ATAC Seq Project Report

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a powerful method used to assess the open chromatin regions across the genome, providing insights into regulatory elements that control gene expression. This project involves the analysis of two ATAC-seq replicates from a human source.

Methodology

Quality Control and Preparation

FastQC v0.12.1 was used to assess the initial quality of the raw sequencing reads, providing insights into potential issues such as contamination or poor quality sequences. MultiQC v1.20 compiled these FastQC reports into a single document, facilitating an overview of quality across all samples to identify any outliers or batch effects.

Read Trimming

Trimmomatic v0.39 trimmed adapters and low-quality sequences from the raw reads, ensuring that only high-quality data was used for alignment and subsequent analyses.

Alignment

The trimmed reads were aligned to the human reference genome (GRCh38) using Bowtie2 v2.5.3, which efficiently handled large datasets to produce BAM files storing aligned read information.

Sorting and Indexing BAM Files

Samtools v1.19.2 was employed to sort the BAM files post-alignment and create index files, facilitating efficient data retrieval and downstream processing.

Shift Reads for ATAC-seq

DeepTools' alignmentSieve v3.5.4 adjusted ATAC-seq read alignments to correct for the Tn5 transposase binding offset, which was crucial for accurate peak calling.

Filtering Mitochondrial DNA

Mitochondrial DNA was excluded using Samtools v1.19.2, which helped to focus analyses on nuclear DNA and improve the signal-to-noise ratio for peak detection.

Peak Calling

MACS3 v3.0.1 identified regions of significant enrichment over the background, indicating open chromatin that was accessible to transcription factors and other regulatory proteins.

Creating Reproducible Peaks
Peaks from different conditions were intersected using BEDTools v2.31.1, ensuring reproducibility and enhancing the reliability of the detected peaks.

Filtering Against Blacklist Regions
Artifactual regions commonly identified in sequencing assays were excluded using BEDTools v2.31.1, which enhanced the specificity of peak detection.

Compute Matrix for Visualization
DeepTools' computeMatrix organized data into a matrix format suitable for visual representation, around the reference point to facilitate comparative analysis.

Generating Coverage Plots
DeepTools' plotProfile visualized signal intensities across regions of interest, elucidating the distribution of open chromatin.

Annotation of Peaks and Chromatin Accessibility
HOMER v4.11 annotated detected peaks with nearby genomic features, providing insights into the regulatory landscapes and potential functions of these elements.

Motif Analysis
HOMER also performed motif analysis to predict transcription factor binding within the peaks, offering clues to the regulatory mechanisms at play.

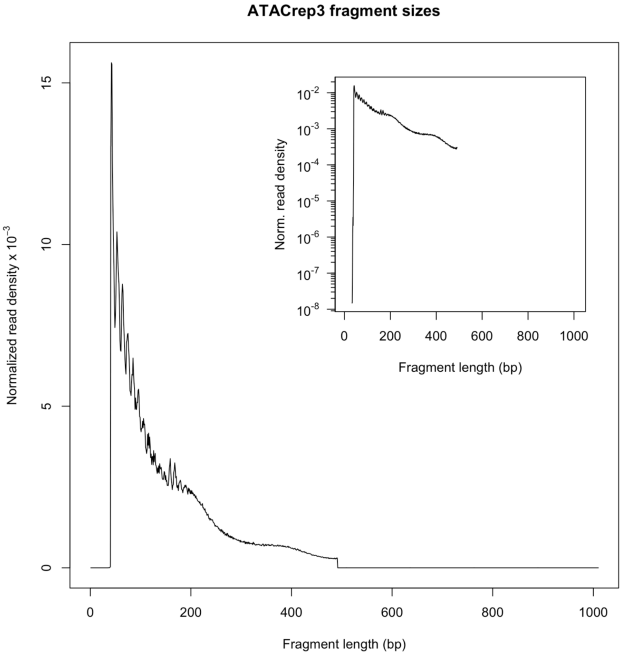
Final Gene Annotation
Peaks were further annotated with gene interactions and potential regulatory roles using HOMER v4.11, enhancing the understanding of how chromatin accessibility impacts gene regulation and expression dynamics.

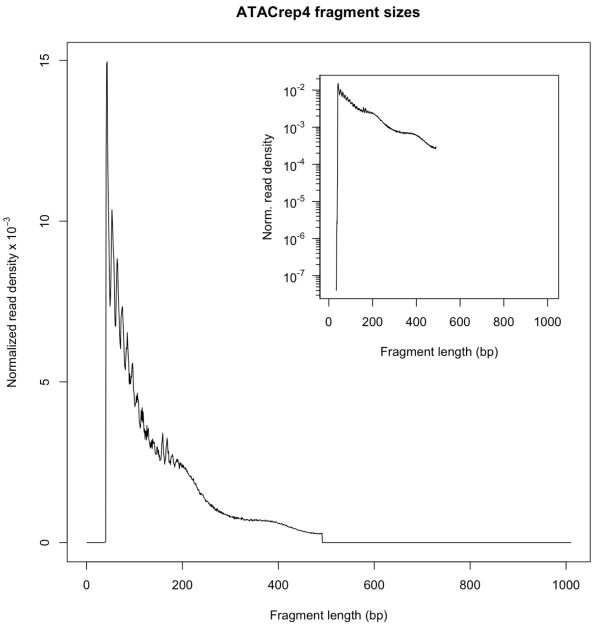
Visualization and Final Analysis
The peaks were visualized with respect to their genomic context by creating coverage plots and signal intensity profiles around transcription start sites (TSS) using DeepTools. This visualization aided in understanding the distribution and characteristics of open chromatin across different genomic regions.

Integration and Interpretation
Data integration involved annotating peaks with additional gene and regulatory information, enriching the dataset with functional insights was performed using DAVID that facilitated a deeper understanding of their implications in cellular processes.

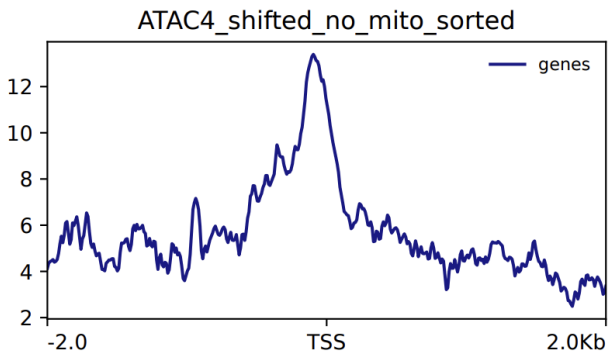
Deliverables

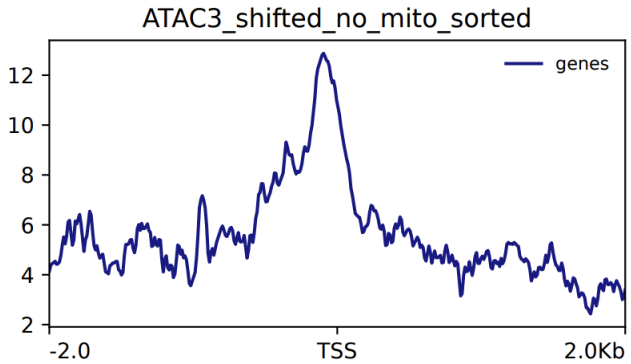
Fragment length distribution plots





TSS Plots (I'm actually not sure if this is right, I did use nbr and nfr bam files)





	ATACrep3	ATACrep4
Total Alignments	34933513	25485688
Alignments after Mitochondrial removal	5443994	4010047
Peaks	42689	41299
Reproducible Peaks	31225	
Filtered Peaks	30484	

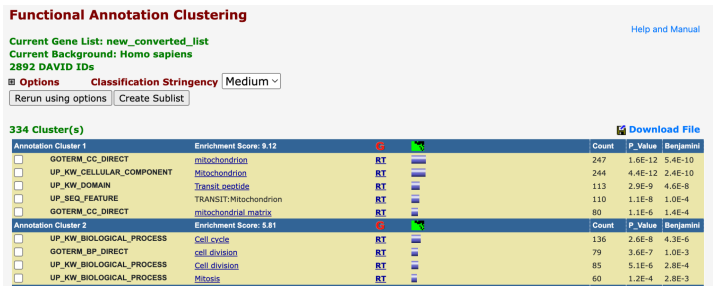
Homer Motifs

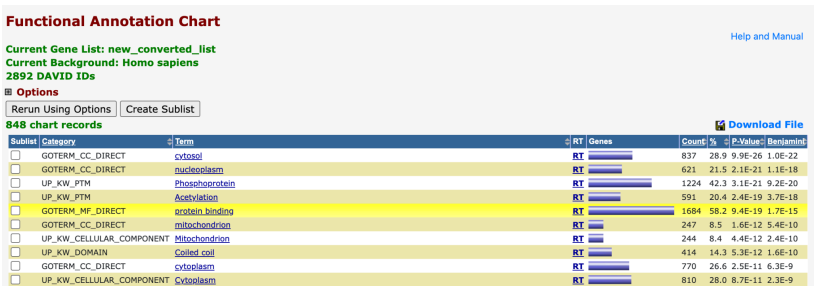
Total Target Sequences = 30360, Total Background Sequences = 29666

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	SVG
1		IRF8(IRF)/BMDM-IRF8-ChIP-Seq(GSE77884)/Homer	1e-1301	-2.996e+03	0.0000	4129.0	13.60%	949.3	3.20%	motif file (matrix)	svg
2		PU.1:IRF8(ETS:IRF)/pDC-Irf8-ChIP-Seq(GSE66899)/Homer	1e-1153	-2.657e+03	0.0000	2815.0	9.27%	486.2	1.64%	motif file (matrix)	svg
3		CTCF(Zf)/CD4+-CTCF-ChIP-Seq(Barski_et_al.)/Homer	1e-1142	-2.632e+03	0.0000	2178.0	7.17%	274.9	0.93%	motif file (matrix)	svg
4		IRF3(IRF)/BMDM-Irf3-ChIP-Seq(GSE67343)/Homer	1e-982	-2.263e+03	0.0000	3413.0	11.24%	842.8	2.84%	motif file (matrix)	svg
5		Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	1e-827	-1.905e+03	0.0000	7363.0	24.25%	3421.2	11.54%	motif file (matrix)	svg
6		PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-796	-1.833e+03	0.0000	3682.0	12.13%	1139.7	3.84%	motif file (matrix)	svg
7		IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer	1e-793	-1.827e+03	0.0000	2093.0	6.89%	390.5	1.32%	motif file (matrix)	svg
8		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer	1e-779	-1.794e+03	0.0000	6535.0	21.53%	2931.3	9.89%	motif file (matrix)	svg
9		IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer	1e-777	-1.790e+03	0.0000	1818.0	5.99%	295.4	1.00%	motif file (matrix)	svg
10		SpiB(ETS)/OCILY3-SPIB-ChIP-Seq(GSE56857)/Homer	1e-722	-1.663e+03	0.0000	2463.0	8.11%	592.0	2.00%	motif file (matrix)	svg
11		PU.1-IRF(ETS:IRF)/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer	1e-693	-1.596e+03	0.0000	7334.0	24.16%	3659.8	12.35%	motif file (matrix)	svg
12		BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-681	-1.570e+03	0.0000	3751.0	12.36%	1302.7	4.39%	motif file (matrix)	svg
13		ETV1(ETS)/GIST48-ETV1-ChIP-Seq(GSE22441)/Homer	1e-673	-1.551e+03	0.0000	7775.0	25.61%	4026.9	13.58%	motif file (matrix)	svg
14		Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	1e-666	-1.534e+03	0.0000	3458.0	11.39%	1153.9	3.89%	motif file (matrix)	svg
15		BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer	1e-665	-1.533e+03	0.0000	2693.0	8.87%	746.2	2.52%	motif file (matrix)	svg

Gene Enrichment - DAVID

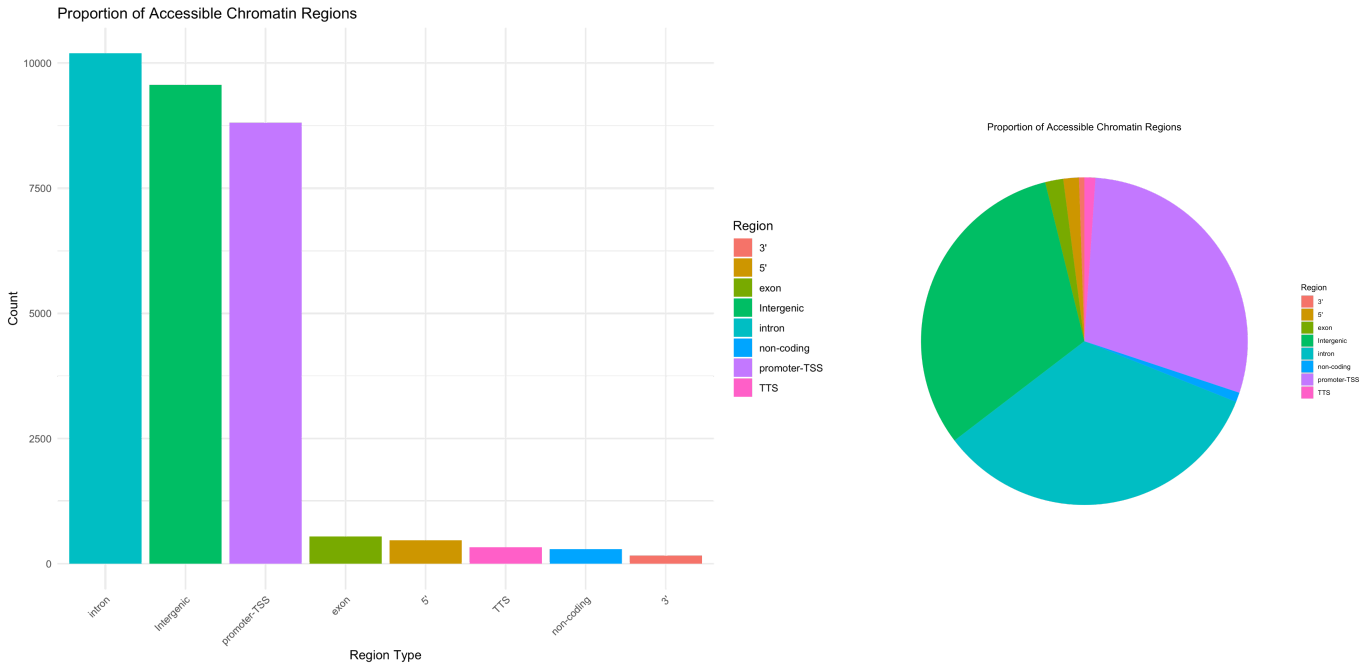
The gene list seems to be particularly enriched in mitochondrial components, as indicated by the clustering and annotation data. This enrichment suggests a strong association of these genes with mitochondrial functions such as energy production and metabolic processes. The prominence of terms related to the mitochondrion, alongside high enrichment scores for mitochondrial transport and cellular components, underscores the potential regulatory roles these genes may play in mitochondrial dynamics and function.





Additionally, the analysis points to significant involvement of the genes in cell cycle regulation and DNA repair mechanisms, highlighting their importance in maintaining genomic stability. The presence of multiple genes related to the cell cycle, DNA replication, and repair processes, as detailed in the functional annotation chart, supports the hypothesis that these genes are crucial for cell division and responding to genomic stress. These findings could guide further research into the molecular mechanisms underlying these processes, potentially identifying new targets for therapeutic intervention in diseases characterized by mitochondrial dysfunction and genomic instability.

Proportions of regions that appear to have accessible chromatin



References

- Yan, F., Powell, D.R., Curtis, D.J. et al. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 21, 22 (2020). https://doi.org/10.1186/s13059-020-1929-3
- Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
- Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. Nucleic Acids Research (2016). doi:10.1093/nar/gkw257.
- Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol (2008) vol. 9 (9) pp. R137
- Aaron R. Quinlan, Ira M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, Volume 26, Issue 6, March 2010, Pages 841–842, https://doi.org/10.1093/bioinformatics/btq033
- Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432
- ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data." BMC Genomics, 19(1), 169. ISSN 1471-2164, doi:10.1186/s12864-018-4559-3, https://doi.org/10.1186/s12864-018-4559-3.