CUT&RUN analysis

Conceptual overview of the analysis:  
1) map raw reads to the genome and identify peaks

2) find differential peaks between different conditions

3) identify super-enhancers and other regulatory elements

4) generate genomics tracks

5) generate deeptools heatmaps summarizing the bw signal

Most of the CUT&RUN data that we use in our CD8 paper are in this folder:

/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023

Unless stated otherwise, everything is mapped to mm10 for mouse and hg38 for human

for testing purposes, one can use the dummy files that are deposited in

/share/lab\_avram/HPC\_Cluster/user/tomas/backup\_testing\_files\_for\_omics\_pipelines

just copy these files to the desired location, run the test, and then you can re-create them from this backup location such as in:

cp -R /share/lab\_avram/HPC\_Cluster/user/tomas/backup\_testing\_files\_for\_omics\_pipelines/\* /share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/test/

1) Run the basic pipeline on HPC - you can navigate to the folder and then run the script

cd /share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/test/M1

sbatch /share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/01\_CUTRUN\_Pipeline.sh

or add the target directory containing fastq.gz files or symbolic links to them to the command like this:

sbatch --chdir=/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/test/M1 /share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/01\_CUTRUN\_Pipeline.sh

This script will map the sequencing reads to the genome and will call MACS2 and SEACR peaks. At this stage, it has to be run for each sample separately. But it can be run in a loop:

samples=(M1 M2 M3 M4 )

for sample in "${samples[@]}"; do sbatch --chdir=/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/test/"$sample" /share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/01\_CUTRUN\_Pipeline.sh; done

2) Identify differential peaks between two conditions (not applicable to Bcl11b datasets). First, it is necessary to create a unique set of peaks common for all the samples. This is now done within the script. Modify the sample.csv file to contain sample names corresponding to the names of the directories containing data from step 1. Add the conditions such as WT/KO. Refer to the sample.csv file in the command as well as to the default reference condition - such as WT (it is default) in this case. This time the chdir should point to the parental directory containing directories with several samples to be analyzed for differential CUR&RUN signal. Depending on the desire to analyze the SEACR or MACS peaks, use the specific script. -s is to specify the sample.csv and -c to specify the reference condition

sbatch --chdir=/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/test --export=s=/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/samples.csv,c=WT /share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/02\_CUTRUN\_DESeq2\_for\_seacr.sh

sbatch --chdir=/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/test --export=s=/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/samples.csv,c=WT /share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/02\_CUTRUN\_DESeq2\_for\_macs2.sh

# For most downstream application I stick to the SEACR peaks which quite nicely correlate with the signal as verified in genome browser. Note that I tested several filtering conditions to identify peaks, but ultimately we care about differential peaks for the histone marks so I left the least stringent SEACR p value (0.01) for the peak identification step and we regulate identification of differentially modified sites based on DESeq2 p values. Here I also teste multiple paramaters and CUT&RUN data often do not behave like ChIPseq or other datasets. So eventually I called differential peaks if the 1) the DESeq2 P value was <0.05 or or 2) if this combination applies: P<0.3 & |Log2FC|>0.2

# Note that for Bcl11b CUT&RUN we only have the WT set so there is no differential analysis - in this case the 0.01 SEACR P value filtering had too many peaks and we ended up using 0.005 filtering.

3) Next, use the data generated so far and use it to identify enhancers, super-enhancers, and poised enhancers. This is done in conjunction with the results from scRNA-seq data and it was mostly run as individual commands saved in this file:

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/03\_identification\_of\_regulatory\_elements.sh

Note that after obtaining the list of regulatory elements, I repurposed the 02\_CUTRUN\_DESeq2\_for\_seacr.sh script which requires a tiny adaptation as described in the 03\_identification\_of\_regulatory\_elements.sh script.

The run using algorithm ROSE was performed in the directory

/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/rose-docker

4) In this folder, I performed some basic characterization of data and peaks for each modification which we don't plan to use for publication at this point

/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/Roverlaps\_intersections

**More importantly,** here I performed the analyses subsequent to the TE/SE/PE identification. This also includes all the R readme file with scripts and the plots selected for publication

/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/ROSE\_downstream\_analyses/SE\_withRNAseq

5) To generate genomics tracks, I switched to my favorite pygenometracks which can generate svg files which are easily modifiable in Inkscape and it is highly customizable. Svg generate in IGV or from the GenomeProteinPaint are pain to work with. These tracks will be used for publication

/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/pygenometracks

Here I also have input folder which contains hopefully all the bw and bed files that one may need to re-create any of the genomics tracks or use them for further analysis and data deposition

To actually run it, the commands are deposited here, so run them as individual commands. You need to have the conda environment installed with all the requirements - information on how can also found in the script file:

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/04\_pygenometracks.sh

6) To create heatmaps using deeptools use this script

/share/lab\_avram/HPC\_Cluster/user/tomas/bin/cut\_and\_run\_pipeline/05\_deeptools\_heatmaps.sh

and the data can be found here where I also symlinked the input folder from the pygenometracks directory. These will likely be used in publication

/share/lab\_avram/HPC\_Cluster/user/tomas/08\_cut\_run\_5-30-2023/deeptools\_plots