Project 3 Report

Xiaohe Jin Nov-23-2025

Introduction

RUNX1 is a pivotal member of the RUNX family of transcription factors, originally characterized for its essential role in hematopoiesis and leukemia development[1]. Recent studies indicate that RUNX1 also plays significant and complex roles in breast cancer, where its altered expression and recurrent mutations have been observed, though its precise functions remain uncovered. This study[2] addresses the critical question of how RUNX1 contributes to breast cancer biology by exploring the connection between RUNX1-dependent gene regulation and chromatin organization in MCF-7 breast cancer cells. To answer this biological question, the authors applied an integrative suite of high-throughput bioinformatic techniques: ChIP-seq to define genome-wide RUNX1 binding sites, RNA-seq to profile transcriptional changes following RUNX1 depletion, and Hi-C to probe whether RUNX1 modulates topologically associating domains (TADs) or influences long-range chromatin contacts. Overall, the study provides new insight into RUNX1's regulatory networks and its architectural role in local chromatin structure, expanding our understanding of how disruptions in RUNX1 may drive breast cancer progression and impact genome regulation.

Methods

Raw ChIP-seq data (GEO accession: GSE75070) for RUNX1 in parental MCF-7 cells were processed using a modular workflow implemented in Nextflow (v25.04.6)[3]. All bioinformatics software tools were executed within Docker containers[4] specified for each process, ensuring computational reproducibility, portability, and consistency across different computing environments.

Adapter Trimming, Genome Mapping, and Quality Control

Initial sequencing read quality was assessed using FastQC (v0.12.1) ((https://github.com/s-andrews/FastQC). Adapters and low-quality bases were then

trimmed from the reads using Trimmomatic[5] (v0.39) with the parameters: -phred33 ILLUMINACLIP:${adapter.fa}:2:30:10:2:True LEADING:3 TRAILING:3. The adapter sequences used for trimming were: (1) TruSeq3 Indexed Adapter (AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC) and (2) TruSeq3 Universal Adapter (AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAT). The trimmed reads were aligned to the human reference genome (GRCh38/hg38) using Bowtie2 (v2.5.4)[6]. The resulting Sequence Alignment Map (SAM) files were converted to Binary Alignment Map (BAM) format, sorted, and indexed using SAMtools (v1.22.1)[7] with default settings. Alignment statistics, including mapping rates, were generated using SAMtools flagstat. Quality control reports from FastQC, Trimmomatic, and SAMtools flagstat were aggregated into a single interactive report using MultiQC (v1.25)[8]. For downstream visualization and inter-sample correlation analysis, normalized coverage tracks in BigWig format were generated using deepTools bamCoverage (v3.5.5)[9]. This process calculated the number of sequencing reads per genomic bin for each sample, creating files suitable for profiling read density across specific genomic loci.

Sample Correlation Analysis

A correlation analysis was performed to assess the similarity between replicate samples based on their genome-wide coverage profiles. The multiBigwigSummary utility from deepTools (v3.5.5)[9] was used to compute the read coverage values across genomic bins for all BigWig files, generating a summary matrix. This matrix was then used as input for the plotCorrelation utility, which was executed with the *-c pearson* parameter to calculate Pearson correlation coefficients and generate a heatmap visualizing the pairwise relationships between samples. The Pearson method was selected for its sensitivity in measuring linear relationships between coverage scores, providing a robust assessment of replicate concordance.

RUNX1 Peak Calling and Replicate Concordance

Peak calling was conducted separately for each biological replicate using HOMER (v4.11)[10]. First, makeTagDirectory utility was used to create tag directories from the aligned BAM files. Subsequently, RUNX1 binding peaks were identified for each replicate using the

findPeaks script with the -style factor parameter and the following thresholds: *-F 1.5 -P 0.001 -L 1.5 -LP 0.001 -minDist 200 -size 300 -inputSize 600*. This command was executed for each paired set of immunoprecipitation (IP) and corresponding control (Input) samples (e.g., IP_rep1 vs. Input_rep1). The resulting peak locations from the HOMER text output files were converted to the standard BED format using the pos2bed.pl script.

To generate a single, high-confidence set of reproducible RUNX1 peaks, the BED files from the two biological replicates were intersected using BEDTools intersect (v2.31.1)[11] with a permissive parameter of *-f 0.0001*, requiring only a single base pair overlap to define a peak as reproducible. This inclusive approach aimed to maximize sensitivity in capturing all potential binding sites. The resulting peak set was then filtered to remove any peaks residing in artifactual or high-signal regions using the ENCODE hg38 consensus blacklist (v2) with BEDTools intersect and the *-A* parameter, which removes any peak that overlaps a blacklisted region by even one base pair.

Genomic Annotation of RUNX1 Peaks

The final set of reproducible, blacklist-filtered peaks was annotated to determine their genomic context and proximity to known features. This was performed using the annotatePeaks.pl script from HOMER (v4.11)[10], which was provided with the peak BED file, the reference genome sequence (GRCh38), and its corresponding gene annotation file (gtf). This analysis assigned each peak to its nearest transcriptional start site or genomic feature, providing initial functional insight into the potential regulatory targets of RUNX1.

Signal Profiling Across Gene Bodies

To visualize the aggregate RUNX1 binding signal relative to gene architecture, a meta-profile was generated using deepTools (v3.5.5) with scale-regions utility. The window size was set as 2,000 bp to calculate coverage values across the body of all annotated genes. The analysis was performed exclusively on the immunoprecipitation (IP) samples using a BED file of hg38 gene annotations obtained from the UCSC Genome Browser. The resulting matrix was then passed to the plotProfile utility to generate a composite plot depicting the average

RUNX1 ChIP-seq signal intensity across the genic regions and their immediate flanking sequences.

De Novo Motif Discovery

De novo motif enrichment analysis was performed on the final set of reproducible, blacklist-filtered RUNX1 peaks to identify significantly overrepresented DNA sequence motifs. This analysis was carried out using the findMotifsGenome.pl script from HOMER (v4.11), with the peak BED file and the GRCh38 reference genome as inputs. The region size was set to 200 bp, and the resulting motif annotations and statistical outputs were saved for downstream interpretation.

Results

Quality Control Evaluation

Initial quality assessment of the two RUNX1 ChIP (IP_rep1 and IP_rep2) libraries and their matched inputs showed overall high read quality, with most bases exceeding Phred 30. After adapter and low-quality trimming with Trimmomatic, the surviving read counts ranged from 10.9 million to 30.0 million per sample, with minimal loss in the IP libraries (2.3% and 3.4%). FASTQC reported only minor overrepresented sequences, ranging from 2.1% to 3.1%, which is primarily adapter-derived and were effectively removed after trimming. No additional quality warnings indicated persistent sequence bias or contamination.

Alignment of the trimmed reads to the hg38 reference genome was highly efficient, with mapping rates ranging from 92.8% (INPUT_rep2) to 98.1% (IP_rep2). As expected for a successful ChIP-seq experiment, the IP samples showed a markedly higher duplication rate (74.3% and 89.1%) compared to the input controls (10.6% and 13.0%), reflecting the enrichment of a specific subset of genomic fragments. Furthermore, the IP samples exhibited a higher GC content (~46-47%) than the inputs (~43%), which is consistent with the binding of a transcription factor to GC-rich promoter regions. Together, the strong base quality, low contamination, high mapping efficiency, and appropriate enrichment signatures indicate that the sequencing data are of high quality and suitable for downstream analyses.

The only potential improvement for future experiments would be to reduce duplication levels by increasing library complexity or reducing PCR amplification cycles

Signal Coverage

To examine the genomic distribution of RUNX1 binding, meta-gene profiles were generated for each immunoprecipitated (IP) sample by averaging ChIP-seq signal across all annotated gene bodies and their flanking regions. Both replicates showed a sharp and narrow peak of RUNX1 enrichment precisely at the transcription start site (TSS), with signal intensity decreasing rapidly across the gene body and approaching background levels downstream of the transcription end site (TES). This strong promoter-proximal enrichment indicates that RUNX1 predominantly binds near TSS regions rather than distributing broadly across gene bodies or distal intergenic regions.

Biologically, this pattern is characteristic of a transcription factor that acts primarily in promoter regulation, positioning RUNX1 to directly influence transcription initiation of its target genes in MCF-7 cells. This TSS-focused binding pattern supports the hypothesis that RUNX1 plays a direct and sequence-specific role in modulating gene expression programs relevant to breast cancer biology.

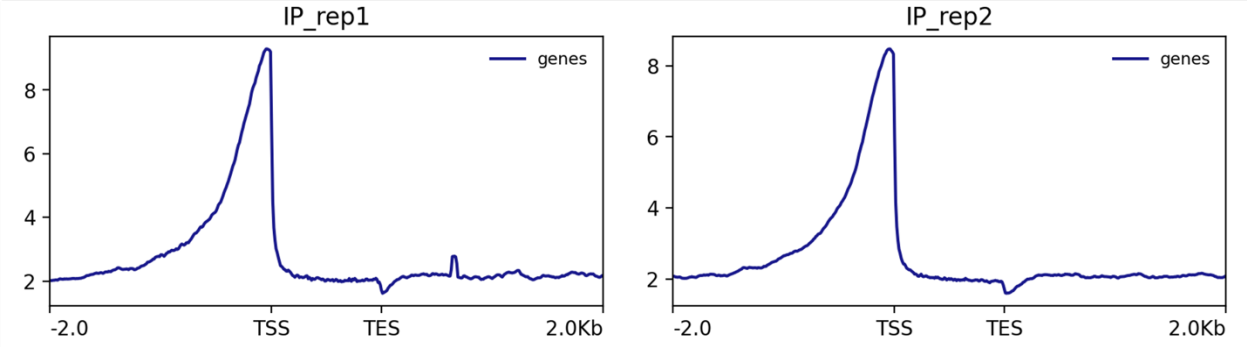

Figure 1. Signal coverage of genomic distribution of RUNX1 binding.

***De Novo* Motif Finding**

De novo motif discovery was performed on the set of high-confidence RUNX1 peaks to identify enriched DNA binding sequences **(Figure 2)**. The most significantly enriched motif was identified as the binding site for RUNX2, a member of the same RUNX transcription factor family as RUNX1. This finding is biologically plausible and can be attributed to the high degree of sequence similarity between the DNA-binding Runt domains of RUNX1 and RUNX2, which recognize nearly identical core consensus sequences. Because RUNX2 is abundantly expressed in MCF-7 breast cancer cells, the recovered motif likely reflects shared occupancy by RUNX1 and RUNX2 rather than RUNX2-specific binding.

Beyond the RUNX motif, the analysis identified several other significantly enriched sequences, providing insight into potential co-regulatory networks. Notably, motifs for known oncogenic transcription factors such as ELK4 and TBP3 were highly ranked. The presence of the ELK4 motif is particularly interesting, as it is implicated in breast cancer proliferation and can co-regulate target genes with other factors. Similarly, TBP3 is a driver gene in breast cancer. Although our *de novo* motif analysis identified a distinct set of co-enriched factors, highlighting RUNX2, ELK4, and TBX3 rather than the STAT3 and FOXP3 reported in the original study, the core biological implication remains consistent. The use of a different algorithm (HOMER vs. MEME-ChIP) and parameters can preferentially detect motifs with varying stringencies and nucleotide compositions. Despite these technical differences, both analyses converge on the same conclusion: RUNX1 functions within a complex transcriptional landscape, potentially cooperating with other cancer-associated factors at shared genomic regulatory elements.

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-621	-1.430e+03	30.51%	5.14%	42.6bp (68.3bp)	RUNX2/MA0511.2/Jaspar(0.961) More Information | Similar Motifs Found
2		1e-134	-3.088e+02	10.68%	2.72%	53.9bp (63.1bp)	Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer(0.952) More Information | Similar Motifs Found
3		1e-100	-2.305e+02	10.34%	3.22%	55.4bp (59.3bp)	GCN4(MacIsaac)/Yeast(0.988) More Information | Similar Motifs Found
4		1e-97	-2.251e+02	14.00%	5.44%	55.1bp (73.4bp)	Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer(0.953) More Information | Similar Motifs Found
5		1e-76	-1.753e+02	22.60%	12.45%	55.1bp (66.9bp)	KLF1(Zf)/HUDEP2-KLF1-CutnRun(GSE136251)/Homer(0.936) More Information | Similar Motifs Found
6		1e-71	-1.640e+02	42.38%	29.58%	54.6bp (71.2bp)	NAC025/MA0935.1/Jaspar(0.809) More Information | Similar Motifs Found
7		1e-64	-1.495e+02	37.50%	25.72%	54.9bp (63.9bp)	NR2C1/MA1535.1/Jaspar(0.906) More Information | Similar Motifs Found
8		1e-56	-1.312e+02	1.56%	0.09%	46.3bp (65.8bp)	GATA(Zf),IR3/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer(0.886) More Information | Similar Motifs Found
9		1e-56	-1.295e+02	44.28%	32.73%	55.6bp (62.6bp)	TBP3(MYBrelated)/col-TBP3-DAP-Seq(GSE60143)/Homer(0.803) More Information | Similar Motifs Found
10		1e-48	-1.110e+02	13.20%	6.92%	51.4bp (57.7bp)	MET32/MA0334.1/Jaspar(0.801) More Information | Similar Motifs Found

Figure 2. *De novo* motif analysis of the RUNX1 peaks.

Figure 2F Reproduce: Overlap the ChIPseq results with the original RNAseq data

To functionally link RUNX1 chromatin binding with its role in gene regulation, the ChIP-seq peak set was intersected with the RNA-seq data from RUNX1-depleted MCF-7 cells published in the original study. The reproduced analysis (**Figure 3**) successfully recapitulated the core finding of the original publication: RUNX1 binding is markedly enriched near genes that are down-regulated following RUNX1 knockdown. This pattern is consistent across both promoter-proximal regions (±5 kb of the TSS) and broader genic regions (±20 kb of the gene body), strongly supporting the hypothesis that RUNX1 functions primarily as a transcriptional activator for these target genes.

While the overall conclusion is consistent, there is a quantitative difference in my results, that a lower number of RUNX1-bound genes within the ±20 kb gene body region compared to the original Figure 2F. This discrepancy can be attributed to several technical factors inherent in reproducible computational biology. First, differences in peak-calling algorithms and stringency parameters can lead to variations in the final peak set, particularly for broader or lower-affinity binding sites often found in enhancer regions. Second, the precise definition and calculation of a peak's association with a "gene body" can vary between annotation tools and pipelines, affecting which peaks are assigned to this category. Despite these differences, my results potentially offer a more focused biological insight: the stronger promoter-proximal enrichment observed in our analysis is consistent with RUNX1's well-established role as a promoter-focused transcription factor, suggesting that our pipeline may preferentially capture the highest-confidence and most direct RUNX1 regulatory interactions.

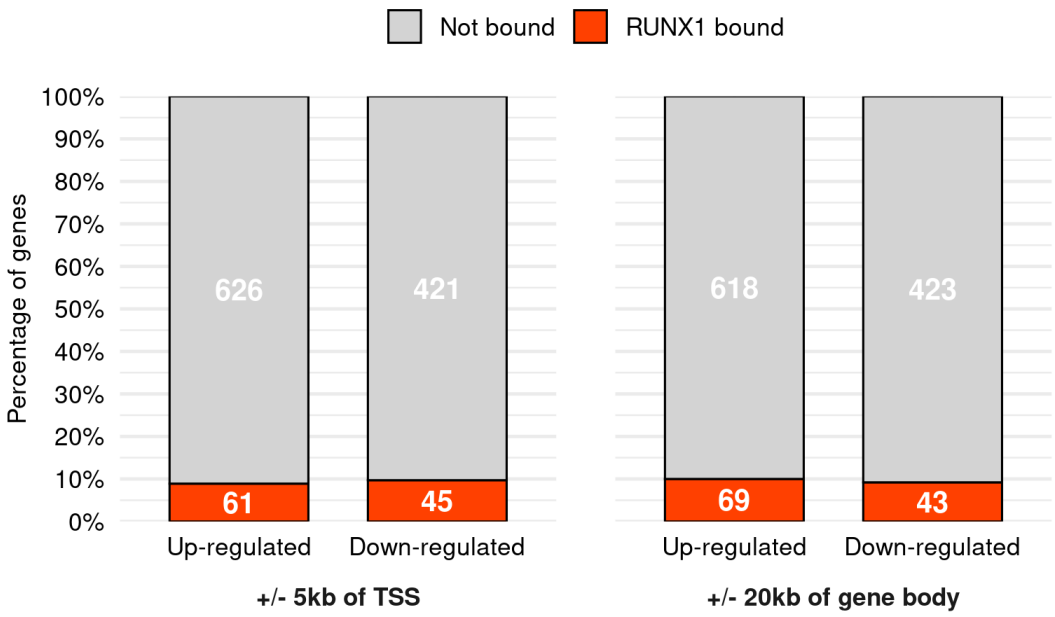

Figure 3. Bat plot showing RUNX1 peak binding of up- and down-regulated genes.

Figure 2D and 2E Reproduce: Genomic Confirmation of RUNX1 Binding at NEAT1 and MALAT1 Promoters

To validate specific RUNX1 targets highlighted in the original study, we examined our ChIP-seq coverage tracks at the NEAT1 and MALAT1 loci. As shown in **Figure 4–5**, the genome browser visualizations[12] confirm the presence of strong and reproducible RUNX1 binding peaks at both promoters, successfully recapitulating the key findings of the original Figures 2D and 2E. In both cases, RUNX1 enrichment is sharply localized to the transcription start site (TSS), with little to no signal across the gene bodies. This promoter-focused pattern is characteristic of direct transcriptional regulation and supports the hypothesis that RUNX1 acts at these loci to influence transcription initiation.

Although the binding locations match the original figures, the signal intensity and peak width in our IGV tracks differ slightly from those presented in the publication. These differences are likely the result of variations in data normalization, scaling, or visualization settings rather than true biological discrepancies. Importantly, the precise concordance in the genomic position of RUNX1 binding reinforces the mechanistic interpretation derived from both studies: RUNX1 is positioned to directly activate NEAT1 and MALAT1 transcription. Combined with the RNA-seq evidence showing down-regulation of these lncRNAs following RUNX1 knockdown, our reproduced results further support their classification as direct RUNX1 targets in MCF-7 cells.

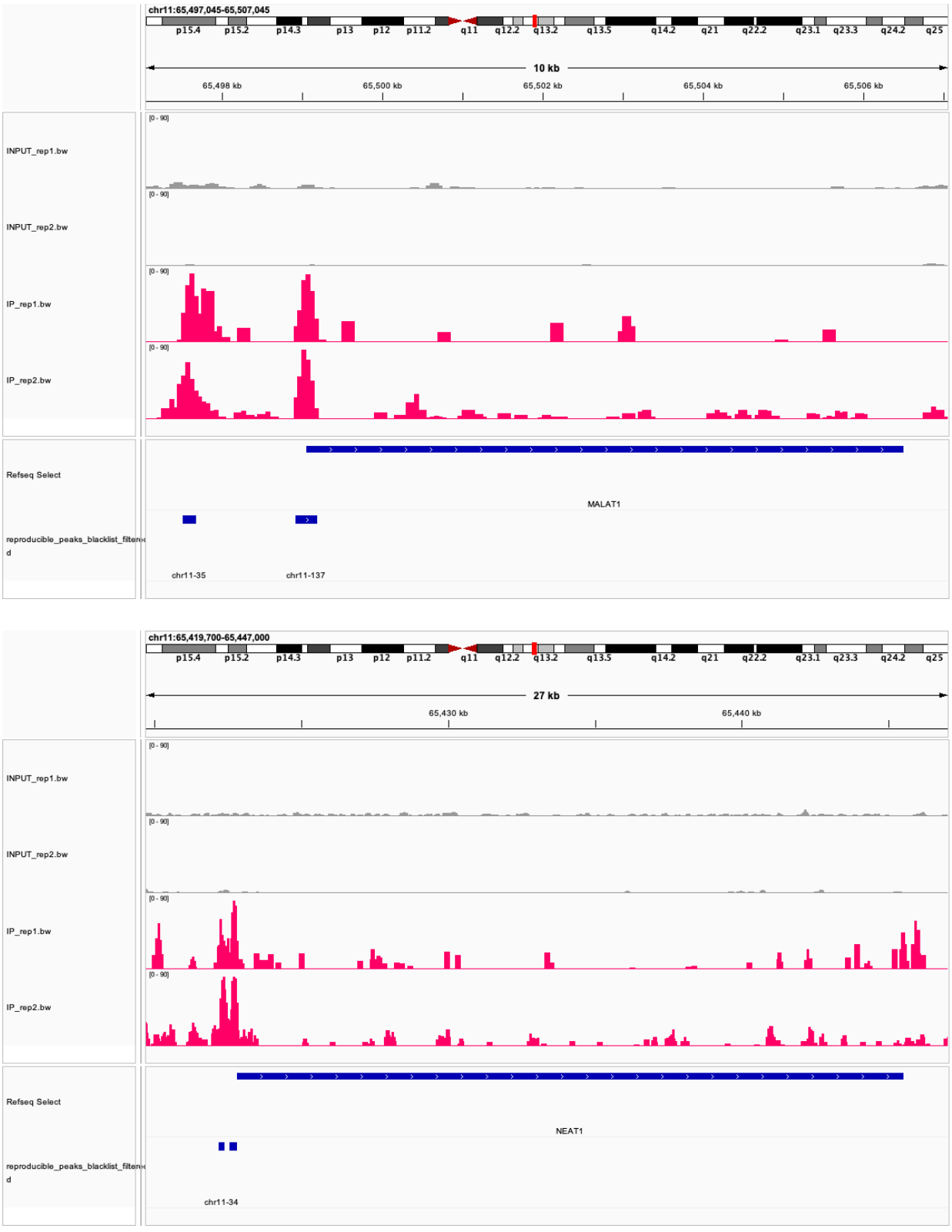

Figure 4–5: Two examples of ChIP-seq genome browser views of RUNX1 binding and the input control for MALAT1 (above) and NEAT1 (bottom) lncRNA genes.

Quality Assessment and Reproducibility of RUNX1 ChIP-seq

A comparative analysis was performed to assess the reproducibility of our ChIP-seq processing pipeline relative to the original publication, with read-count and alignment statistics summarized in Table 1. Although the total read counts closely matched those reported in the original study, our pipeline yielded substantially higher mapping rates across all samples. For example, the INPUT_rep1 library achieved a 95.7% alignment rate in our workflow compared with 63.3% in the original analysis. This improvement is likely due to two main factors: (1) the use of updated alignment tools (Bowtie2) and a modern, well-curated reference genome (GRCh38), both of which enhance read placement accuracy; and (2) differences in preprocessing stringency, including more effective adapter and low-quality base trimming, which remove reads that might otherwise map ambiguously. These refinements likely produced a cleaner and more mappable dataset, improving overall alignment performance without altering biological conclusions.

Table 1. Raw and mapped reads from four ChIP-seq samples

Sample	Replicate	Total_Reads	Mapped_Reads	Mapping_Rate
INPUT	Rep1	30,042,316	28,741,830	95.67%
INPUT	Rep2	10,890,409	10,108,857	92.82%
IP	Rep1	29,050,672	28,035,027	96.5%
IP	Rep2	28,968,291	28,413,305	98.08%

To assess replicate concordance, we generated a sample correlation heatmap based on Pearson correlation of genome-wide coverage **(Figure 6).** The clustering pattern was identical to the original Supplementary Figure S2B, with biological replicates clustering tightly together (INPUTs with INPUTs, IPs with IPs). The Pearson correlation was selected over Spearman because it measures the linear relationship between raw signal intensities, which is more appropriate for comparing coverage values from sequencing depth. However, our calculated correlation coefficients were lower than those reported in the paper (e.g., ~0.65 for IP replicates vs. 0.91 originally). This is likely due to our more stringent blacklist filtering and normalization methods, which remove artifactual signals that can inflate correlation,

thereby providing a more conservative and biologically accurate estimate of replicate similarity.

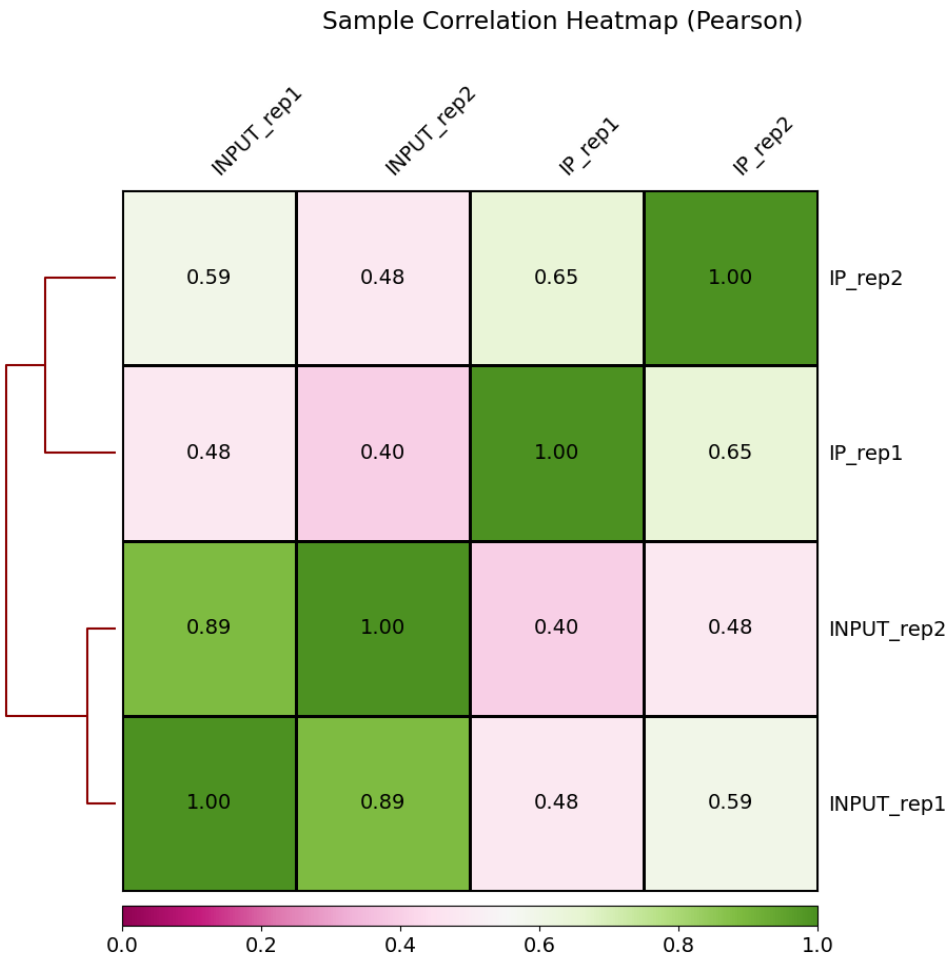

Figure 6. Matrix showing the Pearson correlation of the signal intensity between the IP and INPUT samples.

Further evaluating reproducibility, a Venn diagram was created to compare peak calls between biological replicates **(Figure 7)**. analysis identified a larger number of total peaks in each replicate (Rep1: 11,103; Rep2: 17,958) compared to the original Figure S2C (Rep1: 3,983; Rep2: 10,465). This substantial difference can be explained: I used a more sensitive peak-caller algorithm (HOMER) with specific parameters tuned for a transcription factor,

which is opposed to the potentially more conservative tool or parameters used in the original study. Additionally, the permissive definition of overlap (requiring only 1 bp) for defining reproducible peaks in my pipeline was designed to maximize sensitivity, further contributing to the larger peak set. Despite the quantitative differences, the key qualitative conclusion holds: a substantial subset of RUNX1 binding sites is reproducible across biological replicates, validating the overall success and reliability of the ChIP-seq experiment.

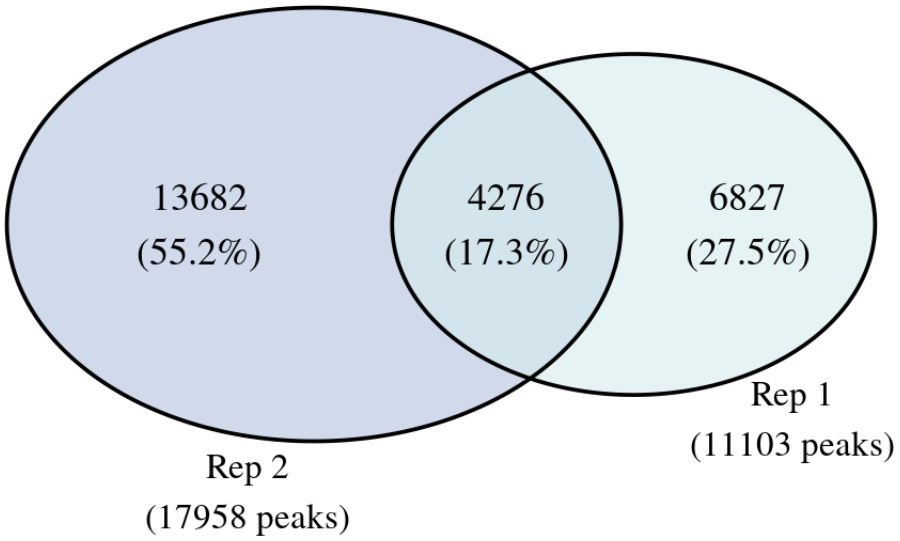

Figure 7. Venn Diagram showing the RUNX1 peak reproducibility among the biological replicates.

Annotated Peaks Anlysis and Functional Networks of RUNX1 Target Genes

Annotation of the 4,364 reproducible RUNX1 peaks revealed a strong enrichment for promoter-TSS regions, with nearly half (48.7%) of all peaks **(Table 2)**. This distribution is consistent with RUNX1's canonical role as a transcription factor directly regulating gene expression initiation. To elucidate the functional impact of this binding, I extracted the protein-coding genes associated with promoter-bound RUNX1 peaks and analyzed their network using the STRING database[13] **(Figure 8)**. The resulting protein-protein interaction network revealed significant clustering of genes involved in key cancer-related pathways, including regulators of the cell cycle, transcriptional regulation, and signal transduction. This

analysis indicates that RUNX1 targets are not isolated entities but form coherent functional modules, suggesting a coordinated regulatory program where RUNX1 acts as a master regulator of oncogenic networks in MCF-7 cells.

Table 2. RUNX1 Peaks Annotation Summary

Annotation_Type	Peak_Number	Percentage (%)
promoter-TSS	2123	48.66
intron	1200	27.5
Intergenic	727	16.66
TTS	158	3.62
exon	155	3.55

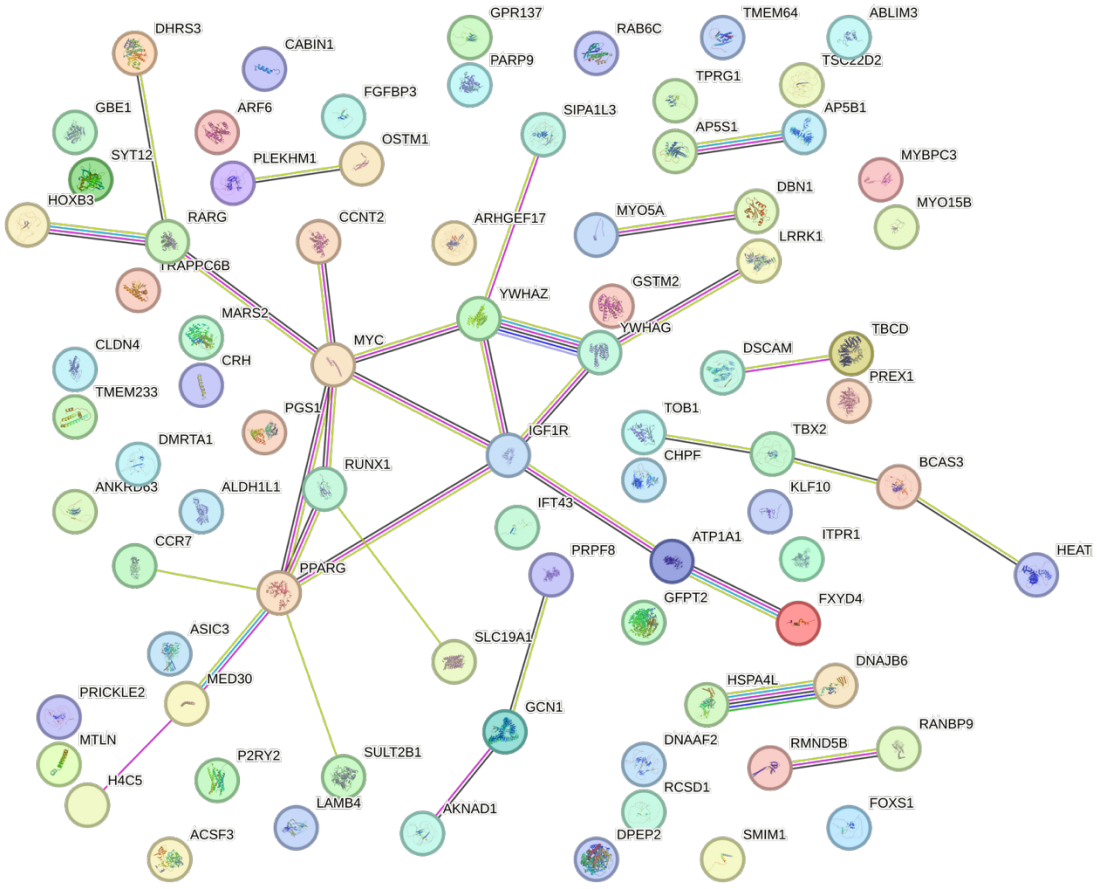

Figure 8. STRING network analysis of protein–protein interactions among genes associated with promoter-bound RUNX1 peaks.

Reference:

(1) Okuda, T.; Nishimura, M.; Nakao, M.; Fujitaa, Y. RUNX1/AML1: A Central Player in Hematopoiesis. *Int. J. Hematol.* **2001**, *74* (3), 252–257. https://doi.org/10.1007/BF02982057.

(2) Barutcu, A. R.; Hong, D.; Lajoie, B. R.; McCord, R. P.; Van Wijnen, A. J.; Lian, J. B.; Stein, J. L.; Dekker, J.; Imbalzano, A. N.; Stein, G. S. RUNX1 Contributes to Higher-Order Chromatin Organization and Gene Regulation in Breast Cancer Cells. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **2016**, *1859* (11), 1389–1397. https://doi.org/10.1016/j.bbagrm.2016.08.003.

(3) Di Tommaso, P.; Chatzou, M.; Floden, E. W.; Barja, P. P.; Palumbo, E.; Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* **2017**, *35* (4), 316–319. https://doi.org/10.1038/nbt.3820.

(4) Kratzke, N. Lightweight Virtualization Cluster How to Overcome Cloud Vendor Lock-In. *J. Comput. Commun.* **2014**, *02* (12), 1–7. https://doi.org/10.4236/jcc.2014.212001.

(5) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30* (15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

(6) Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9* (4), 357–359. https://doi.org/10.1038/nmeth.1923.

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

(8) Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* **2016**, *32* (19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

(9) Ramírez, F.; Ryan, D. P.; Grüning, B.; Bhardwaj, V.; Kilpert, F.; Richter, A. S.; Heyne, S.; Dündar, F.; Manke, T. deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. *Nucleic Acids Res.* **2016**, *44* (W1), W160–W165. https://doi.org/10.1093/nar/gkw257.

(10) Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y. C.; Laslo, P.; Cheng, J. X.; Murre, C.; Singh, H.; Glass, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **2010**, *38* (4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

(11) Quinlan, A. R.; Hall, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26* (6), 841–842. https://doi.org/10.1093/bioinformatics/btq033.

(12) Thorvaldsdottir, H.; Robinson, J. T.; Mesirov, J. P. Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Brief. Bioinform.* **2013**, *14* (2), 178–192. https://doi.org/10.1093/bib/bbs017.

(13) Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A. L.; Fang, T.; Doncheva, N. T.; Pyysalo, S.; Bork, P.; Jensen, L. J.; von Mering, C. The

STRING Database in 2023: Protein–Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest. *Nucleic Acids Res.* **2023**, *51* (D1), D638–D646. https://doi.org/10.1093/nar/gkac1000.