ATACseq_jfaulx

2024-04-27

BF528: Jackson Faulx ATAC-seq Final Project

Methods:

First, QC was performed on the paired-end reads using fastqc (v0.12.1) and the reads were trimmed using trimmomatic (v0.39). The reads were then aligned to the hg38 genome using bowtie2 (v2.5.3) keeping all unique reads that are less than 2kb in size. The reads were then sorted and indexed, and mitochondrial alignments were removed using samtools (v1.19.2) sort and view respectively. The reads were then sorted and indexed again before being shifted using alignmentSieve from deeptools (v3.5.4). Samtools flagstat and multiqc (v1.20) were performed before and after removing mitochondrial alignments. Read size distribution was then visualized using the ATACSeqQC (v1.26.0) package from bioconductor in R. Peak calling was then performed by MACS3 (v3.0.1) with default parameters. Peak intersection and blacklist region filtering was then performed for the samples using bedtools (v2.31.1) intersect. After the intersecting peaks have been found, homer (v4.11) was then used to generate a file of annotated peaks as well as generate a report of motifs. Signal coverage plots were generated for NFR and NBR regions respectively using the homer tools bamCoverage, computeMatrix, and plotProfile.

Workflow and Results

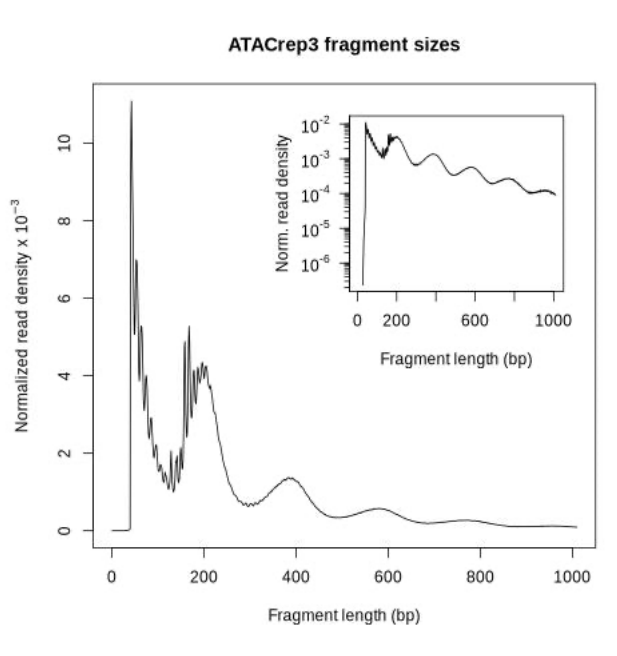
After downloading the fastq files and performing the fastqc check, I glanced over the metrics for each read for each sample and did not see any significantly issues with the reads. Sequence duplication was flagged by fastqc as an issue, however high levels of sequence duplication are expected in ATAC-seq due to potential conserved binding motifs in non-coding regions. I was able to identify the types of adapters used in the sample preparation and accounted for them when performing trimming. Additional quality control was performed following alignment to the reference, which additionally showed that there are no issues with any of the samples, since all of the reads aligned appropriately. 162,679,456 and 121,196,558 reads aligned for samples 3 and 4 respectively. After indexing and sorting the alignments, I found that 25,997,440 and 20,278,186 reads aligned to the mitochondrial genome in each sample. These reads were filtered out before continuing the analysis.

Number of Alignments

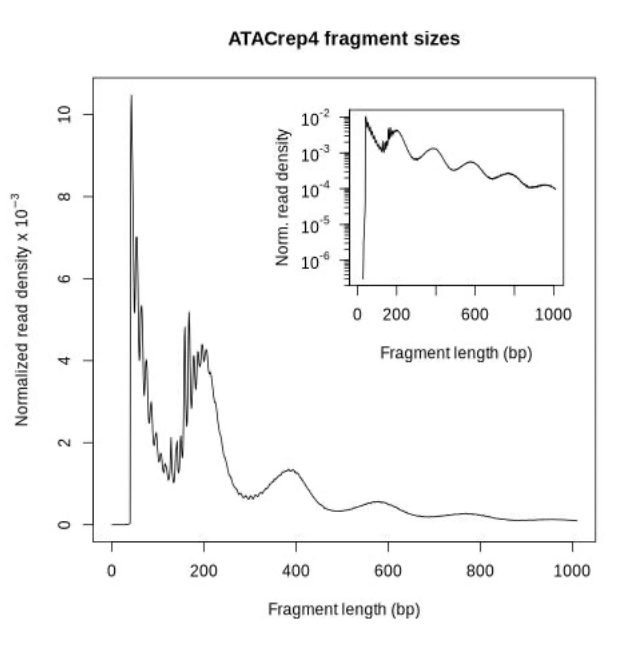
Sample	Pre_Filtering	Post-Filtering	Mitochondrial Alignments
sample 3	162,679,456	25,997,440	136,682,016
sample 4	121,196,558	20,278,186	100,918,372

I then generated a read length distribution plot using ATACseqQC for each sample, which is shown below. There is a large spike at length < 100, corresponding to nucleosome free regions of DNA. The smaller spikes at ~200 and ~400 bp lengths are mononucleosome and dinucleosome bound regions respectively, which is exactly what is expected in an ATAC-seq analysis.

```
knitr::include_graphics("results/ATACrep3_fragplot.jpeg")
```




```
knitr::include_graphics("results/ATACrep4_fragplot.jpeg")
```



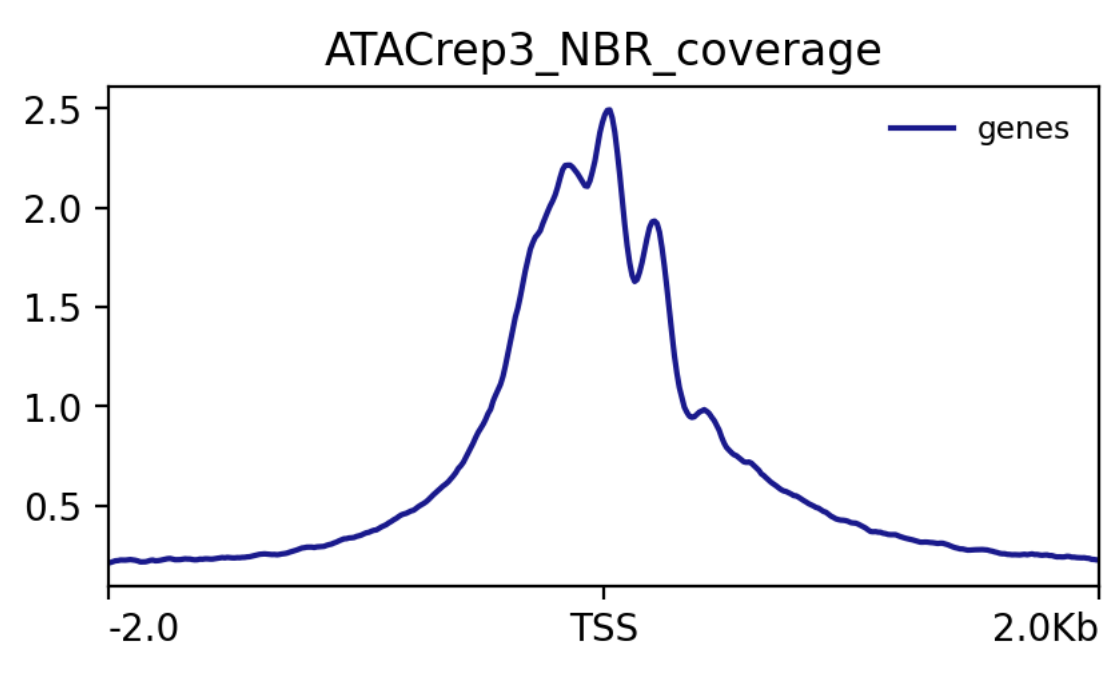
Peak calling was performed on each sample, producing a list of peaks. Sample 3 had 37,332 peaks while sample 4 had 33,931 peaks. In order to find the reproducible peaks, bedtools intersect was performed on the set of peaks files, generating a single file with 22,775 peaks. Peaks falling in blacklist regions were then filtered out, leaving a total of 22,771 reproducible peaks.

Number of Peaks

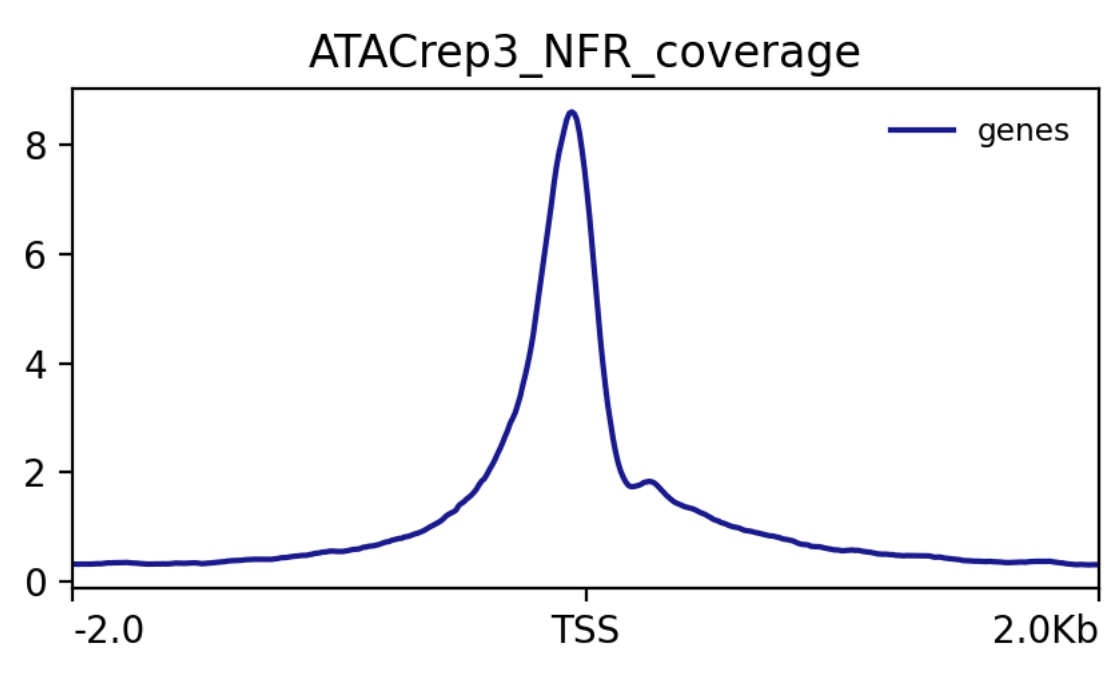
Sample	Number of peaks	Total Reproducible Peaks
sample 3	37,332	22,771
sample 4	33,931	22,771

The final bam files were also used to generate signal coverage plots for the unbound and mononucleosome bound regions in each sample as seen below. Generally, the nucleosome free regions have a higher more consistent peak at the TSS compared to the nucleosome bound regions. This is because generally the nucleosome free regions are ~50 bp long, while the bound region fragment length depends on what protein is bound in that region, which explains the wider peak in the NBR graphs.

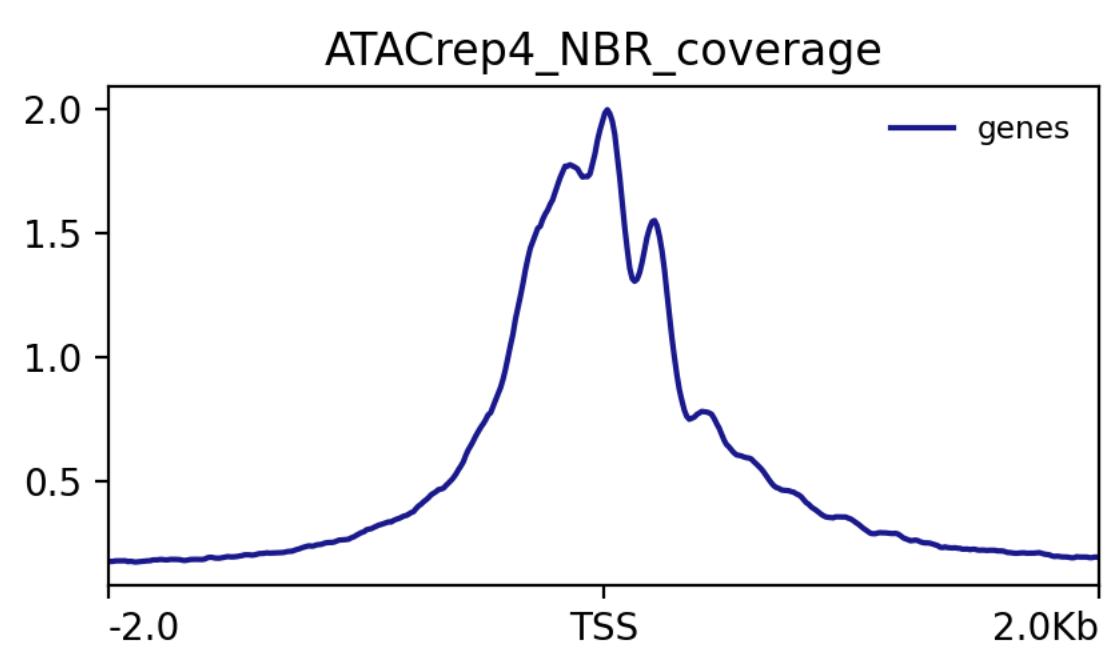
```
knitr::include_graphics("results/matrixplot_ATACrep3_NBR.png")
```



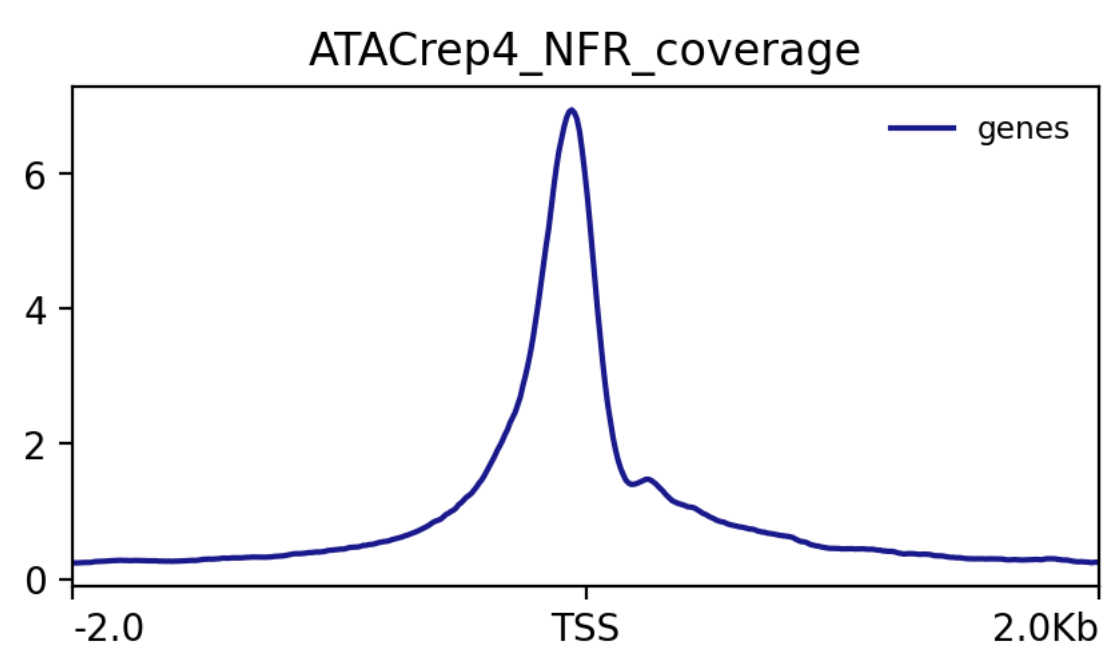

```
knitr::include_graphics("results/matrixplot_ATACrep3_NFR.png")
```




```
knitr::include_graphics("results/matrixplot_ATACrep4_NBR.png")
```

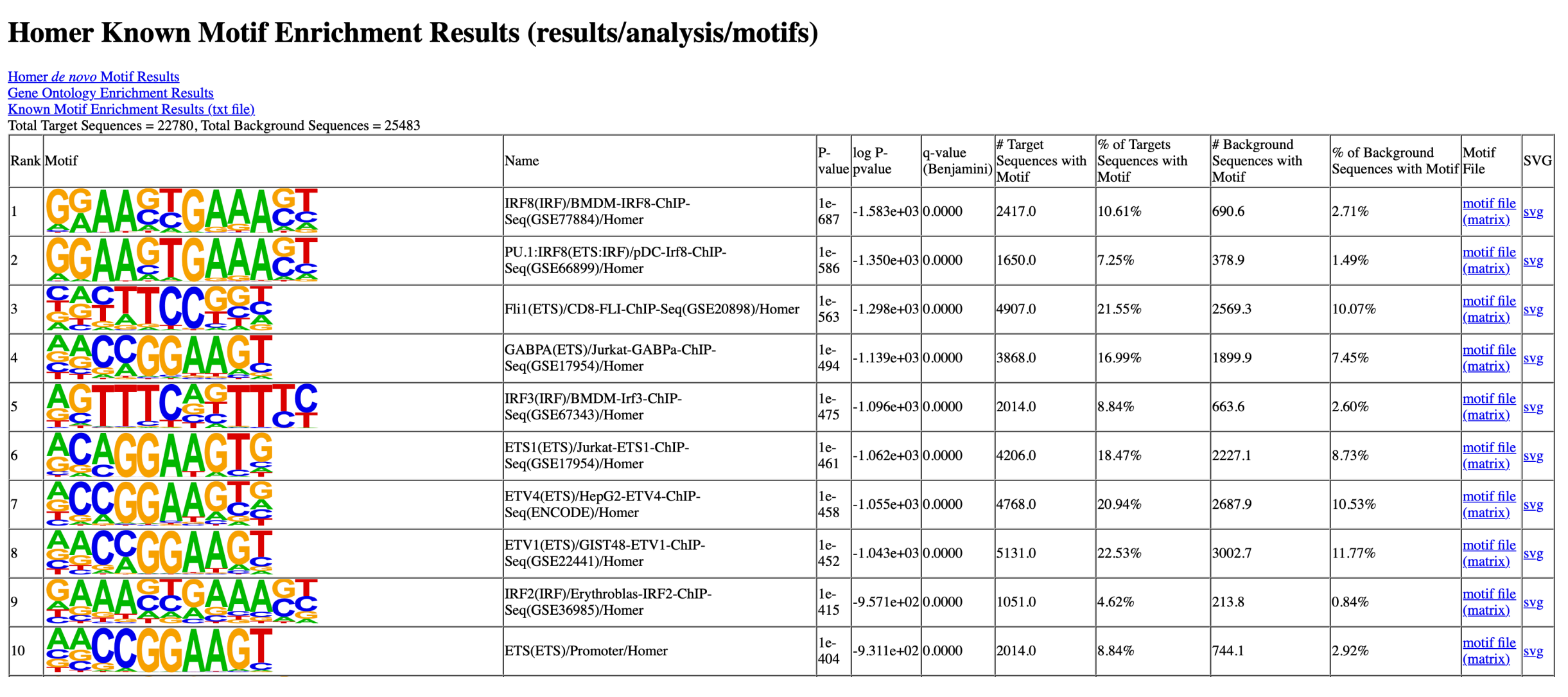



```
knitr::include_graphics("results/matrixplot_ATACrep4_NFR.png")
```



A motif analysis was then run of the set of reproducible peaks. The top 10 most enriched motifs are shown below. The top 10 consists of binding sequences for IRF, ETS, and ETV family transcription factors. IRF transcription factors are involved in regulating immune response-related genes, with IRF-8 - the most enriched motif in the data - is involved in the development of granulocyte/monocyte lineage cells. ETS transcription factors are very common in all organisms and have very wide ranging regulatory functions. Lastly, ETV family transcription factors have been linked to FGF mediated growth signaling pathways. These immune and growth related pathways make sense in the context of the sample cells, lymphoblastoids, which are actively growing and dividing cells involves in the immune response.

```
knitr::include_graphics("results/knownMotifResults.png")
```



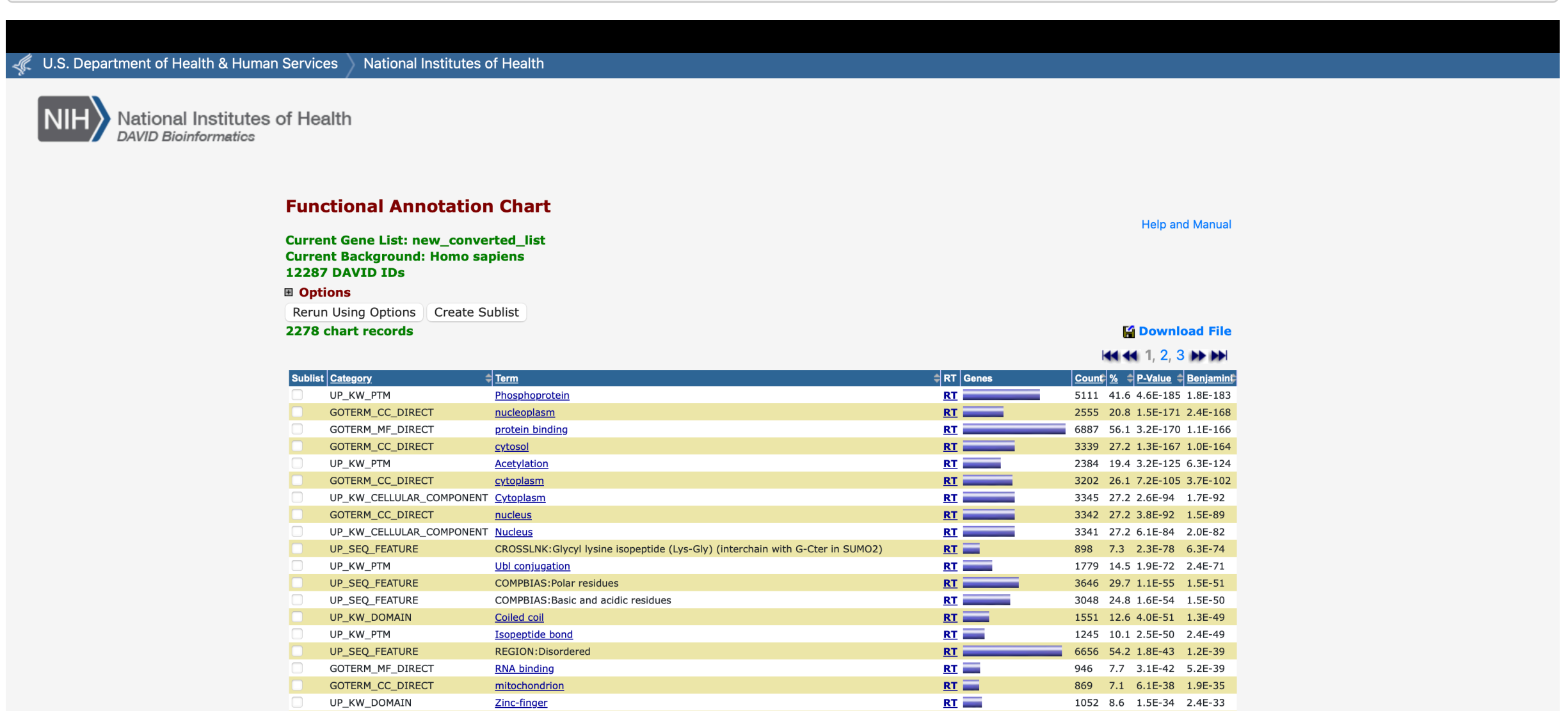
Homer Known Motif Enrichment Results (results/analysis/motifs)

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 22780, Total Background Sequences = 25483

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-687	-1.583e+03	0.0000	2417.0	10.61%	690.6	2.71%	motif file (matrix)	svg
2		PU.1:IRF8(ETS:IRF)/pDC-Irf8-ChIP-Seq(GSE66899)/Homer	1e-586	-1.350e+03	0.0000	1650.0	7.25%	378.9	1.49%	motif file (matrix)	svg
3		Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	1e-563	-1.298e+03	0.0000	4907.0	21.55%	2569.3	10.07%	motif file (matrix)	svg
4		GABPA(ETS)/Jurkat-GABPa-ChIP-Seq(GSE17954)/Homer	1e-494	-1.139e+03	0.0000	3868.0	16.99%	1899.9	7.45%	motif file (matrix)	svg
5		IRF3(IRF)/BMDM-Irf3-ChIP-Seq(GSE67343)/Homer	1e-475	-1.096e+03	0.0000	2014.0	8.84%	663.6	2.60%	motif file (matrix)	svg
6		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer	1e-461	-1.062e+03	0.0000	4206.0	18.47%	2227.1	8.73%	motif file (matrix)	svg
7		ETV4(ETS)/HepG2-ETV4-ChIP-Seq(ENCODE)/Homer	1e-458	-1.055e+03	0.0000	4768.0	20.94%	2687.9	10.53%	motif file (matrix)	svg
8		ETV1(ETS)/GIST48-ETV1-ChIP-Seq(GSE22441)/Homer	1e-452	-1.043e+03	0.0000	5131.0	22.53%	3002.7	11.77%	motif file (matrix)	svg
9		IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer	1e-415	-9.571e+02	0.0000	1051.0	4.62%	213.8	0.84%	motif file (matrix)	svg
10		ETS(ETS)/Promoter/Homer	1e-404	-9.311e+02	0.0000	2014.0	8.84%	744.1	2.92%	motif file (matrix)	svg

After annotating the file of reproducible peaks, a list of genes was generated. I then wrote this list of genes to a file and submitted it to DAVID, a tool for finding enriched functional pathways from a list of genes. DAVID returned the top functional pathways found in the list of genes and sorted them by increasing p-value. The top hits mainly contained basic cellular function pathways including cytoplasm, nucleus, protein binding, metabolism, and replication/translation machinery related pathways. Since the study was essentially a sampling of all of the open chromatin across the genome of a cell, it makes sense that most of the most enriched pathways are related to general cellular functions. Additionally, since lymphoblastoids are actively dividing, that could be the main reason why a lot of cytoplasm, cell structure, and metabolism pathways are enriched in this study.

```
knitr::include_graphics("DAVID_results.png")
```



U.S. Department of Health & Human Services 〉 National Institutes of Health

NIH National Institutes of Health
DAVID Bioinformatics

Functional Annotation Chart

Help and Manual

Current Gene List: new_converted_list
Current Background: Homo sapiens
12287 DAVID IDs
⊞ **Options**
Rerun Using Options | Create Sublist
2278 chart records

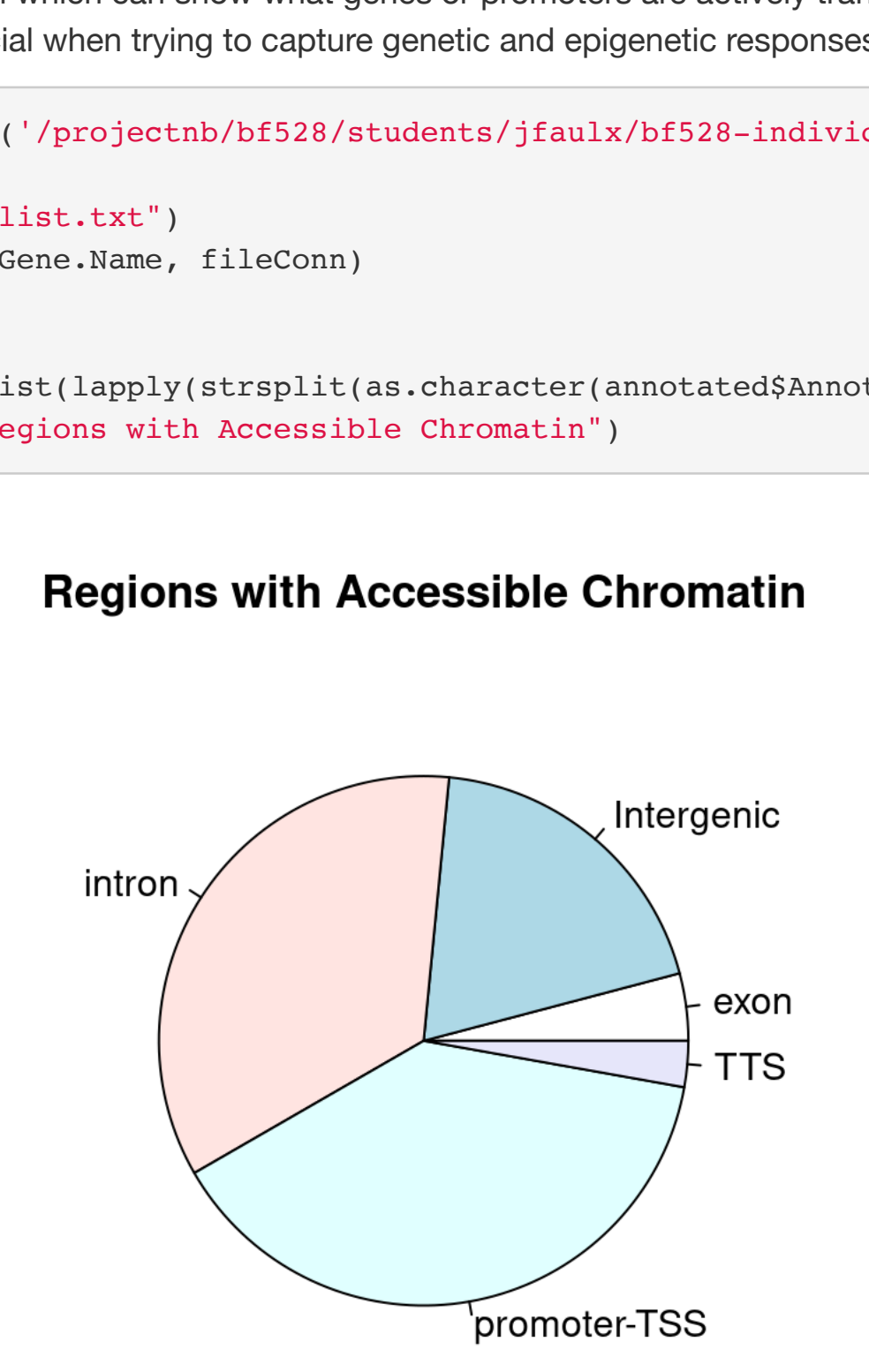
Download File
1, 2, 3

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	UP_KW_PTM	Phosphoprotein	RT		5111	41.6	4.6E-185	1.8E-183
	GOTERM_CC_DIRECT	nucleoplasm	RT		2555	20.8	1.5E-171	2.4E-168
	GOTERM_MF_DIRECT	protein binding	RT		6887	56.1	3.2E-170	1.1E-166
	GOTERM_CC_DIRECT	cytosol	RT		3339	27.2	1.3E-167	1.0E-164
	UP_KW_PTM	Acetylation	RT		2384	19.4	3.2E-125	6.3E-124
	GOTERM_CC_DIRECT	cytoplasm	RT		3202	26.1	7.2E-105	3.7E-102
	UP_KW_CELLULAR_COMPONENT	Cytoplasm	RT		3345	27.2	2.6E-94	1.7E-92
	GOTERM_CC_DIRECT	nucleus	RT		3342	27.2	3.8E-92	1.5E-89
	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		3341	27.2	6.1E-84	2.0E-82
	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	RT		898	7.3	2.3E-78	6.3E-74
	UP_KW_PTM	Ubl conjugation	RT		1779	14.5	1.9E-72	2.4E-71
	UP_SEQ_FEATURE	COMPBIAS:Polar residues	RT		3646	29.7	1.1E-55	1.5E-51
	UP_SEQ_FEATURE	COMPBIAS:Basic and acidic residues	RT		3048	24.8	1.6E-54	1.5E-50
	UP_KW_DOMAIN	Coiled coil	RT		1551	12.6	4.0E-51	1.3E-49
	UP_KW_PTM	Isopeptide bond	RT		1245	10.1	2.5E-50	2.4E-49
	UP_SEQ_FEATURE	REGION:Disordered	RT		6656	54.2	1.8E-43	1.2E-39
	GOTERM_MF_DIRECT	RNA binding	RT		946	7.7	3.1E-42	5.2E-39
	GOTERM_CC_DIRECT	mitochondrion	RT		869	7.1	6.1E-38	1.9E-35
	UP_KW_DOMAIN	Zinc-finger	RT		1052	8.6	1.5E-34	2.4E-33
	GOTERM_CC_DIRECT	centrosome	RT		439	3.6	8.1E-33	2.1E-30
	UP_KW_MOLECULAR_FUNCTION	Transferase	RT		1224	10.0	2.2E-29	2.7E-27
	UP_KW_BIOLOGICAL_PROCESS	Protein transport	RT		451	3.7	3.8E-27	8.1E-25
	GOTERM_MF_DIRECT	ATP binding	RT		927	7.5	8.3E-27	9.2E-24
	GOTERM_CC_DIRECT	macromolecular complex	RT		444	3.6	4.9E-26	1.1E-23
	GOTERM_CC_DIRECT	Golgi apparatus	RT		652	5.3	8.6E-25	1.7E-22
	UP_KW_CELLULAR_COMPONENT	Mitochondrion	RT		818	6.7	1.1E-24	2.5E-23
	UP_KW_BIOLOGICAL_PROCESS	Cell cycle	RT		465	3.8	6.9E-24	7.3E-22
	UP_KW_DOMAIN	Repeat	RT		2496	20.3	1.9E-23	2.1E-22
	UP_KW_BIOLOGICAL_PROCESS	DNA damage	RT		313	2.5	2.6E-22	1.8E-20
	GOTERM_BP_DIRECT	cell division	RT		268	2.2	1.3E-21	1.3E-17
	GOTERM_CC_DIRECT	nuclear body	RT		240	2.0	5.0E-21	8.6E-19
	UP_KW_CELLULAR_COMPONENT	Cytoskeleton	RT		828	6.7	6.2E-21	1.0E-19
	UP_KW_BIOLOGICAL_PROCESS	Host-virus interaction	RT		456	3.7	7.6E-21	4.0E-19
	UP_KW_DOMAIN	Transit peptide	RT		361	2.9	1.3E-20	1.1E-19
	INTERPRO	ARM-type fold	RT		252	2.1	3.4E-20	4.1E-16
	UP_KW_BIOLOGICAL_PROCESS	DNA repair	RT		265	2.2	4.2E-20	1.8E-18
	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2); alternate	RT		242	2.0	5.2E-20	2.8E-16

Finally, the proportions of regions that had accessible chromatin was graphed below. Promoter regions were the majority, since by nature any active promoter sites will have bound TFs. Surprisingly introns were the second largest proportion of open chromatin and exons had the second fewest. Mathematically this makes sense since only 1% of the genome is protein coding however it is interesting that introns still have so much accessible chromatin and adds further complexity to the true functions of introns. ATAC-seq allows us to understand which areas of chromatin are accessible in a given cell which can show what genes or promoters are actively transcribed/bound. Pairing ATAC-seq with other types of NGS analysis can also be beneficial when trying to capture genetic and epigenetic responses in an experiment.

```
annotated<-read.delim('/projectnb/bf528/students/jfaulx/bf528-individual-project-jfaulx/results/analysis/annotated.txt',sep='\t')
fileConn<-file("gene_list.txt")
writeLines(annotated$Gene.Name, fileConn)
close(fileConn)

pietable <- table(unlist(lapply(strsplit(as.character(annotated$Annotation), " \\("),"[[",1)))
pie(pietable, main="Regions with Accessible Chromatin")
```



Citations

1. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
2. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
3. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.
4. Twelve years of SAMtools and BCFtools Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008
5. 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.
6. MultiQC: Summarize analysis results for multiple tools and samples in a single report Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller. Bioinformatics (2016)
7. 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.
8. 2008: Model-based Analysis of ChIP-Seq (MACS)
9. 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
10. 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