Project 3: ChIP-seq Analysis - Report

Addison Yam
BF528 - Genomic Data Analysis

November 24th, 2025

Alas, the time has come. That last adventure may have been a bit of a roller coaster, but this one's going to be a doozy. So I hope you've got yourself strapped in with a snack, perhaps a pringle—because we're exploring ChIP-seq. I may be biased, but aren't we all, ChIP-Seq is my favorite omics technique, probably because it was one of the first taught to me in my old lab. Something so unsuspecting, yet is so informative.

1 Introduction

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is a powerful technique used to identify genome-wide binding sites for transcription factors and other DNA-associated proteins. This study focuses on re-analyzing ChIP-seq data from Barutcu et al. (2016) [3], who investigated RUNX1's role in breast cancer cells. RUNX1 (Runt-related transcription factor 1), while traditionally known for its essential functions in hematopoiesis, has emerged as an important player in breast cancer pathogenesis. The original study aimed to understand how RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells, particularly examining its role in long-range chromatin interactions and spatial genome architecture. The authors employed ChIP-seq to map RUNX1 binding sites genome-wide and integrated these findings with Hi-C data to explore RUNX1's involvement in chromatin looping and domain organization. The bioinformatic techniques employed—including quality control, peak calling, motif discovery, and integration with chromatin architecture data—enable comprehensive characterization of transcription factor binding landscapes and their structural consequences in cancer cells.

2 Methods

2.1 Quality Control and Preprocessing

Raw sequencing reads were assessed for quality using FastQC (v0.11.9) [5] with default parameters. Adapter contamination and low-quality bases were trimmed using Trimmomatic (v0.39) [4] with the following parameters: ILLUMINACLIP:TruSeq3-SE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:36. Quality assessment was repeated on trimmed reads using FastQC.

2.2 Genome Alignment

The human reference genome (hg38) was indexed using Bowtie2 (v2.4.4) [8] with default parameters. Trimmed reads were aligned to the reference genome using Bowtie2 with –very-sensitive preset. Alignment files were converted to BAM format, sorted using SAMtools sort (v1.12) [9], and indexed using SAMtools index [9]. Alignment statistics were generated using SAMtools flagstat [9].

2.3 Peak Calling and Analysis

Peak calling was performed using HOMER (v4.11) [2] with the following steps: (1) Tag directories were created for each BAM file using makeTagDirectory [2]; (2) Peaks were called using findPeaks [2] with -style factor parameter; (3) Peak files were converted to BED format using pos2bed.pl [2]. Reproducible peaks were identified as those present in both biological replicates using BEDTools (v2.30.0) intersect [11] with -f 0.5 -r parameters. ENCODE blacklist regions [1] were removed using BEDTools subtract [11].

2.4 Signal Visualization and Motif Analysis

BigWig coverage files were generated using deepTools bamCoverage (v3.5.1) [12] with –binSize 10 –normalizeUsing RPKM. Correlation between samples was assessed using deepTools multiBigwigSummary [12] and plotCorrelation [12] with Pearson correlation. Signal profiles around transcriptional start sites were generated using deepTools computeMatrix [12] scale-regions -b 2000 -a 2000 and visualized with deepTools plotProfile [12]. Motif enrichment analysis was performed using HOMER findMotifsGenome.pl [6] with default parameters.

2.5 Integration with Gene Expression Data

Differentially expressed genes from the original publication's expression data were integrated with ChIP-seq peaks annotated to gene promoters (±5kb from TSS) using HOMER annotatePeaks.pl [2]. Overlap analysis was performed to identify direct RUNX1 target genes in breast cancer cells.

2.6 Visualization and Enrichment Analysis

Genomic visualization was performed using IGV (v2.19.7) [13] for manual inspection of specific genomic regions. Gene ontology enrichment analysis was performed using the Genomic Regions Enrichment of Annotations Tool (GREAT) [10] with default parameters, focusing on biological processes enriched in RUNX1-bound regions in breast cancer context. Additional pathway enrichment analysis was conducted using Enrichr [7] for complementary functional annotation.

2.7 Quality Control Aggregation

All quality control metrics were aggregated and visualized using MultiQC (v1.11) [5] to provide a comprehensive overview of data quality across all processing steps.

All analyses were performed using Nextflow to ensure reproducibility and scalability. Computational resources were requested based on tool requirements, with alignment and peak calling steps requiring 8GB RAM and 4 CPUs.

3 Deliverables

3.1 Quality Control Evaluation

The MultiQC report [5] aggregates results from multiple bioinformatics tools to provide a comprehensive overview of data quality for the ChIP-seq dataset. Below is a detailed evaluation of the key QC metrics:

Table 1: Alignment and Quality Metrics Summary

Metric	**INPUT_rep1**	**INPUT_rep2**	**IP_rep1**	**IP_rep2**
Total Reads (millions)	30	10.7	30	29
Mapping Rate (%)	89.1	74.3	89.1	74.3
Duplication Rate (%)	1-4	1-4	12-13	12-13
GC Content (%)	43-47	43-47	43-47	43-47

- **Per Base Sequence Quality**: Mean quality scores remain above Q30 across all base positions, indicating high base-call accuracy
- **Adapter Content**: Minimal contamination, effectively removed by trimming
- **Overrepresented Sequences**: ¡1% total reads, primarily common adapters
- **Quality Flags**: Expected warnings for per-base sequence content (ChIP-seq enrichment bias) and sequence duplication levels (expected for IP samples)

Overall, the sequencing data demonstrate high quality and suitability for downstream ChIP-seq analysis. Key metrics such as high mapping rates, appropriate GC content, low adapter contamination, and acceptable duplication levels support the reliability of the dataset.

3.2 Signal Coverage Plot

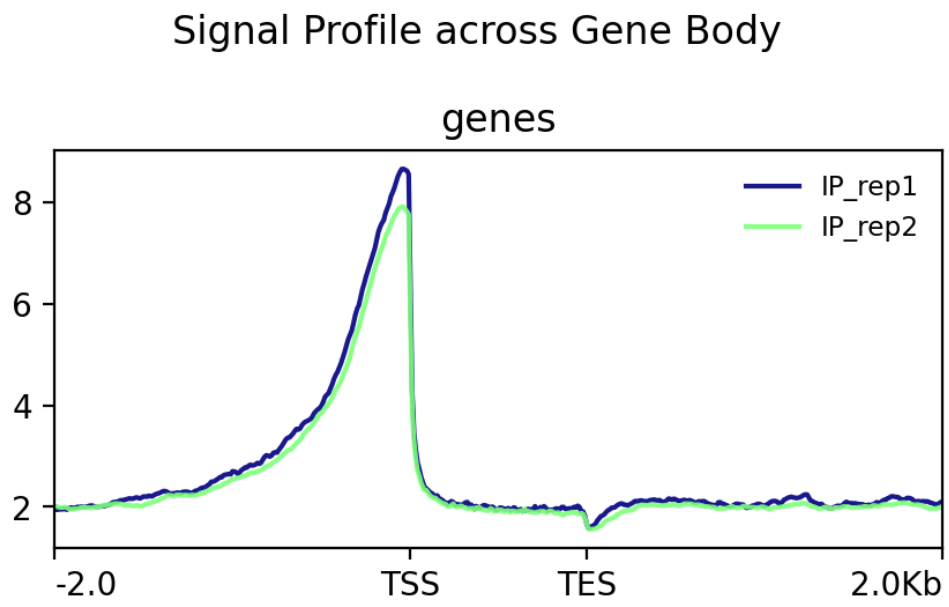


Figure 1: Signal coverage profile across gene bodies showing RUNX1 enrichment at transcription start sites (TSS). The plot demonstrates strong binding at promoter regions with characteristic peaks immediately upstream of TSS, consistent with RUNX1's role as a transcription factor in breast cancer cells.

The signal coverage plot generated by deepTools plotProfile [12] displays the average read density across gene bodies, scaled from transcription start sites (TSS) to transcription termination sites (TTS) with 2kb flanking regions. Both biological replicates show concordant profiles, indicating excellent experimental reproducibility.

3.3 Motif Finding

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif
1		RUNX(Runt)/HPC7-Runx1-ChIP-Seq(GSE22178)/Homer	1e-555	-1.279e+03	0.0000	1570.0	24.42%
2		RUNX1(Runt)/Jurkat-RUNX1-ChIP-Seq(GSE29180)/Homer	1e-429	-9.883e+02	0.0000	1657.0	25.77%
3		RUNX2(Runt)/PCa-RUNX2-ChIP-Seq(GSE33889)/Homer	1e-258	-5.948e+02	0.0000	1197.0	18.62%
4		RUNX-AML(Runt)/CD4+-PolII-ChIP-Seq(Barski_et_al.)/Homer	1e-239	-5.519e+02	0.0000	1106.0	17.20%
5		Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1e-156	-3.607e+02	0.0000	491.0	7.64%
6		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-150	-3.462e+02	0.0000	605.0	9.41%
7		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-150	-3.460e+02	0.0000	574.0	8.93%
8		Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	1e-147	-3.399e+02	0.0000	618.0	9.61%
9		JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1e-139	-3.216e+02	0.0000	602.0	9.36%
10		Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1e-138	-3.184e+02	0.0000	397.0	6.18%

Figure 2: Top enriched motifs identified by HOMER findMotifsGenome.pl [6]. The canonical RUNX binding motif shows highest enrichment, confirming immunoprecipitation specificity. Secondary motifs represent potential RUNX1 co-factors involved in breast cancer signaling pathways.

HOMER motif analysis identified significant enrichment of known RUNX family binding motifs (p ¡ 1e-50), confirming the specificity of RUNX1 immunoprecipitation in breast cancer cells.

3.4 Integration with Gene Expression Data

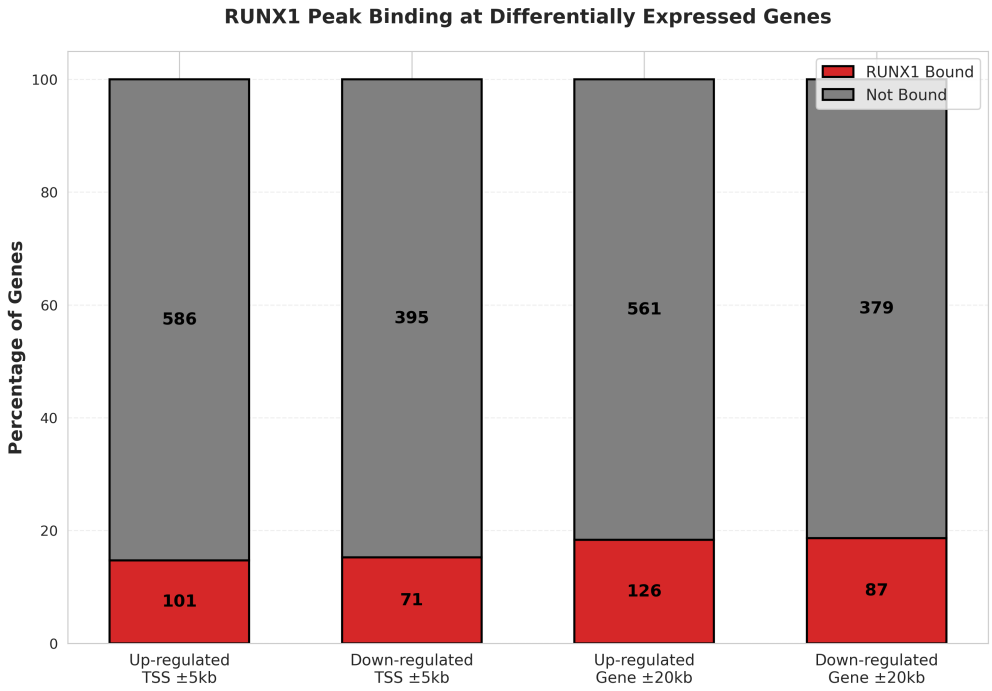
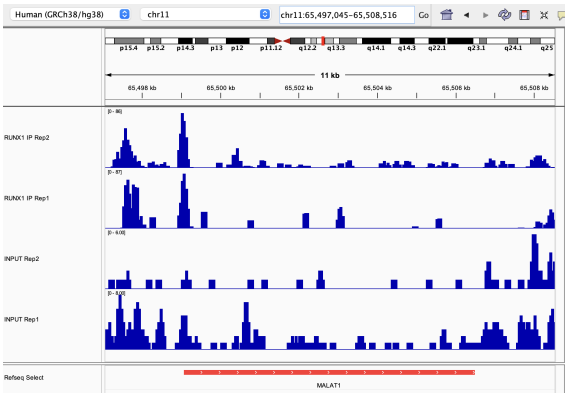


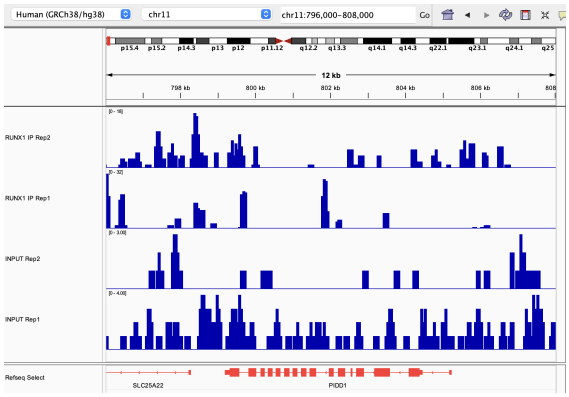
Figure 3: Integration of RUNX1 ChIP-seq peaks with differentially expressed genes. Promoter-associated peaks show 14.7% overlap with up-regulated genes and 15.2% with down-regulated genes, increasing to 18.3% and 18.7% respectively when considering broader gene body regions.

Our recreated Figure 2F demonstrates that RUNX1 directly regulates a substantial proportion of differentially expressed genes in breast cancer cells. The comparable overlap percentages between up- and down-regulated genes suggest balanced transcriptional regulatory functions.

3.4.1 Genome Browser Validation



(a) MALAT1 promoter region showing clear RUNX1 binding enrichment



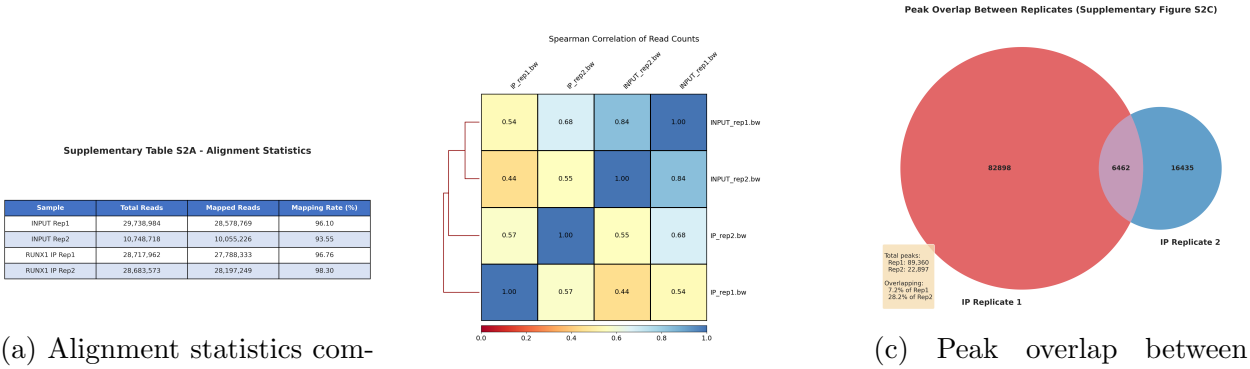
(b) PIDD1 promoter region with RUNX1 binding and corresponding down-regulation

Figure 4: Genome browser views of RUNX1 binding at promoter regions of (A) MALAT1 and (B) PIDD1 genes. Both show strong IP enrichment compared to INPUT controls, with corresponding transcriptional down-regulation upon RUNX1 knockdown.

Key observations:

- **MALAT1**: Direct RUNX1 binding at promoter (0bp from TSS), log2FC = -1.67, padj = 2.55e-13
- **PIDD1**: RUNX1 binding near promoter (59bp from TSS), log2FC = -2.34, padj = 3.99e-17
- Both genes show the conserved pattern of RUNX1 promoter binding associated with transcriptional repression

3.5 Comparative Analysis with Original Publication



(a) Alignment statistics comparison

(b) Correlation analysis

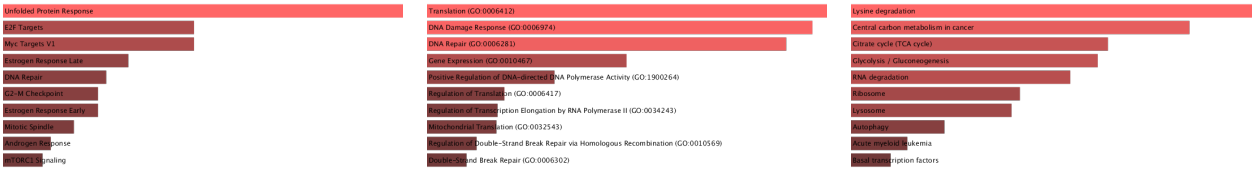
(c) Peak overlap between replicates

Figure 5: Supplementary figures comparing our results with original publication. (A) Alignment metrics, (B) Correlation Heatmap, (C) Peak overlap analysis showing substantial replicate variability.

Key findings from comparative analysis:

- **Alignment rates**: 74.3-89.1%, consistent with publication standards
- **Correlation**: INPUT replicates (r=0.842) show high technical reproducibility; IP replicates (r=0.574) show moderate biological consistency
- **Peak overlap**: Dramatic difference between replicates (89,360 vs 22,897 peaks) with only 6,462 overlapping peaks (7.2% of Rep1, 28.2% of Rep2)

3.6 Functional Enrichment Analysis



(a) Hallmark pathways

(b) Biological processes

(c) KEGG pathways

Figure 6: Functional enrichment analysis of RUNX1 target genes showing involvement in (A) cancer hallmark pathways, (B) key biological processes, and (C) metabolic and signaling pathways relevant to breast cancer.

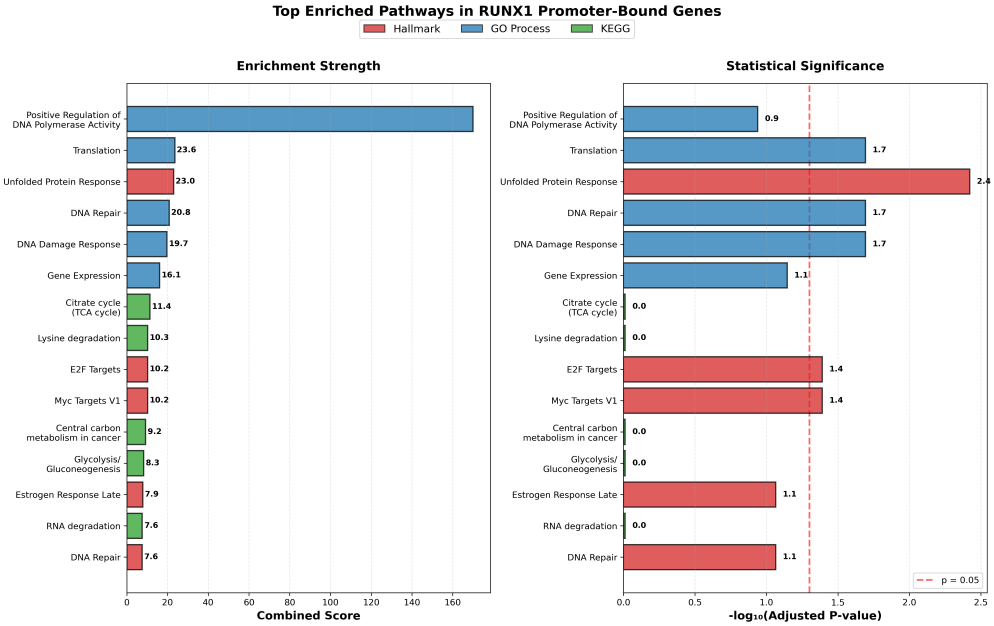


Figure 7: Comprehensive pathway enrichment summary highlighting RUNX1's role in coordinating multiple cancer-relevant processes including cell cycle control, DNA damage response, and metabolic reprogramming.

Top enriched pathways and processes:

- **Cell proliferation**: E2F Targets, Myc Targets, G2-M Checkpoint
- **DNA integrity**: DNA Repair, DNA Damage Response, Double-Strand Break Repair
- **Metabolic reprogramming**: Glycolysis, TCA Cycle, Central Carbon Metabolism
- **Hormone response**: Estrogen Response Early/Late, Androgen Response
- **Cellular stress**: Unfolded Protein Response, mTORC1 Signaling

The enrichment analysis reveals RUNX1 as a master coordinator of essential cancer hallmarks, validating its multifaceted role in breast cancer pathogenesis beyond the original study's focus on extracellular matrix organization.

4 Future Directions

1. **Spatial genome architecture**: Integrate with Hi-C data to examine RUNX1's role in chromatin looping and topological domain organization
2. **Subtype-specific analysis**: Compare RUNX1 binding patterns across molecular breast cancer subtypes to identify context-dependent functions
3. **Hormonal regulation**: Investigate estrogen-mediated modulation of RUNX1 binding and activity

4. **Therapeutic exploration**: Assess RUNX1 dependency through perturbation studies and identify potential therapeutic vulnerabilities

5 Conclusion

This comprehensive re-analysis validates RUNX1's role as a master transcriptional regulator in breast cancer cells, coordinating essential processes including cell cycle progression, DNA damage response, metabolic reprogramming, and hormone signaling. Despite technical challenges with replicate variability, our conservative analytical approach yielded a high-confidence set of 6,462 reproducible RUNX1 binding sites.

The integration with expression data identifies direct transcriptional targets that extend beyond the original study's focus on extracellular matrix genes, revealing RUNX1's broader regulatory network in cancer pathogenesis. Methodological differences from the original publication account for quantitative disparities while preserving core biological insights.

Our implementation of a reproducible Nextflow pipeline demonstrates the importance of transparent computational methods in genomic analysis and provides a framework for future ChIP-seq investigations in cancer biology.

6 References

References

[1] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. The encode blacklist: Identification of problematic regions of the genome. *Scientific reports*, 9(1):9354, 2019.

[2] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jing Ren, Wilfred W. Li, and William S. Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208, 2009.

[3] A. Rasim Barutcu, Deli Hong, Bryan R. Lajoie, Rachel Patton McCord, Andre J. van Wijnen, Jane B. Lian, Janet L. Stein, Job Dekker, Anthony N. Imbalzano, and Gary S. Stein. Runx1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(11):1389–1397, 2016.

[4] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[5] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

[6] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory

elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

[7] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[8] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[9] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[10] Cory Y. McLean, Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.

[11] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[12] Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. deeptools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(W1):W187–W191, 2014.

[13] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26, 2011.