## June 2017

# Exercise 9: Analyzing ChIP-Seq data

## Dena Leshkowitz

### Introduction

# In this workshop we will learn how to analyse ChIP-Seq data. The data is taken from the article Dicken at al. Transcriptional reprogramming of CD11b+Esam(hi) dendritic cell identity and function by loss of Runx3. PLoS One. 2013 Oct 15;8(10). We will use two biological replicate ChIP-Seq experiments that were conducted for detection of Runx3-bound genomic regions using in-house anti Runx3 Ab and 30x106 positive CD11c MACS isolated (Miltenyi Biotec) classical dendritic cells (DC). For further details please see the [article](http://www.plosone.org/article/info%3Adoi/10.1371/journal.pone.0077490). The BAM files we will be using were done with only a subset of the reads present in the study.

General remark- the commands you need to type in the command line are *in italic* format.

### Instructions

### Accessing the Wexac server

Open the MobaXterm as in previous exercises and open the terminal icon.

You now have the ability to write commands in Linux which run on the [WEXAC](http://www.weizmann.ac.il/DIS/high-performance-computing-wexac) cluster.

### Run MACS command

For more details on the MACS tool look at the site

<https://github.com/taoliu/MACS>

*mkdir chip-seq-runx*

*cd chip-seq-runx*

*module load python/bio-2.7.13*

*module load R*

*module load bedtools*

Typing the macs command with the -h will give an explanation on how to run this tool

*macs2 callpeak -h*

Following is the command to run macs on the first replicate (the sign ~ is a shortcut to your home directory)

*bsub -q bio-guest -R "rusage[mem=10000]" -o ./mac1fs\_job.txt -e ./mac1fs\_error.txt -J mac1fs "macs2 callpeak -t ~/course\_2017/chip-seq/bam\_full/IP1.bam -c ~/course\_2017/chip-seq/bam\_full/input1.bam -n macs2\_input1\_IP1"*

Question 1: How many files were created by macs2?

Question 2: From the information written to the output ending with ‘xls’ (which unlike its name is just a tab delimited file), what is the predicted fragment size (d)?

Within the outputs produced, one of the files gives us all the peaks detected in a text file, write the following to see the top lines in the file

*head macs2\_input1\_IP1\_peaks.narrowPeak*

To find the number of peaks, count the number of lines in the *macs2\_input1\_IP1\_peaks.narrowPeak* file using the command wc -l

*wc -l macs2\_input1\_IP1\_peaks.narrowPeak*

Question 3: How many peaks were detected in replicate 1?

Run macs2 analysis on the second replicate.

Question 4: How many peaks were detected in the second replicate experiment?

To find the amount of overlap between the two replicate peak files we will use the program intersectBed. This program comes from [BEDtools suite](http://bedtools.readthedocs.org/en/latest/)

For the program explanation first type –

*intersectBed -h*

Question 5: What does the option -wa do?

We will run intersectBed to find the peaks overlapping between the two files to wc -l in order to count the lines of overlapping peaks.

*intersectBed -wa -a macs2\_input1\_IP1\_peaks.narrowPeak -b macs2\_input2\_IP2\_peaks.narrowPeak | wc -l*

Question 6: What is the amount of overlap between the replicates?

Question 7: What is the amount of overlap if we switch the order of the files (or the -a and -b)?

For good reproducibility we expect to have greater than 50% overlap.

### Analyse peaks using [CEAS: Enrichment of Genome Features](http://liulab.dfci.harvard.edu/CEAS/usermanual.html)

 CEAS provides statistics on ChIP enrichment at important genome features such as specific chromosome, promoters, gene bodies, or exons, and infers genes most likely to be regulated by the binding factor.

The narrowPeaks outputs we have do not comply with the BED format needed to run CEAS. Therefore, we will first make a new file containing the peaks that overlap between the two experiments and that is BED format (by cutting the first four columns of the narrowPeak file using the following command):

*intersectBed -wa -a macs2\_input1\_IP1\_peaks.narrowPeak -b macs2\_input2\_IP2\_peaks.narrowPeak | cut -f 1-4 > macs2\_input\_IP.bed*

*ceas -b macs2\_input\_IP.bed --name=input\_IP\_ceas -g /shareDB/CEAS-DB/mm9.refGene*

### In order to view the CEAS pdf output file look at the sftp pane in the mobaXterm (see picture below), you might need to check in the “Follow terminal folder” and if you still do not see the output folder files, change the path (red arrow). Once you are in the right folder and can see the CEAS pdf file, double click on it in order to open it.



Question 8: Does our transcription factor - Runx3 preferably bind to promoters?

### Analyse Peaks using GREAT

Download the peak BED file *macs2\_input\_IP\_peaks.bed* using the download feature of the sftp (see green arrow above). Create a folder in disk D and save it to this folder. Open [GREAT](http://bejerano.stanford.edu/great/public/html/) and upload the BED file. You should select the mm9 mouse built and click submit.



Without going to deep into the researched biological question, the DC are involved in immune reactions.

Question 9: How many times is the word immune found in the enriched terms? How many peaks and how many genes are associated with the first immune term in the MSigDB Pathway output?

Question 10: Looking at the Region-Gene Association Graphs – what is the most frequent binned by orientation distance from the peak to TSS?

We can view the genes associated with the peaks if we expand the **Jobs Description** (top of the report; press on the +)

Select - View **all genomic region-gene associations**.



The Following window will open - 

For the next assignment select a gene which has a peak next to it.

### Browsing the peaks with a genome browser

We will use the Integrative genomics browser - [IGV](https://www.broadinstitute.org/igv/) (see exercise 1) to view the mapped reads and the peaks. Once the application opened load the mm9 genome. If it is not available in the scroll button, got to Genomes and select “Load genomes from browser” (see below).



Select from File -> “Load from File” (see red arrow below) and select the BAM files and the bed file you created.

##

After loading the bam files you can insert the name of the gene you selected from GREAT and write it in the window where there is a red arrow in the picture below.



Question 11: Where is the peak in regards to the gene (near TSS, upstream…)? Are you convinced that there is enrichment in binding at this location when considering the reads in both the IP experiments as well as in the controls? Are there other peaks near this genes – you can zoom out by clicking on the minus sign in the left corner (see blue arrow).

Answers available here - http://dors.weizmann.ac.il/course/AnswersExercise9.docx

## Supplementary

### Basic Linux commands:

man (command) ...... shows help on a specific command

ls ................. show directory, in alphabetical order

logout ............. logs off system

mkdir .............. make a directory

rmdir .............. remove directory (rm -r to delete folders with files)

rm ................. remove files

cd ................. change current directory

more .............. views a file, pausing every screenful

grep ............... search for a string in a file

head ............... show the first few lines of a file

tail ............... show the last few lines of a file

df ................. shows disk space available on the system

du ................. shows how much disk space is being used up by folders

chmod .............. changes permissions on a file

cut ............... print selected parts of lines

cp ............... copy file

mv ............... move file

wc –l .............. print the number of lines

sort ............. sort lines of text files