BF528 Project 3: ChIPseq analysis of the human transcription factor RUNX1

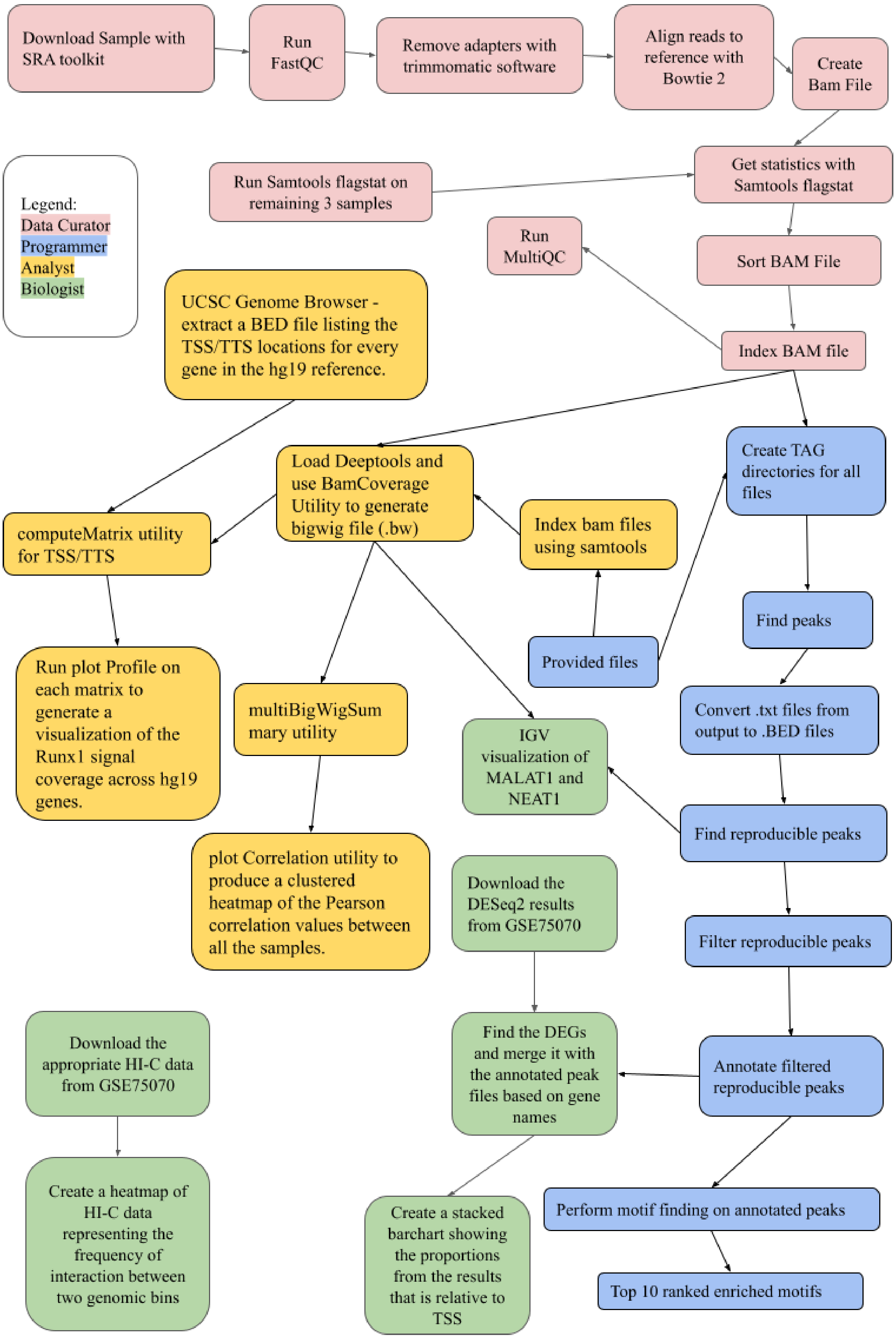
Authors: Manasa Rapuru (Data Curator), Pragya Rawat (Programmer), Pooja Savla (Analyst), Vrinda Jethalia (Biologist)

Date: 5th April, 2023

Introduction:

The RUNX protein family consists of critical transcription factors that regulate a wide range of biological processes to mobilize proper cell fate determination (Chuang et al, 2012). RUNX1 is a transcription factor that has been observed in breast cancer cells to perform both the roles of an oncogene and a tumor suppressor depending on its interacting partners (Barutcu et al, 2016). It does so by participating in pathways that alter chromatin structure in cooperation with chromatin modifiers and remodeling enzymes. In this paper, the authors attempted to investigate the role of RUNX1 in breast cancer cells by suppressing its expression in the human MCF-7 breast cancer cell line and a combination of genome-wide chromatin conformation capture (Hi-C) and ChIP-seq techniques to analyze the effects (Barutcu et al, 2016). They found that RUNX1 is enriched at regions of the genome that are involved in long-range chromatin interactions and that the depletion of RUNX1 leads to changes in the chromatin landscape. Moreover, the researchers identified a set of genes that are regulated by RUNX1, many of which are involved in cellular processes such as cell cycle progression and DNA damage response. We attempted to recreate the results of this paper using the replicate ChIP-seq data for our analysis. The workflow describes our process which is explained further in the methods section.

Workflow: Figure describing steps taken to process and analyse data


Download Sample with SRA toolkit
Run FastQC
Remove adapters with trimmomatic software
Align reads to reference with Bowtie 2
Create Bam File
Get statistics with Samtools flagstat
Run Samtools flagstat on remaining 3 samples
Sort BAM File
Run MultiQC
Index BAM file
Legend:
Data Curator
Programmer
Analyst
Biologist
UCSC Genome Browser - extract a BED file listing the TSS/TTS locations for every gene in the hg19 reference.
Load Deeptools and use BamCoverage Utility to generate bigwig file (.bw)
Create TAG directories for all files
computeMatrix utility for TSS/TTS
Index bam files using samtools
Find peaks
Provided files
Run plot Profile on each matrix to generate a visualization of the Runx1 signal coverage across hg19 genes.
multiBigWigSummary utility
IGV visualization of MALAT1 and NEAT1
Convert .txt files from output to .BED files
Find reproducible peaks
plot Correlation utility to produce a clustered heatmap of the Pearson correlation values between all the samples.
Download the DESeq2 results from GSE75070
Filter reproducible peaks
Download the appropriate HI-C data from GSE75070
Find the DEGs and merge it with the annotated peak files based on gene names
Annotate filtered reproducible peaks
Create a heatmap of HI-C data representing the frequency of interaction between two genomic bins
Create a stacked barchart showing the proportions from the results that is relative to TSS
Perform motif finding on annotated peaks
Top 10 ranked enriched motifs

Data:

The samples in this study come from RUNX-1 depleted and MCF-7 control breast cancer cells.

Methods:

Data acquisition

The data of two ChIP–seq experiments was used in this project. Each pulldown sample along with the corresponding input controls were used in this project. The processing of the fastq files for the control samples of replicate 1 and replicate 2, along with the corresponding pulldown sample for replicate 2 were already processed the steps for alignment. It was the task of the data curator to obtain and process the pulldown sample of replicate one (SRR2919475, 'MCF-7 wildtype RUNX1 ChIPseq Replicate1 Pulldown') .
The first step was to download this sample from GEO Accession. It was downloaded using SRA toolkit's prefetch and fastq-dump modules. Next, the fastq files were processed using a snakemake workflow. The steps of this workflow were to run the following programs in this order:FASTQC, Trimmomatic, Bowtie2, Samtools flagstat, Samtools sort, and Samtools index. FASTQC is a software that measures various metrics that is used to understand the quality of a file((2015), "FastQC," https://qubeshub.org/resources/fastqc). Trimmomatic is used to remove adapters from the sides of reads (Bolger et al, 2014).The adapters removed from reads are 'AGATCGGAAGAGCGTCG
TGTAGGGAAAGAGTGTA' and 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'. Bowtie2 was used to align and map reads to the reference sequence (Langmead et al,2012). Sorting and indexing of a genome file is a method that allows for a more efficient search and extraction of reads. It is also used in later visualization steps such as IGV (Robinson et al, 2011). And SAMTOOLS flagstat is used to tally the amount of reads in a file that have each of the flag types (Li et al, 2009). Lastly, MultiQC was used to create a condensed report and graphs of the earlier steps (Ewels, et al 2016).

Reproducible peak finding and filtering

The alignments generated in the previous step were used to perform peak calling using Homer (Heinz et al, 2010). Each of the four sorted BAM files (two for each replicate) were used to create TAG directories since they provide several useful functions for peak calling including providing useful metadata such as information about the read quality, mapping quality, and other associated metadata (Heinz et al, 2010). Tag directories were generated using the default parameters. Peak calling is a statistical procedure that allows us to identify regions of the genome where the RUNX1 protein is bound to the DNA (Akalin et al, 2016). For this analysis, we used the positive control Inp files as reference to find the peaks for each replicate, generating one peak file for each replicate. For this purpose, we used the findPeaks utility in HOMER with the -style factor mode. The output files from this step were converted into a BED file for downstream analysis using pos2bed, another function of Homer. From these files 5070 reproducible peaks were identified (Fig4.B) using the intersect function available in the Bedtools package (Quinlan et al, 2010). 3 of these peaks were filtered out as they appeared in the blacklist file provided using the subtract function in Bedtools.

Peak annotation and motif finding

The filtered reproducible peaks were annotated using the annotatePeaks.pl utility in Homer to their nearest genomic feature. The goal of peak annotation is to map peaks to gene symbols such as promoter, TSS and so on (10x Genomics). Next, motifs were discovered using the findMotifsGenome.pl function in Homer since the authors did not disclose the exact parameters that they used to conduct motif finding using MEME-ChIP. Motif discovery is a key step in identification of Transcription Factor Binding Sites (TFBSs) that help in learning the mechanisms for regulation of gene expression (Hashim et al, 2019). Therefore, this was a critical step in the author's and our analysis of RUNX1 as a transcription factor.

Generating bigWig file for correlation analysis and visualization

In the next stage of our ChIP-seq analysis, we utilized DeepTools 3.5.1, a collection of powerful utilities designed specifically for ChIP-seq and genome-wide sequencing technologies (Ramírez et al, 2016). Prior to generating bigWig files using BamCoverage, we indexed the bamfiles using samtools. Of the four bamfiles, one indexed bamfile was already present as a result (sorted_bam_SRR2919475_file.bam.bai), so we indexed the remaining three (inp_rep1_sorted.bam, inp_rep2_sorted.bam, runx1_rep2_sorted.bam).

We then used the bamCoverage utility to calculate the coverage of reads in the BAM files and convert them into bigwig files. The generated bigwig files contained information about read coverage across the genome for each of the four samples and could be used for downstream analysis and visualization. Once the bigWig files had been generated as a batch job for each of the BAM files provided resulting in inp_rep1_sorted_output.bw, inp_rep2_sorted_output.bw, runx1_rep2_sorted_output.bw, sorted_bam_SRR2919475_file_output.bw , we then used the multiBigWigSummary utility using the bins option and the four bigWig files for the input replicates and Runx1 replicate, as well as the sorted BAM file. Next, plotCorrelation utility to produce a clustered heatmap of Pearson correlation values between all the samples was used to evaluate the similarity of the samples and explore potential sources of variation in the data. The output plot and correlation matrix could be further analyzed to draw insights and conclusions about the data (Ramírez et al, 2016).

The UCSC Table Browser is then used to extract a BED file listing the transcription start site (TSS) and transcription termination site (TTS) locations for every gene in the hg19 reference. The bigwig files for the IP samples and the BED file of hg19 genes are then used to run the computeMatrix utility in DeepTools in the scale-regions mode twice (once for each IP sample) to generate two matrices of values. Finally, the plotProfile utility is run on each matrix to generate a visualization of the Runx1 signal coverage across hg19 genes using the TSS and TTS as reference points (Kent et al, 2002).

The aim of this analysis is to generate a heatmap of clustered correlation metrics between all the samples, coverage tracks for visualization, and the signal coverage plot across the TSS and TTS of all hg19 genes.

Identification of differentially expressed genes and its involvement with RUNX1

Upon downloading the DESeq2 results from GSE75070, the significant differentially expressed genes were identified based on the following cutoffs - adjusted p-value <= 0.01 and absolute value of log2 foldchange > 1.

Annotated peak genes and differentially expressed genes were merged and a dataframe with significant genes either being bound to RUNX1 or not was created. Based on this, we obtained a stacked bar chart representing the proportions that are relative to TSS (<= 5000).

IGV Visualization

The 4 bigwig files along with the reproducible peaks BED file were loaded onto IGV (Robinson et al, 2011). We select GRCh38/hg38 as the genome. We also autoscaled the bigwig files by right clicking on the left-hand side panel and checking the Autoscale option for each track. We then use the search box to look for MALAT1 and NEAT1 genes.

HI-C Data Visualization

The HI-C data matrix was downloaded from the GEO page GSE75070, specifically for chromosome 10 (GSE75070_HiCStein-MCF7-shGFP_hg19_chr10_C-40000-iced.matrix.gz). Data preprocessing was required which included converting NaN values to zeros and applying log2 transformation. We added 1 to each value before log transformation to avoid taking the log of zeros or negative values, which are not defined. After the data was preprocessed, we used the base-R heatmap() function to generate a heatmap. We specifically turned off the default clustering effect that is incorporated in the function.

<u>Results:</u>

Data Curator:

The results of a FASTQC report are summarized by basic statistics information (such as the total number of sequences, how many were flagged as poor quality, average sequence length), per base sequence amity, per sequence quality, the distribution of sequence lengths, sequence duplication levels, overrepresented sequences and the content of commonly used adapters. From the FASTQC run of the original sample the statistics showed a total of 29734121 sequences with a sequence length of 101 bp. The mean of per base sequence quality reduced towards the end suggesting that some filtering was needed. Additionally the levels of sequence duplication and overrepresented sequences was higher than acceptable by the program.

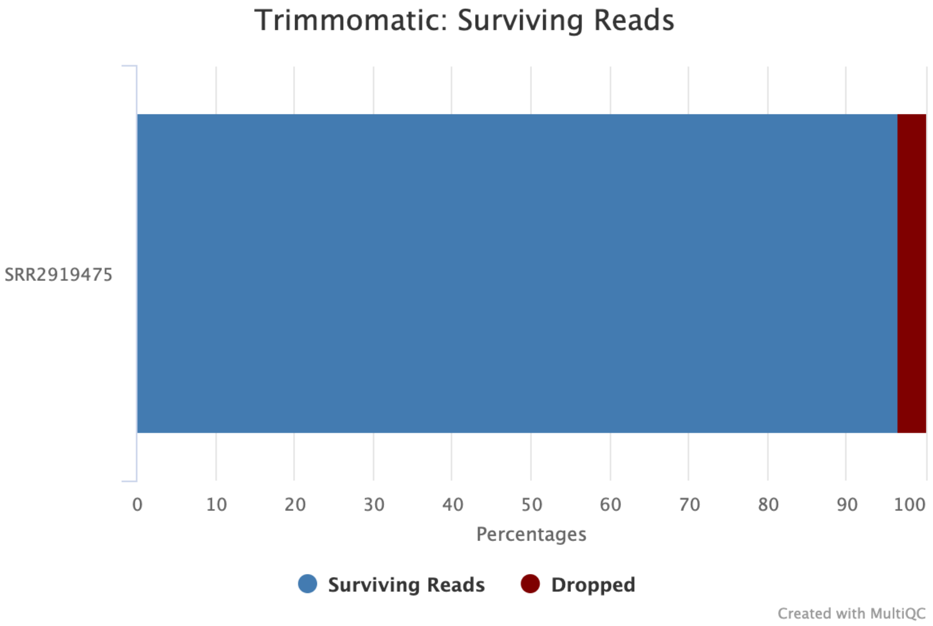


Fig1: Percentage of survived reads from Trimmomatic graphed by MultiQC

Trimmomatic is a software used to trim adapter sequences out of the reads. 3.42% of the original reads were dropped. This graph was generated with MultiQC and it depicts the percentage of reads that survived and the percentage of reads that were dropped.

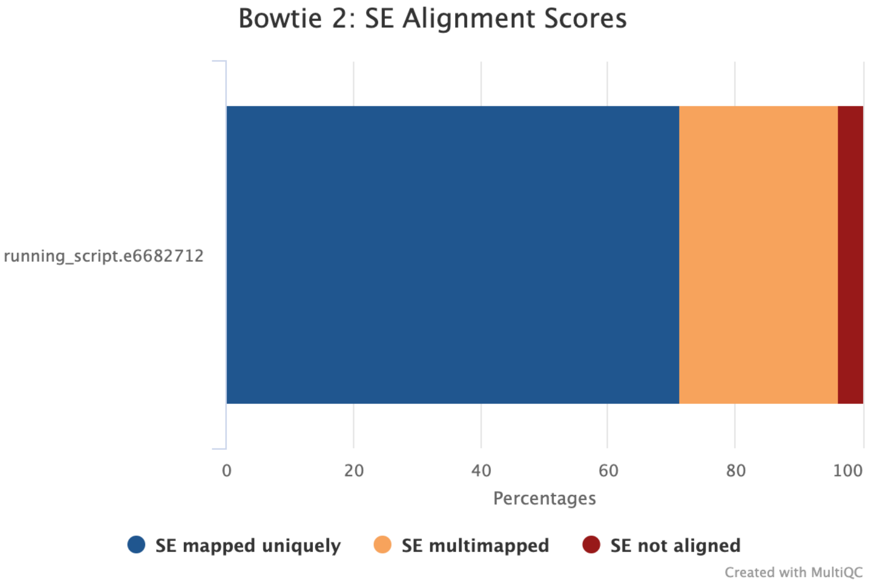


Fig2: MultiQC Bowtie2 Percentage of Alignment Scores

The aligner that was used to map the reads to the hg19 genome was Bowtie2. According to the statistics output 71.3% of the single reads aligned once and 24.98% of the reads aligned in more than 1 location with a total alignment rate of 96.29%

	runx1_rep1 (data curator processed)	runx1_rep2	inp_rep1	inp_rep2
total	28717962 + 0	28968174 + 0	30041540 + 0	10890224 + 0
Secondary	0+0	0+0	0+0	0+0
Supplementary	0+0	0+0	0+0	0+0
Duplicates	0+0	0+0	0+0	0+0
Mapped	27653241 + 0 mapped (96.29% : N/A)	28369845 + 0 (97.93% : N/A)	28590558 + 0 mapped (95.17% : N/A)	10051798 + 0 mapped (92.30% : N/A)
Paired in sequencing	0+0	0+0	0+0	0+0
Read 1	0+0	0+0	0+0	0+0
Read 2	0+0	0+0	0+0	0+0
Properly Paired	0+0	0+0	0+0	0+0
With itself and mate mapped	0+0	0+0	0+0	0+0
Singletons	0+0	0+0	0+0	0+0
With mate mapped to different chr	0+0	0+0	0+0	0+0
With mate mapped to different chr	0+0	0+0	0+0	0+0

(mapQ>=5)				

Table1: Comparison of Samtools stat flags results among all four samples

This table allows one to compare how many reads were mapped to the genome for each of the samples. Because these reads are single end reads they have no paired mates and hence most of the other flags in the table have 0 reads for they are useful if these were paired end sequences. Runx1_rep2 had mapped the best

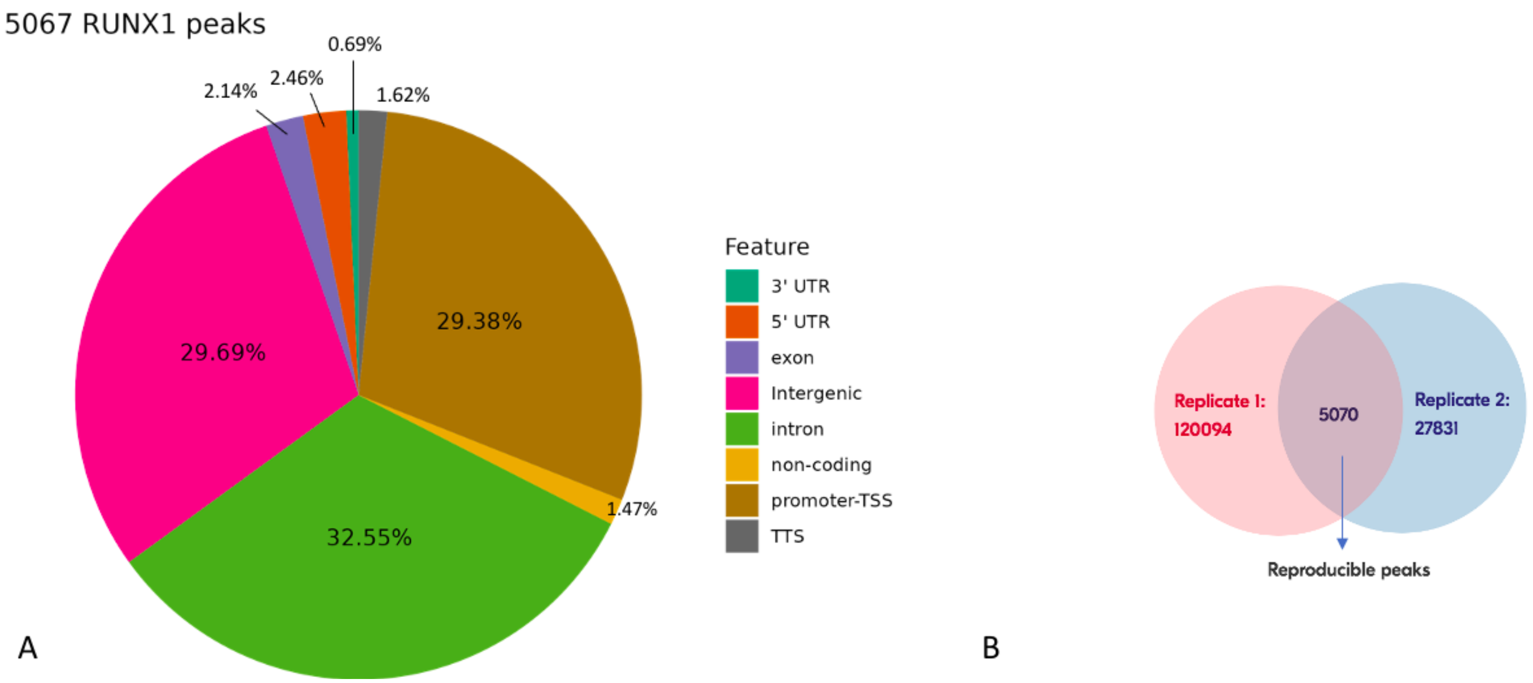


Fig3: A. Pie chart for each genomic feature represented in the annotated reproducible peaks, B. Number of peaks per replicate with reproducible peaks as intersection.

After the filtering step, 120094 peaks were identified in replicate 1 and 27831 peaks were identified in replicate 2. Total reproducible peaks after filtering were found to be 5067 (Fig3.B). As mentioned earlier, 3 peaks were removed from the filtering process. Annotation was performed following the filtering step and majority of the peaks (32.55%) were found to be in the intron region (Fig3.A). A significant number of peaks were also found in the Intergenic (29.69%), and promoter-TSS (29.38%) regions. Very little peaks occurred in the 5' UTR (2.46%), exon (2.14%), TTS (1.62%), non-coding (1.47%), and 3' UTR (0.69%) regions.

Rank	**Motif**	**Best Match**	**P-value**

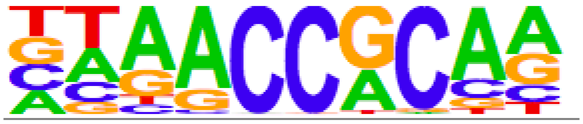
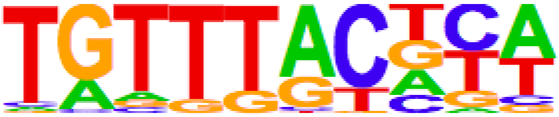
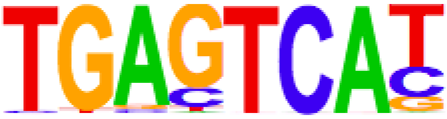
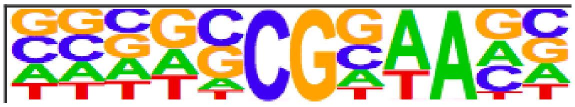
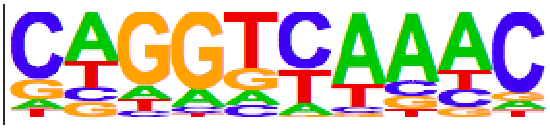
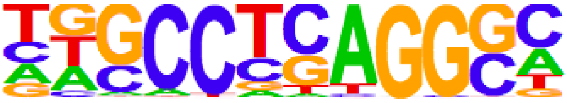
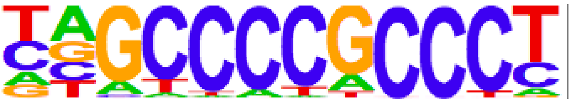
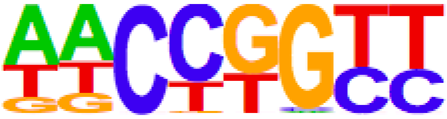
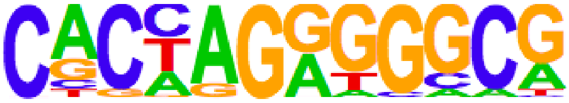
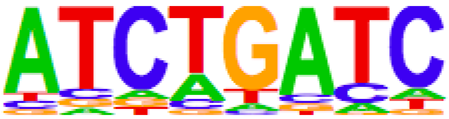
1		RUNX2	1e-764
2		FOXA1	1e-161
3		GCN4	1e-160
4		Elk1	1e-109
5		WRKY20	1e-95
6		AP-2alpha	1e-85
7		Sp2	1e-83
8		TFCP2	1e-67
9		CTCFL	1e-60
10		GATA19	1e-56

Table2: HOMER de novo motif analysis of the RUNX1 peaks. The peaks are ordered by significance from top to bottom.

The Homer *de novo* motif analysis revealed 24 statistically significant motifs, with the last 2 having a high possibility of being false positives. The top match with a p-value of 1e-764

(Table2) was RUNX2. FOXA1 (p-value = 1e-161) and GCN4 (p-value = 1e-160) were also significantly enriched.

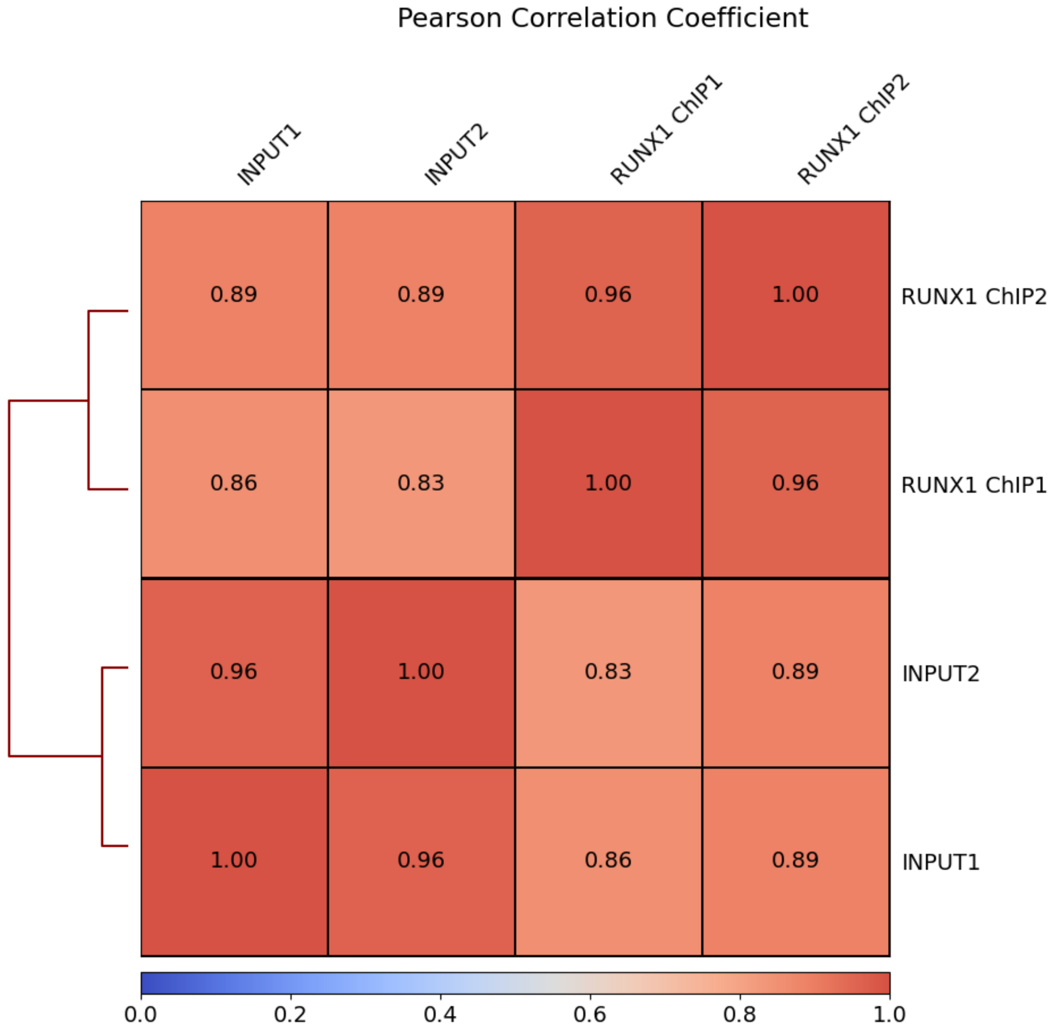


Fig 4 : Heatmap illustrating the Pearson correlation between input and IP samples.

The heatmap in Figure 4 displays the Pearson correlation between input and IP samples. Initially, it was expected that the pull-down samples would be highly similar to each other and the input samples would also be similar. However, there was less similarity observed between IP and input samples. The clustering and values in the heatmap support this observation. A matrix in the paper represents the similarity between samples using the Pearson correlation coefficient, with red indicating strong positive correlation and blue representing no correlation. Each square displays the Pearson correlation coefficient between RUNX1 binding site signal intensities and gene expression. Interestingly, inputs 1 and 2 show high similarity while RUNX1 replicates 1 and 2 have identical scores, indicating that the experiment was reliable and successful without any outliers among the replicates.

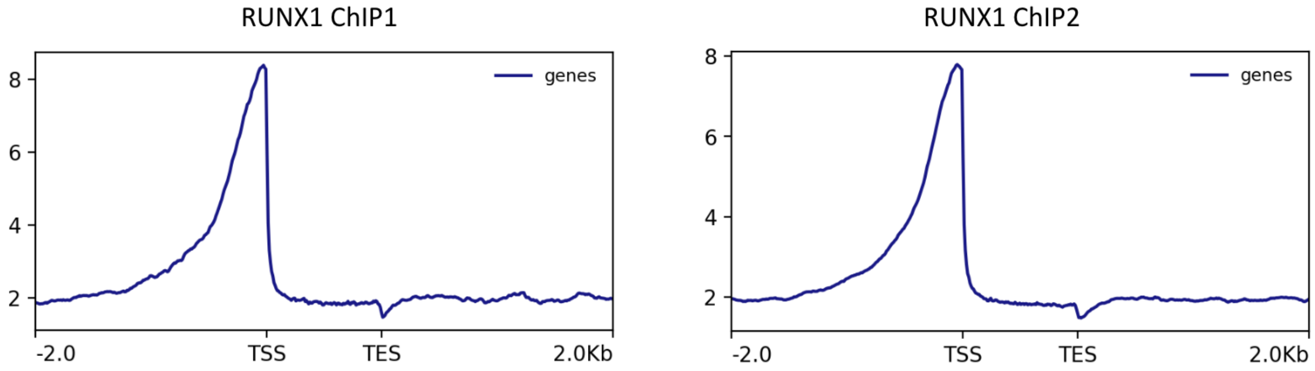


Fig 5 : (A) Normalized RUNX1 ChIP-seq signal coverage intensity plot across UCSC hg19 genes for first replicate (B) Normalized RUNX1 ChIP-seq signal coverage intensity plot across UCSC hg19 genes for second replicate

The purpose of the analysis was to visualize the signal of genes using antibodies that target specific proteins, specifically RUNX1, in order to gain insights into where most genes are binding. The data was obtained from a bigwig file, which represents the number of genes read, and a bed file that provides the structure of the genes, including chromosome, start, and stop locations. These files were used to generate a matrix that allowed the signal to be plotted onto a graph, and the gene sizes were normalized to ensure that every gene was of the same size.

The plot profile, generated using the bigwig files for the IP samples and the BED file of hg19 genes, depicted the coverage of the Runx1 signal across hg19 genes (Fig 5A, Fig 5B). The analysis was conducted using the DeepTools software, specifically the computeMatrix utility in scale-regions mode. The analysis was performed twice, once for each IP sample, to generate two matrices of values. The plot profile showed the average enrichment of Runx1 binding in regions surrounding the transcription start site (TSS) of hg19 genes. It used a 2kb window up- and downstream of the TSS and TTS to ensure that the enriched regions were within a reasonable distance of the gene's transcriptional start and stop sites.Through the analysis, it was observed that transcription factors tend to bind to the promoters of genes and that there was a significant enrichment of regions before the transcription start site (TSS). Although the specific gene binding to RUNX1 could not be determined, it was confirmed that the binding occurred before the TSS.

Overall, the figure provided a clear visualization of the distribution of Runx1 binding across hg19 genes. It supported the author's findings and provided insight into the mechanisms of gene regulation by Runx1. Finally, the ChIP-seq analysis in the parental MCF-7 cells showed that the RUNX1-dependent differences in gene expression observed by RNA-seq were directly related to

RUNX1 binding. The strongest RUNX1 signal was observed at the promoter regions, consistent with the engagement of RUNX1 in both transcriptional activation and repression, both directly and indirectly, suggesting an involvement of RUNX1 in long-range gene regulation.

After applying the cutoff values to find differentially expressed genes, we obtain 1153 differentially expressed genes which match the number indicated in the paper. 466 downregulated genes and 687 upregulated genes were identified following RUNX1 knockdown in MCF-7 cells. Knockdown of RUNX1 led to a decrease in the expression of two extensively studied long non-coding RNAs (lncRNAs), NEAT1 and MALAT1, which are known to play a role in the organization of SC-35 nuclear speckles. From our results, we saw that NEAT1 is downregulated with a fold change decrease of 1.95 and MALAT1 is downregulated with a fold change decrease of 1.67.

To assess the possibility of increased RUNX1 binding at the promoters of the genes that showed up- or down-regulation upon RUNX1 knockdown, we generated a stacked bar chart (Fig 8) of RUNX1 peak binding in the region spanning +/− 5kb around the promoters of the differentially expressed genes. According to our plot, 129 up regulated DEGs and 84 down regulated DEGs are RUNX1 bound. In contrast to the paper, where the count is 59 and 48 for up and down regulated respectively, our results have doubled. The results indicate that there is indication of RUNX1 in transcriptional activation and repression, which means that RUNX1 is involved in gene regulation.

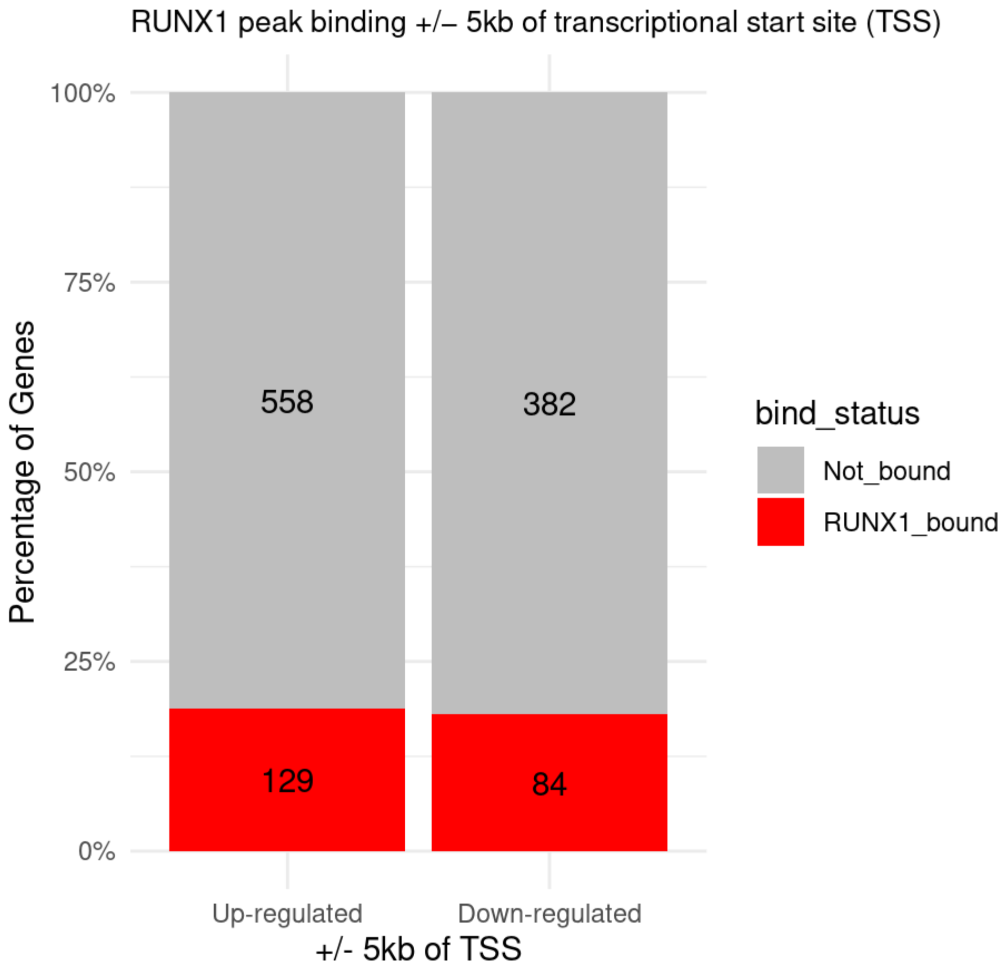


Fig 8: Stacked bar chart representing RUNX1 peak binding +/- 5kb of transcriptional start site (TSS)

Figures 9 and 10 indicate the alignment of input and RUNX1 binding bigwig files to the human genome. We are interested in the MALAT1 and NEAT1 genes, and we can see that there is binding of RUNX1 to the promoter regions of these genes. The input tracks show more peaks and noise since there is non-specific binding. This is reduced in the RUNX1 tracks. It makes sense for specific peaks to show up in the RUNX1 binding tracks for MALAT1 and NEAT1 as these two genes are downregulated when RUNX1 was knocked out in the RNA-Seq analysis. This indicates that RUNX1 binds to the promoter regions of MALAT1 and NEAT1 and is essential for the activation of these two lncRNAs. Based on these alignments, we agree with the conclusions made by the authors of the paper.

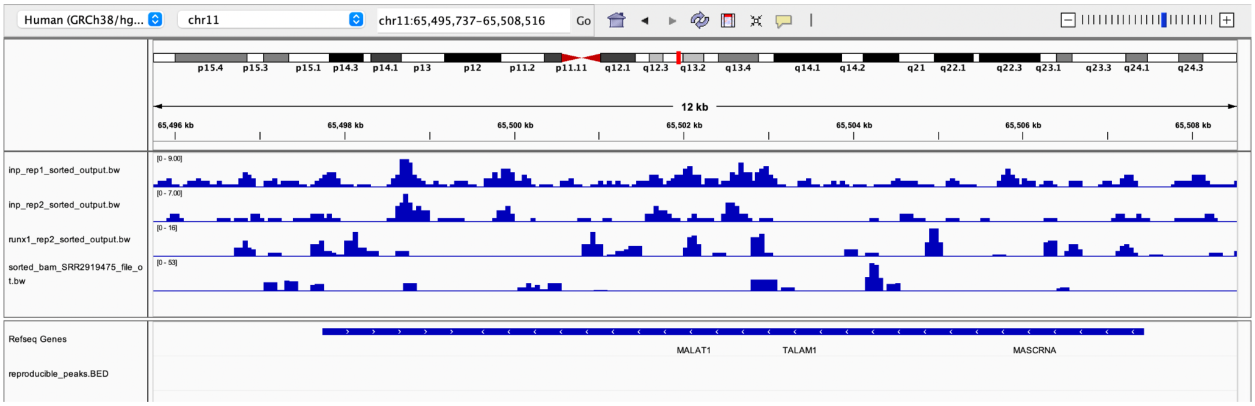

Human (GRCh38/hg...
chr11
chr11:65,495,737-65,508,516
Go
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.11 q12.1 q12.3 q13.2 q13.4 q14.1 q14.2 q21 q22.1 q22.3 q23.1 q23.3 q24.1 q24.3
12 kb
65,496 kb
65,498 kb
65,500 kb
65,502 kb
65,504 kb
65,506 kb
65,508 kb
inp_rep1_sorted_output.bw
[0 - 9.00]
inp_rep2_sorted_output.bw
[0 - 7.00]
runx1_rep2_sorted_output.bw
[0 - 16]
sorted_bam_SRR2919475_file_out.bw
[0 - 53]
Refseq Genes
MALAT1
TALAM1
MASCRNA
reproducible_peaks.BED

Fig 9: IGV visualization of RUNX1 binding and input control for MALAT1

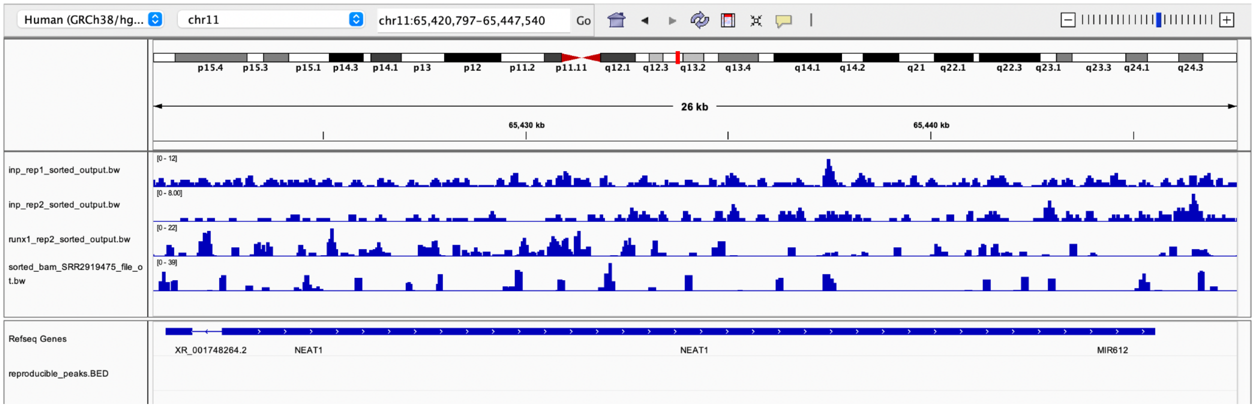

Human (GRCh38/hg...
chr11
chr11:65,420,797-65,447,540
Go
p15.4 p15.3 p15.1 p14.3 p14.1 p13 p12 p11.2 p11.11 q12.1 q12.3 q13.2 q13.4 q14.1 q14.2 q21 q22.1 q22.3 q23.1 q23.3 q24.1 q24.3
26 kb
65,430 kb
65,440 kb
inp_rep1_sorted_output.bw
[0 - 12]
inp_rep2_sorted_output.bw
[0 - 8.00]
runx1_rep2_sorted_output.bw
[0 - 22]
sorted_bam_SRR2919475_file_out.bw
[0 - 38]
Refseq Genes
XR_001748264.2
NEAT1
NEAT1
MIR612
reproducible_peaks.BED

Fig 10: IGV visualization of RUNX1 binding and input control for NEAT1

An HI-C pairwise interaction matrix provides a comprehensive view of the 3D spatial organization of chromatin in the nucleus, revealing the structural features such as compartments, topologically associating domains (TADs), loops, and chromatin domains that are important for genome function and regulation. In Fig 11, we look at the visualization of HI-C data of the first 5.3 megabases of the p arm of chromosome 10. We see 3 different TAD regions based on the color intensity of the plot. The dark regions of the chromosome indicate frequent interactions with RUNX1 which suggests that RUNX1 could function at TAD boundaries.

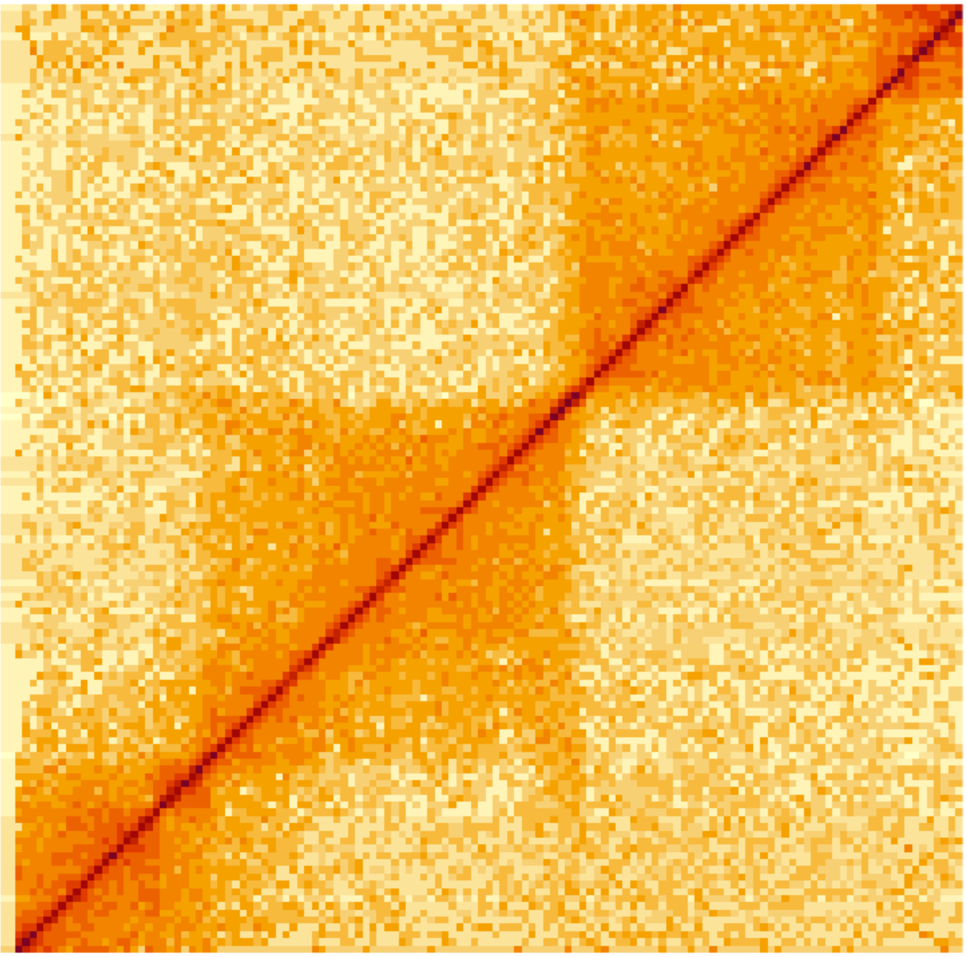


Fig 11: Heatmap of HI-C data for the first 5.3mb on chromosome 10

Discussion:

There is agreement between the authors and the current analysis. Both studies identified binding of Runx1 to the promoter regions of genes and its role in regulating gene expression. While our analysis showed a higher occurrence of reproducible peaks in the intron region (Fig 3.A), it is not significantly different from the proportion of reproducible peaks in the promoter region. We were also able to identify a greater number of peaks in general as well as the reproducible peaks than the authors did in the paper (Fig3.B). This inconsistency arises from the fact that the paper lacks several details such as parameters used for motif finding, and the normalization method used which are required to reproduce their findings. Due to these reasons, we also see an increase in the number of differentially expressed genes bound to RUNX1 with +/- 5 kb of TSS (Fig 8).

The nuclear speckle is a subnuclear structure that is involved in the storage and processing of pre-mRNA, and NEAT1 has been shown to be a structural component of this organelle (Clemson et al, 2009). Additionally, MALAT1 has been found to co-localize with nuclear speckles and has been implicated in regulating RNA splicing and processing within these structures (Tripathi et al, 2013). To investigate the mechanism by which RUNX1 regulates NEAT1 and MALAT1 expression, a future direction could be to investigate the formation of SC-35 speckles in cells where RUNX1 is depleted. This could help shed light on the functional consequences of RUNX1-mediated regulation of these lncRNAs in nuclear structure. Furthermore, chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments could be conducted to identify additional genomic regions bound by RUNX1, and to investigate the potential effects of RUNX1 binding to distal regulatory elements. Such experiments would provide valuable insights into the transcriptional and regulatory roles of RUNX1 in breast cancer cells, and help to elucidate the mechanisms underlying its effects on nuclear structure and cancer progression. In addition, we see similarities between figure 9 and 10 in our report and figures 2d and 2e in the paper. We can conclude that RUNX1 is involved in regulating NEAT1 and MALAT1 expression.

To further understand the role of Runx1 in genomic architecture, we could use CRISPR/Cas9 to delete or mutate Runx1 binding sites at TAD boundaries followed by analysis of the effects of Runx1 depletion on TAD structure and gene expression. This could provide evidence for a causal role of Runx1 in TAD organization. Hi-C analysis could be conducted on this data to analyze changes in 3D chromatin architecture following Runx1 depletion. Revealing if Runx1 plays a role in the formation or maintenance of TAD boundaries. (Fig 11)

To further investigate the role of Runx1 in genomic architecture, additional experiments could be performed. For instance, one could perform Hi-C experiments to investigate the changes in chromatin interactions following Runx1 depletion. Furthermore, additional ChIP-seq experiments could be conducted to identify the specific regions where Runx1 is binding at TAD boundaries and to investigate the potential role of Runx1 in maintaining genomic architecture.

While the original paper focused on the role of Runx1 at promoters, it is appreciated that binding in intergenic and intronic regions can have important functional and regulatory effects. To investigate the potential effects of Runx1 binding to distal regulatory elements, one could perform additional ChIP-seq experiments to identify the specific regions where Runx1 is binding. Additionally, functional assays such as luciferase reporter assays could be performed to investigate the regulatory effects of Runx1 binding to these distal regions.

In conclusion, this study sheds light on the significance of RUNX1 in gene regulation and chromatin organization in breast cancer cells. The results indicate that targeting RUNX1 could be a promising therapeutic approach for breast cancer treatment. While we were able to replicate the study's results using the methods provided, the lack of detailed information on the data processing techniques may hinder the reproducibility of this study. Future studies that provide more information on data processing methods would be beneficial in enhancing the reproducibility of the findings.

Works Cited

Aaron R. Quinlan, Ira M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, Volume 26, Issue 6, March 2010, Pages 841–842, https://doi.org/10.1093/bioinformatics/btq033

Akalin A, Uyar B, Franke V, Ronen J. Computational Genomics with R, 2016. Github, https://github.com/compgenomr/book

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Barutcu AR, Hong D, Lajoie BR, McCord RP, van Wijnen AJ, Lian JB, Stein JL, Dekker J, Imbalzano AN, Stein GS. RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells. Biochim Biophys Acta. 2016 Nov;1859(11):1389-1397. doi: 10.1016/j.bbagrm.2016.08.003. Epub 2016 Aug 9. PMID: 27514584; PMCID: PMC5071180.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Chuang, L.S.H., Ito, K. and Ito, Y. (2013), RUNX family: Regulation and diversification of roles through interacting proteins. Int. J. Cancer, 132: 1260-1271. https://doi.org/10.1002/ijc.27964

Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Molecular cell, 33(6), 717-726. https://doi.org/10.1016/j.molcel.2009.01.026

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Hashim FA, Mabrouk MS, Al-Atabany W. Review of Different Sequence Motif Finding Algorithms. Avicenna J Med Biotechnol. 2019 Apr-Jun;11(2):130-148. PMID: 31057715; PMCID: PMC6490410.

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

"Peak Annotations." 10x Genomics, n.d., https://support.10xgenomics.com/single-cell-multiome-atac-gex/software/pipelines/latest/output/peak-annotations#:~:text=The%20annotation%20procedure%20is%20as,distal%20peak%20of%20that%20gene. Accessed 30 Mar. 2023.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. Nucleic Acids Research

(2016). doi:10.1093/nar/gkw257. https://deeptools.readthedocs.io/en/develop/content/about.html

Tripathi, V., Shen, Z., Chakraborty, A., Giri, S., Freier, S. M., Wu, X., ... & Prasanth, S. G. (2013). Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS genetics, 9(3), e1003368. https://doi.org/10.1371/journal.pgen.1003368

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011)

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006. (2002) https://genome.ucsc.edu/cite.html