

**Insight from RNA-DNA
interaction for regulation of
activity of repeat elements and
genes**

by
Neetesh Chauhan

Under the supervision of
Dr. Vibhor Kumar

Submitted in partial fulfillment of the
requirements for the degree of Master of
Technology, Computational Biology



Center for Computational Biology Indraprastha
Institute of Information Technology - Delhi May,
2022

Certificate

This is to certify that the thesis titled “*Insight from RNA-DNA interaction for regulation of activity of repeat elements and genes*” being submitted by **Neetesh Chauhan** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May,2022

Dr Vibhor Kumar
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgments

This entire thesis was possible due to help from a multitude of people. The biggest contributor is my supervisor, Dr. Vibhor Kumar, without whose help, guidance, and patience this thesis would have been completed. It's through his advice that this work was able to reach this level.

Throughout this work members of the Reggen lab were always there to lend a hand and help me in exploring uncharted areas of knowledge and I would also like to extend my heartfelt gratitude to my teachers at IIT Delhi, my seniors, and my batchmates. I would also like to thank the Department of Biotechnology, the Government of India, for the student fellowship and support to the MTech (Computational Biology) program.

Last but not the least, I express my gratitude to my family members who were always there as a pillar of support.

Abstract

The RNA world is yet to be understood properly, and the same goes for repeat regions as there has been much research going on to understand their regulatory roles. Here, we have analyzed the RNA-DNA interaction and their relationship with the regulation to better understand this unexplored data. We explored the possibility of relationships among different repeat families through the mechanism of RNA-DNA interaction. We also added the dimension of histone modifications to understand the effect of many non-coding RNA through binding to DNA. We explored the different pathways affected by the binding of a few non-coding RNA to DNA. Our analysis highlighted the bias in the DNA binding pattern of GAS5 known to be involved in various diseases, including diabetes mellitus, cancer, and bone disorders. Our analysis also highlighted transcription factors that could be involved in the binding of a few non-coding RNA to DNA.

Contents

1	Introduction	11
1.1	Different regulatory mechanisms in cell	11
1.1.1	Covalent histone modifications	12
1.2	RNA-DNA interactions	14
1.3	Related work	15
1.4	Purpose of this thesis	15
2	Analysis of RNA-DNA interaction	16
2.1	Preliminary Analysis	16
2.2	Overlap with exons, promoters, repeat regions, histone modifications	17
2.2.1	Tools used	17
2.2.2	Overlap with promoter regions	18
2.2.3	Overlap with exon regions	19
2.2.4	Overlap of RNA and DNA with repeat regions	19
2.2.5	Overlap with histone modifications	20
2.3	Detailed Analysis of Repeat Sequences	21
2.3.1	Data Processing	21
2.3.2	Network formation and analysis	22
3	Validation and Biological Inferences	48
3.1	Histone modification plots with exons	48
3.1.1	HEK	49
3.1.2	HFF	53
3.1.3	HUVEC T3d	57
3.1.4	HUVEC T7d	61
3.2	Gene Ontology	65
3.2.1	HEK	65
3.2.2	HFF	66
3.2.3	HUVEC T3d	67
3.2.4	HUVEC T7d	67
3.3	Motif finding using HOMER tool	68
3.3.1	HEK	69

3.3.2	HFF	70
3.3.3	HUVEC T3d	71
3.3.4	HUVEC T7d	72
3.4	LncRNA Literature	73
4	Conclusion	74

List of Figures

1.1	Central Dogma of molecular biology	12
2.1	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HEK cell line, where activation is done through H3K4me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	24
2.2	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HEK cell line, where activation is done through H3K27ac histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	25
2.3	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HEK cell line, where repression is done through H3K27me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	26
2.4	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HEK cell line, where repression is done through H3K9me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	27
2.5	Degree distribution of sub-network for both activating and repressive histone modification of HEK cell line	28
2.6	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HFF cell line, where activation is done through H3K27ac histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	29

2.7	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HFF cell line, where repression is done through H3K27me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	30
2.8	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HFF cell line, where activation is done through H3K4me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	31
2.9	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HFF cell line, where repression is done through H3K9me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	32
2.10	Degree distribution of sub-network for both activating and repressive histone modification of HFF cell line	33
2.11	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T3d cell line, where activation is done through H3K27ac histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	34
2.12	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T3d cell line, where repression is done through H3K27me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	35
2.13	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T3d cell line, where activation is done through H3K4me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	36
2.14	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T3d cell line, where repression is done through H3K9me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	37
2.15	Degree distribution of sub-network for both activating and repressive histone modification of HUVEC T3d cell line	38

2.16	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T7d cell line, where activation is done through H3K27ac histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	39
2.17	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T7d cell line, where repression is done through H3K27me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	40
2.18	Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T7d cell line, where activation is done through H3K4me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	41
2.19	Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T7d cell line, where repression is done through H3K9me3 histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA	42
2.20	Degree distribution of sub-network for both activating and repressive histone modification of HUVEC T7d cell line	43
3.1	Boxplot of number of reads(for activating histone modification H3K27ac) at DNA binding sites of RNA of different genes for HEK cell line . . .	49
3.2	Boxplot of number of reads(for repressive histone modification H3K27me3) at DNA binding sites of RNA of different genes for HEK cell line . . .	50
3.3	Boxplot of number of reads(for activating histone modification H3K4me3) at DNA binding sites of RNA of different genes for HEK cell line . . .	51
3.4	Boxplot of number of reads(for repressive histone modification H3K9me3) at DNA binding sites of RNA of different genes for HEK cell line . . .	52
3.5	Boxplot of number of reads(for activating histone modification H3K27ac) at DNA binding sites of RNA of different genes for HFF cell line . . .	53
3.6	Boxplot of number of reads(for repressive histone modification H3K27me3) at DNA binding sites of RNA of different genes for HFF cell line . . .	54
3.7	Boxplot of number of reads(for activating histone modification H3K4me3) at DNA binding sites of RNA of different genes for HFF cell line . . .	55
3.8	Boxplot of number of reads(for repressive histone modification H3K9me3) at DNA binding sites of RNA of different genes	56
3.9	Boxplot of number of reads(for activating histone modification H3K27ac) at DNA binding sites of RNA of different genes for HUVEC T3d cell line	57

3.10	Boxplot of number of reads(for repressive histone modification H3K27me3) at DNA binding sites of RNA of different genes	58
3.11	Boxplot of number of reads(for activating histone modification H3K4me3) at DNA binding sites of RNA of different genes for HUVEC T3d cell line	59
3.12	Boxplot of number of reads(for repressive histone modification H3K9me3) at DNA binding sites of RNA of different genes for HUVEC T3d cell line	60
3.13	Boxplot of number of reads(for activating histone modification H3K27ac) at DNA binding sites of RNA of different genes for HUVEC T7d cell line	61
3.14	Boxplot of number of reads(for repressive histone modification H3K27me3) at DNA binding sites of RNA of different genes for HUVEC T7d cell line	62
3.15	Boxplot of number of reads(for activating histone modification H3K4me3) at DNA binding sites of RNA of different genes for HUVEC T7d cell line	63
3.16	Boxplot of number of reads(for repressive histone modification H3K9me3) at DNA binding sites of RNA of different genes for HUVEC T7d cell line	64
3.17	Gene Ontology enrichment for DNA binding location for GAS5 RNA in HEK cell line	65
3.18	Gene Ontology enrichment for DNA binding location for COL1A2 RNA in HFF cell line	66
3.19	Gene Ontology enrichment for DNA binding location for THBS1 RNA in HFF cell line	66
3.20	Gene Ontology enrichment for DNA binding location for GAS5 RNA in HFF cell line	67
3.21	Gene Ontology enrichment for DNA binding location for MALAT1 RNA in HUVEC T7d cell line	67
3.22	HOMER results for top ten known motifs of HEK cell line	69
3.23	HOMER results for top ten known motifs of HFF cell line	70
3.24	HOMER results for top ten known motifs of HUVEC T3d cell line	71
3.25	HOMER results for top ten known motifs of HUVEC T7d cell line	72

List of Tables

2.1	HEK cell line threshold value for Reads	22
2.2	HFF cell line threshold value for Reads	23
2.3	HUVEC T3d cell line threshold value for Reads	23
2.4	HUVEC T7d cell line threshold value for Reads	23
2.5	Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HEK cell line	44
2.6	Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HFF cell line	45
2.7	Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HUVEC T3d cell line	46
2.8	Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HUVEC T7d cell line	47

Chapter 1

Introduction

1.1 Different regulatory mechanisms in cell

All organisms are made up of cells. Cells carry out all of our body's functions at a minuscule level which, when collected as a whole maintain our well-being. Cells are responsible for being the building blocks, as well as for the intake of nutrients, their assimilation, excretion, and most importantly, for carrying our genetic material and reproduction(1). Cells are generally divided into prokaryotes and eukaryotes i.e. those without a properly defined nucleus and those with one respectively. It is with the nucleus of a eukaryotic cell that we are concerned, for it contains a vast unexplored RNA world.

Cell cycle is an ordered process through which a single parent cell divides into two daughter cells. This is generally done through Mitosis, in which two identical daughter cells are produced, and Meiosis, in which daughter cells receive only half of their parent's genetic material(2). Cell cycle is generally divided into four stages : G1 phase, S phase, G2 phase, M phase.

In 1953 Watson and Crick explained the structure of DNA double helix using X-ray diffraction data of R. Franklin and M. Wilkins. DNA is the hereditary material in humans and through transcription, it forms RNA which is translated into proteins. DNA normally exists in form of Chromatin and gets into a condensed form during M-phase called a Chromosome. Histones(3) are small basic proteins that form the nucleosome around which DNA is wound and it helps in both DNA replication and regulation of genes.

However when transcribed many eukaryotic genes contain non-coding regions in-between the coding regions called introns and exons respectively. These regions are joined together to create a mRNA for that specific gene(4). In an mRNA however there may be regions that don't code for proteins as in Human beings there are almost

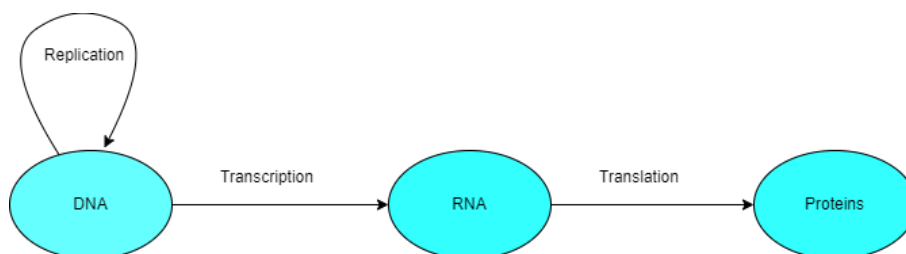


Figure 1.1: Central Dogma of molecular biology

21,000 protein coding genes.

Gene expression is controlled in organisms via positive regulation, negative regulation and co-regulation. Prokaryotes regulate gene expression and cell via operons but eukaryotes use different methods for the same. Broadly eukaryotes regulate through transcriptional control, translational control(5), post-translational control . Translational control refers to, if and how many mRNA are translated into proteins and post-translational control describes if the protein formed is active/inactive or stable/degraded. In context of this thesis we are primarily concerned with transcriptional changes.

These are the changes occurring at stage of transcription to control gene expression(5). This can be achieved by affecting chromatin structure through covalent histone modifications and nucleosome remodelling.

1.1.1 Covalent histone modifications

These refer to changes at the time of replication and transcription. The N-terminal tail of histone proteins can be introduced to various post-translational covalent modifications(6), majorly including acetylation, methylation, phosphorylation.

Acetylation

Histone Acetyltransferases (HATs) add lysine aminon acid to histone tails. This diminishes the positive charge of histone an affinity for DNA and interaction between nucleosomes(7). This acetylation is reversible through the action of histone deacetylases (HDACs) which work through removing acetyl groups. Generally histones is euchromatin are acetylated.

Mehylation

Histone methyltransferases(HMTs) are responsible for histone methylation and histone demethyltransferases (HDMs) work antagonistically to them. Lysine and argi-

nine groups undergo methylation at N-terminal histone tails. They may be mono- or dimethyl for arginine and mono-, di- or trimethyl groups for lysine. Methylation of different histones like H3 and H4 are linked with both activation and repression(7). Methylation of lysine at fourth position of H3 histone (H3K4) is linked with transcriptional activation and methylation of lysine at ninth position of H3 histone too (H3K9) is linked with repression, gene silencing. Histone methylation poses no change to structure of histones.

Phosphorylation

This takes place generally on Serine, Threonine and Tyrosine residues. These are regulated by addition of phosphate groups by kinases and their removal by phosphatases(7). These are mostly effective in cell signalling.

Gene expression is regulated by networks of sequence-specific transcription factors that bind to DNA, and these transcription factors influence regulatory elements, which further pass on the response to transcriptional machinery. Regulatory elements are cis-acting sequences that work with trans-acting transcription factors(6). These are broadly divided into proximal promoter elements, which are located near the transcription start site, including the CAAT box, the GC box, and long-range regulatory elements that work mostly in eukaryotes and include enhancers and silencers.

Those mRNA that doesn't code for a protein or Noncoding RNA (ncRNA) were thought of as junk DNA with very small or no functional importance. However, recent research has pointed out that these parts of noncoding mRNA have a high value in genome regulation which is still not fully understood. LncRNA (long noncoding RNA), which is referred to as transcriptional noise, is actually regulating gene expression in transcriptional, post-transcriptional, or translational phases. In comparison to proteins, lncRNAs(8) are seen as less expensive and can act locally without the need to be transported from cell to nucleus like with proteins. lncRNAs are highly specific in nature, with novel ones getting discovered constantly. Through multi-omics studies, it has been observed that they are also participating in various diseases and several types of cancer. While we can predict short ncRNA but the prediction of pathway and function of lncRNA has yet to be explored properly.

LncRNAs interact with DNA, RNA and proteins mostly through cis-regulation, i.e. proximal protein-coding genes. They can also regulate gene expression through trans-regulation through complex pathways. Some functions of lncRNAs are recruiters for chromatin remodeling protein factors, tethers, scaffolds, decoys, and repress gene activity, coregulators, mRNA processing, translation, encoding peptides(8).

1.2 RNA-DNA interactions

There are several sequencing methods for mapping RNA-DNA interactions via different approaches. ChIRP-seq, CHART-seq, and RAP-seq are done by one RNA versus the genome methodology. ChIRP and RAP do not need knowledge of the RNA of the target. For them, probes that encompass the entire RNA are employed. This increases the likelihood of expressing the whole length of fragmented RNA. While MARGI, ChAR-seq, and GRID-seq are done via all RNAs versus the genome approach(9). RNA and DNA are linked by proximity ligation using a bivalent and biotinylated linker having both ssRNA at one end and dsDNA on another end. In ChAR-seq and GRID-seq, these steps are performed in intact nuclei while in MARGI, RNA-DNA protein complexes are fixed on solid surface of streptavidin beads(10).

Chromatin-associated RNAs are useful along with DNA and histone modifications for providing epigenomic information. From MARGI, a new technology called iMRAGI was innovated(11), which reduces the number of input cells by 100-fold to 5million cells. They used this technology to map RNA-DNA interactions in HEK (human embryonic kidney) and HFF (human foreskin fibroblast) cells.

LncRNAs interact with chromatin with different mechanisms like RNA-DNA binding through protein complexes, triplexes through direct RNA-DNA hybridization, RNA binding to R-loops, and co-transcriptional RNA-RNA interactions(12). They investigated two goals, identification of lncRNAs that function by interacting directly with their targets, and predicted the most plausible interaction may it be co-transcriptional or post-transcriptional for identified lncRNAs. The AntiSense Search Approach Tool was used in this study.

To know the exact functions of lncRNAs, it was necessary to know where they are acting. This requires the development of technology similar to CHIP for lncRNAs(13). For this purpose, CHART (Capture Hybridization analysis of RNA targets) was developed, which can be used to map genomic binding sites for local RNAs. Analogous to CHIP, this technique helped in the identification of endogenous RNA.

CREs (cis-regulatory elements) are non-coding regions of DNA that help in the regulation of genes, and although there is much data available, there is a dire need to properly identify CREs genome wide(14). They have used machine-learning based approaches for CRE predictions using unsupervised learning methods, supervised methods, and deep learning methods. Several tools that help in finding lncRNA are CPC2, lncRNA-MFDL, lncScore, LncADeep, DeepLNC, LncRNAnet, lncRScan-SVM, longdist, and LncFinder.

1.3 Related work

Different methods of RNA-chromatin and RNA-DNA interactions have been discovered. They can interact in a cis manner, near their transcription sites or in trans, on regions in separate chromosomes. Some lncRNAs interact with few genomic locations, while some are promiscuous in their nature and interact at multiple locations on the genome(9). Some act as activators and others as repressors of gene expression.

For characterizing caRNAs and their interaction location, iMARGGI was developed, with the main difference being that ligation is carried out in situ. They put forward an RNA-poise model which can explain trans RNA-DNA interactions. Firstly, trans interactions could occur by caRNA targeting specific genome sequences, which can be arbitrated by tethers, or due to spatial proximity of genes in 3D space, it could bring nascent transcripts of a gene in close contact with another gene(11). This RNA poise model combines both submodels, providing how RNA and DNA interact.

LncRNAs are known to affect cellular and organismal homeostasis, but their regulation in affecting diseases and cellular complications is being explored. Various gene expressions that are affected by lncRNAs also are active in metabolism, cancer, and cell death. Regulation of lncRNA are known to have an effect on breast cancer, bladder cancer, Huntington's disease, Parkinson's disease, Alzheimer's disease, and cardiovascular disease(15). Some examples of such lncRNAs are H19, XIST, MALAT1, SNHG16, TUG1, UCA1, TINCR, MEG3, BACE1.

A sure hypothesis for all types and functions of RNA-DNA interaction is still missing. Apart from regulation of gene expression at transcriptional and post-transcriptional level, lncRNAs are also surfacing as catalytic enzymes. The same transcript can be non-coding or coding depending and this needs to be further investigated to study gene expression(8). There is still a wide gap between detection of noncoding transcripts, unravelling their mechanism and their impact in disease pathogenesis(16).

1.4 Purpose of this thesis

The purposes of this thesis is to estimate the proportion of cis and trans effect in RNA-DNA, and to know repressive or activating effect of histone modifications and if we can use RNA-DNA interaction to study regulatory effects on repeat elements

Chapter 2

Analysis of RNA-DNA interaction

2.1 Preliminary Analysis

RNA-chromatin interactions are an important part of transcriptional regulation of genes and transposable elements. Chromatin-associated RNAs (caRNAs) are thought of as a new side of the epigenome. caRNA interaction with chromatin are necessary for various cellular and molecular functions like X chromosome silencing, homology-directed repair of telomeres. Chromatin interactions can work as biomarkers, or regulated via CRISPR therapy. Chromatin interactions can regulate gene expression by bringing distal regulatory elements, to close spatial proximity of promoters. For the data collection, we downloaded the iMARGI file for HEK cell line and for eQTL data we used Gtex portal link "<https://gtexportal.org/home/datasets>". In here we searched for HEK cell line eQTL data and we used single tissue cis-eQTL data.

We processed the HEK iMARGI file and the eQTL file using pgltools. The intersection of RNA-DNA and Chromatin-Chromatin interaction data results in 1233 loci. The intersection of these 1233 loci with the processed eQTL kidney-specific tissue results in the 32 genes that were found in common.

We got one match in chromosome 10, three matched in chromosome 11, four matches in chromosome 12, four matches in chromosome 14, two matches in chromosome 15, ten matches in chromosome 2, one match in chromosome 3 and 5, three matches in chromosome 8 and one match in chromosome 9. In most of the genes that we get from the final interaction we found most of them are also expressing in the HEK cell line that is in the kidney. The presence of eQTL at the overlapping regions of RNA-DNA and chromatin interaction indicates that RNA could be involved in the regulation of the respective phenotypes via chromatin folding

2.2 Overlap with exons, promoters, repeat regions, histone modifications

In this section, interaction of iMARGI files was checked with different datasets through various tools in LINUX terminal. Most of the tools used are from Bedtools(17), which were first released in 2009. They were developed because fast, flexible tools were required to compare large sets of genomic features. The existing tools were either too slow or the way they presented or computed their results wasn't efficient. The web based tools were also unmanageable and this lead to producing Bedtools from scratch. The tools work fast enough to finish in seconds even for large datasets and even allow control on how the output is produced. Mostly used data format for representation of genome features are BED (Browser Extensible Data) and GFF (General Feature Format) formats. While initially used for BED format, Bedtools also usage on GFF and VCF files. This is extremely helpful because existing annotations can be retrieved in form of BED and GFF format from UCSC genome browser or from Ensemble Genome Browser.

2.2.1 Tools used

Bedtools intersect

This tool allows us to check if in two different sets of genomic features if there are any overlapping regions or same features. It gives fine control over how those overlapping regions are reported and this tool can take input from both BAM and BED/GFF/VCF files. It's usage command is

```
bedtools intersect [OPTIONS] -a < FILE > -b < FILE1, FILE2, ..., FILEN >
```

Bedtools map

This tool allows us to map for overlapping features in one file onto features in another file and apply operations/statistics onto those features. It can be used to calculate sum, mean, median, max, min for all features that overlap in an interval. Even multiple operations can be used at same time and by default it computes sum. It's usage command is

```
bedtools map [OPTIONS] -a < bed/gff/vcf > -b < bed/gff/vcf >
```

Liftover

LiftOver is a tool by UCSC and is used to bring all genomic coordinates to the same assemblies. Mostly used in situations where we have coordinates for a particular genome, and we want to convert it to another genome for the same species, like bringing all data to hg38 build from hg18 and hg19. It can also be used to convert

build of dbSNP rs number and for converting both genome positions and dbSNP rs over different versions. UCSC liftOver tool can be used to lift files in BED format. Some coordinates that are unlifted are saved in an unlifted.bed file.

```
liftOver input.bed hg18ToHg19.over.chain.gz output.bed unlifted.bed
```

wigToBigWig

BigWig format is helpful when displaying dense, continuous data as a graph in the UCSC genome browser. Using this command, bigWig files are created from wig type files. These files are presented in the format of indexed binary. Because only those regions that are required to display a region are transferred to the genome browser server. This is helpful with large datasets as it provides a faster display performance. BigWig files can also be created from bedGraph files.

```
wigToBigWig input.wig chrom.sizes output.bw
```

bigWigToBedGraph

BedGraph format helps in display of continuous value data in a track format. This type of data is helpful while visualising probability scores and transcriptomic data. Since, bigWig files are in indexed binary format it becomes difficult to extract any data from them. Converting them into bedGraph files helps in this problem. UCSC also provides this tool.

```
bigWigToBedGraph input.bw output.bedgraph
```

2.2.2 Overlap with promoter regions

First of all, iMARGI files were downloaded from NCBI GEO of cell lines, HEK(11), HFF(11), HUVEC T3d(18), and HUVEC T7d(18). All these cell lines were present in bedGraph format with the 1,2,3 and 9 columns as genomic coordinates of RNA and 4,5,6,10 columns as genomic coordinates of DNA. Another column of numbers was added to index the corresponding RNA and DNA with each other, and afterward, separate .bedfiles were created for RNA, using columns 1,2,3, and for DNA using columns 4,5,6 along with their respective indices as the fourth column.

Rows with negative values or zeroes were removed, and then using the DNA .bed file, sorting was done because it is imperative that all chromosome locations should be sorted before using any bedtools. Afterwards *bedtools intersect* tool was used with the DNA file of all cell lines and the promoter regions, which were sorted beforehand. This gives the output file, and rows with 0 overlaps were removed to give resultant file for all four cell lines.

2.2.3 Overlap with exon regions

The exon file was downloaded from GENCODE for the hg38 version of humans. However, this GTF file contained other information as well, and to select only the required data; we made another dataset in .bed format. From this file using *grep* command, all the terms containing ‘exons’ were processed in a file containing their chromosomal locations, transcript id, gene type, and gene name. This is the file containing all information pertaining to exons.

From the RNA file in section 2.2.2, we remove all the negative and zero values and sort this according to the chromosomal locations. Afterwards, we used the *bedtools intersect* tool with the RNA location files of all four cell lines and the exonic locations. All those locations which didn’t yield any output in intersections were removed, and we had datasets containing RNA locations corresponding to their locations for exons, their type, along with their ID.

2.2.4 Overlap of RNA and DNA with repeat regions

The UCSC Repeat browser contains a full set of human repeat reference sequences that were derived from the RepeatMasker program. It provides filtering for DNA sequences with interspersed repeats and low complexity DNA sequences. It gives a detailed annotation for repeats that are found in query sequences along with a modified version where annotated repeats have been masked for query sequences. At present, 56 percent of human genomic sequences have been identified through this program. The comparisons are made by several popular search engines like nhmmer, and RMBlast. For repeat sequences, it also uses curated libraries and supports Dfam and Repbase, which the Genetic Information Research Institute provides.

In addition to this, UCSC Repeat Browser can also be used to map the human genome tracks, get an alignment from the human genome, and can show all of this as a fully-fledged interface to handle repeat elements. There are also processed tracks for a variety of open-source datasets related to repeat elements like ChIP-seq datasets for KZNFs. This track can display around ten different classes for repeats.

- Short interspersed nuclear elements (SINE)
- Long interspersed nuclear elements (LINE)
- Long terminal repeat elements (LTR), like retroposons
- DNA repeat elements (DNA)
- Simple repeats (micro-satellites)
- Low complexity repeats

- Satellite repeats
- RNA repeats (including RNA, tRNA, rRNA, snRNA, scRNA, srpRNA)
- Other repeats, which has class like RC (Rolling Circle)
- Unknown

From RepeatMasker, we download the location of a list of repeat elements containing their chromosomal location and their Repeat name, and Repeat class. RepeatMasker also contains information about the Repeat family, but it would be too broad of a spectrum, so we didn't use it. Now, this file is processed by sorting it according to the chromosomal locations and removing any rows with non-positive or bad data.

From section 2.2.2, we take the files containing information about RNA and DNA locations, and after sorting, these are ready for processing. We take the file containing information on chromosomal location for repeat name and repeat the class, and we use *bedtools intersect* to intersect it with our processed files possessing RNA information for HEK, HFF, HUVEC T3d, and HUVEC T7d cell lines. We repeat the same process for DNA files of all cell lines, so now we possess details on intersecting data for repeats with RNA and DNA locations.

2.2.5 Overlap with histone modifications

Data collection

All these files were downloaded in bigWig or Wig format.

HEK ChIP-seq data were downloaded from four histone modifications, H3K4me3(19), H3K27ac(19), which are activating histone modifications, and H3K9me3(20), H3K27me3(21), which are repressive in nature from GEO.

HFF ChIP-seq was downloaded from repressive histone modifications H3K9me3(22), H3K27me3(22) and for activating histone modifications H3K4me3(22), H3K27ac(22) for the HFF cell line.

HUVEC For both HUVEC T3d and HUVEC T7d, we use the same data of ChIP-seq for repressive histone modifications H3K9me3(23), H3K27me3(23), and activating histone modifications H3K4me3(24), H3K27ac(23) from GEO.

Data Processing

From section 2.2.2, files containing information on DNA locations for all cell lines are used. To the coordinates, we have added and subtracted 400 base pairs to get a window of 800 base pairs, and we get a four-column file.

(Chromosome number) (start location - 400) (start location + 400) (Index)

The histone files, which were extracted in the form of bigWig files, were first all unzipped, and then they were converted to wig format using the command `wigToBigWig`. These were once again converted to bedGraph format using the command `bigWigToBedGraph`. We are doing this because all these files are either of hg19 or hg18 genome build and the RNA-DNA interaction files we have (iMARGI files) are of hg38 build. So we will have to use the *liftOver* tool. However *liftOver* tool is used with bedGraph files, and that's why once we get all histone modification files, both of activation and repression, we use the *liftOver* tool to convert them to that of the hg38 genome build; we sort them according to chromosomal location and convert them to .bed file format.

Now, we use the processed DNA interaction files, and along with processed histone files in bed format for all cell lines, we run the *bedtools map* command. We use the default functionality of this tool which gives us a sum of all the reads that are present in that specific 800 base-pair windows of DNA coordinates. We remove the negative values, and our output is for HEK, HFF, HUVEC T3d, and HUVEC T7d cell lines; we have histone reads corresponding to their DNA coordinates.

2.3 Detailed Analysis of Repeat Sequences

To do an analysis of repeat elements, we have done a network analysis of the data. For this purpose, we have used Networkx, a python language package, to create, interact and analyze the structure and function of complex networks. We can represent networks in the form of nodes and edges and store these to use for later analysis or even convert them into different file formats for analysis in different software. We can make different types of networks, be it in a circular, planar, shell, circular, spring, or random layout.

2.3.1 Data Processing

These steps are performed for all cell lines HEK, HFF, HUVEC T3d, and HUVEC T7d. The DNA files intersected with repeat elements are used, and only the DNA coordinates of cell lines, index numbers, and repeat names are processed in a file using the *awk* command. Similarly, the RNA coordinates intersected with repeats are processed so that they also contain only the RNA coordinates, index number,

and repeat names. Next, using the index number, we match the DNA and their corresponding RNA as was done to keep track of those RNA and DNA locations that were in the same row. We process the DNA coordinates along with the index number and the repeat names that are lying on the corresponding DNA and RNA. Now, we have output files containing the DNA locations of cell lines, their row index number, and the DNA and RNA repeat name.

Next, we used the histone modification files for all cell lines from section 2.2.5 and further processed it to remove blank spaces and take only the columns containing reads value and index. These files are sorted according to the index. Now, we use the DNA-RNA repeat names file and the Reads file for all histone modifications and matches this according to the index values. Next, we remove the unnecessary columns so that it contains only repeat names of the DNA-RNA pair, and now we use python libraries to calculate the median value of Reads for every single repeat name. Here every DNA-RNA pair is treated as one and this enables us to get median values for all the reads of every unique DNA-RNA pair. We use these files and form a network, as explained in the next section.

2.3.2 Network formation and analysis

To construct a network, we used various python libraries like pandas, matplotlib, and networkx. These files were further processed, and prefix ‘DNA’ and ‘RNA’ was added to the DNA and RNA columns, respectively. This is to differentiate as to which nodes originate from DNA and which are from RNA, and the third column was of Read values. This was done for all histone modifications H3K4me3, H3K27ac, which are activating, and H3K9me3, H3K27me3, which are repressing in nature. Since it is a very big file, we put a threshold value to select only those rows that have a higher reads value . From these many rows, we construct a directed network in Networkx, and then we export it in the form of a gml file for further analysis in Cytoscape.

These are the threshold values used for constructing a dense network for analysis and sparse network for visual representation.

Table 2.1: HEK cell line threshold value for Reads

	Dense Network	Sparse Network
H3K27ac	2500	20000
H3K27me3	1500	17000
H3K4me3	100	7000
H3K9me3	120	300

Table 2.2: HFF cell line threshold value for Reads

	Dense Network	Sparse Network
H3K27ac	1000	2300
H3K27me3	800	1600
H3K4me3	2100	9000
H3K9me3	279	580

Table 2.3: HUVEC T3d cell line threshold value for Reads

	Dense Network	Sparse Network
H3K27ac	1100	5000
H3K27me3	300	1000
H3K4me3	37000	52000
H3K9me3	550	350

Table 2.4: HUVEC T7d cell line threshold value for Reads

	Dense Network	Sparse Network
H3K27ac	1000	5000
H3K27me3	300	1000
H3K4me3	35000	50000
H3K9me3	330	520

HEK

Networks and degree distribution plots were made for all histone modifications, H3K4ME3, H3K9me3, H3K27ac, H3K27me3. All the degree distribution plots show small world nature, this further validates our findings as biological based networks show small world nature.

H3K4me3 The denser network has 3 nodes that act as hubs, tRNA, LSU, and ERVL. HUERS also is connected to many nodes. However we observed, that all these repeat nodes are originating from DNA repeat name regions. From the sparser network, we can also detect presence of hubs.

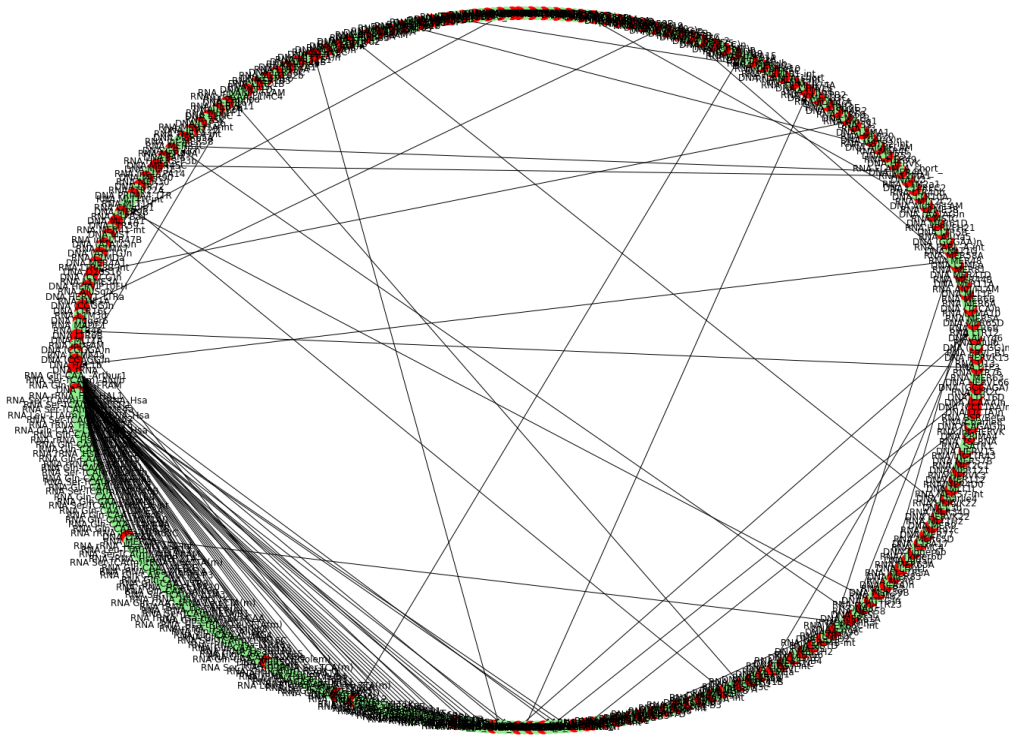


Figure 2.1: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HEK cell line, where activation is done through **H3K4me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K27ac The denser network has more of a mish-mash structure, with very few nodes acting as hubs, which are tRNA and ERVL. Even these nodes are linked to DNA repeat name regions. In the sparse network, we can observe a highly linked network.

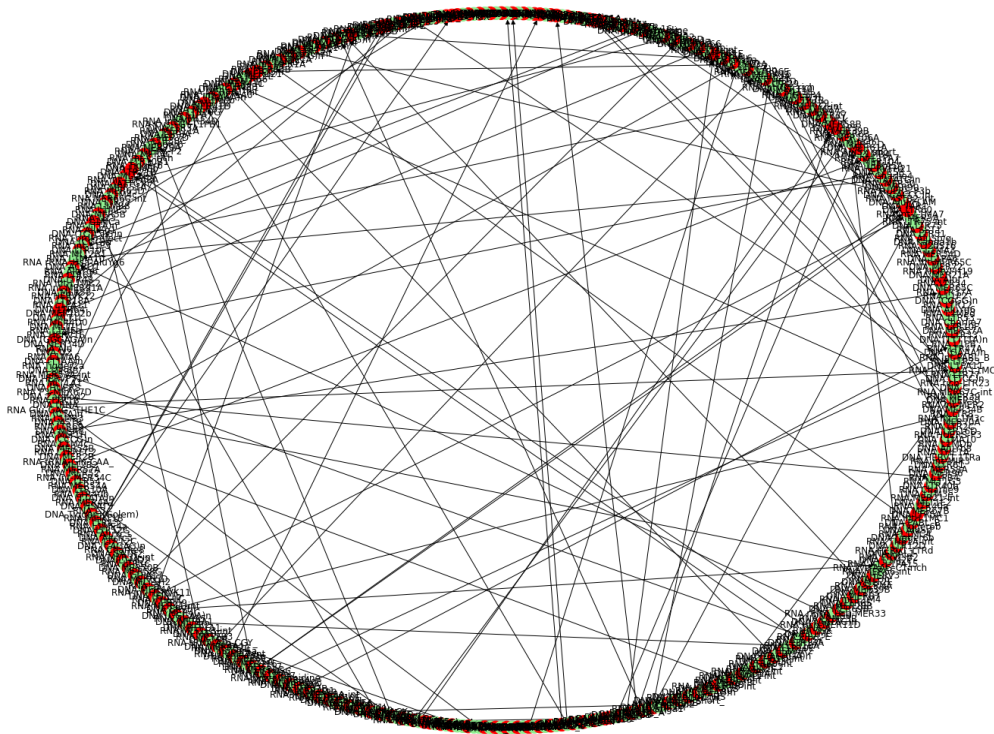


Figure 2.2: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HEK cell line, where activation is done through **H3K27ac** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K27me3 The denser network has 4 nodes that act as hubs tRNA, LSU, ERVL, and HERVL. We can observe that all these repeat nodes are originating from DNA repeat name regions. From the sparser network, we can also detect presence of hubs.

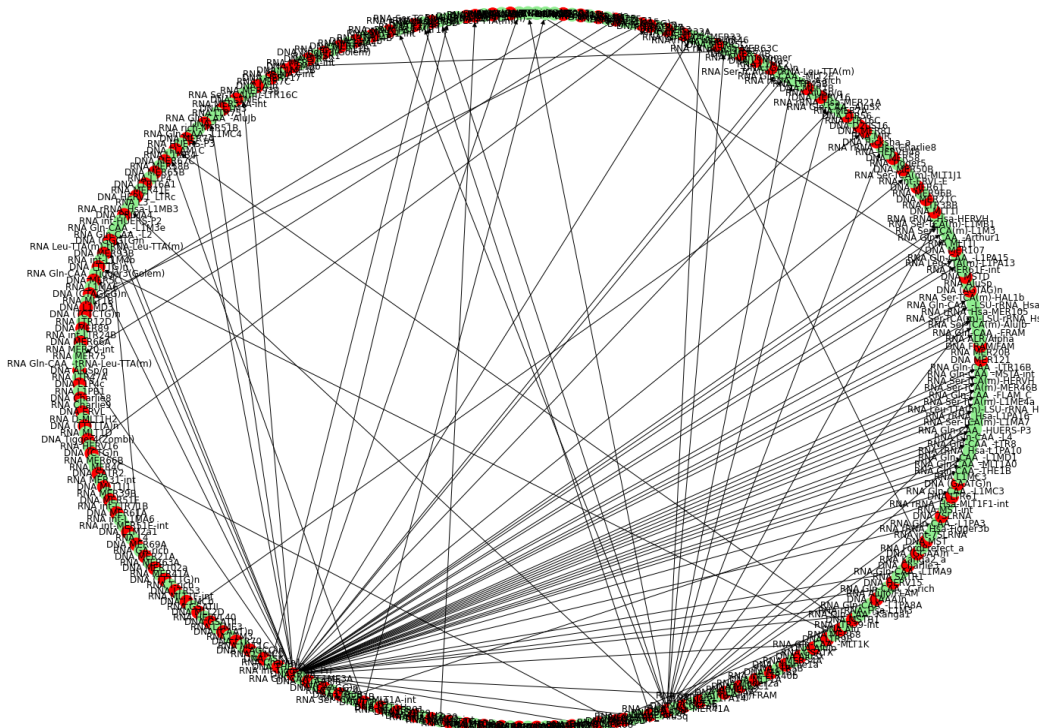


Figure 2.3: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HEK cell line, where repression is done through **H3K27me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K9me3 The denser network has 4 nodes that act as hubs, tRNA which is the most connected to other nodes and can be observed even from the sparser network. Other hub regions are ERVL, HERVL, and LSU. Even in this case, all nodes are related to DNA repeat name regions. In the sparse network, we can clearly detect hubs.

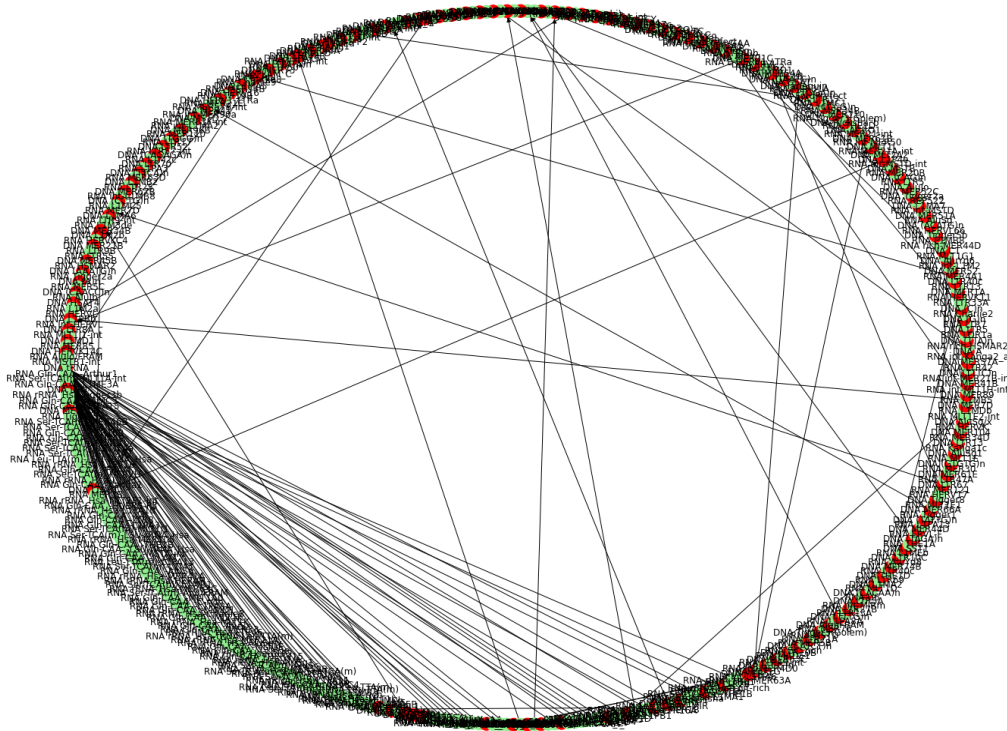
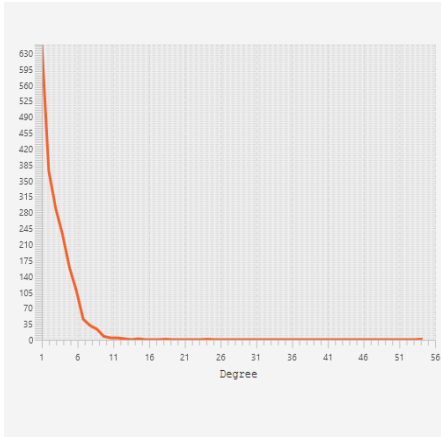
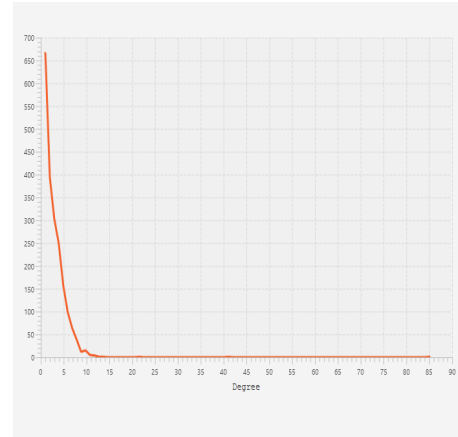


Figure 2.4: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HEK cell line, where repression is done through **H3K9me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

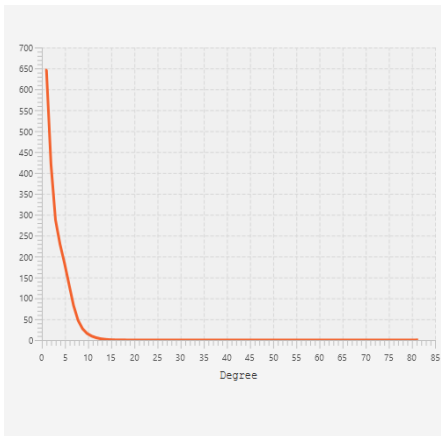
These are the degree distribution plots(25) for histone modification of both activation and repression. There is no significant difference between them. However, these do show a small world nature as is observed in biological network which validates them.



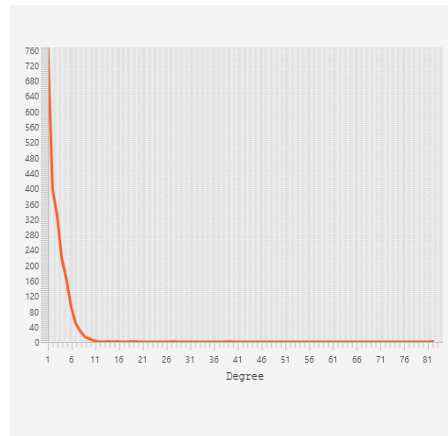
(a) Activating histone modification **H3K27ac**



(b) Activating histone modification **H3K4me3**



(c) Repressive histone modification **H3K27me3**



(d) Repressive histone modification **H3K9me3**

Figure 2.5: Degree distribution of sub-network for both activating and repressive histone modification of HEK cell line

HFF

Networks and degree distribution plots were made for all histone modifications, H3K4ME3, H3K9me3, H3K27ac, H3K27me3. All the degree distribution plots show small world nature, this further validates our findings as biological based networks show small world nature.

H3K27ac The denser network has 2 nodes that act as hubs ERVL, and tRNA. HERVK3 also is connected to many nodes. However we observed, that all these repeat nodes are originating from DNA repeat name regions. From the sparser network, we can also detect presence of a few hubs.

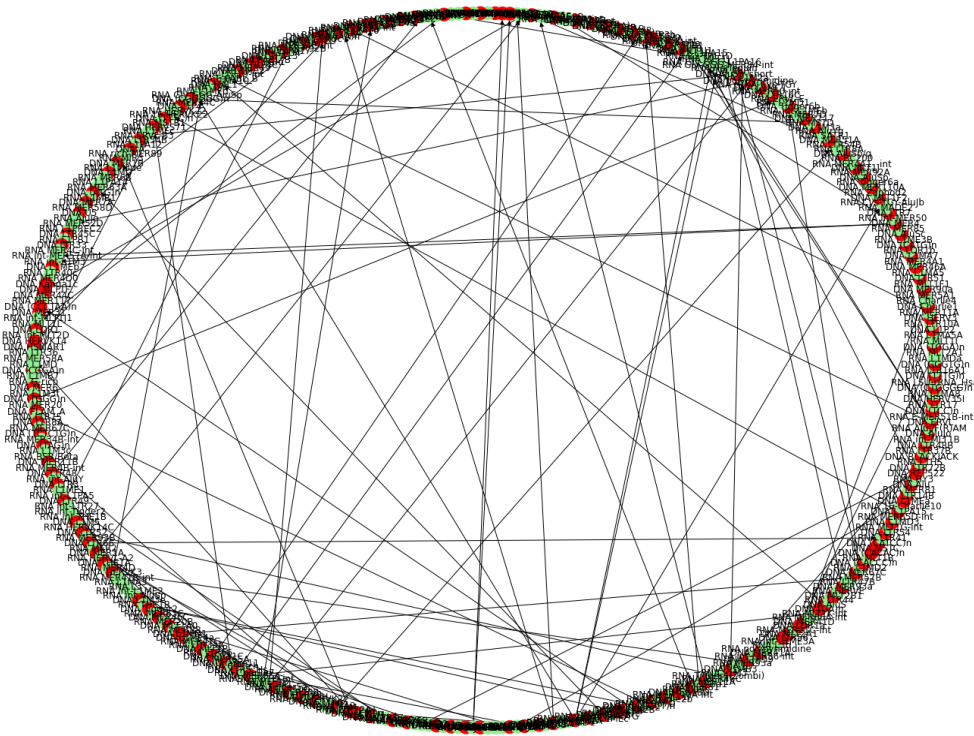


Figure 2.6: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HFF cell line, where activation is done through **H3K27ac** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K27me3 The denser network has 4 nodes that act as hubs MER51A, tRNA, MER21B, and ERVL. All these repeat nodes are originating from DNA repeat name regions. From the sparser network, we can also detect presence of less hubs, and it is more well-connected graph.

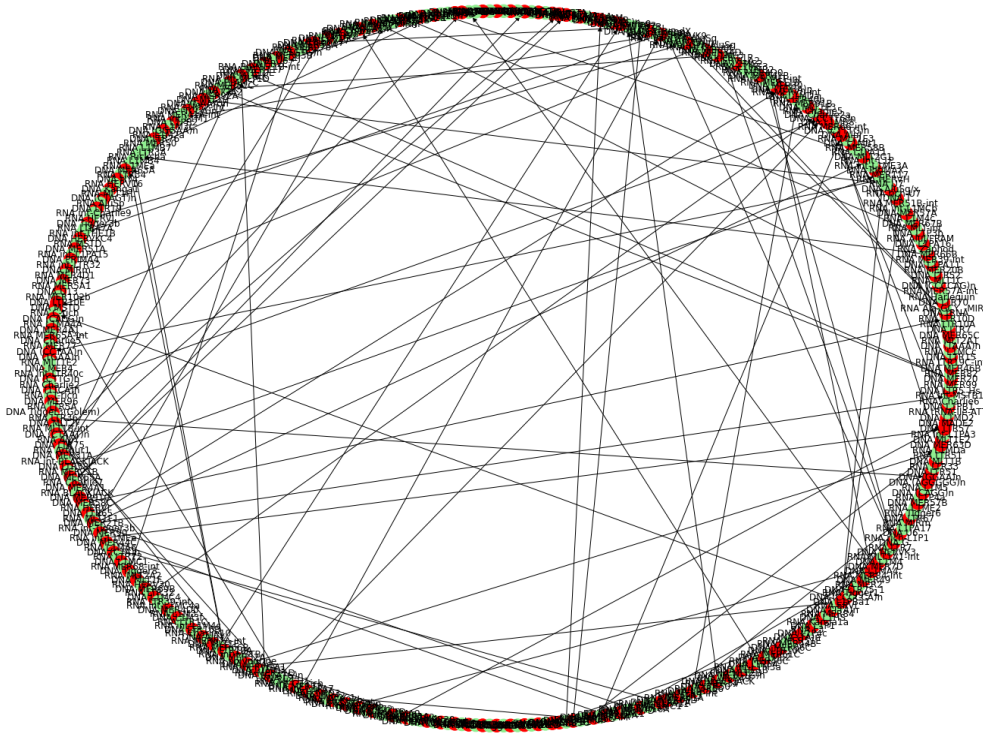


Figure 2.7: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HFF cell line, where repression is done through **H3K27me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K4me3 The denser network has 2 nodes that act as hubs ERVL, and tRNA. MLT1H also is connected to many nodes. However we observed, that all these repeat nodes are originating from DNA repeat name regions. From the sparser network, we can observe few hub regions.

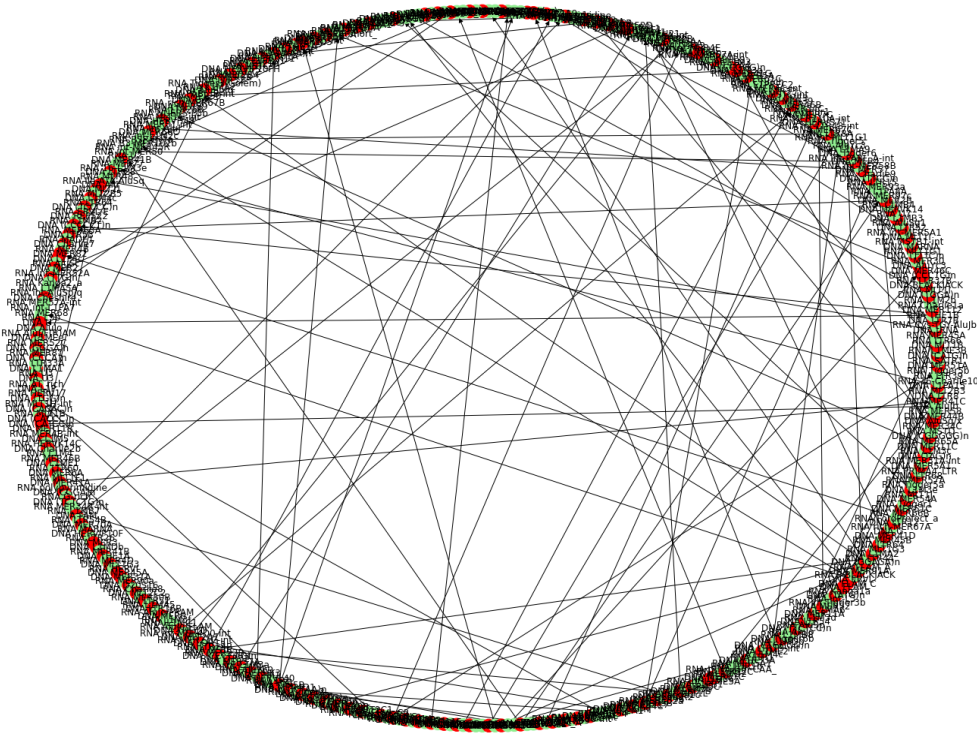


Figure 2.8: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HFF cell line, where activation is done through **H3K4me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K9me3 The denser network has 4 nodes that act as hubs ERVL, ALR/Alpha, HERVL, and BSR/Beta. HUERS repeat is also connected with many nodes. All these repeat nodes are originating from DNA repeat name regions. From the sparser network, we can also detect presence of hub regions.

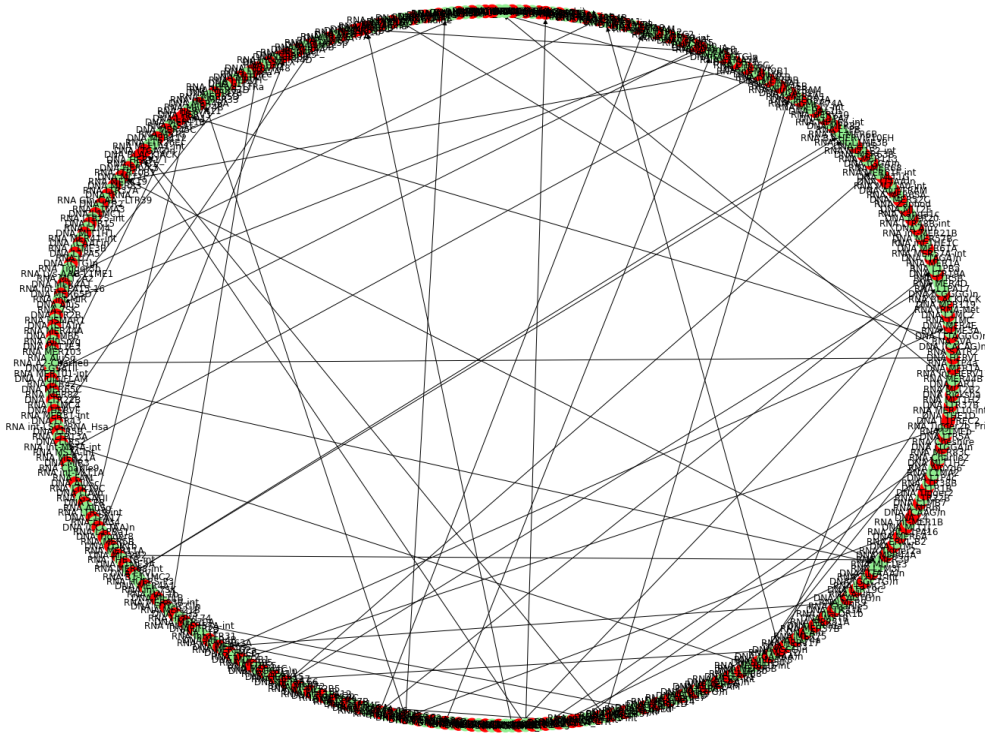
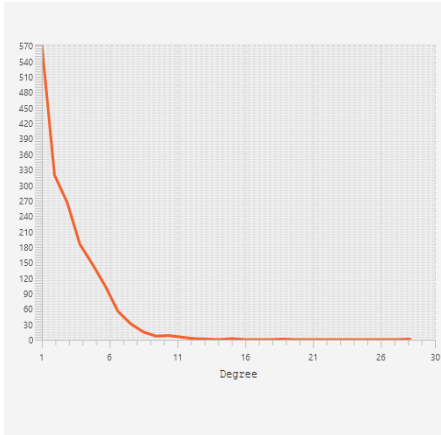
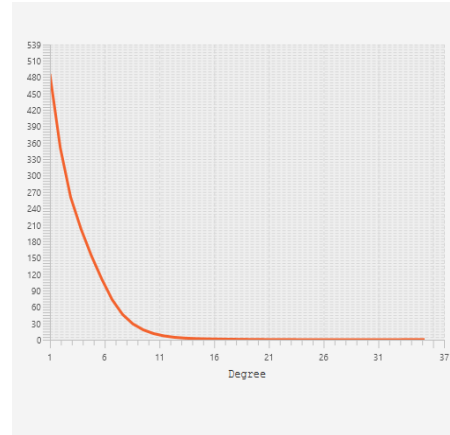


Figure 2.9: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HFF cell line, where repression is done through **H3K9me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

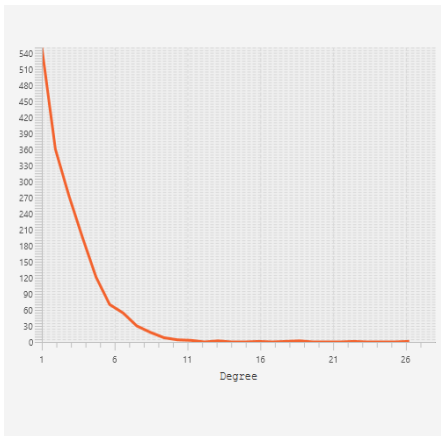
These are the degree distribution(25) plots for histone modification of both activation and repression. There is no significant difference between them. However, these do show a small world nature as is observed in biological network which validates them.



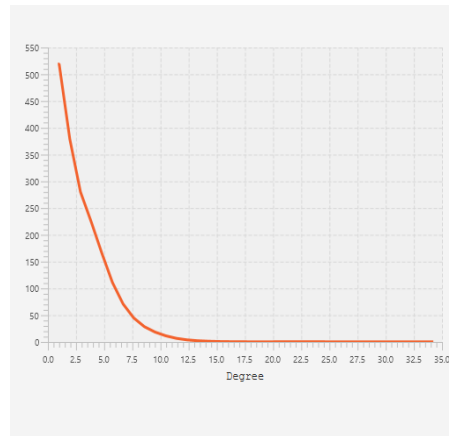
(a) Activating histone modification **H3K27ac**



(b) Activating histone modification **H3K4me3**



(c) Repressive histone modification **H3K27me3**



(d) Repressive histone modification **H3K9me3**

Figure 2.10: Degree distribution of sub-network for both activating and repressive histone modification of HFF cell line

HUVEC T3d

Networks and degree distribution plots were made for all histone modifications, H3K4ME3, H3K9me3, H3K27ac, H3K27me3. All the degree distribution plots show small world nature, this further validates our findings as biological based networks show small world nature.

H3K27ac The denser network has 3 nodes that act as hubs LTR5A, MSTB, and MER4A1. All these repeat nodes are originating from DNA repeat name regions. From the sparse network, hubs can be observed.

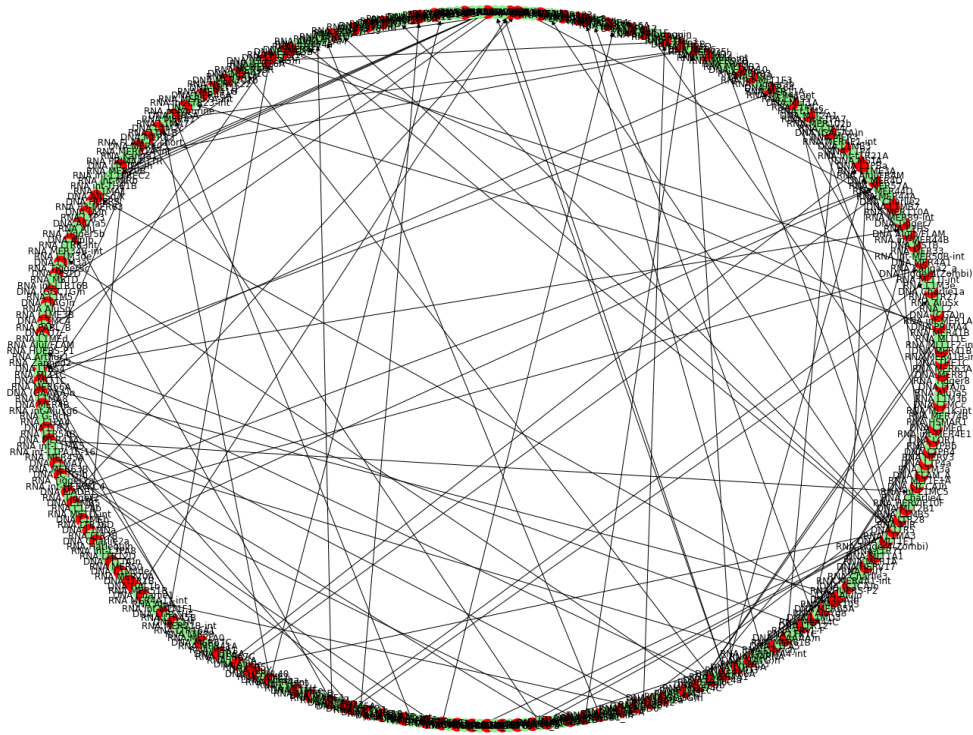


Figure 2.11: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T3d cell line, where activation is done through **H3K27ac** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K27me3 The denser network has 3 nodes that act as hubs ERVL, MER51B, and MER66B. All these repeat nodes are originating from DNA repeat name regions. From the sparse network, we can detect these highly connected hubs.

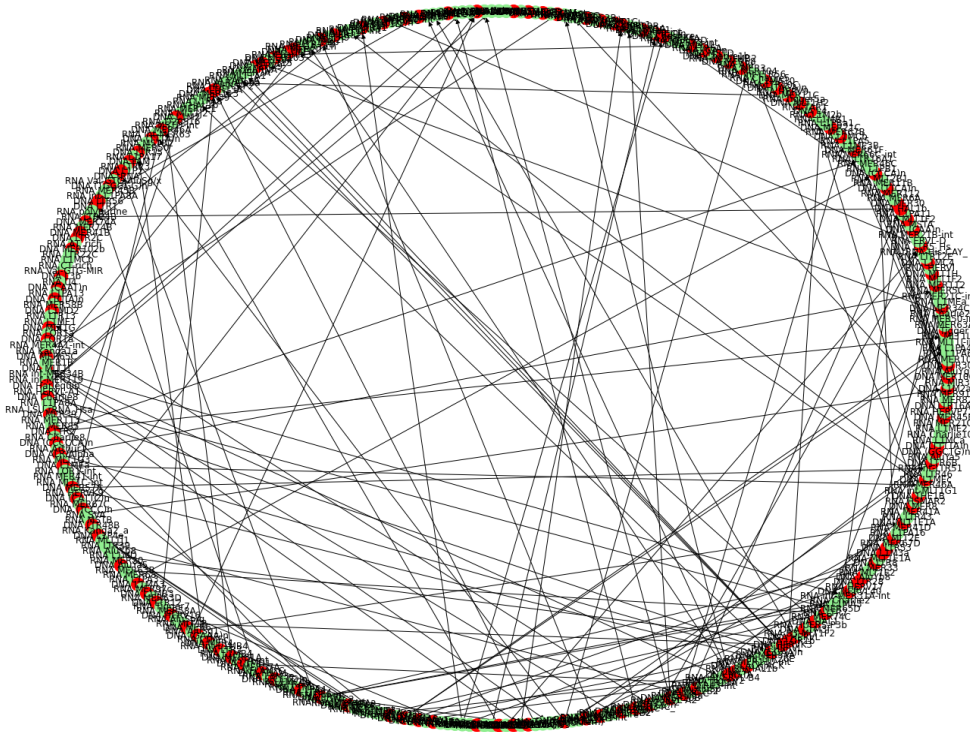


Figure 2.12: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T3d cell line, where repression is done through **H3K27me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K4me3 The denser network has 4 nodes that act as hubs L1ME3B, AluJo, AluJb, and AluSx. All these repeat nodes are originating from DNA repeat name regions. From the sparse network, we can even visually observe these highly connected hubs.

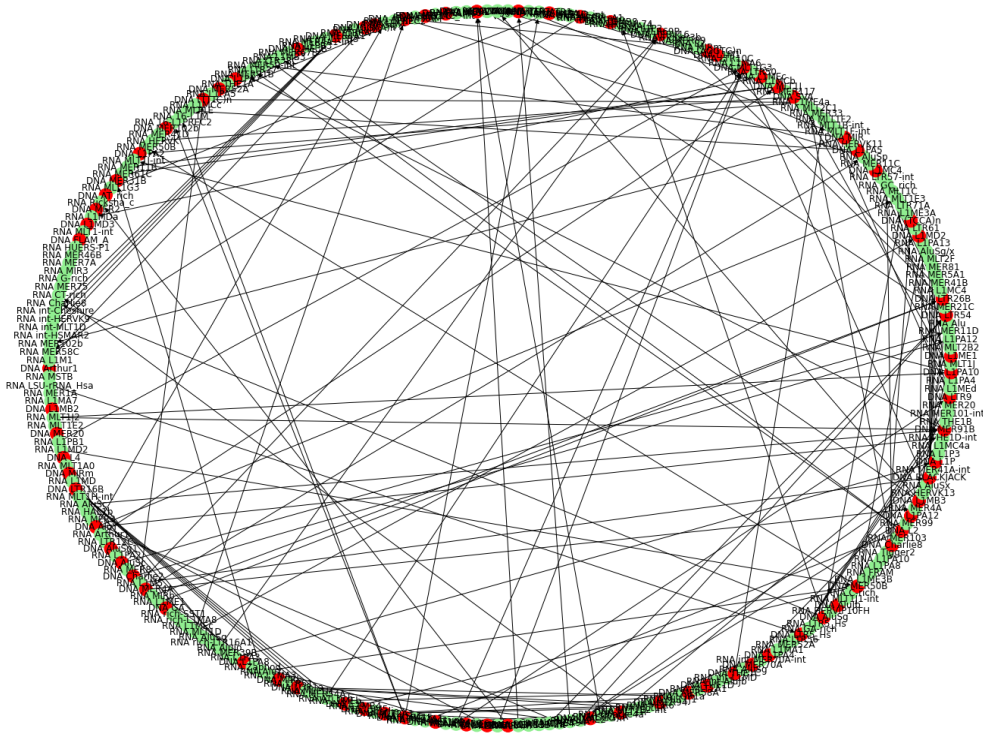


Figure 2.13: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T3d cell line, where activation is done through **H3K4me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K9me3 The denser network has 2 nodes that are highly linked ALR/Alpha, and MER46A. All these repeat nodes are originating from DNA repeat name regions. From the sparse network, we can even visually observe these highly connected hubs.

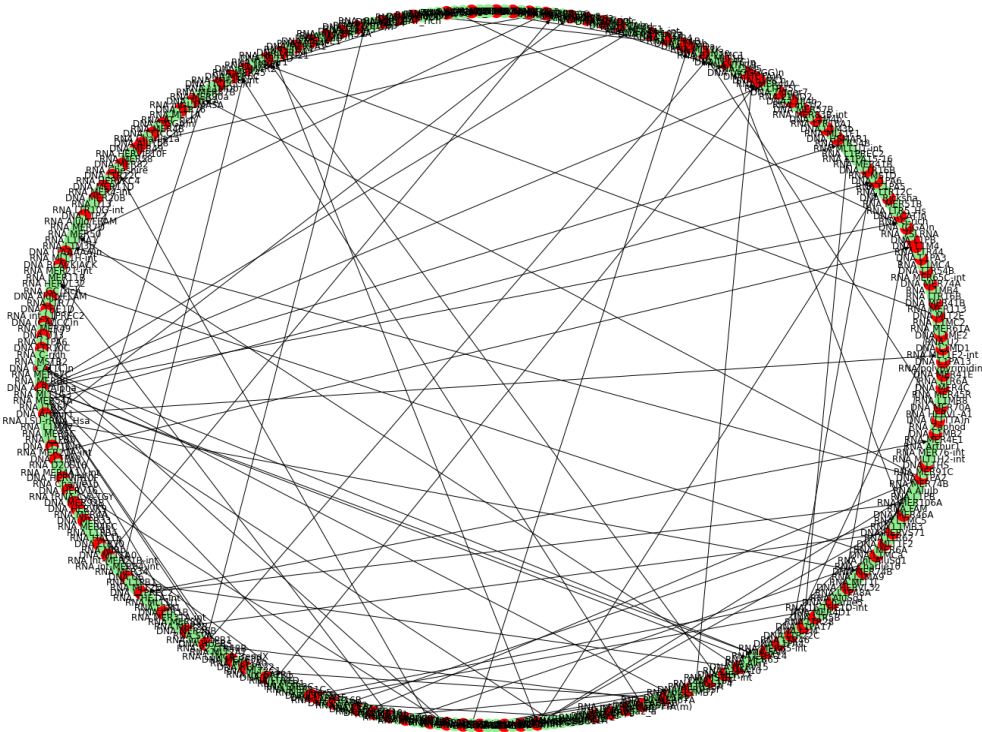
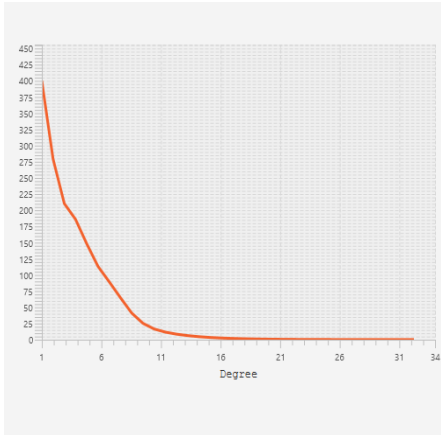
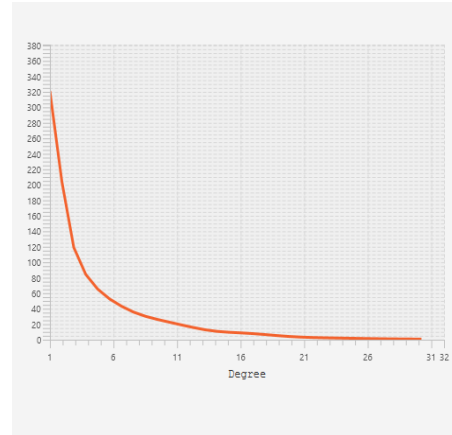


Figure 2.14: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T3d cell line, where repression is done through **H3K9me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

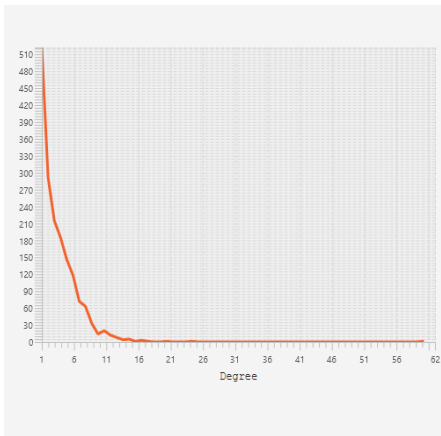
These are the degree distribution(25) plots for histone modification of both activation and repression. There is no significant difference between them. However, these do show a small world nature as is observed in biological network which validates them.



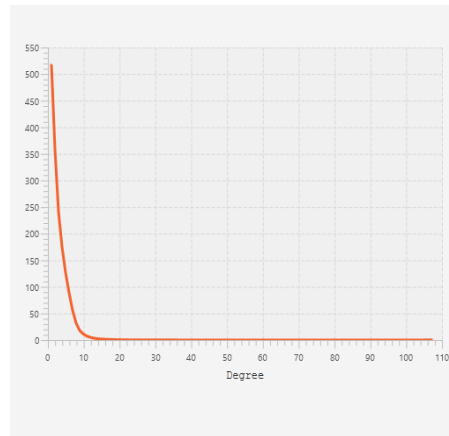
(a) Activating histone modification **H3K27ac**



(b) Activating histone modification **H3K4me3**



(c) Repressive histone modification **H3K27me3**



(d) Repressive histone modification **H3K9me3**

Figure 2.15: Degree distribution of sub-network for both activating and repressive histone modification of HUVEC T3d cell line

HUVEC T7d

Networks and degree distribution plots were made for all histone modifications, H3K4ME3, H3K9me3, H3K27ac, H3K27me3. All the degree distribution plots show small world nature, this further validates our findings as biological based networks show small world nature.

H3K27ac The dense network consists of 3 nodes that can be observed as hubs MSTA, LTR5A, and MSTB. These nodes as well originate from DNA repeat regions. We can visualise these hubs in the sparse network.

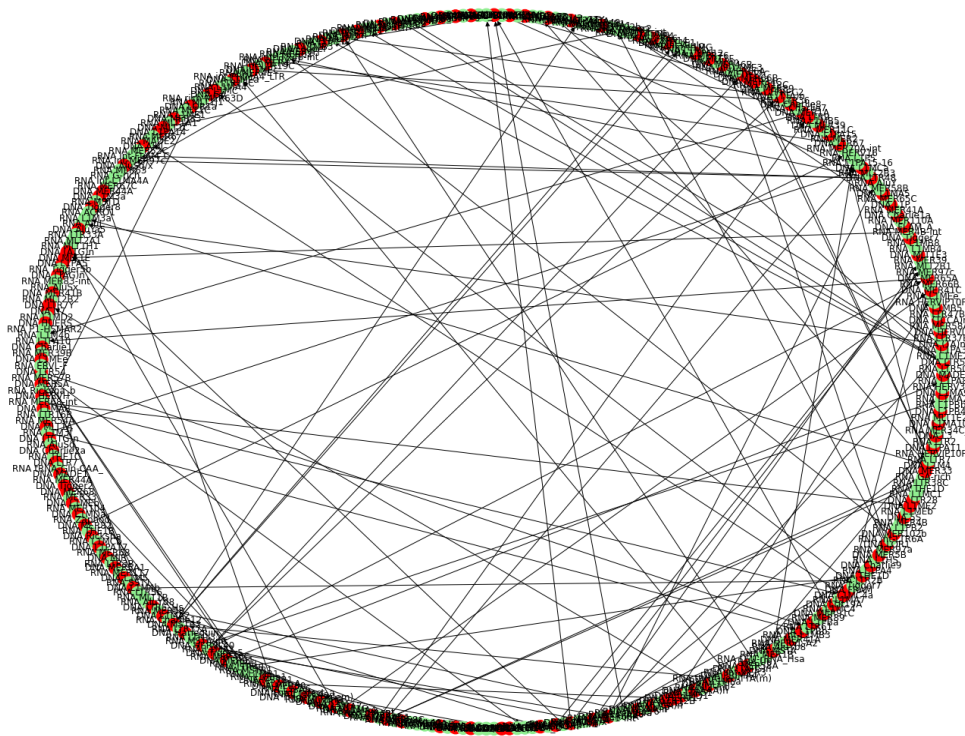


Figure 2.16: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T7d cell line, where activation is done through **H3K27ac** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K27me3 The denser network has 4 nodes that act as hubs ERVL, HERVL, THE1C, and MER51B. All these repeat nodes are originating from DNA repeat name regions. From the sparse network, we can detect these hubs

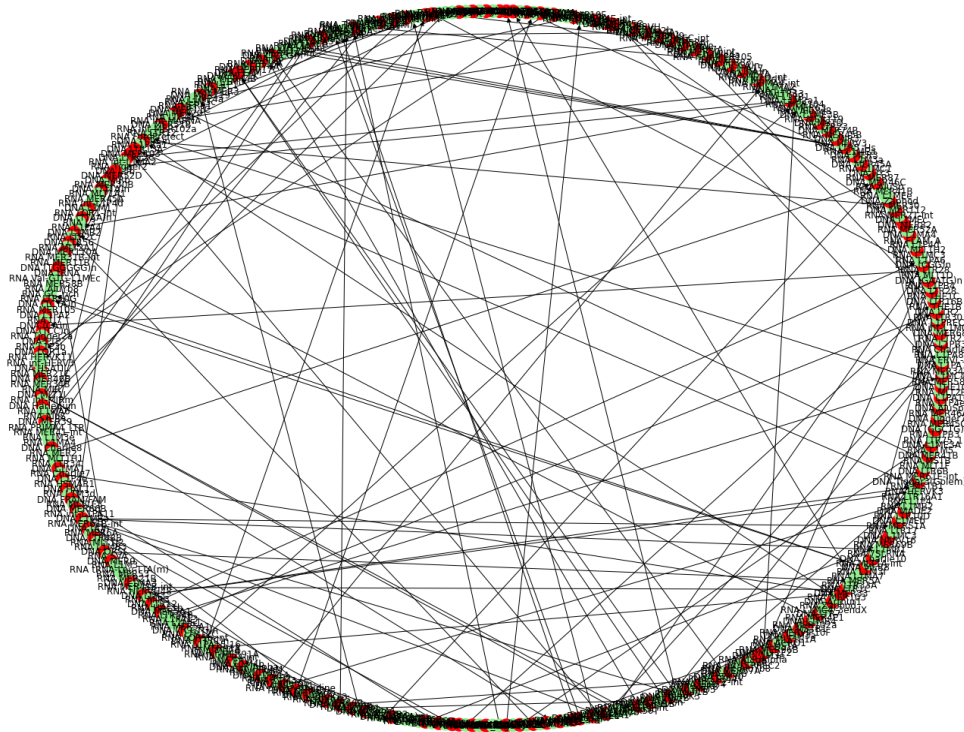


Figure 2.17: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T7d cell line, where repression is done through **H3K27me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K4me3 The dense network has 4 nodes that are highly connected AluJb, AluSx, L2 which originate from DNA repeat regions and L2 which originates from RNA repeat regions. We can observe these hubs in the sparse network.

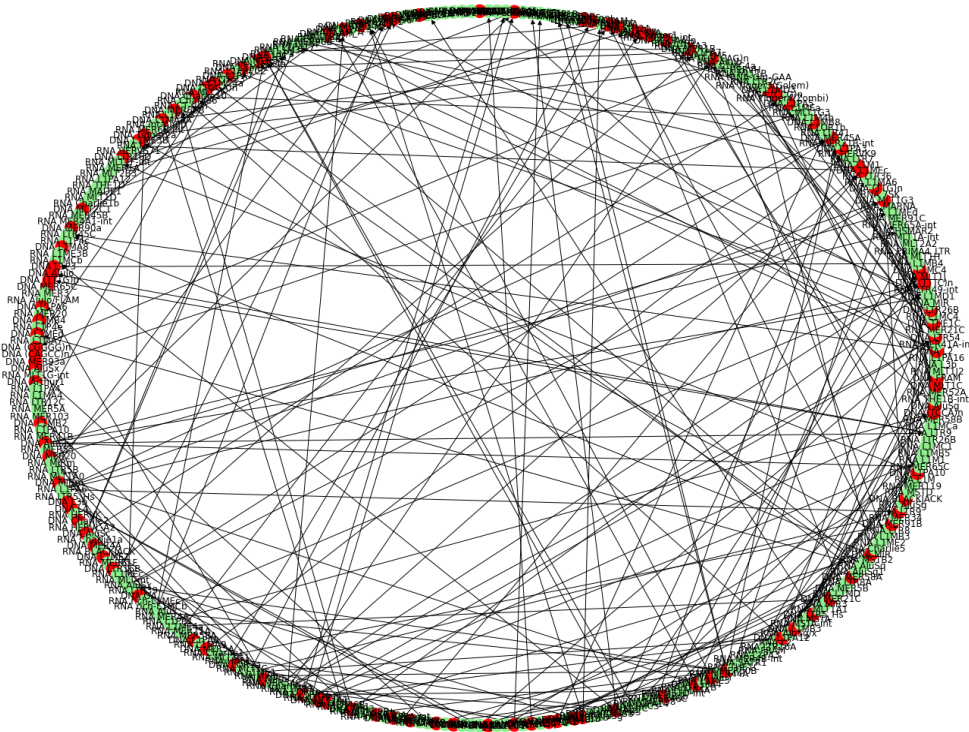


Figure 2.18: Sub-network among all RNA-DNA Repeat interactions showing activating connections for HUVEC T7d cell line, where activation is done through **H3K4me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

H3K9me3 The dense network has a single node that acts as a major hub ALR/Alpha which belongs to DNA repeat regions. We can visualise this network in the sparse network.

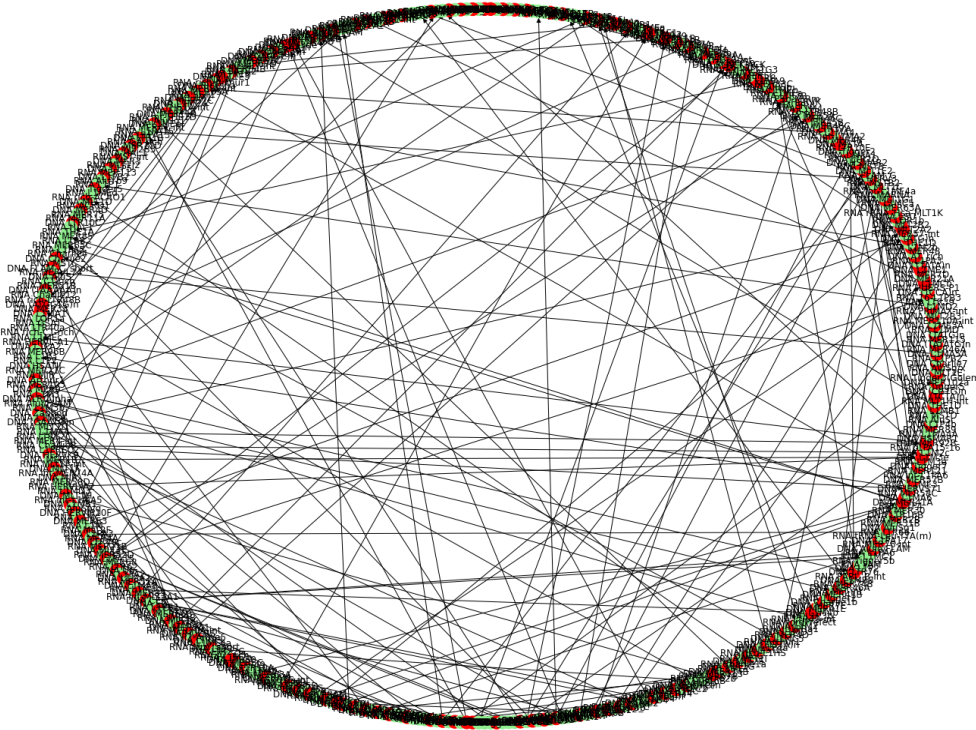
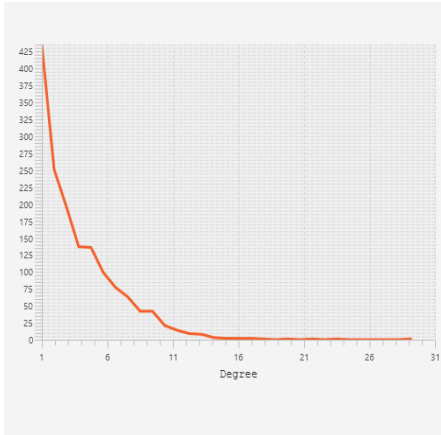
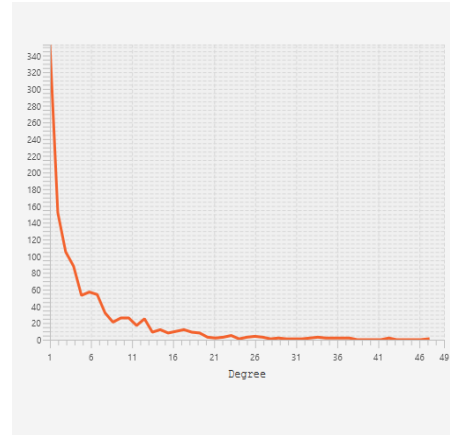


Figure 2.19: Sub-network among all RNA-DNA Repeat interactions showing repressive connections for HUVEC T7d cell line, where repression is done through **H3K9me3** histone modification. The green nodes refer to repeats interacting with RNA and red nodes refer to repeats interacting with DNA

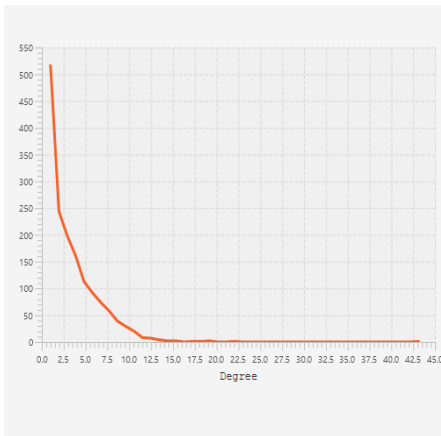
These are the degree distribution(25) plots for histone modification of both activation and repression. There is no significant difference between them. However, these do show a small world nature as is observed in biological network which validates them.



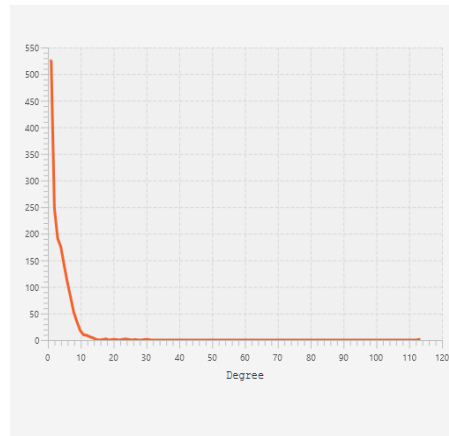
(a) Activating histone modification **H3K27ac**



(b) Activating histone modification **H3K4me3**



(c) Repressive histone modification **H3K27me3**



(d) Repressive histone modification **H3K9me3**

Figure 2.20: Degree distribution of sub-network for both activating and repressive histone modification of HUVEC T7d cell line

Finally, the median values of all these reads were put together in one table for all the histone modifications, and these were normalized by dividing each column by their mean value. This gives a better picture of which DNA-RNA pair may cause upregulation and may be responsible for downregulation.

Table 2.5: Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HEK cell line

DNA Repeat Names	RNA Repeat Names	H3K27ac	H3K4me3	H3K27me3	H3K9me3
LTR6B	MER57A-int	463.186	50.505	0.026	0.13
HERVP71A	MLT1A1	444.763	567.56	0.034	0.033
HUERS	P3-MER67D	441.968	393.177	0.238	0
Zaphod2	LTR1B	439.378	190.454	0.745	0.277
tRNA	Glu-GAG_-THE1C	430.695	147.777	0.136	0.033
(CTA) _n	L1MB3	425.763	371.491	0.273	0
MSTB	LTR69	419.49	227.872	0.145	0.033
MER8	MLT1H	414.524	173.906	1.805	0.033
(CCCG) _n	L1M4c	412.688	64.636	0.221	0.163
MER54B	L1PREC2	411.423	57.72	0.17	0.489
MER2B	tRNA-Gln-CAA_	410.019	70.988	0.23	0.033
MLT2B2	MER57A	406.561	13.789	0.579	0
MER52	int-MER34C	406.283	202.312	0.128	0.033
LTR54	MER54A	405.24	17.945	0.085	0.163
LTR10A	LTR40b	401.205	234.099	0.06	0.294
(TTATA) _n	MER4A1	397.063	0.75	0.443	0.065
HSAT5	L3	396.81	39.68	0.043	0.033
Tigger3(Golem)	MER11B	396.672	81.69	0.06	0.424
LTR13_	L1ME4a	394.046	3.26	0.281	0.13
AluYc3	LTR12C	393.538	11.533	0.434	0
LTR12C	AluYc3	393.538	11.533	0.434	0
(CAGAG) _n	Charlie8	390.496	261.075	0.034	1.142
SATR2	MER31-int	389.995	890.941	720.588	0.457

Table 2.6: Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HFF cell line

DNA Repeat Names	RNA Repeat Names	H3K27ac	H3K4me3	H3K27me3	H3K9me3
(CGGA)n	L1MB7	31.259	74.85	3.351	1.04
(CGGA)n	G-rich	31.216	74.986	3.594	1.07
MER6	L1M3f	29.221	69.776	6.277	1.427
(TGGG)n	MER20	27.182	67.266	2.764	0.773
FLAM_A	LTR75	25.584	65.809	1.954	0.505
LTR8A	MER67C	24.839	59.231	3.169	0.595
(TCTCTG)n	MER34B-int	24.583	74.968	3.088	0.713
(TAG)n	L1M3c	24.214	61.955	3.857	1.13
LTR8A	BSR/Beta	23.958	54.873	0.709	2.289
MER11B	MER4B-int	23.306	76.532	5.274	1.992
LTR48	int-AluY	22.104	46.277	1.002	0.446
LTR3	L1ME1	22.017	37.822	0.729	0.743
LTR48	int-L1PA5	21.952	50.149	1.033	0.149
LTR49	int-LTR27	21.36	51.931	1.336	1.07
LTR49	int-Tigger2	21.36	51.931	1.336	1.07
LTR48	int-THE1B	21.197	43.973	0.881	0.476
L1M5	HERVK14C	20.702	75.975	4.657	0.892
LTR52	MER93B	20.631	1.166	0.243	1.248
LTR66	MER2	20.501	2.09	0.456	1.189
MER1A	HERVL-A2	20.327	10.124	0.8	0.297
THE1A	MER4D	20.191	71.434	3.128	0.713
HERVK3	MER41B-int	20.06	47.23	1.853	2.764
LTR3	L1MA3	19.93	37.265	0.952	0.624
LTR3	L3	19.745	36.419	0.962	0.624

Table 2.7: Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HUVEC T3d cell line

DNA Repeat Names	RNA Repeat Names	H3K27ac	H3K4me3	H3K27me3	H3K9me3
Arthur1	MSTB	0.199	59.8	1.169	1.257
Arthur1	LSU-rRNA_Hsa	9.708	58.458	15.744	127.018
Arthur1	L1MA7	9.096	57.542	17.17	126.784
Arthur1	MER1A	0.071	57.542	7.706	1.562
L1MB2	MLT1J2	8.073	55.878	2.583	2.081
L1MB2	MLT1E2	0.161	55.878	0.745	0.557
MER20	L1PB1	0.173	48.036	0.351	0.649
MER20	L1MD2	0.169	47.644	0.42	0.766
L4	MLT1A0	0.243	38.363	0.19	0.884
MIRm	L1MD	0.194	37.34	0.253	0.629
LTR16B	MLT1H-int	0.25	36.598	1.441	0.731
LTR16B	AluSc	0.255	36.598	0.372	0.88
LTR16B	HAL1b	8.876	36.598	0.33	1.24
L1MB2	MER3	0.153	36.404	0.669	0.881
AluY	Arthur1	0.183	35.28	0.654	0.718
MIRm	LTR12C	0.328	35.105	0.408	0.632
AluSg1	L1PA2	0.176	35.024	0.575	0.935
AluSc	MER8	0.303	34.512	0.401	0.788
Charlie2	L1M5	0.205	34.34	0.661	0.854
MER4A	MIRb	0.192	33.883	0.758	1.039
MER4A	L1ME1	0.195	33.667	1.562	0.96
GA	rich-SST1	46.514	33.194	1.697	0.57
GA	rich-L1MA8	0.395	33.194	0.384	0.933
MER4A	L1MEc	0.103	33.093	1.173	0.969

Table 2.8: Median normalized read-count for histone modification at DNA sites with a type of repeat and bound by RNA overlapping another type of repeat for HUVEC T7d cell line

DNA Repeat Names	RNA Repeat Names	H3K27ac	H3K4me3	H3K27me3	H3K9me3
AluSx	MLT1G-int	0.245	73.851	2.678	0.56
Arthur1	L1PA4	0.15	59.439	0.435	0.765
Arthur1	MER103	16.661	54.746	30.392	262.14
Arthur1	LTR12C	10.474	54.746	15.517	131.24
Arthur1	L1MA4	16.661	54.746	15.464	131.449
Arthur1	MER5A	0.131	54.746	0.755	0.817
L1MB2	L1PA10	0.052	53.163	1.503	1.429
L1MB2	MER41B	0.026	53.163	0.107	0.372
AluSp	MER53	0.315	45.158	0.545	0.717
MER20	MIRm	0.263	43.323	0.569	0.873
AluSp	LTR5B	0.372	39.127	3.281	1.18
L1MB2	MLT1A0	0.105	39.127	1.128	0.615
L1MB2	LTR5B	0.413	39.127	0.54	0.631
MIRm	LTR5_Hs	0.131	35.526	1.632	0.353
MIRm	L1PA6	0.118	35.526	0.523	0.827
L3b	MLT1A0	0.154	34.82	0.469	0.431
L4	HERVK	14.388	32.789	0.08	2.059
Charlie2	HERVL-A2	10.413	32.672	0.705	0.628
AluSc	Charlie1a	0.138	32.515	0.755	0.794
MER4A	BLACKJACK	46.967	32.236	1.703	2.082
MER4A	MER61F	46.967	32.236	1.703	2.082
L1ME1	BLACKJACK	0.523	32.236	0.043	0.8
LTR16B	L1MEc	0.183	32.156	0.811	0.553
LTR16B	MLT-int	3.539	32.156	0.541	1.614

Chapter 3

Validation and Biological Inferences

3.1 Histone modification plots with exons

From Section 2.2.3, we use the files containing RNA locations intersected with exons for cell lines HEK, HFF, HUVEC T3d, and HUVEC T7d. Using Python libraries, we find out which exons are repeated the most times. These files are then filtered, and only those rows which contain these specific 20 exons are left. We only extract columns of index and exons from these files.

Histone modifications from section 2.2.5 were then sorted according to chromosomal locations, and these are matched according to the index values to get the corresponding exons. Further processing is done on these files, and we get output as DNA locations of cell lines along with exons for each histone modification. We prepared a boxplot from this data.

3.1.1 HEK

H3K27ac This is an activating histone modification and the median values of RACK1, MT-RNR1, and GAS5 are comparatively higher.

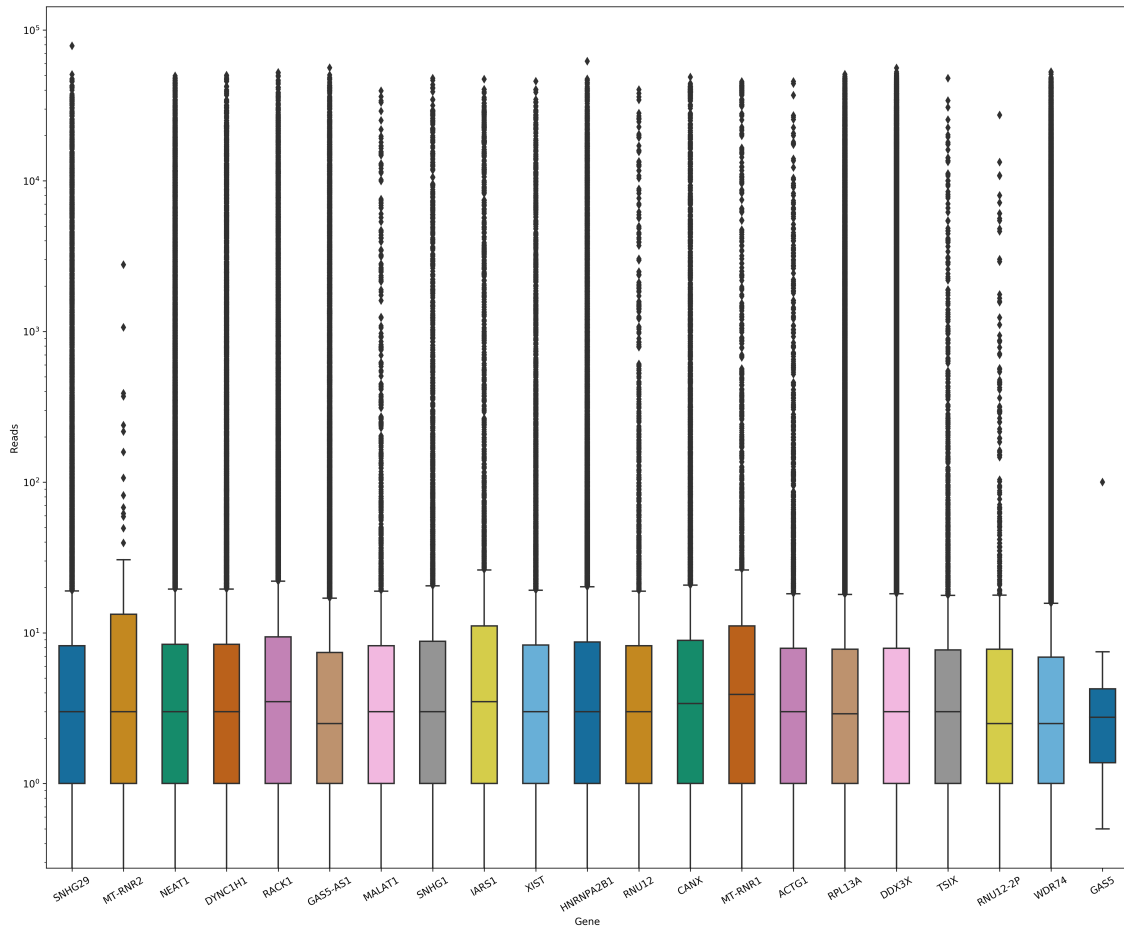


Figure 3.1: Boxplot of number of reads(for activating histone modification **H3K27ac**) at DNA binding sites of RNA of different genes for HEK cell line

H3K27me3 This is a repressing histone modification and the median values of GAS5 and RNU12-P are comparatively higher.

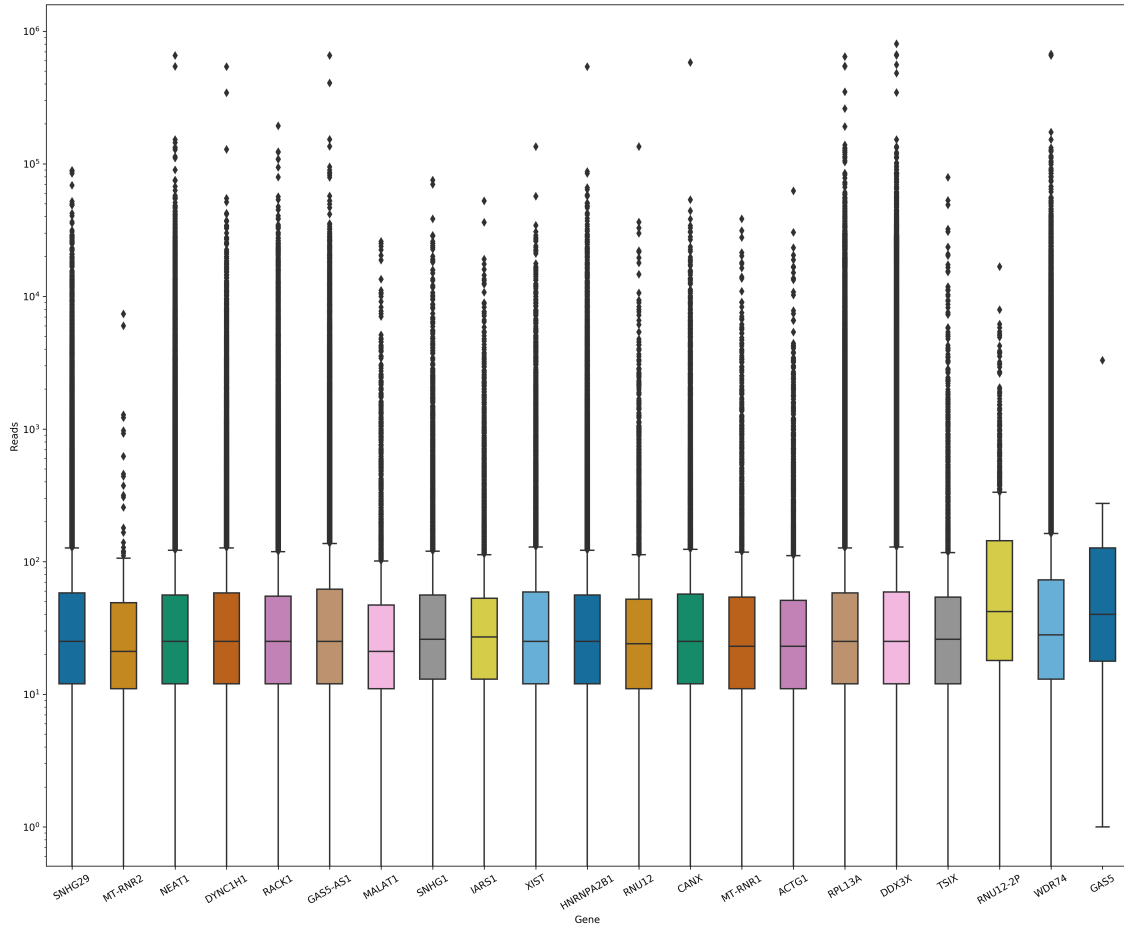


Figure 3.2: Boxplot of number of reads(for repressive histone modification **H3K27me3**) at DNA binding sites of RNA of different genes for HEK cell line

H3K4me3 This is an activating histone modification and the median value of GAS5 is comparatively higher.

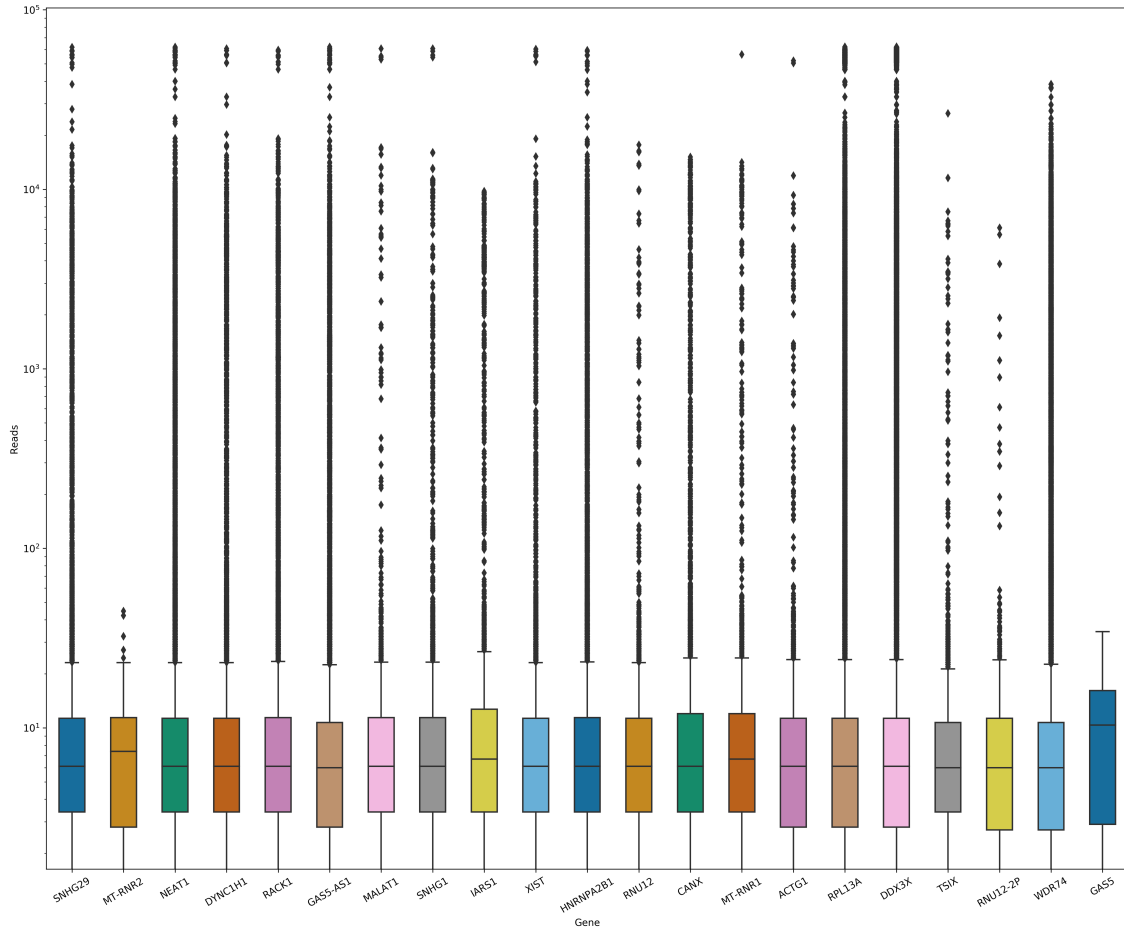


Figure 3.3: Boxplot of number of reads(for activating histone modification **H3K4me3**) at DNA binding sites of RNA of different genes for HEK cell line

H3K9me3 This is a repressing histone modification and the median values of GAS5, RPL13A, and DDX3X are higher.

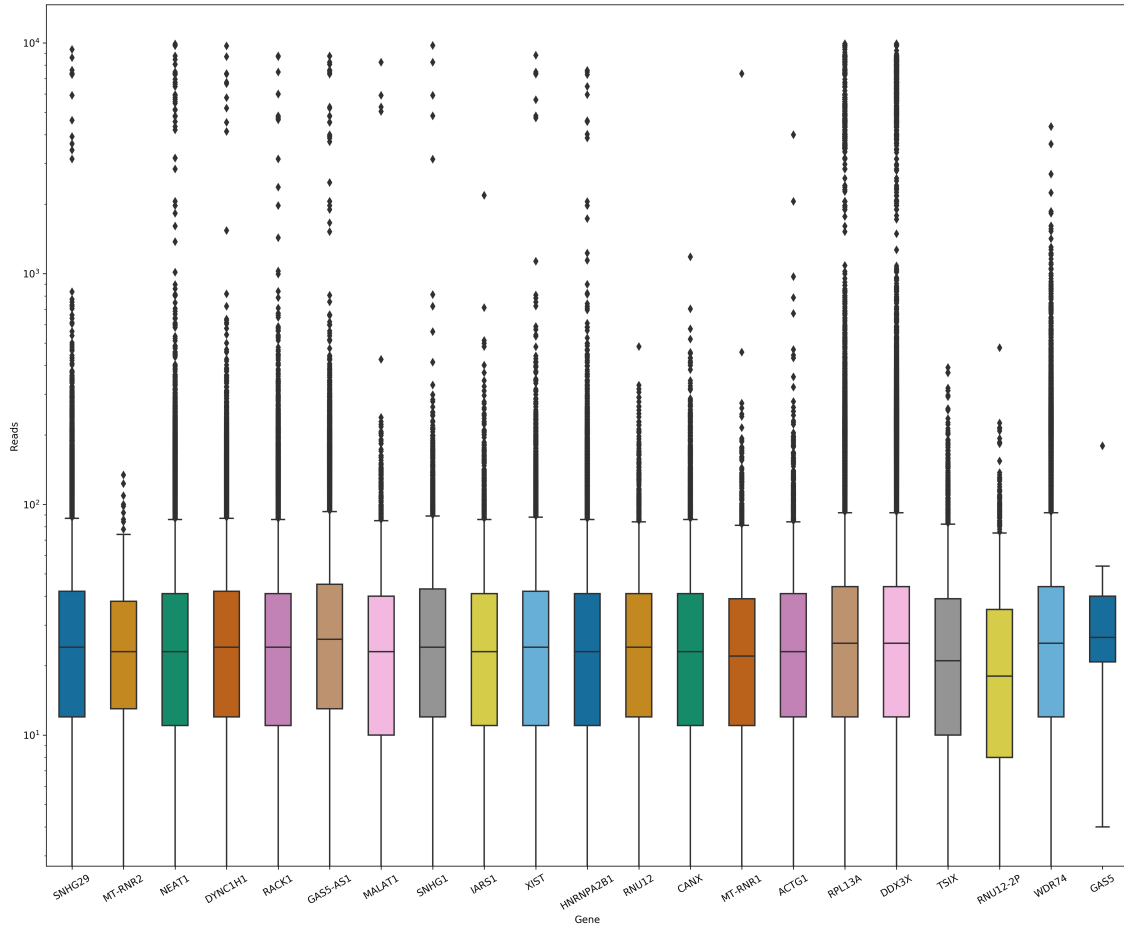


Figure 3.4: Boxplot of number of reads(for repressive histone modification **H3K9me3**) at DNA binding sites of RNA of different genes for HEK cell line

Through the boxplots, we can infer that the expression of GAS5 is relatively high and in the repressing histone modifications, GAS5-AS1 is also expressed in higher numbers.

3.1.2 HFF

H3K27ac This is an activating histone modification and the median values of THBS1, COL1A2, and GAS5 are comparatively higher.

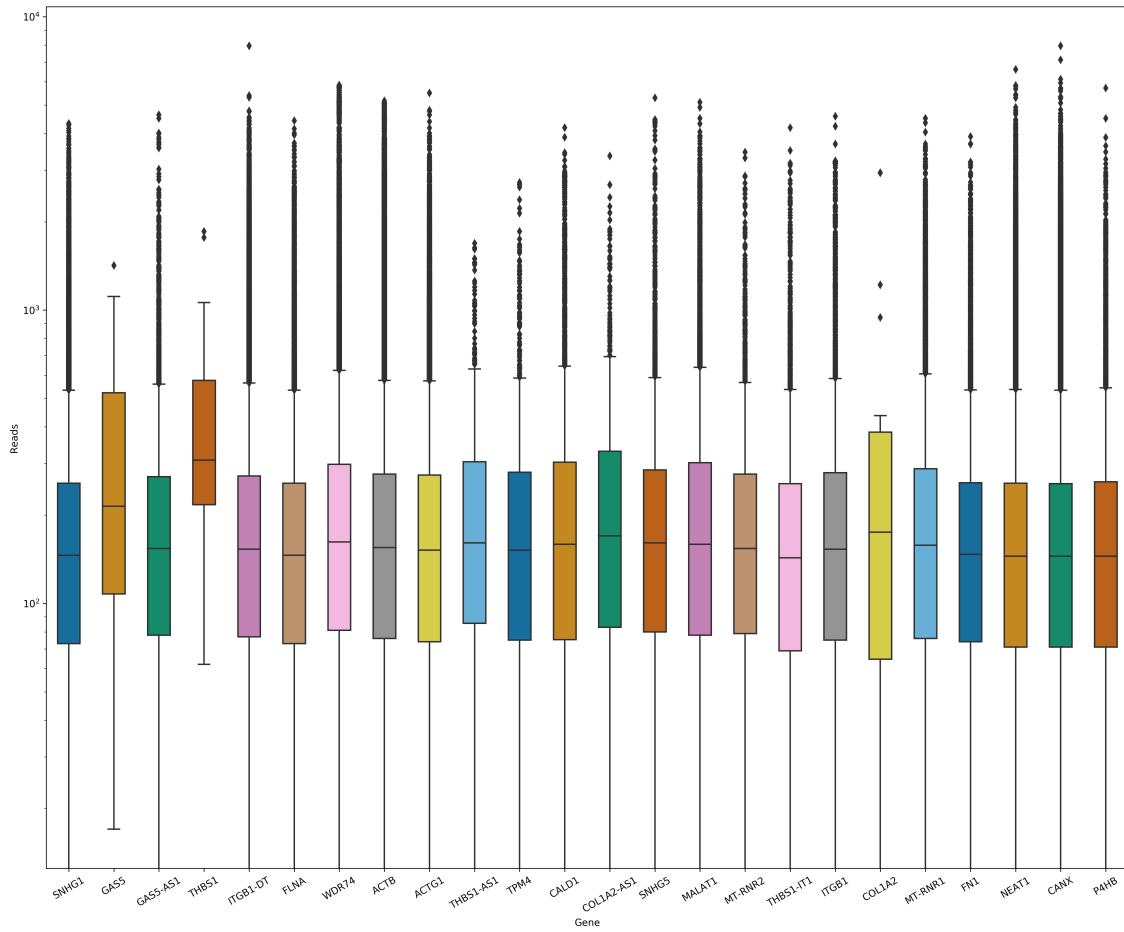


Figure 3.5: Boxplot of number of reads(for activating histone modification **H3K27ac**) at DNA binding sites of RNA of different genes for HFF cell line

H3K27me3 This is a repressing histone modification and the median values of COL1A2 and ITGB1 are comparatively higher.

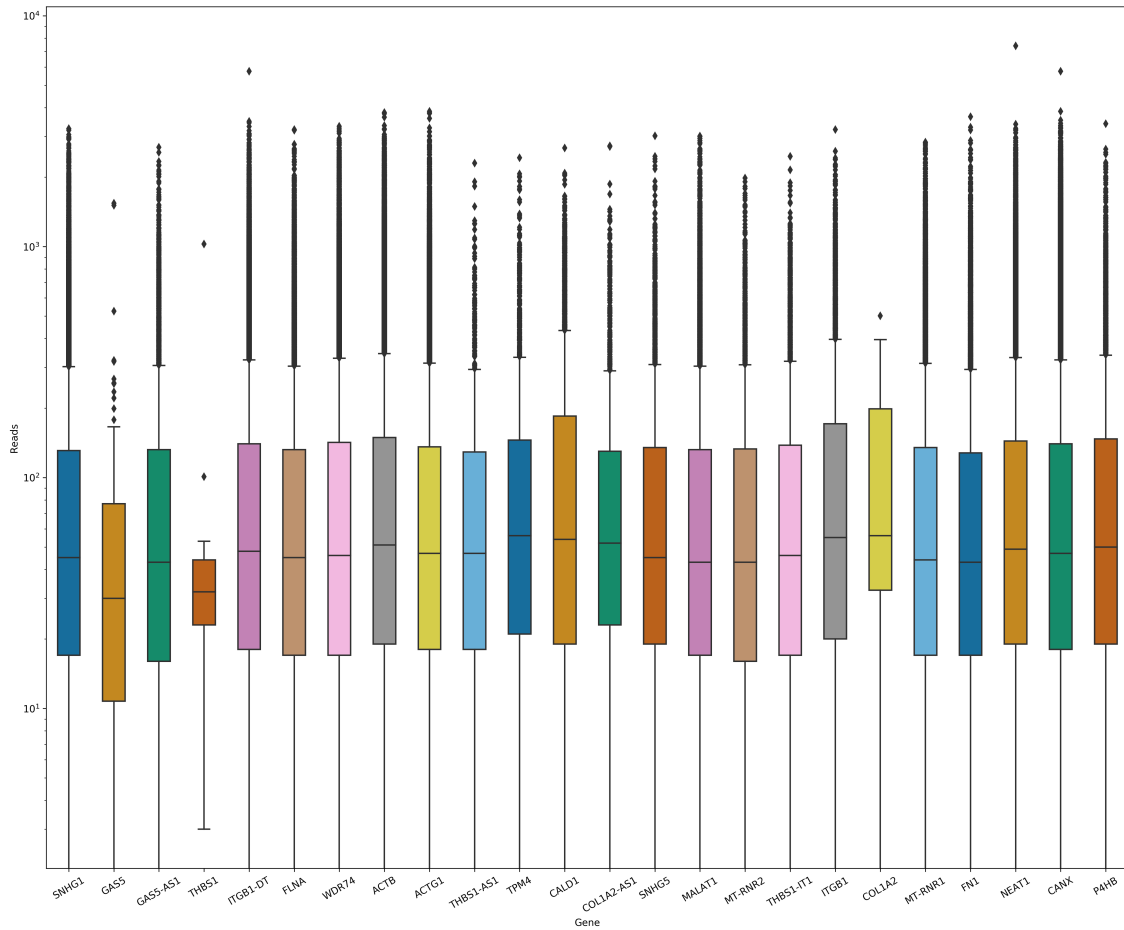


Figure 3.6: Boxplot of number of reads(for repressive histone modification **H3K27me3**) at DNA binding sites of RNA of different genes for HFF cell line

H3K4me3 This is an activating histone modification and the median values of GAS5, THBS1, and COL1A2 are comparatively higher.

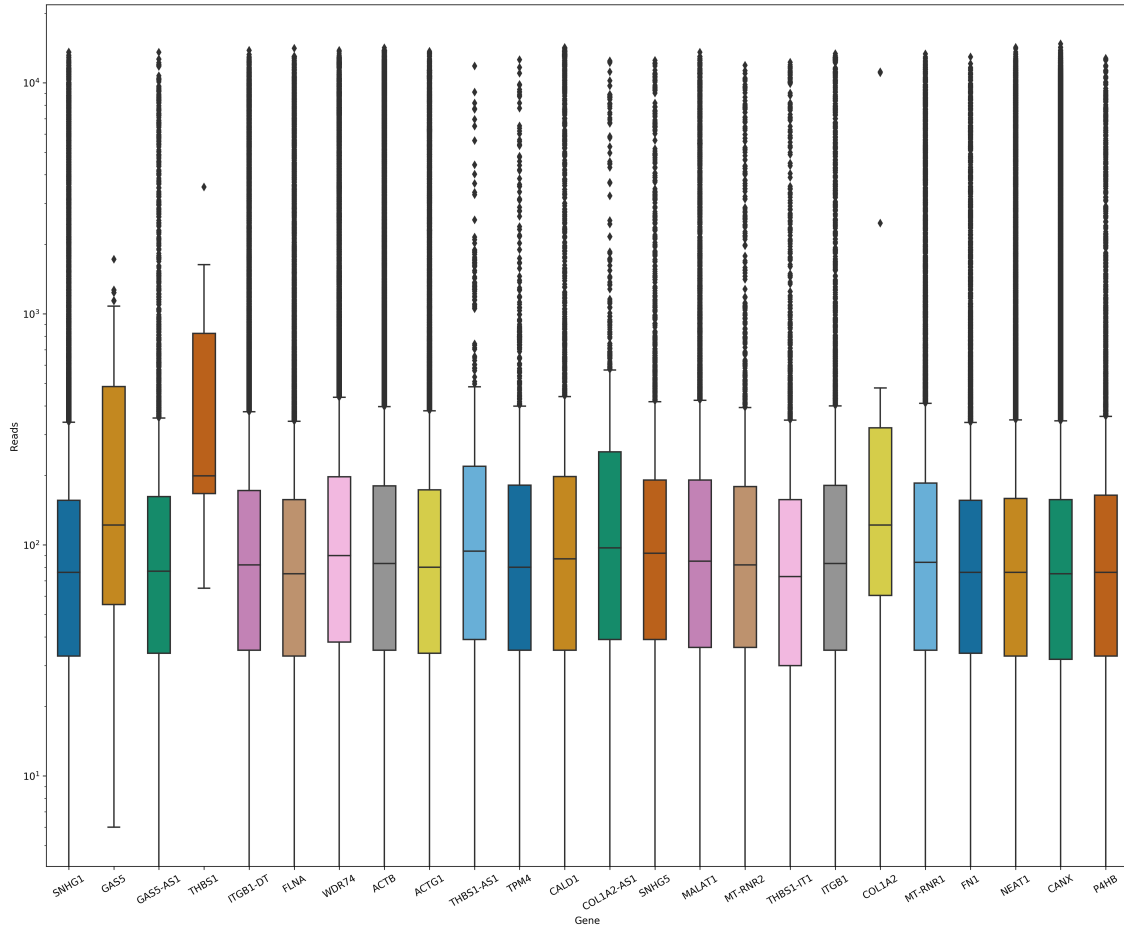


Figure 3.7: Boxplot of number of reads(for activating histone modification **H3K4me3**) at DNA binding sites of RNA of different genes for HFF cell line

H3K9me3 This is a repressing histone modification and the median values of GAS5-AS1, COL1A2-AS1, and NEAT1 are comparatively higher.

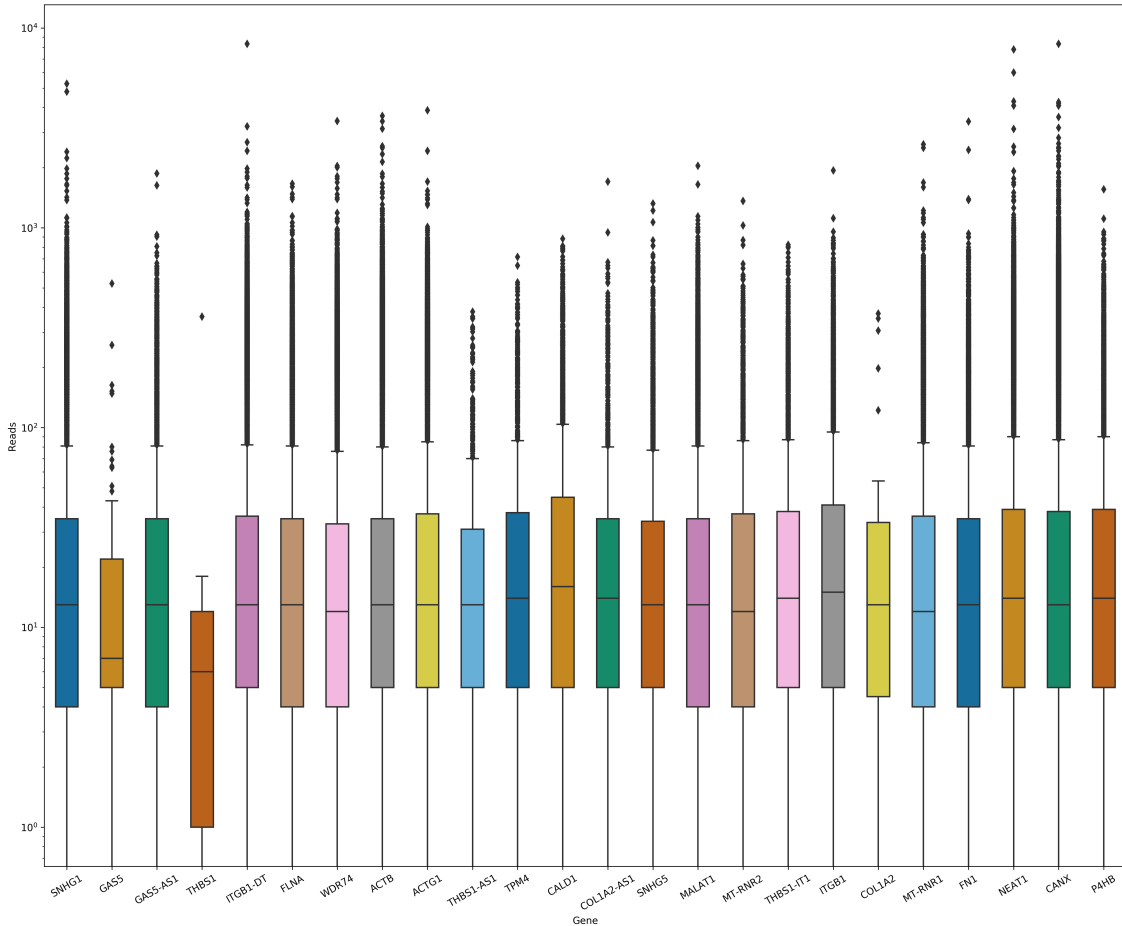


Figure 3.8: Boxplot of number of reads(for repressive histone modification **H3K9me3**) at DNA binding sites of RNA of different genes

In the HFF cell line, we can clearly observe the activating and repressing effect of histone modifications on exons. GAS5 and THBS1 are expressed highly in activating histones H3K27ac, H3K4me3, and their antisense exons GAS5-AS1, and THBS1-AS1 are expressed less. Whereas in the histones having repressive effects H3K27me3, H3K9me3 the exons GAS5, and THBS1 are expressed low and their antisense counterparts are expressed highly in their comparison. COL1A2 and its antisense exon COL1A2-AS1 also exhibit similar behavior to a lesser extent.

3.1.3 HUVEC T3d

H3K27ac This is an activating histone modification and the median values of FLNB, and FLNB-AS1 are comparatively higher.

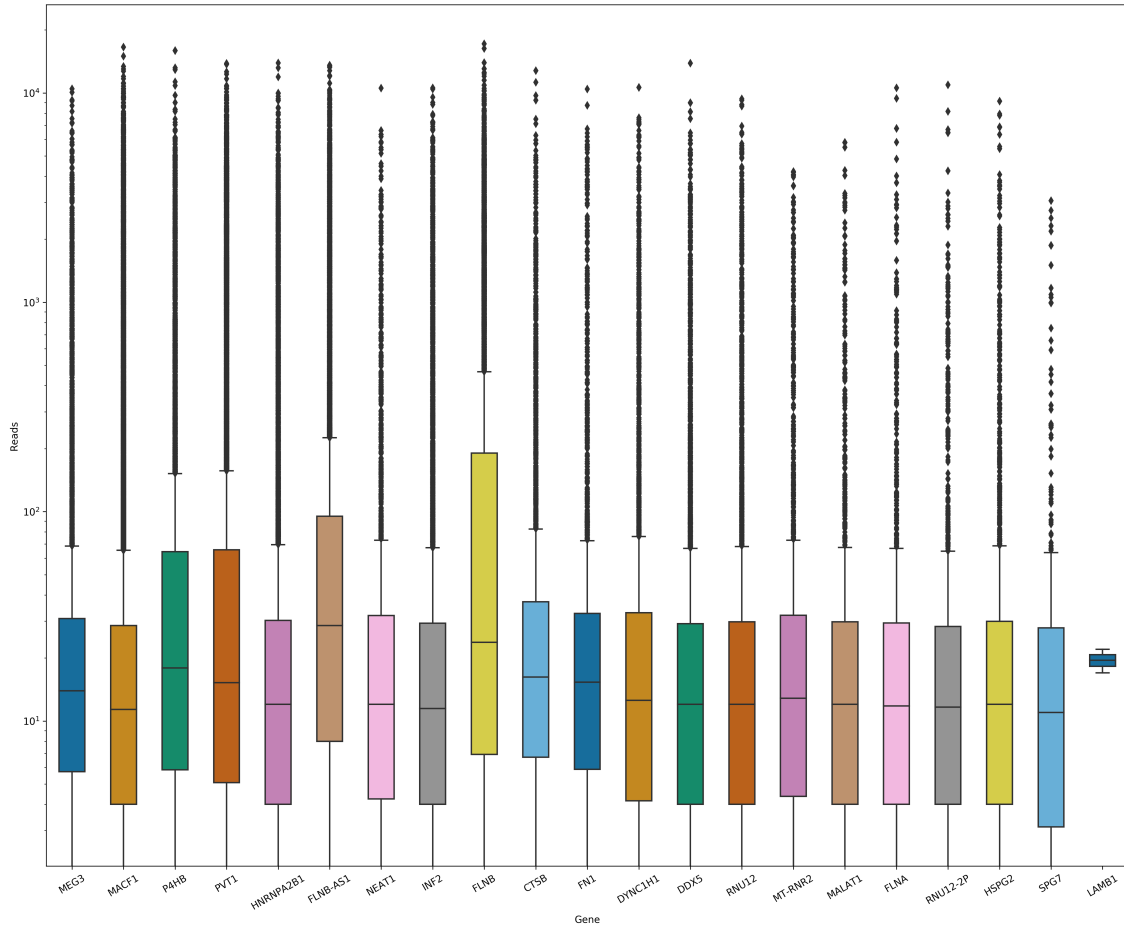


Figure 3.9: Boxplot of number of reads(for activating histone modification **H3K27ac**) at DNA binding sites of RNA of different genes for HUVEC T3d cell line

H3K27me3 This is a repressing histone modification and the median values of LAMB1, PVT1, and DOX5 are comparatively higher.

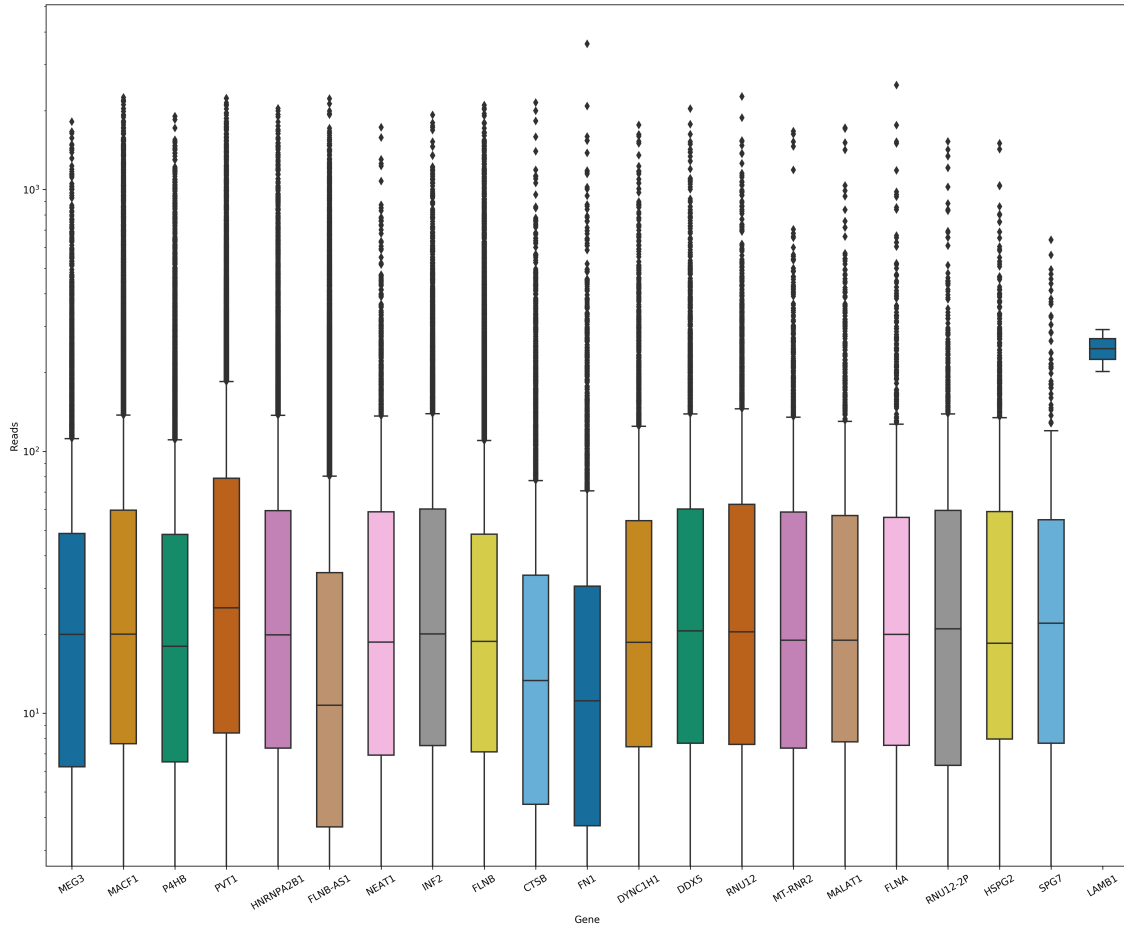


Figure 3.10: Boxplot of number of reads(for repressive histone modification **H3K27me3**) at DNA binding sites of RNA of different genes

H3K4me3 This is an activating histone modification and the mean values of NEAT1, and MALAT1 are comparatively higher.

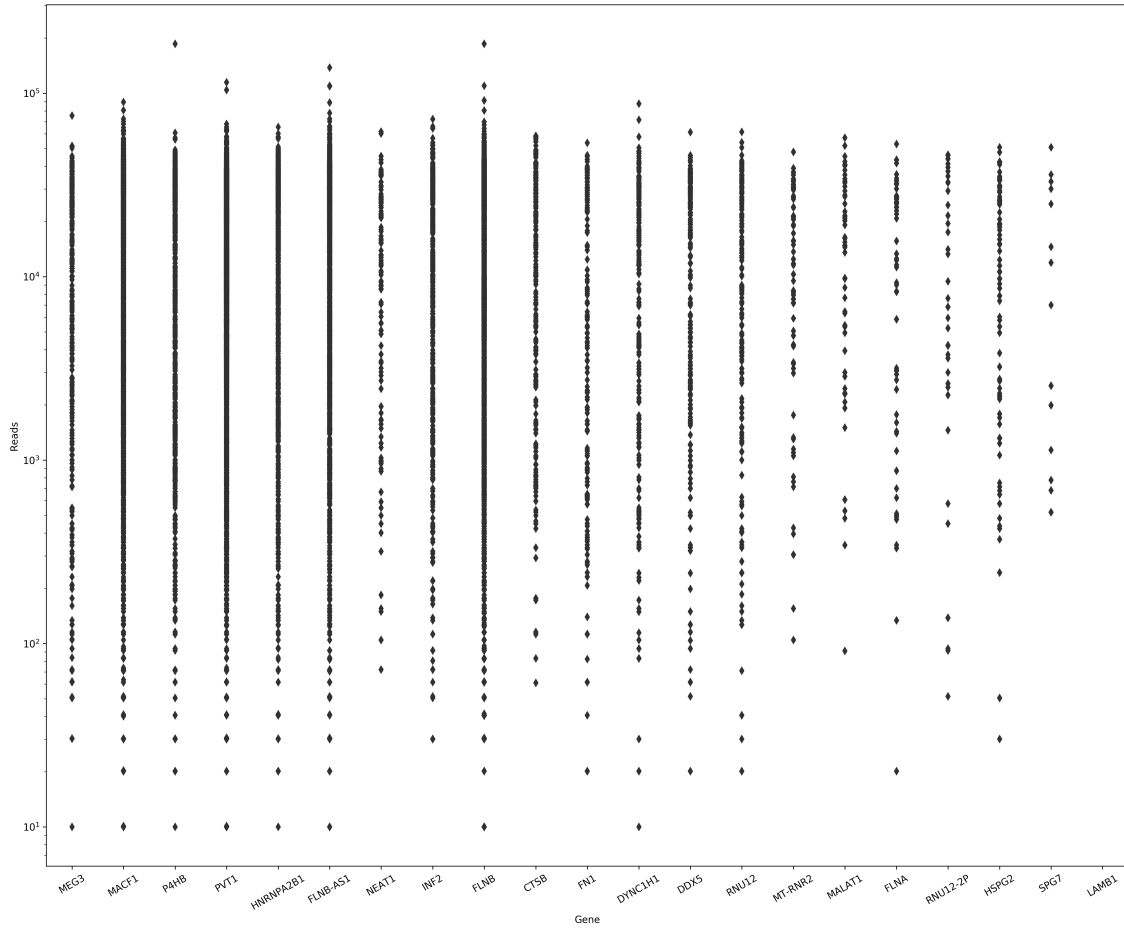


Figure 3.11: Boxplot of number of reads(for activating histone modification **H3K4me3**) at DNA binding sites of RNA of different genes for HUVEC T3d cell line

H3K9me3 This is an activating histone modification and the mean values of NEAT1, and MALAT1 are comparatively higher.

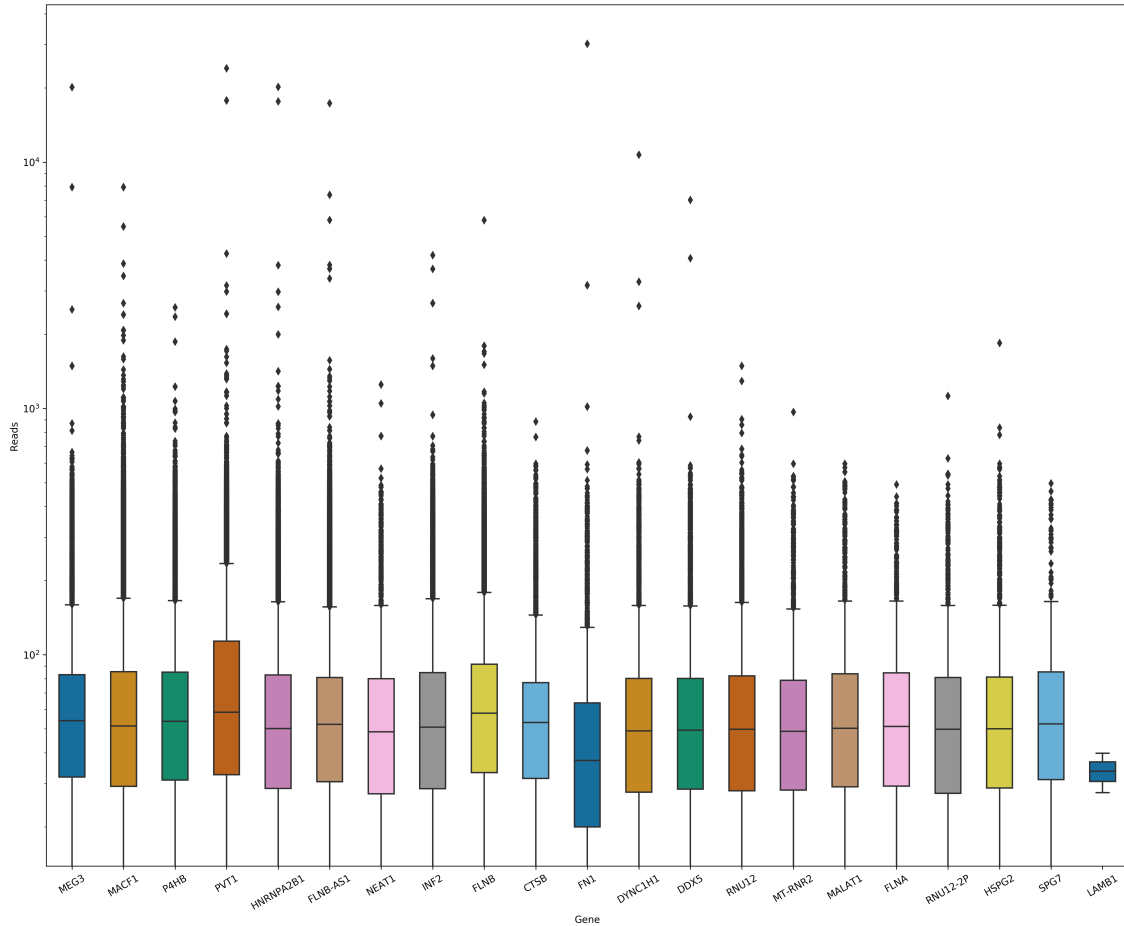


Figure 3.12: Boxplot of number of reads(for repressive histone modification **H3K9me3**) at DNA binding sites of RNA of different genes for HUVEC T3d cell line

Through the boxplots, we can infer that some exons are expressed highly in the repressive histone modifications while others are expressed highly in the activating histone modifications. FN1 and CTSB exons show higher expression levels in the activating histone modifications compared to repressive histone modifications. While FLNB is expressed highly in repressive histone modification and it's antisense counterpart FLNB-AS1 show higher readings in the activating histone modifications.

3.1.4 HUVEC T7d

H3K27ac This is an activating histone modification and the median values of NEAT1, DYNC1H1, and CTSB are comparatively higher.

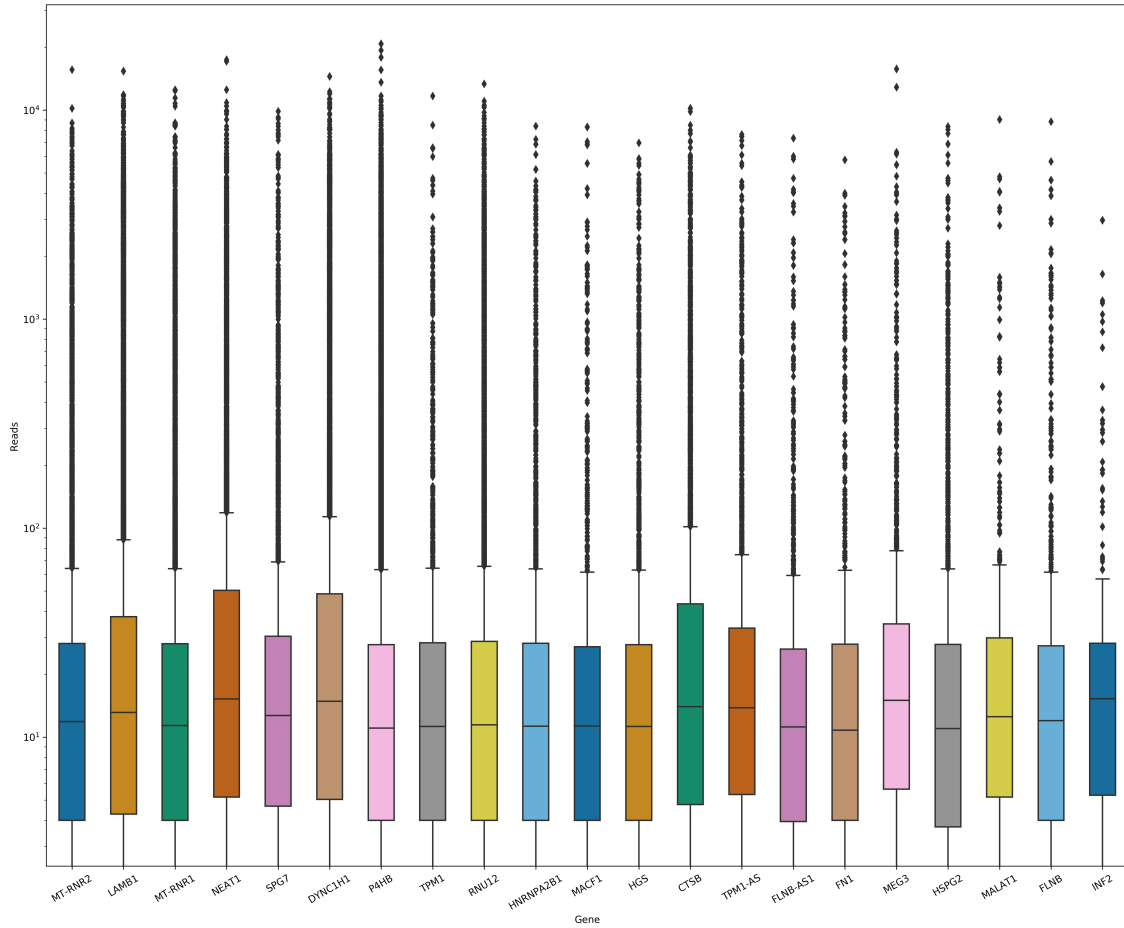


Figure 3.13: Boxplot of number of reads(for activating histone modification **H3K27ac**) at DNA binding sites of RNA of different genes for HUVEC T7d cell line

H3K27me3 This is a repressing histone modification and the median values of LAMB1, and FLNB-AS1 are comparatively higher.

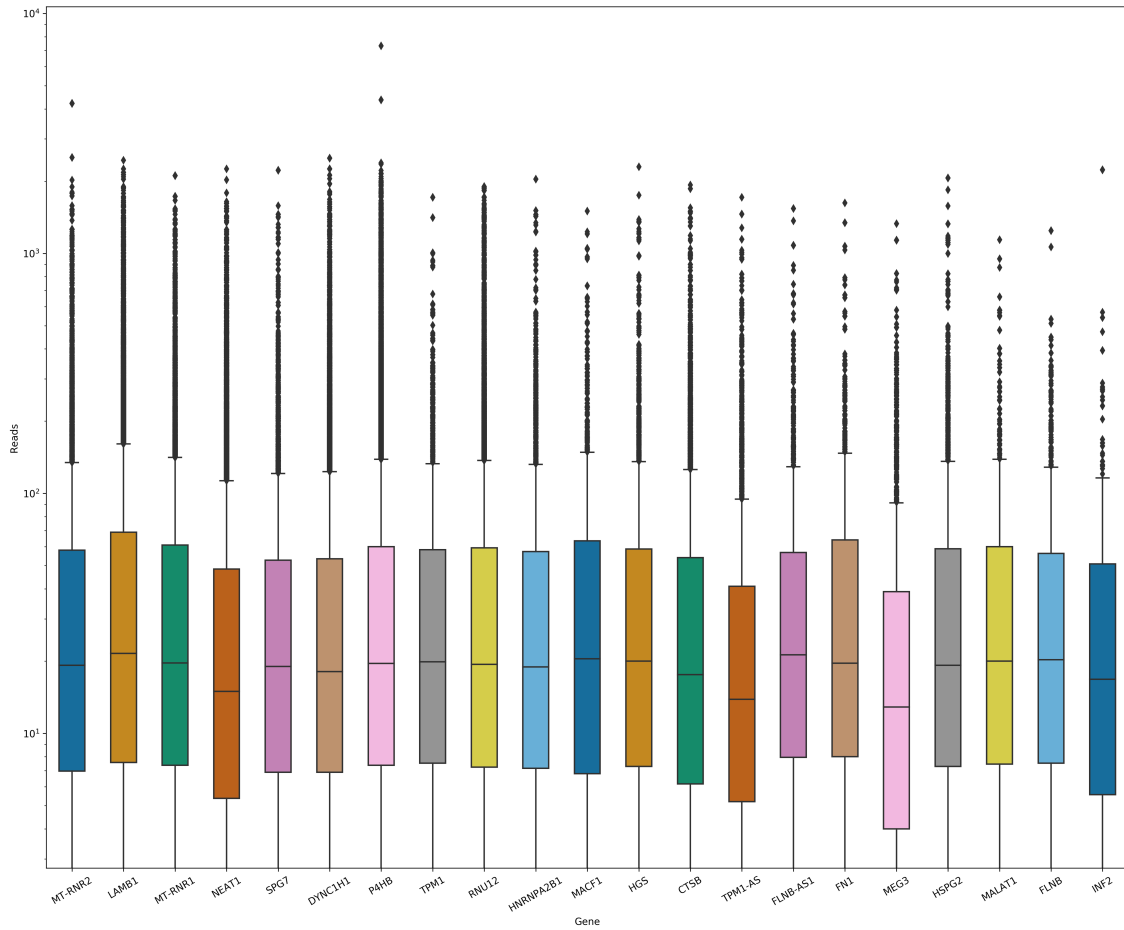


Figure 3.14: Boxplot of number of reads(for repressive histone modification **H3K27me3**) at DNA binding sites of RNA of different genes for HUVEC T7d cell line

H3K4me3 This is an activating histone modification and the mean values of NEAT1, and MALAT1 is comparatively higher.

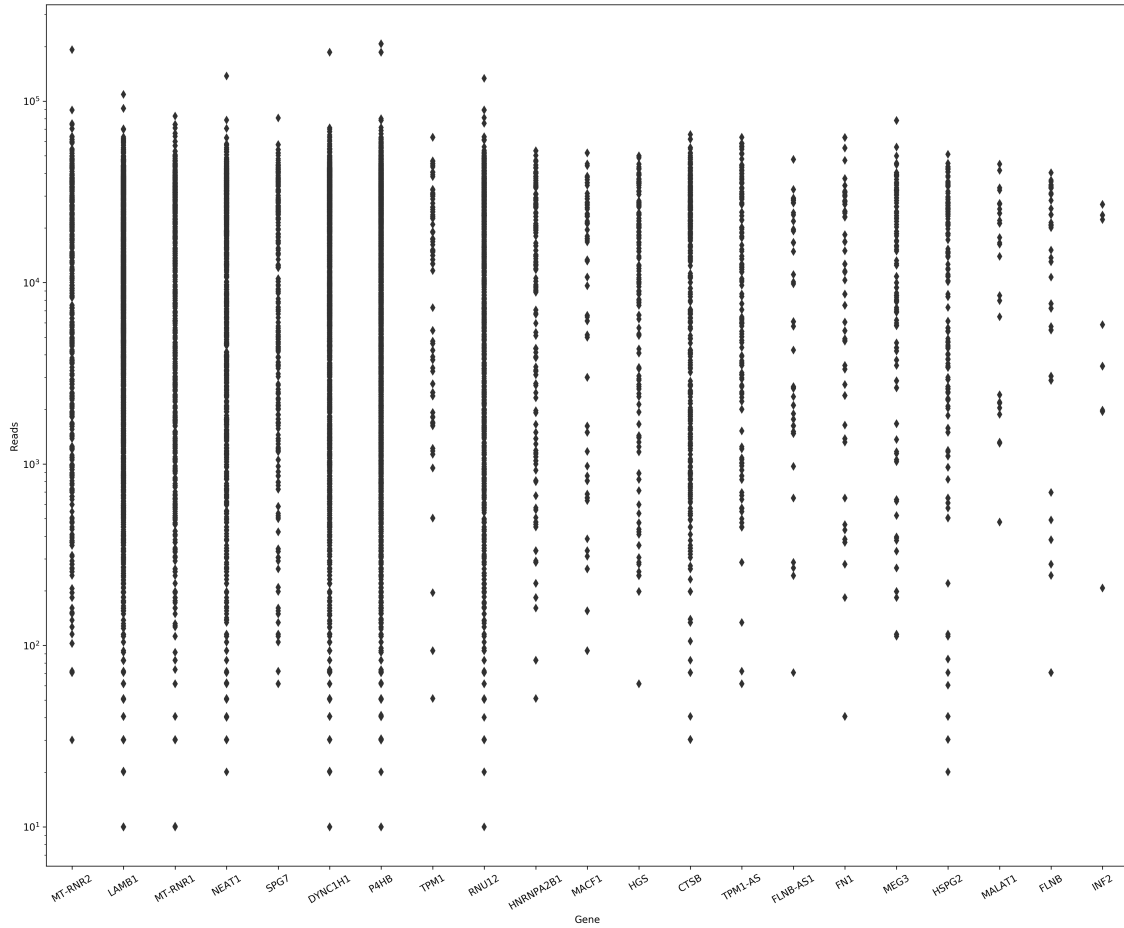


Figure 3.15: Boxplot of number of reads(for activating histone modification **H3K4me3**) at DNA binding sites of RNA of different genes for HUVEC T7d cell line

H3K9me3 This is a repressing histone modification and the median values of MALAT1, and MACF1 are comparatively higher.

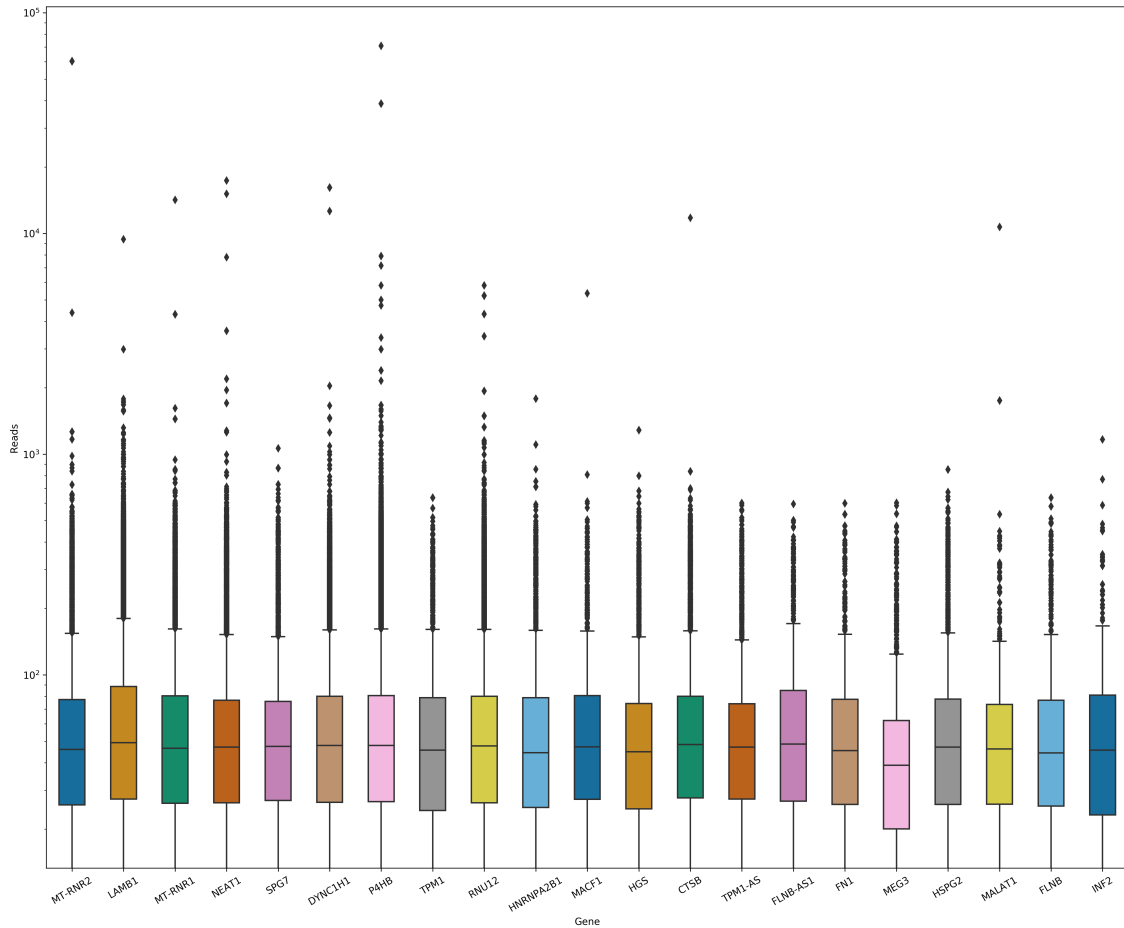


Figure 3.16: Boxplot of number of reads(for repressive histone modification **H3K9me3**) at DNA binding sites of RNA of different genes for HUVEC T7d cell line

Through inference of boxplot, we can summarize that several exons like NEAT1, CTSB, and MEG3 have higher median values with activating histone modifications. MEG3 shows higher values with activating histone modifications and lower values with repressive histone modifications.

3.2 Gene Ontology

Gene Ontology is usually done in three domains, molecular functions, biological processes, and cellular components. To do gene ontology, we have used the GREAT gene ontology tool, which can be used for the prediction of functions for cis-regulatory regions. Using the data we got in section 3.1.1, 3.1.2, 3.1.3, and 3.1.4 we can surmise the exons with high median values visually from the boxplot. Now, we take only those rows which contain only these exons for all histone modifications.

By using the index column, we match it with the original file for all cell lines containing DNA locations and add the exons column to it. After removing the negative value, we upload these files for every exon separately to the GREAT gene ontology tool with parameters genome build as hg38, background region as whole-genome, and association setting as Single nearest gene within 5kb.

While doing Gene Ontology, we found that results for activating or repressive histone modifications for a specific cell line were same.

3.2.1 HEK

Following are the exons with highest median values that are taken for further analysis. GAS5 helps in regulation of metabolic functions and this shows that it is an important regulator in activating and repressing histone modifications.

- H3K4me3: GAS5
- H3K9me3: GAS5, RPL13A, DDX3X
- H3K27ac: RACK1, MT-RNR1
- H3K27me3: GAS5, RNU12-P

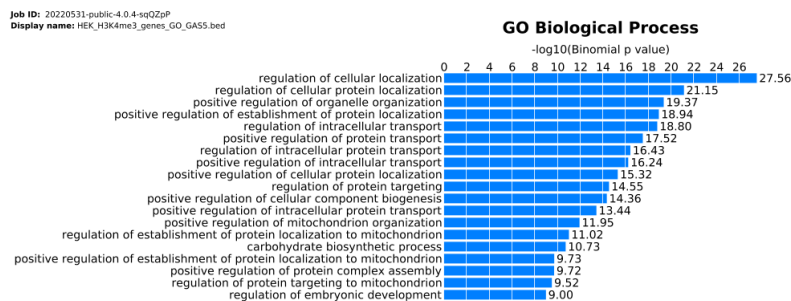


Figure 3.17: Gene Ontology enrichment for DNA binding location for GAS5 RNA in HEK cell line

3.2.2 HFF

Following are the exons with highest median values that are taken for further analysis. GAS5 helps in regulation of metabolic functions and this shows that it is an important regulator in activating and repressing histone modifications.

- H3K4me3: GAS5, THBS1, COL1A2

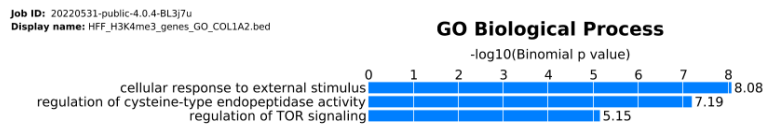


Figure 3.18: Gene Ontology enrichment for DNA binding location for COL1A2 RNA in HFF cell line

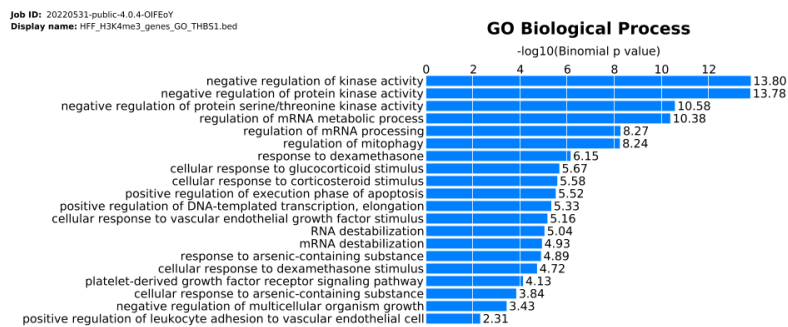


Figure 3.19: Gene Ontology enrichment for DNA binding location for THBS1 RNA in HFF cell line

- H3K9me3: GAS5-AS, COL1A2-AS1, NEAT1
- H3K27ac: GAS5, THBS1, COL1A2
- H3K27me3: ITGB1, COL1A

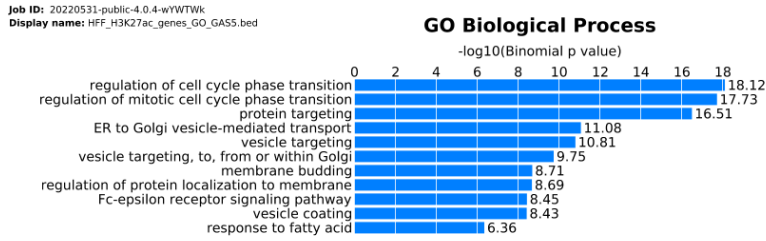


Figure 3.20: Gene Ontology enrichment for DNA binding location for GAS5 RNA in HFF cell line

3.2.3 HUVEC T3d

Following are the exons with highest median values that are taken for further analysis.

- H3K4me3: NEAT1, MALAT1
- H3K9me3: PVT1, FLNB
- H3K27ac: FLNB
- H3K27me3: LAMB1, PVT1, DOX5

3.2.4 HUVEC T7d

Following are the exons with highest median values that are taken for further analysis.

- H3K4me3: NEAT1, MALAT1

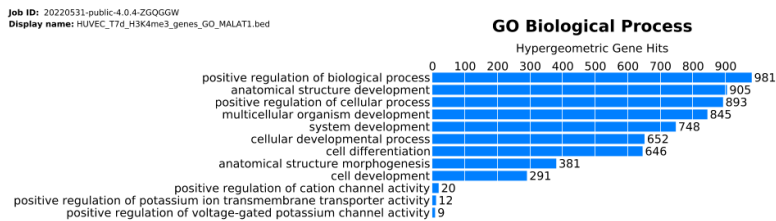


Figure 3.21: Gene Ontology enrichment for DNA binding location for MALAT1 RNA in HUVEC T7d cell line

- H3K9me3: MALAT1, MACF1
- H3K27ac: CTSB
- H3K27me3: LAMB1

3.3 Motif finding using HOMER tool

First, we identified dominant genes, which could be working to bind to DNA sites, through which we identified GAS5, and then we did gene ontology and found bias for certain functions. For example, GAS5 region RNA seems to be interacting with DNA locations which are responsible for metabolic functions. Hence, we had a question, why this is happening and what could be the mode of binding of GAS5. So we did motif finding for GAS5 in HEK, HFF cell line, PVT1 for HUVEC T3d cell line and NEAT1 for HUVEC T7d cell line.

We have used the HOMER tool to determine the known motif regions in our data. Firstly, we used the original cell line data in bedpe format and we extracted the DNA sequences in bed format. Along with it, we added a column of ID index and another column of peak with index number because this is the prescribed format in which HOMER undertakes files.

We match the IDs with the IDs of the RNA file intersected with exons. Afterwards, we sort the file columns in a proper format. However, this is in the hg38 genome build and we reduced it to the hg19 genome build using the LiftOver tool. Now we run the HOMER tool using this command.

```
findMotifsGenome.pl InputFile hg19 OutputFile/ -size 200 -mask
```

Here are the results of top ten known motifs for all cell lines.

3.3.1 HEK

We use rows containing only GAS5 exons.

1		Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738)/Homer
2		Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465)/Homer
3		MyoG(bHLH)/C2C12-MyoG-ChIP-Seq(GSE36024)/Homer
4		Tcf21(bHLH)/ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369)/Homer
5		Myf5(bHLH)/GM-Myf5-ChIP-Seq(GSE24852)/Homer
6		SCL(bHLH)/HPC7-Sc1-ChIP-Seq(GSE13511)/Homer
7		Asc11(bHLH)/NeuralTubes-Asc11-ChIP-Seq(GSE55840)/Homer
8		MyoD(bHLH)/Myotube-MyoD-ChIP-Seq(GSE21614)/Homer
9		E2A(bHLH)/proBcell-E2A-ChIP-Seq(GSE21978)/Homer
10		Ptf1a(bHLH)/Panc1-Ptf1a-ChIP-Seq(GSE47459)/Homer

Figure 3.22: HOMER results for top ten known motifs of HEK cell line

3.3.2 HFF

For the HFF cell line too, we use rows with GAS5 exons.

1		Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738) Homer
2		MyoG(bHLH)/C2C12-MyoG-ChIP-Seq(GSE36024) Homer
3		Tcf21(bHLH)/ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369) Homer
4		Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465) Homer
5		Myf5(bHLH)/GM-Myf5-ChIP-Seq(GSE24852) Homer
6		MyoD(bHLH)/Myotube-MyoD-ChIP-Seq(GSE21614) Homer
7		SCL(bHLH)/HPC7-Sc1-ChIP-Seq(GSE13511) Homer
8		E2A(bHLH)/proBcell-E2A-ChIP-Seq(GSE21978) Homer
9		Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840) Homer
10		Ptf1a(bHLH)/Panc1-Ptf1a-ChIP-Seq(GSE47459) Homer

Figure 3.23: HOMER results for top ten known motifs of HFF cell line

3.3.3 HUVEC T3d

We use rows having PVT1 exons.

1		Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738)Homer
2		Tcf21(bHLH)/ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369)Homer
3		USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)Homer
4		Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)Homer
5		RARg(NR)/ES-RARg-ChIP-Seq(GSE30538)Homer
6		bHLHE41(bHLH)/proB-Bhlhe41-ChIP-Seq(GSE93764)Homer
7		Ptf1a(bHLH)/Panc1-Ptf1a-ChIP-Seq(GSE47459)Homer
8		Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465)Homer
9		RAR:RXR(NR),DR5/ES-RAR-ChIP-Seq(GSE56893)Homer
10		p53(p53)/Saos-p53-ChIP-Seq(GSE15780)Homer

Figure 3.24: HOMER results for top ten known motifs of HUVEC T3d cell line

3.3.4 HUVEC T7d

We use rows having NEAT1 exons.







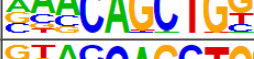



1		Ap4(bHLH) AML-Tfap4-ChIP-Seq(GSE45738) Homer
2		SCL(bHLH) HPC7-Scl-ChIP-Seq(GSE13511) Homer
3		MyoG(bHLH) C2C12-MyoG-ChIP-Seq(GSE36024) Homer
4		Ptf1a(bHLH) Panc1-Ptf1a-ChIP-Seq(GSE47459) Homer
5		MyoD(bHLH) Myotube-MyoD-ChIP-Seq(GSE21614) Homer
6		Tcf12(bHLH) GM12878-Tcf12-ChIP-Seq(GSE32465) Homer
7		Tcf21(bHLH) ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369) Homer
8		Atoh1(bHLH) Cerebellum-Atoh1-ChIP-Seq(GSE22111) Homer
9		Ascl1(bHLH) NeuralTubes-Ascl1-ChIP-Seq(GSE55840) Homer
10		c-Myc(bHLH) mES-cMyc-ChIP-Seq(GSE11431) Homer

Figure 3.25: HOMER results for top ten known motifs of HUVEC T7d cell line

3.4 LncRNA Literature

Growth-arrest specific transcript 5(GAS5) is a lncRNA that has the potential to regulate cell growth, apoptosis, and cell division. The 5'-terminal oligopyrimidine (TOP) sequence is involved in protein synthesis, and the mTOR pathway slightly affects the translation of 5'-TOP RNAs, which controls mRNA expression of GAS5(26). In diabetes patients, this level of mRNA expression of GAS5 has been observed to reduce which increases the potential for the occurrence of diabetes. GAS5 can bind to promoter region of the insulin receptor, and its expression is modified in individuals with type 2 diabetes mellitus.(27) In another study, in qPCR, people with absolute GAS5 less than 10nanogram/microliter are nearly twelve times more at risk of being diagnosed with diabetes(28). This further proves our conclusions that levels of GAS5 expression are correlated with the occurrence of type 2 Diabetes Mellitus.

LncRNA GAS5 was also observed to be related to several age-related diseases(29) like rheumatoid arthritis, atherosclerosis, cancer, osteoarthritis, osteoporosis, etc. Their RNA secondary structures can fold and compete with Glucocorticoid Receptor (GR) for binding. So, GAS5 can regulate cell death by acting as a competitive inhibitor. Disorders related to bone were thought to be due to disorders of calcium and phosphorous absorption and assimilation. Although, recent studies have proved that bone-derived cells and similar chemicals are also involved in the regulation and progression of these diseases. Some lncRNAs like H19 have suspected involvement in osteoporosis and other bone disorders(30). The involvement of GAS5 in bone studies has yet to be fully explored.

Compared to normal cells, GAS5 is upregulated in cancerous and other diseased cells(31)(32). With the increase in the level of sequencing-based methods, hopefully we will be able to understand the additional signaling pathways along with their targets. In another study, Metformin which has an anticancerous effect on several cancers was studied if there is a relation between it, lncRNA, and breast cancer. It was found that Metformin(28) increases the lncRNA GAS5 expression, which inhibits the over-activation of the mTOR pathway, which is directly responsible for inhibiting growth and promoting cell death in Breast Cancer cells, thereby introducing a new treatment for Breast cancer. Therefore, much work is still required to validate the literature observations.

Chapter 4

Conclusion

We have tried to highlight a new dimension of regulation of functions and genes in cells through the analysis of RNA-DNA interaction with exonic regions, promoter regions, repeat regions, and histone modifications. Through our analysis, we have provided a glance at the most interesting questions. However, the most interesting question that we had sought was if we could discern a relationship between RNA-DNA interaction and the regulatory effects on repeat elements. Through network analysis, several repeat name regions were common among the four cell lines which were involved in the regulatory process.

We found certain genes where exonic regions RNA appear to be interacting at many DNA binding sites in our data, and they show variability in activating and repressive histone modifications. Several exons showed a negative correlation with their antisense counterparts in activating and repressing histone modifications, which further validated these findings, like GAS5, THBS1, FLNB, and MEG3.

We identified DNA bound by RNA of a few genes where there were high activating histone modifications, and by selecting these specific genes, we performed gene ontology. Through Gene Ontology, the significance of these lncRNAs was grasped as these are mostly involved in important pathways in cell regulation, like regulation of cellular localization, carbohydrate biosynthetic process, and regulation of intracellular transport with significant p values. Through this, we were alerted to their potential and researched if there was a common motif region where they are binding and affecting regulation. Further studies might be beneficial and verification in laboratories with actual data. There is also the possibility to refine this workflow further to ascertain more significant lncRNAs and remove outliers.

This work presents some novel findings and observations. Firstly, a network of the lncRNAs interacting with RNA and DNA along with their expressions in cells. This provides valuable insights for the analysis of their relationship. Secondly, the relation-

ship of exons and their expression with activating and repressive histone modifications. This is vital to identify the exons that are most prominent in their expression, and through them, we were able to identify for which pathways they are responsible. This has huge potential benefits because we can use the expression of these lncRNAs for disease prediction in real life and personalized healthcare. GAS5 is reported to be responsible for many diseases, including heart diseases and bone diseases, and its expression levels can paint a story of a person's well-being.

Bibliography

- [1] D. H. Wolf and R. Menssen, “Mechanisms of cell regulation–proteolysis, the big surprise,” *FEBS letters*, vol. 592, no. 15, pp. 2515–2524, 2018.
- [2] K. Schafer, “The cell cycle: a review,” *Veterinary pathology*, vol. 35, no. 6, pp. 461–478, 1998.
- [3] L. Mariño-Ramírez, M. G. Kann, B. A. Shoemaker, and D. Landsman, “Histone structure and nucleosome stability,” *Expert review of proteomics*, vol. 2, no. 5, pp. 719–729, 2005.
- [4] J. Harrow, A. Nagy, A. Reymond, T. Alioto, L. Patthy, S. E. Antonarakis, and R. Guigó, “Identifying protein-coding genes in genomic sequences,” *Genome Biology*, vol. 10, no. 1, pp. 1–8, 2009.
- [5] J. W. Hershey, N. Sonenberg, and M. B. Mathews, “Principles of translational control: an overview,” *Cold Spring Harbor perspectives in biology*, vol. 4, no. 12, p. a011528, 2012.
- [6] G. G. Wang, C. D. Allis, and P. Chi, “Chromatin remodeling and cancer, part i: Covalent histone modifications,” *Trends in molecular medicine*, vol. 13, no. 9, pp. 363–372, 2007.
- [7] B. D. Strahl and C. D. Allis, “The language of covalent histone modifications,” *Nature*, vol. 403, no. 6765, pp. 41–45, 2000.
- [8] Y. Ramakrishnaiah, L. Kuhlmann, and S. Tyagi, “Computational approaches to functionally annotate long noncoding rna (lncrna),” 2020.
- [9] T. C. Nguyen, K. Zaleta-Rivera, X. Huang, X. Dai, and S. Zhong, “Rna, action through interactions,” *Trends in Genetics*, vol. 34, no. 11, pp. 867–882, 2018.
- [10] B. Sridhar, M. Rivas-Astroza, T. C. Nguyen, W. Chen, Z. Yan, X. Cao, L. Hebert, and S. Zhong, “Systematic mapping of rna-chromatin interactions in vivo,” *Current Biology*, vol. 27, no. 4, pp. 602–609, 2017.

- [11] Z. Yan, N. Huang, W. Wu, W. Chen, Y. Jiang, J. Chen, X. Huang, X. Wen, J. Xu, Q. Jin *et al.*, “Genome-wide colocalization of rna–dna interactions and fusion rna pairs,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 3328–3337, 2019.
- [12] M. D. Simon, C. I. Wang, P. V. Kharchenko, J. A. West, B. A. Chapman, A. A. Alekseyenko, M. L. Borowsky, M. I. Kuroda, and R. E. Kingston, “The genomic binding sites of a noncoding rna,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 51, pp. 20 497–20 502, 2011.
- [13] I. Antonov and Y. Medvedeva, “Direct interactions with nascent transcripts is potentially a common targeting mechanism of long non-coding rnas,” *Genes*, vol. 11, no. 12, p. 1483, 2020.
- [14] Y. Li, C.-y. Chen, A. M. Kaye, and W. W. Wasserman, “The identification of cis-regulatory elements: A review from a machine learning perspective,” *Biosystems*, vol. 138, pp. 6–17, 2015.
- [15] G. Hu, F. Niu, B. A. Humburg, K. Liao, S. Bendi, S. Callen, H. S. Fox, and S. Buch, “Molecular mechanisms of long noncoding rnas and their role in disease pathogenesis,” *Oncotarget*, vol. 9, no. 26, p. 18648, 2018.
- [16] S. Van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes, “Gene co-expression analysis for functional classification and gene–disease predictions,” *Briefings in bioinformatics*, vol. 19, no. 4, pp. 575–592, 2018.
- [17] A. R. Quinlan and I. M. Hall, “Bedtools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.
- [18] R. Calandrelli, L. Xu, Y. Luo, W. Wu, X. Fan, T. Nguyen, C.-J. Chen, K. Sriram, X. Tang, A. B. Burns *et al.*, “Stress-induced rna–chromatin interactions promote endothelial dysfunction,” *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.
- [19] W. Liu, Q. Ma, K. Wong, W. Li, K. Ohgi, J. Zhang, A. K. Aggarwal, and M. G. Rosenfeld, “Brd4 and jmjd6-associated anti-pause enhancers in regulation of transcriptional pause release,” *Cell*, vol. 155, no. 7, pp. 1581–1595, 2013.
- [20] A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen *et al.*, “Chip-seq accurately predicts tissue-specific activity of enhancers,” *Nature*, vol. 457, no. 7231, pp. 854–858, 2009.
- [21] H. Fan, J. Lu, Y. Guo, D. Li, Z.-M. Zhang, Y.-H. Tsai, W.-C. Pi, J. H. Ahn, W. Gong, Y. Xiang *et al.*, “Bahcc1 binds h3k27me3 via a conserved bah module to mediate gene silencing and oncogenesis,” *Nature genetics*, vol. 52, no. 12, pp. 1384–1396, 2020.

- [22] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker *et al.*, “The nih roadmap epigenomics mapping consortium,” *Nature biotechnology*, vol. 28, no. 10, pp. 1045–1048, 2010.
- [23] E. P. Consortium *et al.*, “An integrated encyclopedia of dna elements in the human genome,” *Nature*, vol. 489, no. 7414, p. 57, 2012.
- [24] M. Tausendschön, M. Rehli, N. Dehne, C. Schmidl, C. Döring, M.-L. Hansmann, and B. Brüne, “Genome-wide identification of hypoxia-inducible factor-1 and-2 binding sites in hypoxic human macrophages alternatively activated by il-10,” *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1849, no. 1, pp. 10–22, 2015.
- [25] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [26] J. L. Esguerra, J. K. Ofori, M. Nagao, Y. Shuto, A. Karagiannopoulos, J. Fadista, H. Sugihara, L. Groop, and L. Eliasson, “Glucocorticoid induces human beta cell dysfunction by involving riborepressor gas5 lincrna,” *Molecular metabolism*, vol. 32, pp. 160–167, 2020.
- [27] G. Carter, B. Miladinovic, A. A. Patel, L. Deland, S. Mastorides, and N. A. Patel, “Circulating long noncoding rna gas5 levels are correlated to prevalence of type 2 diabetes mellitus,” *BBA clinical*, vol. 4, pp. 102–107, 2015.
- [28] Y. Jiang, T. Qian, S. Li, Y. Xie, and M. Tao, “Metformin reverses tamoxifen resistance through the lincrna gas5-mediated mtor pathway in breast cancer,” *Annals of Translational Medicine*, vol. 10, no. 6, 2022.
- [29] Y. Wang, M. Xue, F. Xia, L. Zhu, D. Jia, Y. Gao, L. Li, Y. Shi, Y. Li, S. Chen *et al.*, “Long non-coding rna gas5 in age-related diseases,” *Current medicinal chemistry*, vol. 29, no. 16, pp. 2863–2877, 2022.
- [30] Z. Zhou, J. Chen, Y. Huang, D. Liu, S. Chen, and S. Qin, “Long noncoding rna gas5: A new factor involved in bone diseases,” *Frontiers in Cell and Developmental Biology*, vol. 9, 2021.
- [31] M. R. Pickard and G. T. Williams, “Molecular and cellular mechanisms of action of tumour suppressor gas5 lincrna,” *Genes*, vol. 6, no. 3, pp. 484–499, 2015.
- [32] Q. Gao, H. Xie, H. Zhan, J. Li, Y. Liu, and W. Huang, “Prognostic values of long noncoding rna gas5 in various carcinomas: an updated systematic review and meta-analysis,” *Frontiers in physiology*, vol. 8, p. 814, 2017.