ATACseq

This data was analyzed by PEPATAC pipeline which is well described here https://pepatac.databio.org/en/latest/

All the ATACseq data I have analyzed were aggregated within a single folder:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023

In the data\_input folder, you can find symbolic links to all inputs, results then contain all the results

Specifically, there are data from two human melanoma patient TILs - from 40019 (non-responder) and 40041 (complete responder). These human TILs were CRIPSR modified to remove BCL11B (gBCL11B = KO) and CD14 (gCD14 = WT) as a non-specific control.

Then there are mouse T cells from ovarian mouse model. We have data for sorted CD4 and sorted CD8 T cells from polyclonal system (not OT1/OT2/PMEL like in the scRNAseq experiments). Some T cells were sorted from peritoneal lavage and some T cells are directly from the tumor (the true TILs). We are currently focusing on the tumor samples only.

There are several config files that need to be setup before running the pipeline.

Here there are stored all sample.csv and configuration.yaml files

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/tools

In the sample.csv you setup the sample names and genotype. In the configuration file, you specify parameters of the run - select which tool to use for each step, etc. I compared genrich (preferential peak caller) with macs2 under narrowpeak and broad peak settings. The other parameters can remain intact.

The pipeline allows to use different tools for different steps of the ATACseq sample processing. The two most used for peak calling of ATACseq data are MACS2 and Genrich. The Genrich tools is superior in my opinion and I compared both as well as different parameters.

As shown in the script 01\_pepatac.sh, there are several modules that need to be loaded and python environment to be activated. The python environment is saved here but I just copied it from me home directory so I'm not sure it will work as it is:

source /share/lab\_avram/HPC\_Cluster/user/tomas/bin/soft\_versions/peptac/bin/activate

In addition, there are a few more folder related to this software that were created back then by Roy. I think the other important folder is this one (but again it was copied from home directory so it may not work as it is):

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/soft\_versions/pepatac-master

In this folder, there are a few more config files that can be setup as described in the instructions https://pepatac.databio.org/en/latest/

For example you can modify the parameters for all the tools here

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/soft\_versions/pepatac-master/pipelines/pepatac.yaml

Such as modify the parameters for calling fixed-size macs2 peaks and more:

params: '--shift -75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -p 0.01'

1) The first script to check is:

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/atacseq\_pipeline/01\_pepatac.sh

this can be used to run the pipeline but it contains also additional info on how to run it differently. There are three components that can be run one after another:

looper run tools/configuration\_C8T.yaml ...this analyzes individual samples

looper runp tools/configuration\_C8T.yaml ...this analyzes the project level - i.e. combining the samples, creating a simplified consensus peak file, etc...this output the "summary" folder

looper report tools/configuration\_C8T.yaml ...this will create a report from all the results

**However,** sometime the looper does not work well and in general I found it more efficient to just run the "dry" run of the pipeline:

looper run tools/configuration\_C8T.yaml -d

This generates submission scripts that are deposited here:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/CD8\_cells/tumor/submission

Then, I would just add the HPC header to the sub scripts and run these sub scripts one by one:

# add the following in the sub script

#!/bin/bash

#SBATCH --nodes=1

#SBATCH --cpus-per-task=16

#SBATCH --mem-per-cpu=10GB

#SBATCH -t 48:00:00

#SBATCH --job-name=8ATW

#SBATCH --output=PEPATAC\_mC8ATW.%j.out

#SBATCH --error=PEPATAC\_mC8ATW.%j.err

# then run

sbatch results/CD8\_cells/tumor/submission/PEPATAC\_C8ATK2.sub

This can also be repeated for the project level with

looper runp tools/configuration\_C8T.yaml -d

The results can be found in the "results\_pipeline" such as here:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/CD8\_cells/tumor/results\_pipeline

2) There are different suggested strategies on how to deal with replicates for differential analysis. I compared a few options here, such as after running the pipeline on individual samples, I merged the raw data for all good-quality replicates into single fastq.gz files (found in input folder) and re-run the pipeline on merged samples. These can be found in the results folder as: (m=**m**erged; C4/C8 = **C**D**4**/**C**D**8** T cells; **A**= ATACseq; P=**P**eritoneal; T=intra-**T**umoral; W/K = **W**T/**K**O)

mC4APK mC4APW mC8APK mC8APW mC8ATK mC8ATW

This was ultimately used to prepare bw files for genomic tracks. It could also be used to get a single consensual peak file.

Another option how to get the consensus peak file is by running the project level pipeline and a simplified consensus peak file is then generate within the summary folder.

Alternatively, you can just merge or intersect the peaks as you wish. My preferred option was to use the Genrich tool which generates a consensus file from multiple replicates in a more sophisticate manner. The problem is that the pipeline only allows Genrich peak calling from individual samples. So to overcome this issue, I create an independent script to specifically get the Genrich-based consenus peaks:

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/atacseq\_pipeline/02\_pepatac\_genrich.sh

This will create genotype-specific consensus peak file and then merge them together which can be used in the next step for differential analysis

3) By default function of the pipeline, after running the pipeline for the first time, including at the project level, you'll get the simplified consensus file in the summary folder. This, or any other peak file, can then be set in the configuration.yaml file as:

frip\_ref\_peaks: /share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/CD8\_cells/tumor/genrich/combined\_genrich\_peaks.narrowPeak

Then you need to re-run the pipeline which will use this peak file to generate the reference peak coverage, such as here:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/CD8\_cells/tumor/results\_pipeline/C8ATK2/coverages\_peak\_calling\_mm10/genrich\_ref\_coverage/C8ATK2\_ref\_peaks\_coverage.bed

Then, to perform the differential analysis, you just run this script

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/atacseq\_pipeline/03\_pepatac\_DESeq2\_for\_genrich.sh

which will extract the raw read counts from these coverage files and create a raw read file:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/CD8\_cells/tumor/06\_USE\_DESeq2\_genrich

/raw\_read\_counts\_pepatac\_consensus\_peaks.txt

This will then also run the differential analysis with results shown here:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/CD8\_cells/tumor/06\_USE\_DESeq2\_genrich

Similar steps can be followed for the macs2 derived peaks. Even though I tested different parameters, ultimately, the fixed-size macs2 peaks are the best option

4) The bam and bw files derived from the merged sample were further used to identify motif footprints using TOBIAS. These data are in these two folder:

for ATACseq data in general

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/merged/for\_motif\_tobias\_7-12-2024

for regulatory elements such as super-enhancers and enhancers

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/merged/for\_motif\_tobias\_01-07-2025

The script to run this is here:

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/atacseq\_pipeline/04\_tobias\_motif\_footprint.sh

These results were further processed in R, which is deposited here:

/share/lab\_avram/HPC\_Cluster/user/tomas/06\_ATACseq\_murine\_tumors\_Leo\_Tomas\_02-02-2023/results/merged/local\_R\_analysis\_motifs