Identifying Chromatin Accessibility Changes in Leukemia using ATAC Sequencing

Ramya Harshitha Bolla
Department of EECS
University of Kansas
Lawrence, KS, USA
3158932

Hemika Amilineni
Department of EECS
University of Kansas
Lawrence, KS, USA
3160576

George Steven Muvva
Department of EECS
University of Kansas
Lawrence, KS, USA
3132947

***Abstract*—Chromatin accessibility is a fundamental feature of gene regulation, governing the ability of transcription factors and regulatory proteins to access DNA. Disruptions in chromatin structure are frequently observed in cancer, including leukemia, where they can lead to aberrant gene expression and dysregulated cellular behavior. In this study, we use ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) to perform a genome-wide comparison of chromatin accessibility between leukemic (K562) and healthy (GM12878) human cell lines.**

We developed a fully automated and reproducible analysis pipeline using Nextflow DSL2, integrating essential steps such as quality control, adapter trimming, alignment, peak calling, differential accessibility analysis, gene annotation, and functional enrichment. The pipeline produces interpretable outputs including annotated peak tables, statistical summaries, and visualizations such as volcano plots, heatmaps, PCA plots, and pathway enrichment profiles.

Our analysis revealed hundreds of differentially accessible regions (DARs), many of which are associated with genes involved in hematopoiesis, chromatin remodeling, and cancer signaling. Enrichment analysis highlighted key biological processes and pathways dysregulated in leukemia, including immune signaling, developmental regulation, and oncogenic transcriptional programs.

Together, these results demonstrate the power of ATAC-seq to uncover epigenomic alterations in leukemia and provide a flexible, scalable pipeline that can be extended to other disease models and datasets.

Keywords— ATAC-seq, Leukemia, Chromatin Accessibility, Differential Accessibility, Nextflow, Bioinformatics, Gene Regulation, Functional Enrichment

I. Introduction

In eukaryotic cells, gene regulation is controlled not just by DNA sequences, but by how that DNA is packaged into chromatin. This structure composed of DNA wrapped around histone proteins can either restrict or permit access to different parts of the genome. When chromatin is open and accessible, regulatory proteins like transcription factors can bind and initiate gene expression. When it's tightly packed, gene activity is typically repressed.

Chromatin accessibility plays a vital role in determining which genes are turned on or off in different cell types and under various conditions. Disruptions in this accessibility can have serious consequences, especially in diseases like cancer. In leukemia, for example, mutations and epigenetic changes often lead to abnormal chromatin states, misregulating genes involved in cell growth, differentiation, and survival. Understanding these changes at the chromatin level can offer valuable insights into how leukemia develops and potentially reveal new points of therapeutic intervention.

To study chromatin accessibility on a genome-wide scale, researchers use a technique called ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing). It works by using a special enzyme (Tn5 transposase) to insert sequencing adapters into open regions of DNA places where the chromatin is not tightly packed. These accessible regions are then sequenced, allowing us to map regulatory elements like enhancers and promoters across the genome.

In this project, we use ATAC-seq data from two well-established human cell lines: K562 (a model for chronic myelogenous leukemia) and GM12878 (a healthy B-lymphocyte-derived line). Our goal is to compare their chromatin accessibility landscapes and identify regions that are significantly more or less accessible in leukemia. These differentially accessible regions (DARs) may correspond to important regulatory elements that are disrupted in cancer.

To carry out this analysis, we built a fully automated ATAC-seq pipeline using Nextflow DSL2. The pipeline includes steps for quality control (FastQC), adapter trimming (Trim Galore), genome alignment (Bowtie2), peak calling (MACS2), read quantification (featureCounts), and differential analysis (DESeq2). We also integrated gene annotation and enrichment analysis using Gene Ontology (GO), KEGG, and Reactome databases.

By automating the workflow, we ensure that the entire process from raw data to biological interpretation is reproducible, efficient, and easy to adapt for future studies. The final outputs include differentially accessible peaks, associated genes, and clear visual summaries such as volcano plots, PCA, heatmaps, and enrichment charts that help interpret the results in a biological context.

II. ATAC-seq: Overview and Motivation

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a modern and efficient method used to map genome-wide chromatin accessibility. It gives researchers a snapshot of the regulatory landscape by identifying regions

of DNA that are "open" or accessible to transcription factors and other regulatory proteins.

The method works by using a hyperactive Tn5 transposase enzyme that simultaneously cuts open DNA and inserts sequencing adapters in a single step a process known as tagmentation. After sequencing, these DNA fragments are aligned to a reference genome. Short fragments often indicate nucleosome-free regions, such as active promoters or enhancers, while longer fragments correspond to regions where DNA is wrapped around nucleosomes.

Because it requires minimal input material and has a simple workflow, ATAC-seq has become a preferred tool for profiling epigenetic states in both bulk and single-cell experiments. Its high resolution and sensitivity make it especially useful for identifying regulatory elements involved in cell identity, development, and disease.

A. Why ATAC-seq is Important for Leukemia Research

Leukemia is a diverse group of hematologic cancers characterized by uncontrolled proliferation of blood-forming cells. While many studies have focused on mutations and gene expression profiles, epigenetic regulation - particularly chromatin accessibility has emerged as a key factor in leukemogenesis. Abnormal chromatin structure can activate oncogenes, repress tumor suppressors, and disrupt normal hematopoietic differentiation.

ATAC-seq enables researchers to identify leukemia-specific regulatory elements that may not be evident through RNA-seq or DNA mutation analyses. For instance, it can uncover enhancer reprogramming, aberrant transcription factor binding, or promoter accessibility shifts that precede changes in gene expression. These insights are vital for understanding both disease initiation and progression.

Moreover, because chromatin accessibility is dynamic and reversible, it presents a promising target for therapeutic intervention. Epigenetic drugs such as histone deacetylase (HDAC) inhibitors and DNA methyltransferase (DNMT) inhibitors have already shown promise in treating leukemia, and mapping accessibility changes can help refine such treatments.

B. Study Motivation

Although ATAC-seq has proven valuable in studying chromatin dynamics, many existing studies lack standardized and reproducible computational workflows especially in the context of leukemia. This makes it difficult to compare results across datasets or confidently interpret findings.

Our study is driven by two core needs: to better understand the regulatory landscape of leukemia at the chromatin level, and to create a reproducible pipeline that simplifies ATAC-seq data analysis from start to finish.

C. Objectives

This study set out with two main objectives. The first was biological- to understand how chromatin accessibility differs between leukemic and healthy human cells, and what these differences can reveal about the regulatory mechanisms driving leukemia. The second was computational to develop a reliable, automated pipeline that enables reproducible analysis of ATAC-seq data from start to finish.

Specifically, our goals were to:

- **Identify Differentially Accessible Regions (DARs):**Identify Differentially Accessible Regions (DARs): By comparing chromatin accessibility between K562 (leukemic) and GM12878 (normal) cell lines, we aimed to uncover genomic regions that are significantly more or less accessible in leukemia. These changes could mark regulatory elements involved in disease progression.
- **Interpret Functional Implications:** Annotate significant peaks with their nearest genes and perform enrichment analysis to reveal affected biological processes and pathways involved in leukemogenesis.
- **Build a Reproducible ATAC-seq Pipeline:** Develop a modular and fully automated workflow using Nextflow DSL2, integrating tools such as FastQC, Trim Galore, Bowtie2, MACS2, featureCounts, and DESeq2.
- **Enable Insightful Visualization:** Generate clear visual outputs like volcano plots, heatmaps, PCA plots, and enrichment diagrams to help interpret and communicate the key findings.

Together, these objectives aim to provide both mechanistic insights into leukemia and a reusable pipeline for broader ATAC-seq applications in biomedical research.

III. Methodology

This section describes the computational workflow implemented for chromatin accessibility analysis using ATAC-seq data. The pipeline was developed using Nextflow DSL2, enabling modular design, scalability, and reproducible execution of all analytical steps. Publicly available ATAC-seq datasets were obtained from the ENCODE project, representing two human cell lines: K562, a chronic myelogenous leukemia cell line, and GM12878, a normal lymphoblastoid B-cell line. The workflow includes data preprocessing, alignment, peak calling, quantification, differential analysis, annotation, and downstream visualization.

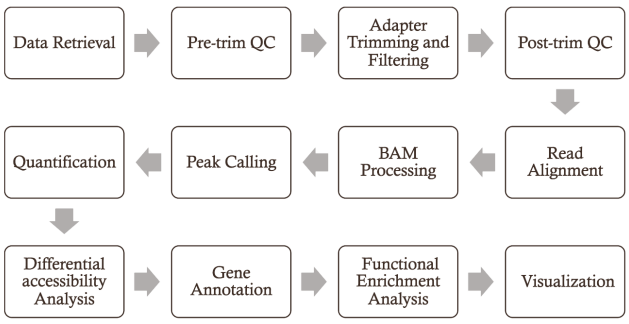

Fig. 1. Overview of the ATAC-seq data analysis workflow.

A. Data Acquisition

Publicly available ATAC-seq datasets were obtained from the ENCODE Project Portal to analyze chromatin accessibility patterns in leukemic and healthy human cell lines. The datasets selected represent well-characterized models for disease and control conditions.

Leukemic Cell Line - K562:

- **Accession IDs:** ENCFF391BFJ, ENCFF186CQZ
- **Library Type:** Single-end ATAC-seq
- **Purpose:** Profile chromatin accessibility in the K562 cell line, a model for chronic myelogenous leukemia (CML).
- **Significance:** Identifies leukemia-associated regulatory changes and open chromatin regions.

Healthy Control – GM12878:

- **Accession IDs:** ENCFF411MVZ, ENCFF736ZWY
- **Library Type:** Single-end ATAC-seq
- **Purpose:** Represents a normal chromatin landscape using the GM12878 lymphoblastoid cell line.
- **Significance:** Serves as a baseline for identifying leukemia-specific chromatin accessibility alterations.

The raw FASTQ files were downloaded directly from the ENCODE portal using the provided accession numbers. A metadata file was manually created to associate each sample with its biological condition (Leukemia or Healthy) and replicate ID. The directory structure was organized to facilitate compatibility with the downstream Nextflow pipeline.

```
FileName,Condition
ENCFF186CQZ_leukemia.bam,Leukemia
ENCFF391BFJ_leukemia.bam,Leukemia
ENCFF411MVZ.bam,Normal
ENCFF736ZWY.bam,Normal
```

Fig. 2. Metadata

- **Reference Genome:** GRCh38, downloaded from the GENCODE project.
- **Annotation File:** GENCODE v43 GTF file, used for Gene-level peak annotation

B. Quality Control using FastQC

Chromatin accessibility profiling by ATAC-seq produces short DNA fragments that are often prone to technical artifacts such as adapter contamination, low-quality base calls, and biased nucleotide composition, particularly at the ends of reads. If not corrected, these issues can adversely impact downstream analyses such as read alignment and peak calling.

To assess the quality of the raw sequencing data, we ran FastQC on each FASTQ file individually. This was done using the following command within a Nextflow pipeline:

fastqc <sample.fastq.gz> -o .

We examined the following metrics in each FastQC report:

- **Per-base sequence quality:** Most reads showed high-quality scores across all positions. Slight quality drop-offs at the ends were observed, which is typical in ATAC-seq datasets.
- **Per-sequence GC content:** The GC distribution matched expectations for human genomic DNA, indicating no major contamination or bias.
- **Adapter content:** A small percentage of reads contained residual adapter sequences, which justified performing a trimming step.
- **Sequence duplication levels:** Moderate duplication levels were detected, likely due to PCR amplification during library preparation.

These results informed our decision to perform adapter and quality trimming in the next step using Trim Galore.

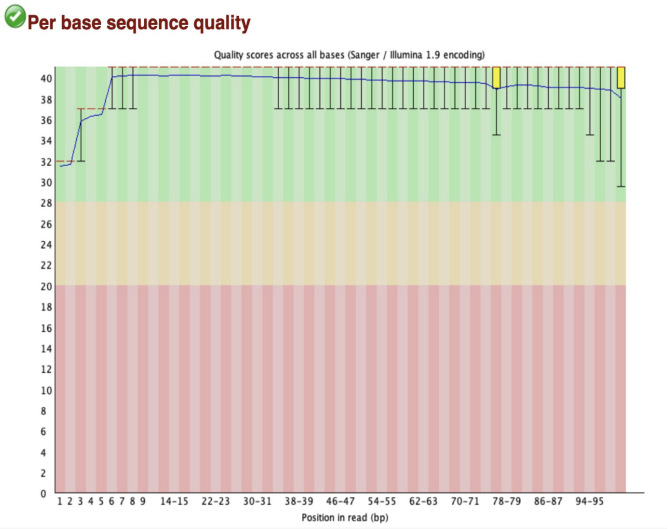

Fig. 3. Per base sequence quality

Basic Statistics

Measure	Value
Filename	ENCFF411MVZ.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	38239941
Total Bases	3.8 Gbp
Sequences flagged as poor quality	0
Sequence length	101
%GC	52

Fig. 4. Basic Statistics

C. Adapter Trimming using Trim Galore

Following initial quality control, the FastQC reports indicated the presence of adapter sequences and reduced base quality towards the ends of reads common artifacts introduced during ATAC-seq library preparation. These issues can negatively affect read alignment and the accuracy of downstream peak calling.

To correct these issues, we used Trim Galore, a widely adopted wrapper around Cutadapt, specifically designed for trimming Illumina adapter sequences and low-quality bases from sequencing reads. This step ensures that only high-quality, adapter-free reads are passed to the alignment stage.

Trim Galore was run automatically on each input file in the pipeline using the following command:

trim_galore <sample.fastq> –gzip –output_dir .

To confirm the effectiveness of trimming, we ran FastQC on the cleaned reads again. The updated quality reports(Fig. 6) showed a complete or near-complete removal of adapter contamination, improved per-base quality scores, lower duplication levels and overall cleaner read profiles.

By trimming adapters and low-quality bases, we improved the integrity of the sequencing data, which subsequently enhanced alignment accuracy and downstream chromatin accessibility analysis.

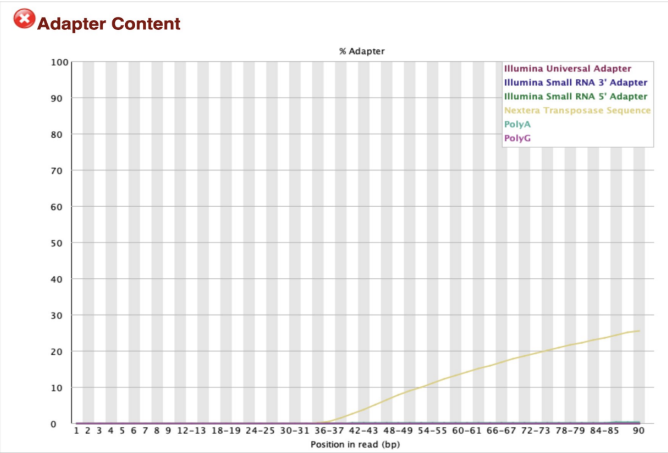

Fig. 5. Adapter Content before Trimming

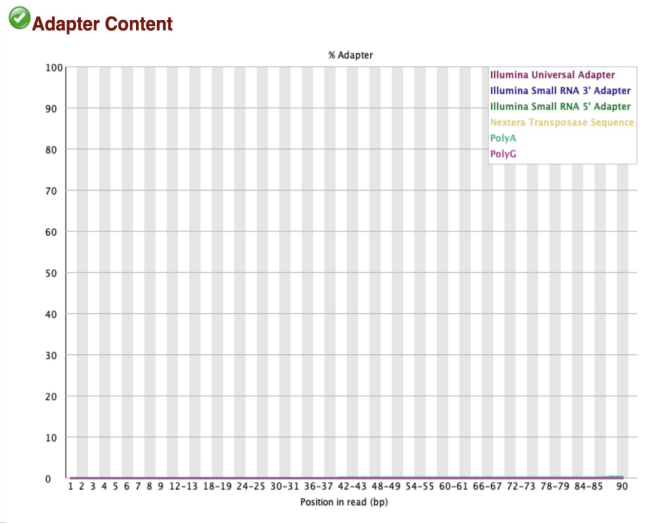

Fig. 6. Adapter Content after Trimming

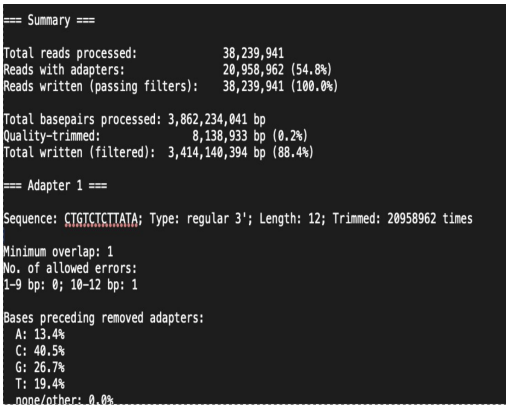

Fig. 7. Trimming Information

D. Alignment to the Reference Genome using Bowtie2

High-quality, adapter-trimmed reads were aligned to the GRCh38 human reference genome using **Bowtie2**, a fast and memory-efficient short-read aligner commonly used in ATAC-seq workflows. The alignment was performed in single-end mode with the –sensitive preset, which provides a good balance between speed and accuracy for genomics applications.

To streamline processing and conserve storage, the Bowtie2 output was directly piped into **SAMtools** to sort the aligned reads by genomic coordinates. Additionally, index files (.bai) were generated for each sorted BAM file to enable efficient querying by downstream tools such as IGV and featureCounts.

The pipeline executed the following alignment command for each sample:

bowtie2 -x <genome_index> -U <sample.fastq> –sensitive -p 4 2>><sample>_align.log | samtools sort -o <sample>.bam - samtools index <sample>.bam

Each alignment also generated a corresponding log file as shown in Fig. 8 containing key statistics like total number of input reads, number and percentage of uniquely mapped reads, number of reads mapped to multiple locations, overall alignment rate.

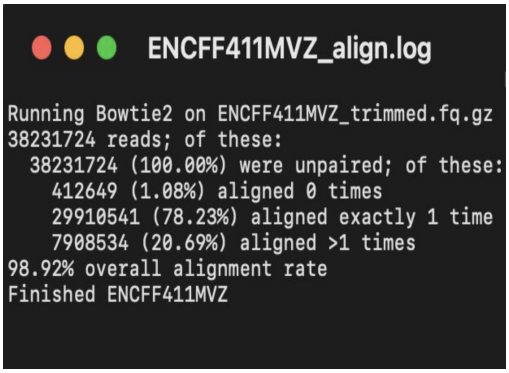

Fig. 8. Alignment log details of 1 sample

E. Peak Calling using MACS2

Following alignment, open chromatin regions were identified using **MACS2** (Model-based Analysis for ChIP-Seq), a standard peak caller used in ATAC-seq workflows. MACS2 is particularly well-suited for identifying sharp, narrow peaks associated with accessible regulatory elements such as promoters and enhancers.

The pipeline executed the following command for each aligned BAM file:

macs2 callpeak -t <sample.bam> -f BAM -g hs –nomodel -n <sample_name>_peaks

The output of this step included .narrowPeak files containing peak coordinates and enrichment scores as shown in Fig. 9,.xls and .bed files with additional peak information.

To create a unified peak reference for quantification, all .narrowPeak files across samples were merged using *Bedtools*, which consolidated overlapping peaks into non-redundant regions.

*cat *.narrowPeak | sort -k1,1 -k2,2n | bedtools merge -i - > merged_peaks.bed*

The merged peak regions were then converted to SAF (Simplified Annotation Format), which is compatible with *featureCounts*. Each SAF row included a unique peak ID, chromosome, start, end, and strand, allowing accurate assignment of reads to regulatory elements in subsequent quantification.

start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)
9908	10759	852	10201	87	68.9659	12.7536	65.5276
11136	11516	381	11266	40	20.0579	5.82386	17.6716
28904	29956	1053	29290	77	58.319	11.5727	55.1146
96382	97082	701	96685	47	36.3563	10.6195	33.6074
180722	181069	348	180921	38	22.2127	6.89046	19.7734

Fig. 9. MACS2 narrowPeak output for detected accessible chromatin regions.

F. Reads Counting using FeatureCounts

Quantification of reads overlapping the merged peak regions was performed using *featureCounts*, a high-performance read summarization tool. The SAF-formatted merged peak file and all aligned BAM files were used to generate a matrix of raw read counts, with rows corresponding to peak regions and columns to individual samples. This matrix served as the input for downstream differential accessibility analysis using DESeq2.

featureCounts -a $saf_file -F SAF -o read_counts.txt $bam_files.join(' ')

	ENCFF186CQZ_leukemia.bam	ENCFF391BFJ_leukemia.bam	ENCFF736ZWY.bam	ENCFF411MVZ.bam
peak_1	393	445	524	588
peak_2	109	101	120	122
peak_3	361	414	381	382
peak_4	16	16	149	151
peak_5	59	62	65	76
peak_6	136	140	157	153
peak_7	58	55	43	45
peak_8	509	544	482	533
peak_9	25	19	30	30
peak_10	77	78	100	97

Fig. 10. Peaks Reads Count

As shown in Fig. 10, peak-level read counts vary substantially across samples, with some peaks showing consistently high signal, indicating strong chromatin accessibility. This count matrix was further processed to match the sample metadata and filtered to remove low-coverage peaks and inconsistently labeled samples, ensuring high confidence in downstream differential accessibility analysis.

G. Differential Peak Analysis using Deseq2

Differential accessibility analysis was performed using DESeq2, an R package designed to model count data and identify statistically significant differences between conditions. The raw count matrix generated by featureCounts, along with sample condition metadata, was used as input. DESeq2 performed normalization and statistical testing to identify differentially accessible chromatin regions between the leukemia and normal samples.

- Adjusted p-value less than 0.001
- Absolute log2 fold change greater than 2

As illustrated in Fig. 11, the significant peaks passing these thresholds were exported as a CSV file, which included log2 fold changes, p-values, and adjusted p-values. These peaks were then used for gene annotation and biological interpretation in subsequent steps.

H. Peak Annotation

The significant peaks identified by DESeq2 (Fig. 12) were annotated with their nearest genes using Bedtools closest. First, the significant peaks were converted to BED format and sorted. A chromosome prefix (”chr”) was added to the peak coordinates to match the format of the GENCODE v43 annotation file. The GTF file was also sorted and filtered to include only gene entries.

Each annotated peak included:

1) The nearest gene.
2) Distance to the gene’s transcription start site.
3) Gene annotation metadata.

bedtools closest -a $chr_peaks -b $genes_gtf -D a > peaks_with_nearest_genes.csv

The output from Bedtools was merged with DESeq2 results to produce a final annotated peak table(Fig. 13), incorporating both statistical significance and gene-level context. This comprehensive table formed the basis for functional enrichment analysis and biological interpretation.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
peak_1	547.504494577151	-1.57446859310051	0.134525858650939	-11.7038360423022	1.21819365291332E-31	1.10767465031798E-30
peak_2	130.30033485863	-1.77845147095745	0.218496533003003	-8.13949515131671	3.96929496164922E-16	2.12204921660192E-15
peak_3	471.814725316068	-1.9720795760457	0.139248930861925	-14.1622600894592	1.56866759572334E-45	1.96982868187067E-44
peak_4	22.029120054417	-2.42646549318689	0.546699464773272	-4.43839010194239	9.06342411109214E-06	2.39472989042493E-05
peak_5	54.7093291077234	1.07176307086977	0.356316549457292	3.00789585132148	0.00263063274275752	0.00499494409886966

Fig. 11. Differential Peak Analysis

chr1	96410	97120	peak_5	chr1	HAVANA	gene	89295	133723	.	-	.	gene_id "ENSG00000238009.6"; gene_type "lncRNA"; gene_name "ENSG00000238009"; level 2; tag "overlapping_locus"; havana_gene "OTTHUMG00000001096.2";	0
chr1	180643	181049	peak_6	chr1	HAVANA	gene	182696	184174	.	+	.	gene_id "ENSG00000279928.2"; gene_type "unprocessed_pseudogene"; gene_name "DDX11L17"; level 2; hgnc_id "HGNC:55080"; havana_gene "OTTHUMG00000191962.1";	1647
chr1	281406	281709	peak_13	chr1	HAVANA	gene	257864	359681	.	-	.	gene_id "ENSG00000228463.10"; gene_type "transcribed_processed_pseudogene"; gene_name "ENSG00000228463"; level 1; tag "pseudo_consens"; havana_gene "OTTHUMG00000002552.3";	0
chr1	355351	355600	peak_14	chr1	HAVANA	gene	257864	359681	.	-	.	gene_id "ENSG00000228463.10"; gene_type "transcribed_processed_pseudogene"; gene_name "ENSG00000228463"; level 1; tag "pseudo_consens"; havana_gene "OTTHUMG00000002552.3";	0
chr1	440980	441425	peak_20	chr1	HAVANA	gene	365389	522928	.	-	.	gene_id "ENSG00000290385.1"; gene_type "lncRNA"; gene_name "ENSG00000290385"; level 2; tag "ncRNA_host"; tag "overlaps_pseudogene";	0

Fig. 12. Annotation of significant peaks with their nearest genes.

PeakID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Gene	Distance
peak_4	202.156030422698	-4.63301654280971	0.359425996420409	-12.890042982285	5.12152590779651E-38	3.79332895188839E-37	ENSG00000238009	0
peak_5	50.3261244303083	-4.95980247400742	0.775737315233975	-6.39366236044925	1.61958849318504E-10	5.67419010825061E-10	ENSG00000238009	0
peak_6	36.9988465114526	-6.54582617055973	1.47624089503511	-4.43411789537508	9.24500279781661E-06	2.27120801936925E-05	DDX11L17	1647
peak_13	53.1700674631786	-4.43439676382513	0.651893513403121	-6.80233300784965	1.02938308155018E-11	3.83760891734195E-11	ENSG00000228463	0
peak_14	234.137966743584	-4.77980985468361	0.34778733477988	-13.7434845282932	5.57277611078798E-43	4.5571200144579E-42	ENSG00000228463	0
peak_20	151.7380834379	-5.76406884145229	0.57349549341418	-10.0507657124508	9.11563360453027E-24	4.99514720700644E-23	ENSG00000290385	0
peak_21	38.0317960295435	-4.54143289804058	0.798878418532605	-5.68476102581713	1.30995703770729E-08	4.10939368261446E-08	ENSG00000290385	0
peak_22	16.3504451539825	-6.82612929615664	1.52736843679341	-4.46920934839244	7.8509261882401E-06	1.94451257326241E-05	ENSG00000230021	0
peak_29	42.0052749921199	-5.11922621768559	0.889971605721996	-5.75212308434557	8.81295953797999E-09	2.79883942095251E-08	ENSG00000230021	0
peak_29	42.0052749921199	-5.11922621768559	0.889971605721996	-5.75212308434557	8.81295953797999E-09	2.79883942095251E-08	MTCO1P12	0
peak_29	42.0052749921199	-5.11922621768559	0.889971605721996	-5.75212308434557	8.81295953797999E-09	2.79883942095251E-08	MTCO2P12	0
peak_29	42.0052749921199	-5.11922621768559	0.889971605721996	-5.75212308434557	8.81295953797999E-09	2.79883942095251E-08	ENSG00000278791	0
peak_31	32.7465990574177	-6.36879260426363	1.48119503724312	-4.29976636710691	1.70978261685472E-05	4.07591343052237E-05	MTCO3P12	0
peak_31	32.7465990574177	-6.36879260426363	1.48119503724312	-4.29976636710691	1.70978261685472E-05	4.07591343052237E-05	ENSG00000230021	0
peak_31	32.7465990574177	-6.36879260426363	1.48119503724312	-4.29976636710691	1.70978261685472E-05	4.07591343052237E-05	MTATP6P1	0

Fig. 13. Significant Peaks with their nearest genes.

I. Visualization

To better interpret the chromatin accessibility differences between leukemia and healthy samples, we employed three main visualization strategies:

i. Volcano Plot

A volcano plot was generated using the DESeq2 results to highlight significantly differentially accessible peaks. Peaks were plotted by their log2 fold change (x-axis) and the -log10 adjusted p-value (y-axis). Peaks with adjusted p-value < 0.0001 and $|\log2FC| > 2$ were marked as significant and colored red. Gene names for the top 100 most significant peaks were overlaid using ggrepel.

This plot helped us visually identify the most biologically relevant regions with strong statistical support, such as SMYD3, KIF26B, and EFCAB2, which showed notable accessibility changes.

ii. Heatmap of Top Variable Peaks

We created a heatmap of the top 100 most variable peaks (based on variance across samples) after variance stabilizing transformation (VST) using DESeq2. The heatmap clearly clustered leukemia and healthy samples into separate groups, demonstrating consistent patterns of chromatin accessibility across conditions. Peaks were scaled row-wise to enhance visualization.

iii. Principal Component Analysis (PCA)

PCA was performed on VST-transformed data to reduce dimensionality and visualize sample-level variation. The first two principal components explained 98% and 2% of variance, respectively. The PCA plot showed a clear separation between leukemia and healthy samples, reinforcing the robustness of the differential accessibility signals.

iv. IGV Genome Browser Snapshots

To enable visual inspection of chromatin accessibility at specific genomic regions, we generated genome browser snapshots using Integrative Genomics Viewer (IGV). An automated script loaded the narrowPeak files, configured the hg38 genome, and navigated to the selected peak region to capture a high-resolution snapshot.

J. Functional Enrichment Analysis

Following peak annotation, we performed functional enrichment analysis to identify biological processes and pathways associated with genes near differentially accessible regions. We used the g:Profiler2 R package to perform overrepresentation analysis.

Enrichment was conducted using three well-curated databases:

Gene Ontology – Biological Process (GO:BP)

Kyoto Encyclopedia of Genes and Genomes (KEGG)

Reactome (REAC)

i. Overview Genes associated with significantly accessible peaks were extracted and filtered to exclude missing or uninformative entries. A total of **N** annotated genes (insert actual number here) were submitted for enrichment analysis. The top 10 enriched terms were visualized using both dot plots and bar charts. Among the most significantly enriched terms were:

- *GO: Regulation of transcription by RNA polymerase II*
- *KEGG: Pathways in cancer*
- *Reactome: Chromatin organization*

ii. Visualization The following plots were programmatically generated as part of the Nextflow workflow:

- **General Dot Plot:** Displays the top 10 enriched pathways across all databases, based on $-\log_{10}$(p-value) and gene intersection size.
- **Source-Specific Dot Plots:** Separate plots for KEGG and Reactome pathways to highlight context-specific enrichment.
- **Bar Plots:** Top 10 pathways overall, and grouped by source (GO, KEGG, REAC), using $-\log_{10}$(p-value) as the significance measure.

IV. Results

This section presents the key outcomes of the ATAC-seq analysis pipeline applied to leukemic (K562) and normal (GM12878) human cell lines. The results are organized to reflect each major stage of the workflow, including quality control, peak identification, differential accessibility analysis, annotation, visualization, and enrichment findings.

A. Genome-wide Peak Coverage Snapshot

To visualize the distribution of accessible chromatin regions, we plotted MACS2-called peak intensities across the entire genome for all four samples (Figure 14). Each track represents a sample, showing the normalized peak signal across chromosomes 1 to Y.

As expected, we observed widespread enrichment in promoter and transcription start site regions, with distinct patterns between healthy (GM12878) and leukemic (K562) cells. Leukemic samples generally exhibited more intense and frequent peaks, indicating globally elevated chromatin accessibility.

Notably, certain chromosomal regions showed consistently stronger signals in leukemia tracks, suggesting potential hotspots of regulatory activity specific to cancerous states. These patterns align with known oncogenic loci and highlight the capacity of ATAC-seq to capture epigenomic dysregulation at a genome-wide scale.

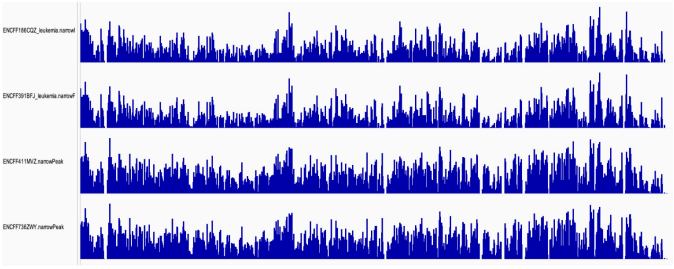

Fig. 14. Genome-wide chromatin accessibility. Each track corresponds to a sample; peak heights represent signal intensity.

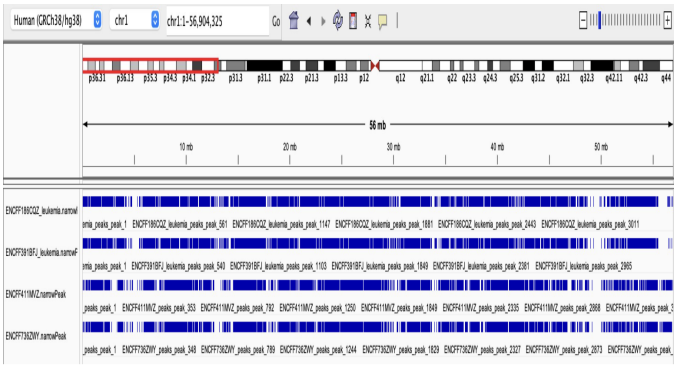

Fig. 15. Genome browser view of MACS2-called peaks across chromosome 1

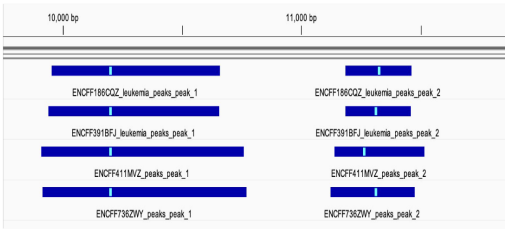

Fig. 16. IGV snapshot showing ATAC-seq peak regions for leukemia (K562) and normal (GM12878) samples.

B. Differential Accessibility Analysis

To identify regions with significant changes in chromatin accessibility between leukemia and healthy samples, we used DESeq2 on the peak count matrix derived from featureCounts. Differentially accessible peaks showing increased accessibility in K562 (leukemia) and in GM12878 (healthy) were identified.

To visualize the statistical significance and effect size of these differences, a volcano plot was generated (Figure 17). The x-axis represents the log2 fold change in accessibility between conditions, while the y-axis shows the $-$log10 of the adjusted p-value. Peaks with adjusted p-value < 0.0001 and absolute $\log_2$ fold change > 2 were highlighted in red as significantly differentially accessible.

Notably, several peaks with extreme fold changes and strong statistical support were annotated with gene names. These included both known leukemia-associated genes and novel candidates, providing insights into potential regulatory drivers of leukemic transformation.

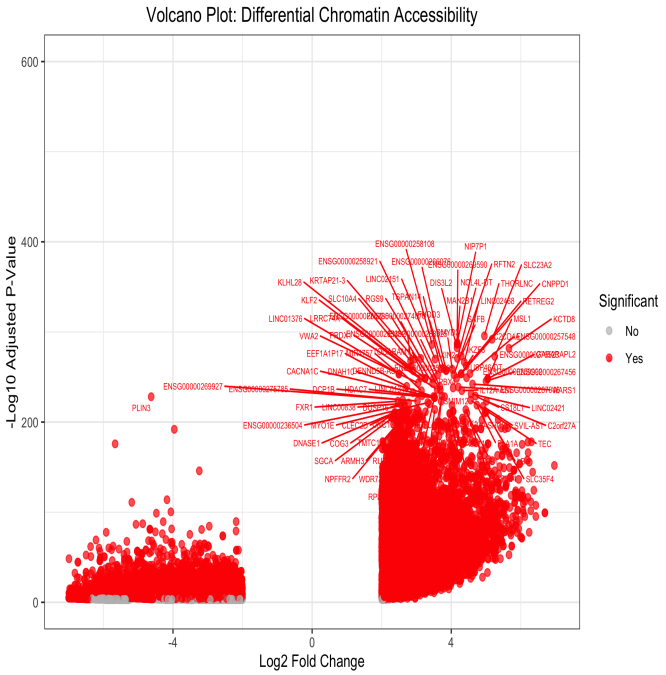

Fig. 17. Volcano plot of differentially accessible peaks p-value < 0.0001 and $|\log_2 \mathrm{FC}| > 2$.

C. Clustering of Variable Peaks via Heatmap

To visualize global patterns in chromatin accessibility, we created a heatmap using the top 100 most variable peaks across all samples. These peaks were selected after applying variance-stabilizing transformation (VST) to the read count matrix, ensuring consistent scaling across samples.

The heatmap (Fig. 18) revealed clear clustering of samples by biological condition. Leukemic samples (K562) and healthy samples (GM12878) formed distinct clusters, reflecting consistent differences in chromatin accessibility profiles.

Each row represents a genomic region (peak), and the values are scaled per row to highlight relative differences. Many peaks show strong accessibility in leukemia but not in healthy cells, and vice versa. This clustering pattern confirms the biological relevance of our differential analysis and suggests that some of these peak regions could play a role in leukemia-specific regulation.

Overall, this heatmap provides a clear visual summary of how chromatin accessibility differs between leukemic and normal cells, reinforcing the robustness of our analysis pipeline.

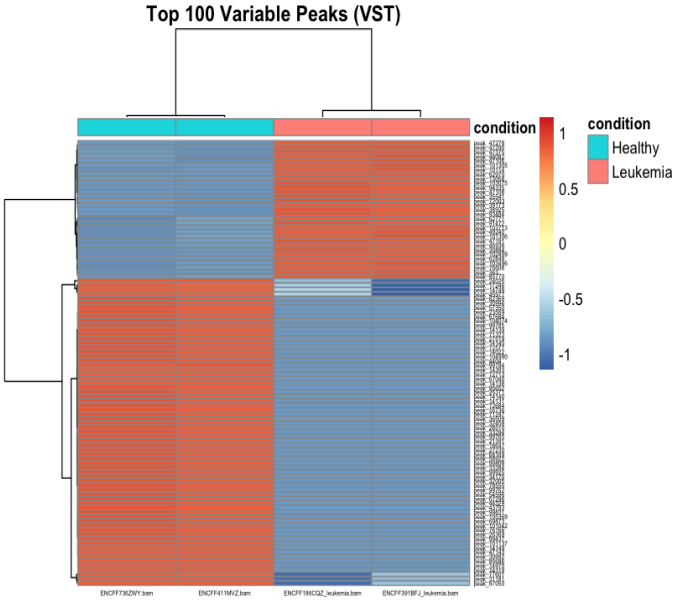

Fig. 18. Heatmap of the top 100 most variable peaks across samples.

D. Principal Component Analysis (PCA)

To evaluate sample-level variation and assess global chromatin accessibility trends, we performed Principal Component Analysis (PCA) on the variance-stabilized (VST) transformed count matrix. The PCA plot (Figure 19) visualizes samples in two dimensions defined by the first two principal components.

PC1 accounted for the majority of the variance, clearly separating leukemia (K562) and healthy (GM12878) samples into two distinct clusters. This separation reflects condition-specific chromatin accessibility profiles and validates the biological relevance of the identified differences.

Importantly, the compact clustering within each group indicates high internal consistency and low technical noise, reinforcing the quality of the data and robustness of the pipeline.

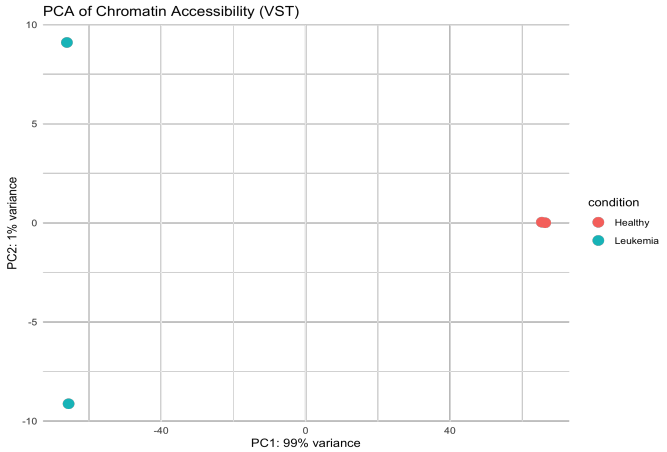

Fig. 19. PCA Plot

E. Top Enriched Pathways - Bar Plot

To understand the biological significance of the differentially accessible regions, we performed functional enrichment analysis on the associated genes. The top 10 enriched pathways are presented in Figure 20, ranked by statistical significance using $-\log_{10}$(adjusted p-value).

The analysis revealed strong enrichment for key biological processes, particularly those related to gene regulation and development. Notable terms included:

- *Positive regulation of biological process*
- *Biological regulation*
- *Developmental process*
- *Response to stimulus*

These results suggest that chromatin accessibility changes in leukemia may be driving widespread shifts in gene expression programs involved in cell differentiation, development, and immune response. The consistent appearance of regulatory terms across the top hits supports the role of epigenetic mechanisms in leukemic transformation.

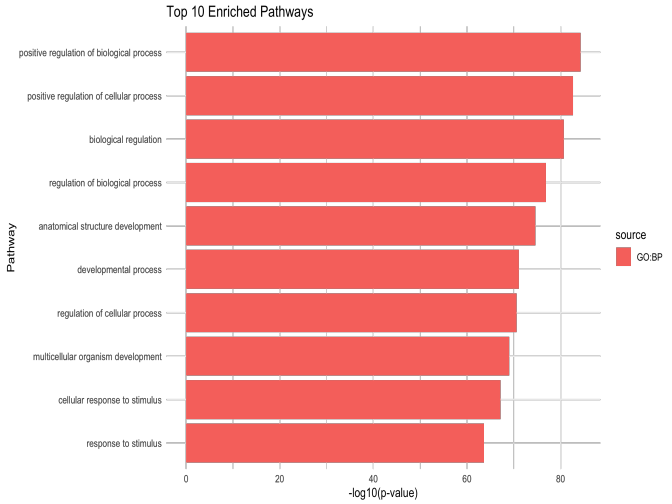

Fig. 20. Top 10 enriched pathways associated with DARs. Bar length indicates pathway significance ($-\log_{10}$(adj p-value)).

F. Top Enriched Pathways - Dot Plot

To complement the bar plot view, we generated a dot plot highlighting the top 10 enriched biological terms based on chromatin accessibility differences. In this visualization (Figure 21), each dot represents a pathway term, with its size proportional to the number of genes associated with that term.

Larger dots indicate pathways with broader gene overlap, while the horizontal axis reflects statistical significance. Pathways like *positive regulation of biological process* and *biological regulation* were not only highly significant but also involved a greater number of genes.

This view adds a layer of interpretation by showcasing both the biological relevance (gene count) and statistical strength of each enriched term, helping prioritize which pathways may be most functionally impacted in leukemia.

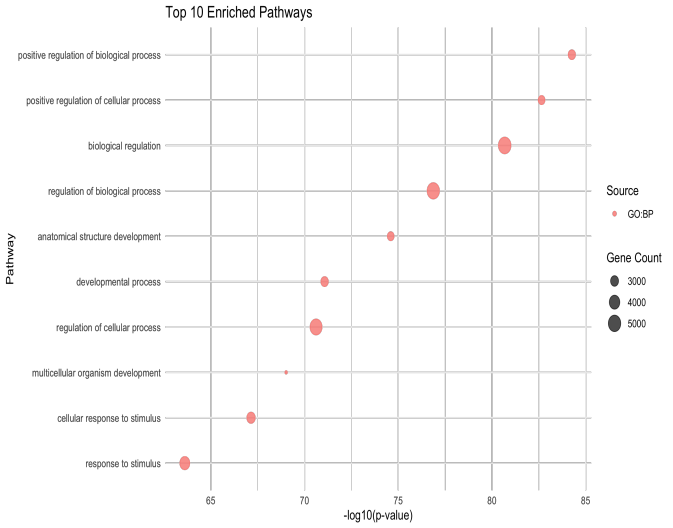

Fig. 21. Dot plot of the top 10 enriched pathways. Dot size reflects the number of overlapping genes, and position indicates $-\log_{10}$(p-value).

G. Top GO Biological Processes

The bar plot in Figure 22 highlights the top enriched biological processes based on Gene Ontology (GO:BP) analysis. The most significant terms include positive regulation of biological process, developmental process, and cellular response to stimulus.

These enriched GO terms indicate that differentially accessible chromatin regions are closely linked to gene regulatory programs controlling development, cell differentiation, and responses to environmental signals. Such functions are highly relevant in the context of leukemia, where abnormal transcriptional regulation and disrupted differentiation pathways are hallmarks of disease progression.

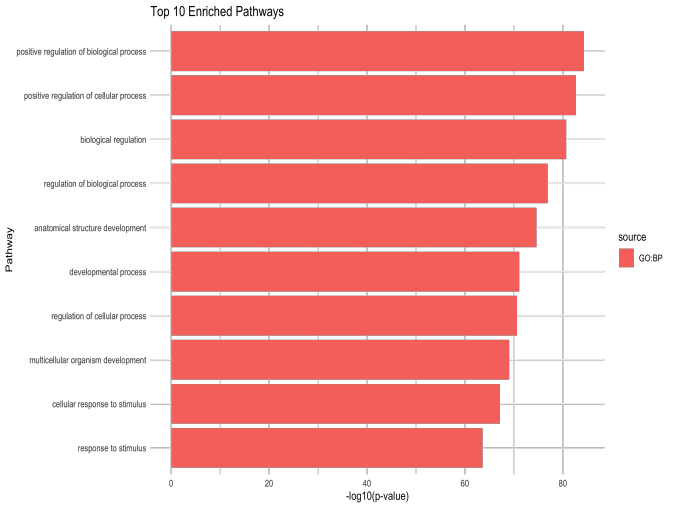

Fig. 22. Top enriched GO biological processes. Bar length indicates $-\log_{10}$(adj p-value) for each term.

H. KEGG Pathway Enrichment

Figure 23 highlights the top 10 KEGG pathways enriched among genes associated with differentially accessible chromatin regions. Each dot represents a pathway, where its position on the x-axis reflects statistical significance (as $-\log_{10}$(p-value)), and the dot size corresponds to the number of overlapping genes.

The most significantly enriched pathway was Pathways in cancer, reflecting the general disruption of oncogenic signaling networks in leukemic cells. Additional enriched pathways included PD-L1 expression and PD-1 checkpoint pathway in cancer, which is central to immune evasion mechanisms, and T cell receptor signaling pathway, suggesting altered immune signaling environments in the leukemia context.

Other notable pathways such as Focal adhesion, Axon guidance, and Platelet activation also surfaced, hinting at possible roles in leukemic cell migration, microenvironment interaction, and cellular communication.

These results reinforce the biological interpretation that chromatin accessibility changes may underpin key immune and cancer-related processes, further validating the epigenomic signatures uncovered in leukemia.

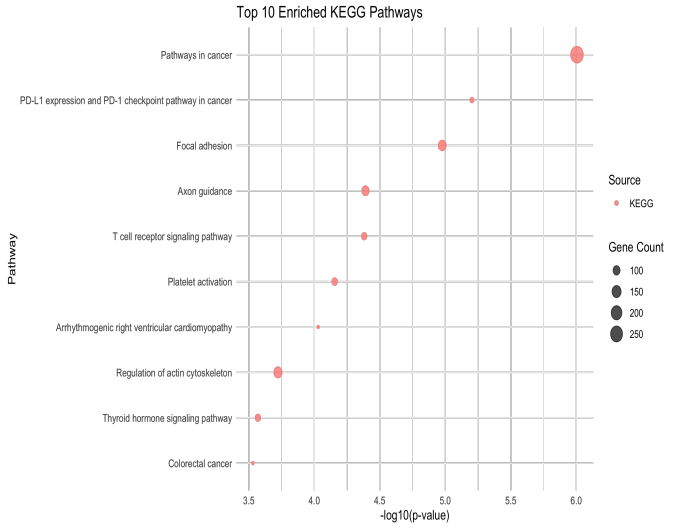

Fig. 23. Dot plot showing top KEGG pathways enriched among genes near differentially accessible peaks. Dot size reflects gene overlap, x-axis shows pathway significance.

I. Reactome Pathway Enrichment

Figure 24 illustrates the top Reactome pathways enriched among genes associated with differentially accessible chromatin regions. Each dot reflects a pathway, with the x-axis showing $-\log_{10}$(p-value) for significance and dot size indicating the number of genes involved.

The most enriched Reactome term was Signal Transduction, suggesting that chromatin accessibility alterations in leukemia may heavily impact cellular signaling cascades. Several small GTPase cycles, including RHO, RAC1, RAC2, and CDC42, were also prominent. These are critical for cell morphology, migration, and intracellular communication functions often dysregulated in cancer.

Other enriched terms included TGF-beta receptor signaling and its downstream complexes, pointing toward potential epigenetic influence on tumor suppressor pathways and cell fate decisions.

Overall, this enrichment highlights how leukemia-associated chromatin remodeling may influence complex signaling networks, reinforcing the idea that accessibility changes are not random but tied to functional cellular outcomes.

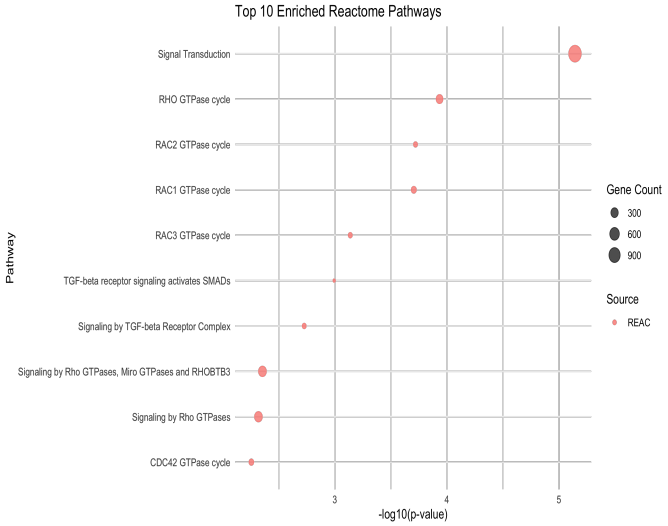

Fig. 24. Reactome-enriched pathways among genes near DARs. Size represents gene overlap; position reflects statistical significance.

V. Limitations

While this study offers valuable insight into chromatin accessibility differences between leukemic and normal human cell lines, several limitations should be considered:

- **Limited sample diversity:** The analysis was conducted using only one leukemia cell line (K562) and one normal control (GM12878), each with two replicates. While these are widely used ENCODE reference lines, they may not capture the full heterogeneity seen across leukemia subtypes or patient-derived samples.
- **Single-end sequencing:** All samples used single-end ATAC-seq data, which lacks the fragment size information provided by paired-end sequencing. This restricts our ability to distinguish between nucleosome-free regions and nucleosome-bound DNA, limiting insights into chromatin architecture.
- **Lack of experimental validation:** Although computational results were statistically robust and supported by functional enrichment, no wet-lab validation was performed to confirm the biological impact of the identified peaks or genes.
- **Annotation limitations:** Peak annotation relied on proximity to the nearest gene using GENCODE GTF data. This approach does not account for long-range regulatory interactions such as enhancer–promoter looping.

Despite these constraints, the study successfully demonstrates the utility of ATAC-seq for uncovering leukemia-related epigenetic alterations.

VI. Future Directions

This project lays the groundwork for further exploration into chromatin accessibility in leukemia. Future extensions could include:

- Incorporate additional leukemia subtypes and patient-derived ATAC-seq samples to improve biological generalizability.
- Integrate matched RNA-seq data to directly link changes in chromatin accessibility with gene expression outcomes.
- Utilize paired-end ATAC-seq data or single-cell ATAC-seq to better capture chromatin architecture and cell-to-cell heterogeneity.
- Perform experimental follow-ups to validate functional roles of key differentially accessible regions.
- Extend the pipeline with additional modules such as transcription factor footprinting, motif enrichment, or 3D genome structure analysis (Hi-C).

VII. Conclusion

This project presents a comprehensive and fully automated ATAC-seq analysis pipeline to investigate chromatin accessibility differences between leukemic (K562) and normal (GM12878) human cell lines. The pipeline, implemented using Nextflow DSL2, integrates a range of trusted bioinformatics tools from preprocessing and peak calling to statistical testing and gene/pathway annotation within a reproducible and modular framework.

Through differential peak analysis using DESeq2, we identified hundreds of genomic regions with significant accessibility changes between leukemia and control samples. These peaks were annotated to genes involved in hematopoiesis, chromatin remodeling, and cancer signaling. Functional enrichment analysis further revealed pathways related to transcription regulation, immune response, and oncogenic signaling like KEGG cancer pathways, Reactome chromatin organization.

The visualization components including volcano plots, heatmaps, PCA, and enrichment plots provided intuitive and biologically meaningful insights into the data. All results were programmatically generated and neatly organized for easy inspection and reuse.

Ultimately, this study not only enhances our understanding of leukemia-associated chromatin dynamics but also delivers a flexible and scalable ATAC-seq pipeline that can be extended to other disease contexts or omics datasets. This lays the groundwork for integrative epigenomics studies and potential discovery of novel regulatory biomarkers in leukemia and beyond.

References

[1] F. C. Grandi, H. Modi, L. Kampman, and M. R. Corces, "Chromatin accessibility profiling by ATAC-seq," *Nature Protocols*, vol. 17, no. 9, pp. 1518–1552, 2022. doi: 10.1038/s41596-022-00764-4.

[2] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nature Methods*, vol. 10, no. 12, pp. 1213–1218, 2013. doi: 10.1038/nmeth.2688.

[3] M. R. Corces, et al., "The chromatin accessibility landscape of primary human cancers," *Science*, vol. 362, no. 6413, eaav1898, 2018. doi: 10.1126/science.aav1898.

[4] F. Yan, D. R. Powell, D. J. Curtis, and N. C. Wong, "From reads to insight: A hitchhiker's guide to ATAC-seq data analysis," *Genome Biology*, vol. 21, no. 1, p. 22, 2020. doi: 10.1186/s13059-020-1929-3.

[5] Center for Research Computing, University of Pittsburgh, "ATACSeq Data Analysis – CRCD User Manual," [Online]. Available: https://crc-pages.pitt.edu/user-manual/advanced-genomics-support/ATACSeq-data-analysis/.

[6] M. R. Corces, J. D. Buenrostro, B. P. Wu, et al., "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution," *Science*, vol. 351, no. 6280, pp. 1394–1399, 2016. doi: 10.1126/science.aab1601.

[7] S. L. Klemm, Z. Shipony, and W. J. Greenleaf, "Chromatin accessibility and the regulatory epigenome," *Nature Reviews Genetics*, vol. 20, no. 4, pp. 207–220, 2019. doi: 10.1038/s41576-018-0081-3.

[8] A. T. Satpathy, J. M. Granja, K. E. Yost, Y. Qi, F. Meschi, G. P. McDermott, B. N. Olsen, M. R. Mumbach, S. E. Pierce, M. R. Corces, et al., "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion," *Nature Biotechnology*, vol. 37, pp. 925–936, 2019. doi: 10.1038/s41587-019-0206-z.

[9] D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure, "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing," *Science*, vol. 348, no. 6237, pp. 910–914, 2015. doi: 10.1126/science.aab1601.