CLIP-SEQ DATA ANALYSIS: From pre-processing to motif detection.

Presented To
Professor Wendy Lee
Department of Computer Science
San José State University

In Partial Fulfillment
Of the Requirements for the
Class CS123B

By

Shwethal Sayeeram Trikannad and Saumi Saurin Shah

May 2024

I. INTRODUCTION

CLIP-seq is the integration of in vivo protein Ultraviolet (UV) crosslinking with RNA immunoprecipitation and high throughput sequencing. This technique is implemented to identify ribonucleic acid (RNA) targets and binding sites for RNA binding proteins (RBPs). It enables comprehensive analysis from preprocessing raw data to detecting regulatory motifs within RNA sequences. However, this method is prone to creation of duplicates during Polymerase Chain Reaction (PCR) amplification. A variation of this technology, eCLIP-seq [1], reduces duplicates by 1000-fold. The data collected after these steps need to be analyzed to achieve conclusive results.

This project presents an entire bioinformatics pipeline, executed on Google Cloud Platform (GCP) [2] applying numerous tools to obtain meaningful results. The dataset is a subset of the RBFOX2 eCLIP data, originating from a study conducted by Nostrand et al. [1] in 2016.The dataset contains an experiment and control paired-end (PE) sequence data obtained from the Hep G2 cancer cell line, specifically treated with eCLIP-seq. The control referred to as input data, is an internal control dataset obtained from the same library well, prior to isolation of experiment data. The input files also contain a hg38 chromosomes sizes text file and a hg38 Gene Transfer File (GTF) annotated file from Ensembl.

RBFOX2 [3] has been implicated for its regulatory role in the Epithelial-Mesenchymal transition (EMT) pathway responsible for cancer progression and metastasis. This RNA binding protein exerts its effects by influencing alternative splicing and other post transcriptional modifications. Literature highlights a conserved RNA element UGCAUG as its primary binding site and transcripts and introns as its major targets. Our objective is to meticulously analyze and validate the datasets to ascertain the presence of this conserved motif as TGCATG on deoxyribonucleic acid (DNA) sequence. Through this process, we aim to elucidate the functional role of RBFOX2 and provide insights into the source and role of the targeted RNA.

II. METHODS AND MATERIALS

A total of 10 steps involving over 19 tools were used to create a comprehensive bioinformatics pipeline to collect and analyze data to procure decisive results and insights. A virtual machine (VM) instance was created on Google Cloud Platform. The project was completed by running commands in a custom conda [4] environment using Secure Shell (SSH) of the VM. Figure 1 displaying the steps and primary tools is given below.

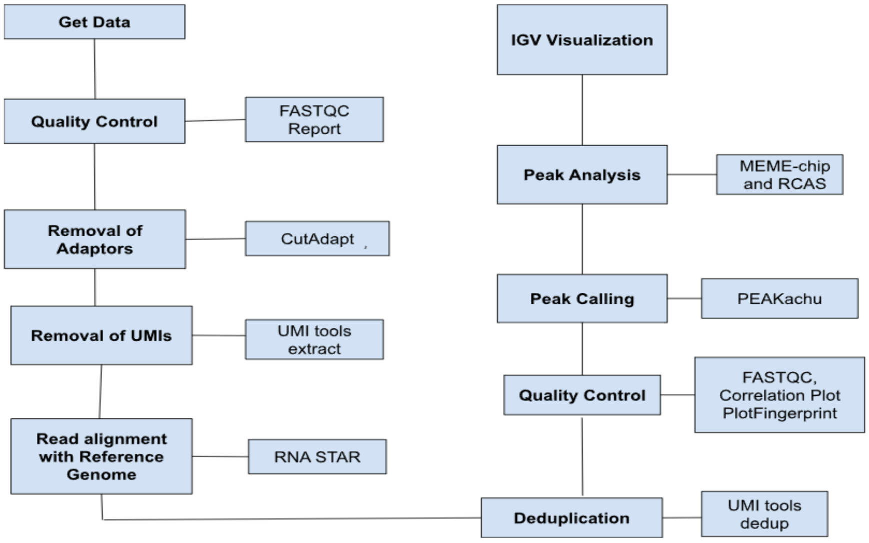
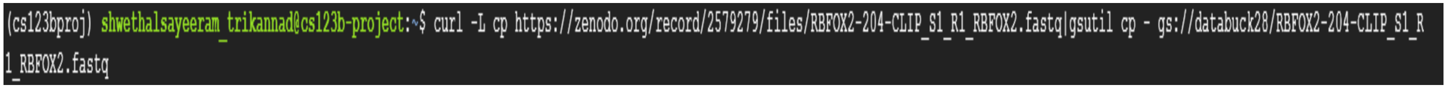


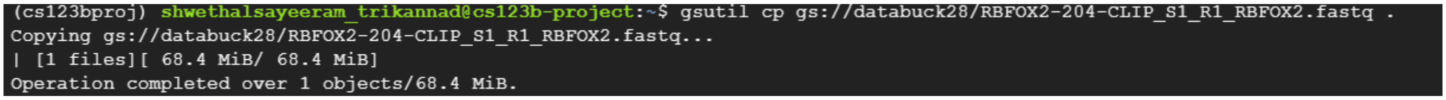
Fig.1 Project workflow with steps and relevant tools

Step 1. Get data:



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ curl -L cp https://zenodo.org/record/2579279/files/RBFOX2-204-CLIP_S1_R1_RBFOX2.fastq|gsutil cp - gs://databuck28/RBFOX2-204-CLIP_S1_R
1_RBFOX2.fastq
```

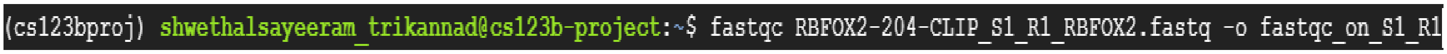
This command utilizes the curl command to get Sample 1 , Read 1 experiment data from the Zenodo website and is piped and copied into custom bucket databuck28.



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ gsutil cp gs://databuck28/RBFOX2-204-CLIP_S1_R1_RBFOX2.fastq .
Copying gs://databuck28/RBFOX2-204-CLIP_S1_R1_RBFOX2.fastq...
| [1 files][ 68.4 MiB/ 68.4 MiB]
Operation completed over 1 objects/68.4 MiB.
```

This command copies the Sample 1 , Read 1 experiment data into local directory. The above steps are repeated to copy Sample 1, Read 2 experiment data, Sample 2 Read 1 and Read 2 control data, hg38 gtf file, hg38 chromosome sizes txt file into local directory.

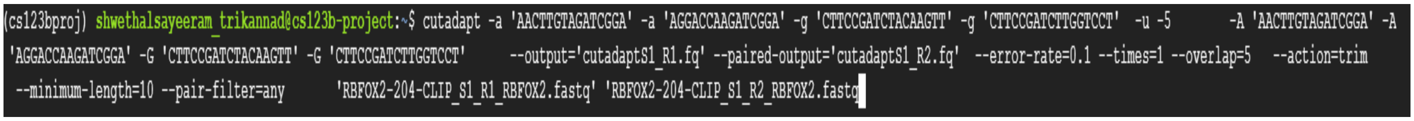
Step 2. Quality Control: Tool used : FastQC



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ fastqc RBFOX2-204-CLIP_S1_R1_RBFOX2.fastq -o fastqc_on_S1_R1
```

This command uses the installed FastQC tool and runs the tool on Sample 1, Read 1 experiment data to obtain a FastQC HTML report. The same command is run on Sample 1, Read 2 experiment data, Sample 2 Read 1 and Read 2 data as well.

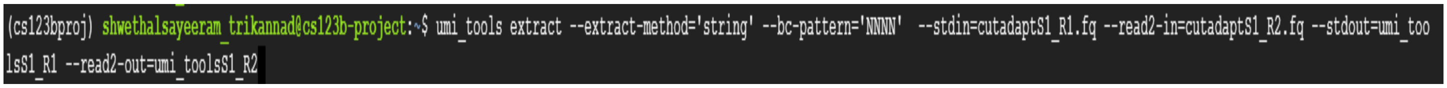
Step 3. Removal of adapters: Tool used : CutAdapt



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ cutadapt -a 'AACTTGTAGATCGGA' -a 'AGGACCAAGATCGGA' -g 'CTTCCGATCTACAAGTT' -g 'CTTCCGATCTTGGTCCT'  -u -5       -A 'AACTTGTAGATCGGA' -A
'AGGACCAAGATCGGA' -G 'CTTCCGATCTACAAGTT' -G 'CTTCCGATCTTGGTCCT'     --output='cutadaptS1_R1.fq' --paired-output='cutadaptS1_R2.fq'  --error-rate=0.1 --times=1 --overlap=5   --action=trim
 --minimum-length=10 --pair-filter=any      'RBFOX2-204-CLIP_S1_R1_RBFOX2.fastq' 'RBFOX2-204-CLIP_S1_R2_RBFOX2.fastq
```

This command removes adapter sequences and trims low-quality bases from sequencing reads, and then outputs the trimmed reads into separate files for forward and reverse reads. The same command is run on Sample 2 , Read 1, and Read 2 control data.

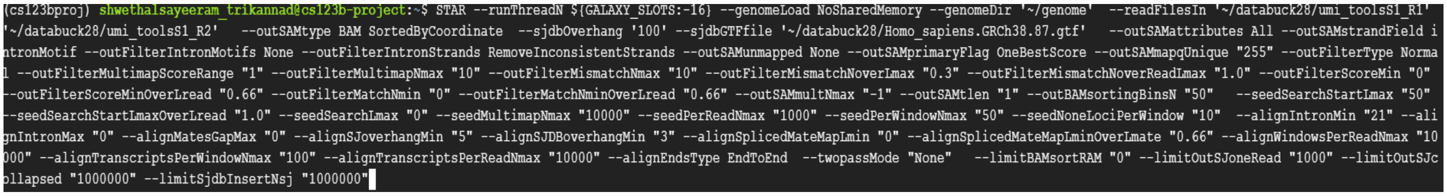
Step 4. Removal of UMIs: Tool used : UMI tools (feature – extract)



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ umi_tools extract --extract-method='string' --bc-pattern='NNNN'  --stdin=cutadaptS1_R1.fq --read2-in=cutadaptS1_R2.fq --stdout=umi_too
lsS1_R1 --read2-out=umi_toolsS1_R2
```

Unique molecular identifiers (UMIs) [5] are short randomly generated nucleotides that are attached to sequence fragments before amplification instead of a barcode. They help in the reduction of PCR duplicates. The above command extracts UMI reads from the cutadapt fastq files, attaches them to read id of the specific sequence and stores all the extracted sequences in a new file. The same is repeated for the control fastq files.

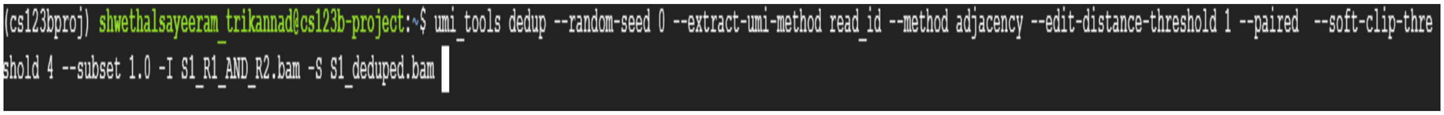
Step 5. Read Alignment with Reference Genome: Tool used :RNA star



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ STAR --runThreadN ${GALAXY_SLOTS:-16} --genomeLoad NoSharedMemory --genomeDir '~/genome'  --readFilesIn '~/databuck28/umi_toolsS1_R1'
'~/databuck28/umi_toolsS1_R2'   --outSAMtype BAM SortedByCoordinate  --sjdbOverhang '100' --sjdbGTFfile '~/databuck28/Homo_sapiens.GRCh38.87.gtf'   --outSAMattributes All --outSAMstrandField i
ntronMotif --outFilterIntronMotifs None --outFilterIntronStrands RemoveInconsistentStrands --outSAMunmapped None --outSAMprimaryFlag OneBestScore --outSAMmapqUnique "255" --outFilterType Norma
l --outFilterMultimapScoreRange "1" --outFilterMultimapNmax "10" --outFilterMismatchNmax "10" --outFilterMismatchNoverLmax "0.3" --outFilterMismatchNoverReadLmax "1.0" --outFilterScoreMin "0"
--outFilterScoreMinOverLread "0.66" --outFilterMatchNmin "0" --outFilterMatchNminOverLread "0.66" --outSAMmultNmax "-1" --outSAMtlen "1" --outBAMsortingBinsN "50"   --seedSearchStartLmax "50"
--seedSearchStartLmaxOverLread "1.0" --seedSearchLmax "0" --seedMultimapNmax "10000" --seedPerReadNmax "1000" --seedPerWindowNmax "50" --seedNoneLociPerWindow "10"  --alignIntronMin "21" --ali
gnIntronMax "0" --alignMatesGapMax "0" --alignSJoverhangMin "5" --alignSJDBoverhangMin "3" --alignSplicedMateMapLmin "0" --alignSplicedMateMapLminOverLmate "0.66" --alignWindowsPerReadNmax "10
000" --alignTranscriptsPerWindowNmax "100" --alignTranscriptsPerReadNmax "10000" --alignEndsType EndToEnd  --twopassMode "None"   --limitBAMsortRAM "0" --limitOutSJoneRead "1000" --limitOutSJc
ollapsed "1000000" --limitSjdbInsertNsj "1000000"
```

RNA star [6] is an alignment tool and was selected for this step as it is splice aware, accepts both genomic and transcriptomic data and has a high mapping rate due to flexibility in soft clipping selection. As this tool is very RAM intensive a new Virtual Machine was created with 64gb RAM and 100gb persistent disk space. A genome index was created with the help of hg38 primary assembly gene file and hg38 gtf file. The above command was then run to align the fastq files from the previous step to the reference genome. The same was repeated for control data.

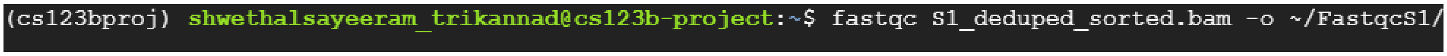
Step 6. Deduplication: Tool used : UMI tools (feature – dedup)



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ umi_tools dedup --random-seed 0 --extract-umi-method read_id --method adjacency --edit-distance-threshold 1 --paired  --soft-clip-thre
shold 4 --subset 1.0 -I S1_R1_AND_R2.bam -S S1_deduped.bam
```

The above command uses the bam output file from RNA star and removes all duplicates by checking the UMI attached on the read id. Samtools sort is then used to sort this output file. The same is carried out for control data.

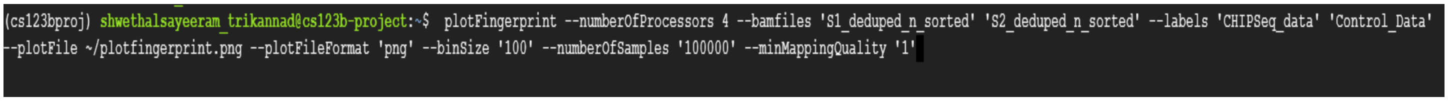
Step 7. Quality Control: Tool used: FastQC



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ fastqc S1_deduped_sorted.bam -o ~/FastqcS1/
```

This command runs fastqc to create an HTML report on the output bam file from the previous step for both experimental and control data.

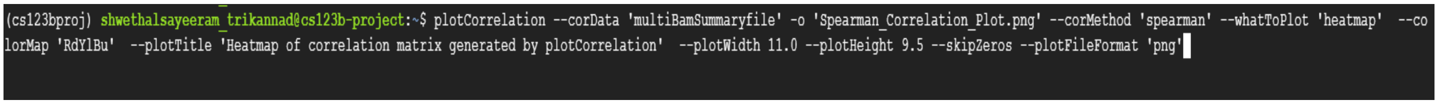
Tool used : plotFingerprint (deeptools)



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$  plotFingerprint --numberOfProcessors 4 --bamfiles 'S1_deduped_n_sorted' 'S2_deduped_n_sorted' --labels 'CHIPSeq_data' 'Control_Data'
--plotFile ~/plotfingerprint.png --plotFileFormat 'png' --binSize '100' --numberOfSamples '100000' --minMappingQuality '1'
```

The above command uses a tool called plotFingerprint from the deeptools suite to create a plot of both experiment and control bam files from the UMI tools deduplication step. The output is a plot in the png format.

Tool used: plotCorrelation (deeptools)



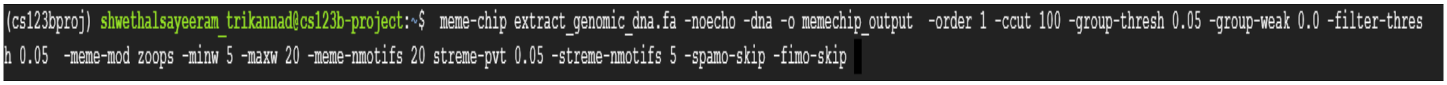
```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$ plotCorrelation --corData 'multiBamSummaryfile' -o 'Spearman_Correlation_Plot.png' --corMethod 'spearman' --whatToPlot 'heatmap'  --co
lorMap 'RdYlBu'  --plotTitle 'Heatmap of correlation matrix generated by plotCorrelation'  --plotWidth 11.0 --plotHeight 9.5 --skipZeros --plotFileFormat 'png'
```

MultiBamSummary is a tool that divides the entire sequence into bins and outputs a bam file. Experiment and control bam files from the UMI tools deduplication step are passed into this tool. The above command uses a tool called plotCorrelation from the deeptools suite to create a heatmap from the multiBamSummary bam file. The correlation plot is created with the spearman correlation method.

Step 8. Peak calling: Tool used: PEAKachu

Peakachu [7] is a peak calling tool that identifies regions of gene enrichment. As there was an error in the source code of the command line version of the tool, with approval from Dr. Lee we used the output files obtained after running the tool on galaxy in our project.

Step 9. Peak Analysis: Tool used: MemeChip



```
(cs123bproj) shwethalsayeeram_trikannad@cs123b-project:~$  meme-chip extract_genomic_dna.fa -noecho -dna -o memechip_output  -order 1 -ccut 100 -group-thresh 0.05 -group-weak 0.0 -filter-thres
h 0.05  -meme-mod zoops -minw 5 -maxw 20 -meme-nmotifs 20 streme-pvt 0.05 -streme-nmotifs 5 -spamo-skip -fimo-skip
```

MemeChip [8] is a suite of tools essential to identify motifs in input sequences. It has three major algorithms meme, streme and dreme with dreme being specifically designed to identify motifs less than 8 bases in length. Dreme however, is deprecated in the command line version of the tool deck. The tabular file output from Peakchu is reformatted to a bed file and then processed with the help of slopBed and getFasta from the bedtools suite to obtain genomic intervals of specific sizes. This processed fasta file is the input file for memechip to extract conserved motifs. The command above runs both meme and streme algorithms on the data.

Tool used: RNA Centric Anotation System (RCAS): This tool is an R package by Bioconductor. It creates a comprehensive analysis report using a hg38 annotated gtf file and the output bed file from slopBed. This tool was run on our local computer on Rstudio with the programming language R to obtain the HTML report.

Additional step. Integrated Genomics Visualization: Tool: Integrated Genomics Visualizer (IGV): After extracted alignment ends, sorting bed file with sortBed, creating a bedgraph and then converting it into a bigwig file for both experiment and control data on the VM the bigwig files are downloaded onto our local computer. IGV is then installed and used to visualize the bigwig files against the hg38 genome and hg38 gtf file. This step is done to visualize signal intensities or peaks.

III. RESULTS

A number of plots and HTML reports will be analyzed to gain insight into the functioning of RBFOX2. Fig 2. represents FastQC reports of experiment and control data prior to trimming. Fig 3. Showcases FastQC reports of experiment and control data after trimming and deduplication. There is a significant reduction in sequence duplication levels along with removal of adapter content. Fig 4. displays the plot obtained after running plotFingerprint. The Y-axis represents a fraction of the total number of reads in a random segment of the gene. The X-axis is the genome coverage area with 0 being the lowest and 1.0 the highest. The plot shows a perfect elbow with both ClipSeq data and control data aligning perfectly on top of each other. This, however, does not mean that both samples are perfectly correlated. It could be because of the small samples size owing to the fact that subsets of actual samples are being used. Using an internal control can also influence plots to behave this way. This result is proved wrong by the correlation plot in Fig 5. which shows two separate heterogenous clusters of each sample showing no positive correlation between the two. An MA plot which is the output file of PEAKachu is highlighted in Fig 6. The MA plot represents differential gene expression. The X axis maps the average expression level of genes for both control and experimental groups and the Y axis maps the log 2 ratio of difference in gene expression levels for both the control and the experimental groups. Each dot on the plot signifies a gene with blue representing control and red experiment data. There

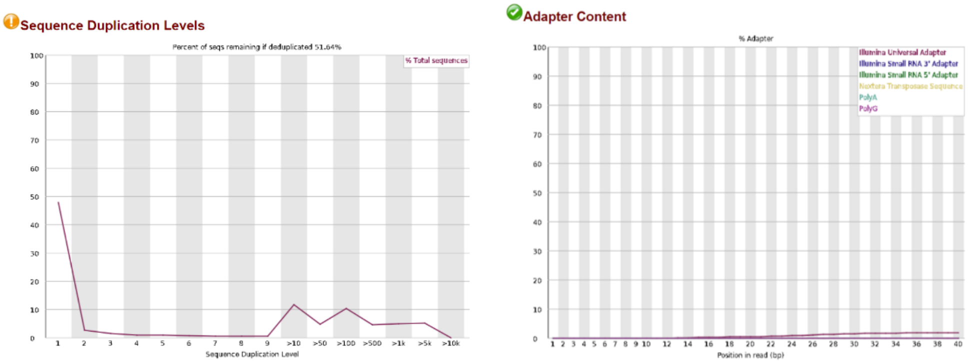


Fig 2. FastQC before trimming and deduplication.

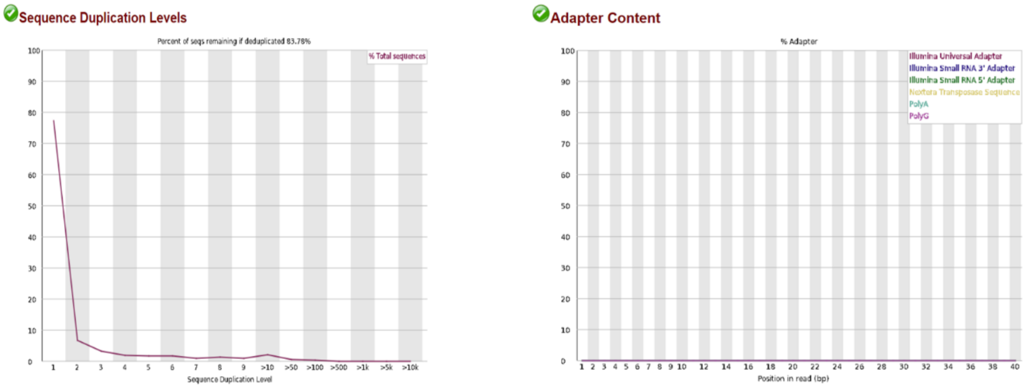


Fig 3. FastQC after trimming and deduplication.

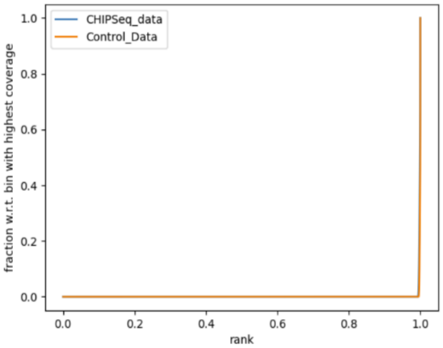


Fig 4. Plotfingerprint

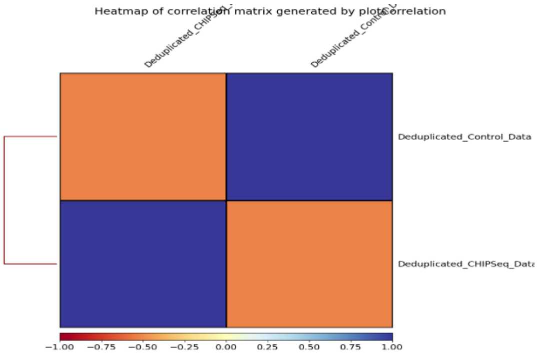


Fig 5. Correlation Plot

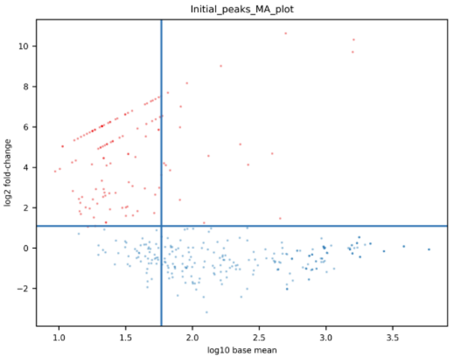


Fig 6. MA plot

is difference in gene expression due to the disparity in arrangement of gene counts with significant upregulation in experiment gene counts. Memechip tools meme and streme were unable to identify conserved motif TGCATG as the deprecated dreme algorithm was responsible for identifying this motif. According to RCAS reports the target regions of RBFOX2 are 81% introns and 98% transcripts. RBFOX2 also appears to favor 3’ end of intron-exon junctions and transcripts. The Functional expression analysis section of the report reveals that target RNA are involved in actin filament formation, actin myoskeleton and folic acid functioning. Fig 7 shows peaks observed in IGV.These peaks belong to ClipSeq data as compared to control which seems relatively quiet. This validates literature and showcases the binding sites of RBFOX2 across the hg38 genome.

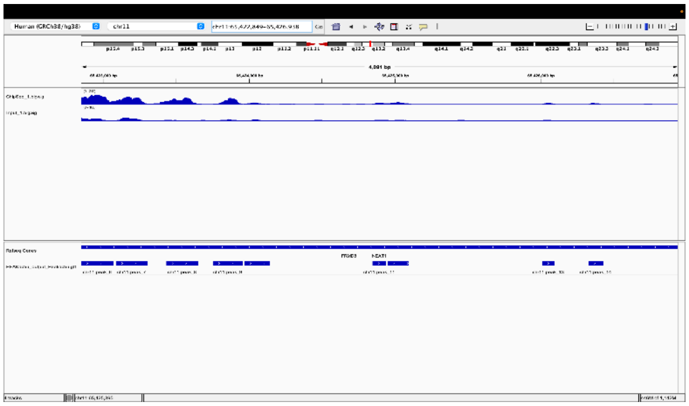


Fig 7. Integrated Genome Visualizer

IV. DISCUSSION

This project aimed to answer three important questions regarding identifying conserved motif TGCATG, function of both RBFOX2 and target RNA. Due to dreme tool deprecation we were unable to find the motif. However, we were successful in elucidating the function of RBFOX2, particularly its binding affinity towards the 3' end of RNA molecules and intron junctions. Notably, our investigation highlighted a predominant enrichment of RBFOX2 targets originating from pivotal cellular components such as the actin cytoskeleton, actin filaments, and folic acid bodies. These findings contribute significantly to our understanding of RBFOX2's regulatory role and provide insights into its functional relevance in cellular processes associated with these structures.

References

In-Text No.	Citation
[1]	Van Nostrand, E., Pratt, G., Shishkin, A. et al.2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods 13, (March 2016), 508–514. https://doi.org/10.1038/nmeth.3810
[2]	Bayan H. Banimfreg. 2023. A comprehensive review and conceptual framework for cloud computing adoption in bioinformatics, Healthcare Analytics, Volume 3, (November 2023). https://doi.org/10.1016/j.health.2023.100190.
[3]	Braeutigam, C., Rago, L., Rolke, A. et al. 2014. The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. Oncogene 33, (February 2014), 1082–1092. https://doi.org/10.1038/onc.2013.50
[4]	Safdar, Nimra.2024. Revolutionizing Bioinformatics Unleashing Conda's Full Potential in Google Colab. (January 2024)
[5]	Tsagiopoulou M, Maniou MC, Pechlivanis N, Togkousidis A, Kotrová M, Hutzenlaub T, Kappas I, Chatzidimitriou A and Psomopoulos F. 2021. UMIc: A Preprocessing Method for UMI Deduplication and Reads Correction. Front. Genet. 12 (May 2021). doi: 10.3389/fgene.2021.660366
[6]	Dobin A, Gingeras TR. 2015. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51:11.14.1-11.14.19. (Sep 2015). doi:10.1002/0471250953.bi1114s51
[7]	Terlouw, Barbara & Vromans, Sophie & Medema, Marnix. 2022. PIKAChU: a Python-based Informatics Kit for Analysing Chemical Units. (January 2022) 10.21203/rs.3.rs-1239072/v1.
[8]	Tran NT, Huang CH. 2014. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. Biol Direct, (February 2014). doi:10.1186/1745-6150-9-4
[9]	Uyar B, Yusuf D, Wurmus R, Rajewsky N, Ohler U, Akalin A. 2017. RCAS: an RNA centric annotation system for transcriptome-wide regions of interest. Nucleic Acids Res. (June 2017);45(10):e91. doi:10.1093/nar/gkx120