (2014), pp. 1531–1541

116. L. Ying, Y. Huang, H. Chen, Y. Wang, L. Xia, Y. Chen, *et al.* "Downregulated MEG3 activates autophagy and increases cell proliferation in bladder cancer" Mol Biosyst, 9 (2013), pp. 407–411
117. Y. Zhu, M. Yu, Z. Li, C. Kong, J. Bi, J. Li, *et al.* "ncRAN, a newly identified long noncoding RNA, enhances human bladder tumor growth, invasion, and survival" Urology, 77 (510) (2011), pp. e1–e5
118. Y. Han, Y. Liu, Y. Gui, Z. Cai "Long intergenic non-coding RNA TUG1 is overexpressed in urothelial carcinoma of the bladder" J Surg Oncol, 107 (2013), pp. 555–559
119. F. Wang, X. Li, X. Xie, L. Zhao, W. Chen "UCA1, a non-protein-coding RNA up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion" FEBS Lett, 582 (2008), pp. 1919–1927
120. Y. Fan, B. Shen, M. Tan, X. Mu, Y. Qin, F. Zhang, Y. Liu "Long non-coding RNA UCA1 increases chemoresistance of bladder cancer cells by regulating Wnt signaling" FEBS J, 281 (2014), pp. 1750–1758
121. P.S. Eis, W. Tam, L. Sun, A. Chadburn, Z. Li, M.F. Gomez, *et al.* "Accumulation of miR-155 and BIC RNA in human B cell lymphomas "Proc Natl Acad Sci U S A, 102 (2005), pp.
122. H.P. Qiao, W.S. Gao, J.X. Huo, Z.S. Yang "Long non-coding RNA GAS5 functions as a tumor suppressor in renal cell carcinoma" Asian Pac J Cancer Prev, 14 (2013), pp. 1077–1082
123. M.A. Frevel, S.J. Sowerby, G.B. Petersen, A.E. Reeve "Methylation sequencing analysis refines the region of H19 epimutation in Wilms tumor" J Biol Chem, 274 (1999), pp. 29331–29340
124. D. Bertozzi, R. Iurlaro, O. Sordet, J. Marinello, N. Zaffaroni, G. Capranico "Characterization of novel antisense HIF-1α transcripts in human cancers" Cell Cycle, 10 (2011), pp. 3189–3197
125. R.R. Pandey, T. Mondal, F. Mohammad, S. Enroth, L. Redrup, J. Komorowski, *et al.* "Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation" Mol Cell, 32 (2008), pp. 232–246
126. H.M. Zhang, F.Q. Yang, S.J. Chen, J. Che, J.H. Zheng "Upregulation of long non-coding RNA MALAT1 correlates with tumor progression and poor prognosis in clear cell renal cell carcinoma" Tumour Biol (Dec 6 2014) http://dx.doi.org/10.1007/s13277-014-2925-6 [Epub ahead of print]
127. T. Kawakami, T. Chano, K. Minami, H. Okabe, Y. Okada, K. Okamoto "Imprinted DLK1 is a putative tumor suppressor gene and inactivated by epimutation at the region upstream of GTL2 in human renal cell carcinoma" Hum Mol Genet, 15 (2006), pp. 821–830
128. H. He, R. Nagy, S. Liyanarachchi, H. Jiao, W. Li, S. Suster, *et al.* "A susceptibility locus for papillary thyroid carcinoma on chromosome 8q24" Cancer Res, 69 (2009), pp. 625–631
129. Y. Wang, Q. Guo, Y. Zhao, J. Chen, S. Wang, J. Hu, *et al.* "BRAF-activated long non-coding RNA contributes to cell proliferation and activates autophagy in papillary thyroid carcinoma" Oncol Lett, 8 (2014), pp. 1947–1952
130. J. Jendrzejewski, H.L. He, H.S. Radomska, W. Li, J. Tomsic, S. Liyanarachchi, *et al.* "The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type" Proc Natl Acad Sci U S A, 109 (2012), pp. 8646–8651
131. P. Baldinu, A. Cossu, A. Manca, M.P. Satta, M.C. Sini, C. Rozzo, *et al.* "Identification of a novel candidate gene, CASC2, in a region of common allelic loss at chromosome 10q26 in human endometrial cancer" Hum Mutat, 23 (2004), pp. 318–326
132. J. Huang, P. Ke, L. Guo, W. Wang, H. Tan, Y. Liang, *et al.* "Lentivirus-mediated RNA interference targeting the long noncoding RNA HOTAIR inhibits proliferation and invasion of endometrial carcinoma cells in vitro and in vivo" Int J Gynecol Cancer, 24 (2014), pp. 635–642
133. K. Yamada, J. Kano, H. Tsunoda, H. Yoshikawa, C. Okubo, T. Ishiyama, *et al.* "Phenotypic characterization of endometrial stromal sarcoma of the uterus" Cancer Sci, 97 (2006), pp. 106–112
134. W. Yu, D. Gius, P. Onyango, K. Muldoon-Jacobs, J. Karp, A.P. Feinberg, *et al.* "Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA" Nature, 451 (2008), pp. 202–206
135. D. Mertens, A. Philippen, M. Ruppel, D. Allegra, N. Bhattacharya, C. Tschuch, *et al.* "Chronic lymphocytic leukemia and 13q14: miRs and more" Leuk Lymphoma, 50 (2009), pp. 502–505
136. Y. Liu, M. Corcoran, O. Rasool, G. Ivanova, R. Ibbotson, D. Grandér, *et al.* "Cloning of two candidate tumor suppressor genes within a 10 kb region on chromosome 13q14, frequently deleted in chronic lymphocytic leukemia" Oncogene, 15 (1997), pp. 2463–2473
137. L. Benetatos, E. Hatzimichael, A. Dasoula, G. Dranitsaris, S. Tsiara, M. Syrrou, *et al.* "CpG methylation analysis of the MEG3 and SNRPN imprinted genes in acute myeloid leukemia and myelodysplastic syndromes" Leuk Res, 34 (2010), pp. 148–153

138. R.C. Thompson, I. Vardinogiannis, T.D. Gilmore "Identification of an NF-κB p50/p65-responsive site in the human MIR155HG promoter" BMC Mol Biol, 14 (2013), p. 24
139. M. Saitou, J. Sugimoto, T. Hatakeyama, G. Russo, M. Isobe "Identification of the TCL6 genes within the breakpoint cluster region on chromosome 14q32 in T-cell leukemia" Oncogene, 19 (2000), pp. 2796–2802
140. A.R. Dallosso, A.L. Hancock, S. Malik, A. Salpekar, L. King-Underwood, K. Pritchard-Jones, *et al.* "Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer" RNA, 13 (2007), pp. 2287–2299
141. A.C. Tahira, M.S. Kubrusly, M.F. Faria, B. Dazzani, R.S. Fonseca, V. Maracaja-Coutinho, *et al.* "Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer" Mol Cancer, 10 (2011), p. 141
142. L. You, D. Chang, H.Z. Du, Y.P. Zhao" Genome-wide screen identifies PVT1 as a regulator of Gemcitabine sensitivity in human pancreatic cancer cells" Biochem Biophys Res Commun, 407 (2011), pp. 1–6
143. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, Sevignani C, *et al.* Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell 2007; 12: 215–229.
144. Pibouin L, Villaudy J, Ferbus D, Muleris M, Prosperi MT, Remvikos Y, *et al.* Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. Cancer Genet Cytogenet 2002; 133: 55–60.
145. Fu X, Ravindranath L, Tran N, Petrovics G, Srivastava S. Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, *PCGEM1*. DNA Cell Biol 2006; 25: 135–141.
146. Guffanti A, Iacono M, Pelucchi P, Kim N, Solda G, Croft LJ, *et al.* A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 2009; 10: 163.
147. Lin R, Maeda S, Liu C, Karin M, Edgington TS. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene 2007; 26: 851–858.
148. Sonkoly E, Bata-Csorgo Z, Pivarcsi A, Polyanka H, Kenderessy-Szabo A, Molnar G, *et al.* Identification and characterization of a novel, psoriasis susceptibility-related noncoding RNA gene, *PRINS*. J Biol Chem 2005; 280: 24159–24167.
149. Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A, *et al.* Identification of a novel non-coding RNA, *MIAT*, that confers risk of myocardial infarction. J Hum Genet 2006; 51: 1087–1099.
150. Pasmant E, Laurendeau I, Heron D, Vidaud M, Vidaud D, Bieche I. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of *ANRIL*, an antisense noncoding RNA whose expression coclusters with ARF. Cancer Res 2007; 67: 3963–3969.
151. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. Nat Med 2008; 14: 723–730.
152. Daughters RS, Tuttle DL, Gao W, Ikeda Y, Moseley ML, Ebner TJ, *et al.* RNA gain-of-function in spinocerebellar ataxia type 8. PLoS Genet 2009; 5: e1000600.
153. Gupta RA, Shah N, Wang KC *et al*. "Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis." Nature 2010;464:1071–1076.
154. Prensner JR, Iyer MK, Balbin OA *et al*. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011;29:742–749.
155. Gutschner T, Diederichs S. The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol 2012;9:703–719.
156. Ji P, Diederichs S, Wang W *et al*. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non–small cell lung cancer. Oncogene 2003;22:8031–8041.
157. Yang L, Lin C, Liu W *et al*. NcRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. Cell 2011;147:773–788.
158. Tripathi V, Ellis JD, Shen Z *et al*. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 2010;39:925–938.
159. Huarte M, Guttman M, Feldser D *et al*. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 2010;142:409–419.
160. R. Qin, Z. Chen, Y. Ding, J. Hao, J. Hu, F. Guo "Long non-coding RNA MEG3 inhibits the proliferation of cervical carcinoma cells through the induction of cell cycle arrest and apoptosis" Neoplasma, 60 (2013), pp. 486–492

161. S. Cao, W. Liu, F. Li, W. Zhao, C. Qin "Decreased expression of lncRNA GAS5 predicts a poor prognosis in cervical cancer" Int J Clin Exp Pathol, 7 (2014), pp. 6776–6783
162. Xie L, Hu Z, Wang X *et al*. Expression of long noncoding RNA MALAT1 gene in human nasopharyngeal carcinoma cell lines and its biological significance. Nan Fang Yi Ke Da Xue Xue Bao 2013;33:692–697.
163. Wheeler TM, Leger AJ, Pandey SK *et al*. Targeting nuclear RNA for *in vivo* correction of myotonic dystrophy. Nature 2012;488111–488115.
164. Ren S, Liu Y, Xu W *et al*. Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. J Urol 2013;190:2278–2287.
165. Fritah, Sabrina, Simone P. Niclou, and Francisco Azuaje. "Databases for lncRNAs: A Comparative Evaluation of Emerging Tools." *RNA* 20.11 (2014): 1655–1665. *PMC*. Web. 19 Apr. 2015.
166. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y. (2013) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Research* [Epub ahead of print].
167. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Reczko M, Maragkakis M, Dalamagas TM, Hatzigeorgiou AG. (2012) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res* [Epub ahead of print].
168. P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, and J. S. Mattick, "LncRNAdb: a reference database for long noncoding RNAs," Nucleic Acids Research, vol. 39, no. 1, pp. D146–D151, 2011.
169. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K, Vandesompele J, Mestdagh P. (2012) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* [Epub ahead of print].
170. Niazi F, Valadkhan S. (2012) Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. *RNA* 18(4):825-43.
171. Yang JH, Li JH, Jiang S, Zhou H and Qu LH. "ChIPBase: A database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data." Nucleic Acids Res. 2013, First published online: November 17, 2012.
172. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Reczko M, Maragkakis M, Dalamagas TM, Hatzigeorgiou AG. (2012) DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res* [Epub ahead of print].
173. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. (2012) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* [Epub ahead of print].
174. Qinghua Jiang; Jixuan Wang; Xiaoliang Wu; Rui Ma; Tianjiao Zhang; Shuilin Jin; Zhijie Han; Renjie Tan; Jiajie Peng; Guiyou Liu; Yu Li; Yadong Wang. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. Nucleic Acids Research 2014; doi: 10.1093/nar/gku1173
175. Gong J, Liu W, Zhang J, Miao X, Guo AY. (2015) lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res* 43(Database issue):D181-6.
176. Ma et al. (2014) LncRNAWiki: Harnessing Community Knowledge in Collaborative Curation of Human Long Non-coding RNAs.
177. Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B, Jain S, Sati S, Sengupta S, Sachidanandan C, Raghava GP, Sivasubbu S, Scaria V. (2013) lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database (Oxford)*2013(0), bat034.
178. Jin J, Liu J, Wang H, Wong L, Chua, N. (2013) PLncDB: Plant Long noncoding RNA Database. *Bioinformatics* [Epub ahead of print].
179. Li et al. Nucleic Acids Res. 2014
180. Li, Aimin (2015): ALDB: a domestic-animal long noncoding RNA database.
181. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*. 2009;10(3):155–159.
182. Khachane AN, Harrison PM. Mining mammalian transcript data for functional long non-coding RNAs. *PLoS One*. 2010;5(4)e10316
183. Yun Xiao, Yanling Lv, Hongying Zhao, et al., "Predicting the Functions of Long Noncoding RNAs Using RNA-Seq Based on Bayesian Network," BioMed Research International, vol. 2015, Article ID 839590, 14 pages, 2015. doi:10.1155/2015/839590
184. Hao, Yibin et al. "Prediction of Long Noncoding RNA Functions with Co-Expression Network in Esophageal Squamous Cell Carcinoma." *BMC Cancer* 15 (2015): 168. *PMC*. Web. 20 Apr. 2015.

185. Khachane AN, Harrison PM. Mining mammalian transcript data for functional long non-coding RNAs. *PLoS One*. 2010;5(4)e10316
186. A. Jeggari, D. S. Marks, and E. Larsson, "miRcode: a map of putative microRNA target sites in the long non-coding transcriptome," Bioinformatics, vol. 28, pp. 2062–2063, 2012.
187. A. Jeggari, D. S. Marks, and E. Larsson, "miRcode: a map of putative microRNA target sites in the long non-coding transcriptome," Bioinformatics, vol. 28, pp. 2062–2063, 2012.
188. M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia, "Predicting protein associations with long noncoding RNAs," Nature Methods, vol. 8, no. 6, pp. 444–445, 2011.
189. Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, and Florence Jaffrézic on behalf of The French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Brief Bioinform 2012 : bbs046v1-bbs046.
190. Mathieu Giraud, Jean-Stéphane Varré, Parallel Position Weight Matrices algorithms, Parallel Computing, Volume 37, Issue 8, August 2011, Pages 466-478, ISSN 0167-8191, http://dx.doi.org/10.1016/j.parco.2010.10.001.
191. Xie, Z., Hu, S.H., Blackshaw, S., Zhu, H. and Qian, J. (2009) hPDI: a database of experimental human protein-DNA interactions, **Bioinformatics**.(in press)
192. Yoon, Je-Hyun, Kotb Abdelmohsen, and Myriam Gorospe. "Functional Interactions among microRNAs and Long Noncoding RNAs." *Seminars in cell & developmental biology* 0 (2014): 9–14. *PMC*. Web. 3 Dec. 2015.
193. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number:147470 updated on:28-July-2015 world wide web URL:http://omim.org/entry/147470?search=IGF2&highlight=igf2
194. Online Entrez Gene National Center for Biotechnology Information (NCBI) Gene ID: 2162 updated on 20-Dec-2015: World Wide Web URL:http://www.ncbi.nlm.nih.gov/gene/2162
195.Online Entrez Gene National Center for Biotechnology Information (NCBI) Gene ID:1277 updated on 20-Dec-2015: World Wide Web URL http://www.ncbi.nlm.nih.gov/gene/1277
196. The UniProt Consortium**UniProt: a hub for protein information** Nucleic Acids Res. 43: D204-D212 (2015).
    a. World Wide Web URL: http://www.uniprot.org/uniprot/Q9Y5S9
197. PMID:12624257 "Vertebrate microRNA genes" Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP Science. 299:1540(2003).
198. PMID:15965474 "Identification of hundreds of conserved and nonconserved human microRNAs" Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z Nat Genet. 37:766-770(2005).
199. The UniProt Consortium**UniProt: a hub for protein information** Nucleic Acids Res. 43: D204-D212 (2015). World Wide Web URL: http://www.uniprot.org/uniprot/P98179
200. Usha K. Muppirala, Vasant G. Honavar, Drena Dobbs. Predicting RNA-protein Interactions using Only Sequence Information. *BMC Bioinformatics* 2011, **12**:489
201. Muppirala UK, Lewis BA, Dobbs D (2013) Computational tools for investigating RNA-protein interaction partners. *J comput Sci Syst Biol* **6**:182-187
202. Online Entrez Gene National Center for Biotechnology Information (NCBI) Gene ID: 7057 updated on 13-Dec-2015: World Wide Web URL: http://www.ncbi.nlm.nih.gov/gene/7057
203. Molecular Signature Database (MSigDB) and also Xie et al. (2005, Nature 434)
204. Seppälä, Janne , Novellette: an RNA-sequencing data analysis pipeline for detecting novel transcripts, Master's thesis, Tampere University of Technology, 2013-12-04, http://URN.fi/URN:NBN:fi:tty-201312191514
205. Online Entrez Gene National Center for Biotechnology Information (NCBI) Gene ID: 5919 updated on 10-Jan-2016: World Wide Web URL: http://www.ncbi.nlm.nih.gov/gene/5919
206. Online RNA-seqlopedia, RR032670 (NIH, National Center for Research Resources), University of Oregon. Accessed on 15-Dec-2015: World Wide Web URL: http://rnaseq.uoregon.edu/

9 Appendix

In this appendix, the R script and other files that are used in the data analysis are introduced. The files are available for download at https://github.com/insilicolife/Thesis_Files.git gitHub repository. The details of the files in the repository are explained in the table below.

Files	Description
DEAnalyzer.r	This R script performs normalization, & preprocessing of raw count data, differential expression analysis, enrichment and association analysis for gene expression
DEquartileAnalyzer.r	A function to analyze the differential expression based on quartile.
GSEAcalculater.r	A function that analyses the GSEA enrichment analysis
LincRNASeqHg19.fasta	A fasta sequence file extracted from hg19 genome assembly for the lincRNA sequences discovered by Novellette algorithm.
MIcalculater.r	A mutual information calculation function for the given two datasets.
PWMSoreSerial.r	R script for DNA-protein and RNA-protein motif scanning
corCalculater.r	R script for calculating the correlation between two datasets.
enrichment.r	R function that analyzes the gene-list enrichment analysis.
msigdb	Molecular signature database for enrichment analysis from board institute.
lincRNAControlSeq.fa	A fasta file that is used as a control sequence for protein-DNA interaction significance test.
hPDI	Directory that holds the human protein-DNA interaction motifs in PFM form
53_Novel_Transcript_from_Novellette_algorithm.xlsx	Descriptions of lincRNAs that are identified using Novellette algorithm.
miRanda-3.3a	An open source software package from Memorial Sloan-Kettering Cancer Center, New York for microRNA target

	scanning on the reference RNA using the dynamic programming and thermodynamic approaches.
miRNASeqmiRbaseDB.fasta	A fasta file containing highly differentially expressed miRNA sequence
reverseInteractionInDataFrameToxslx.xlsx	The interaction motif scanning result of RBP that have common binding motifs with the reverse strand promoter regions of highly differentially expressed genes and the lincRNAs.
forwardInteractionInDataFrameToxslx.xlsx	The interaction motif scanning result of RBP that have common binding motifs with the forward strand promoter regions of highly differentially expressed genes and the lincRNAs.
bothStrandInteractionInDataFrameToxslx.xlsx	The interaction motif scanning result of RBP that have common binding motifs with both forward and reverse strand promoter regions of highly differentially expressed genes and the lincRNAs.