ChIP-seq and ATAC-seq peack calling and QC

1. **Prepare the working directory.**
   1.1 Create a working directory "/workdir/$USER".

Copy all data files for this exercise from "/shared_data/epigenomics/exercise2/" into the working directory.

```
mkdir -p  /workdir/$USER/exercise2
cd /workdir/$USER/exercise2
cp /shared_data/epigenomics/exercise2/* ./
ls -l
```

1.2 Install Filezilla client on your laptop

Unnessessary if you still have this installed from last week.

Filezilla is a sftp client software. If you do not have any sftp client software on you laptop, download Filezilla client from this page. Double click to install.

https://filezilla-project.org/download.php?show_all=1

*** The installer might prompt you to install other additional software, e.g. virus protector, always click "no" to decline.**

1.3 Install IGV on your local workstation

Unnessessary if you still have this installed from last week.

Go to the IGV web site (https://software.broadinstitute.org/software/igv/), click “Download”. Double click the IGV installation tool to install IGV. On Windows computer, the software is installed in the directory C:\Program Files.

2. **Visualize ChIP-seq and ATAC-seq data in IGV**

   2.1 Download files to your laptop.

   Use Filezilla to download:

   ChIP-seq_SRR5179211_Rep1.bam

   ChIP-seq_SRR5179211_Rep1.bam.bai

   ATAC-seq_Rep1_SRR7696734.bam

   ATAC-seq_Rep1_SRR7696734.bam.bai

   to a folder on your computer. The *.bai files are index files that are companion files to BAM files.

   2.2 Load sacCer3 genome into IGV

   Click the dropdown menu in the upper left and select 'more...'. Filter for 'sacCer3' and select this genome. It will automatically initialize on your system.

   2.3 Open BAM files

   Open ChIP-seq_SRR5179211_Rep1.bam and ATAC-seq_Rep1_SRR7696734.bam and view the following genomic coordinates:

chrV:175,594-181,774
chrXII:666,176-671,951
chrXIV:449,980-454,181

This data represents the raw aligned data from the experiments. Processing of these files can yield BED/GFF peak files and continuous data such as bedGraph and bigWig files.

***Depending on the RAM and CPU of your computer, consider only viewing one coordinate. You may also need to restart IGV regularly if you are running with default memory allocations**

3. **Call ChIP-seq peaks with MACS2**

   3.1 Load MACS2 into environment

   Initialize MACS2 in the BioHPC enviroment

```
export PYTHONPATH=/programs/macs2-2.2.7.1/lib64/python3.6/site-packages

export PATH=/programs/macs2-2.2.7.1/bin:$PATH
```

   3.2 How to call ChIP-seq peaks using MACS2

   You are provided with ChIP-seq BAM files in replicate. Call peaks using the below command.

```
cd /workdir/$USER/exercise2

macs2 callpeak -t ChIP-seq_SRR5179211_Rep1.bam -f BAM -g 1.2e7 -n CS_Rep1 -B -q
0.05 --nomodel --extsize 100 --keep-dup all --call-summits

macs2 callpeak -t ChIP-seq_SRR5179212_Rep2.bam -f BAM -g 1.2e7 -n CS_Rep2 -B -q
0.05 --nomodel --extsize 100 --keep-dup all --call-summits

ls -l
```

   Inspect the contents of the newly generated files.

```
less CS_Rep1_peaks.narrowPeak
# press "space" to move to the next page; press "q" to exit
```

   3.3 Filter peaks with blacklist file

Remove peaks that overlap with our blacklist regions.

```
export PATH=/programs/bedtools2-2.29.2/bin:$PATH

cd /workdir/$USER/exercise2

bedtools intersect -v -a CS_Rep1_peaks.narrowPeak -b sacCer3_blacklist.bed >
CS_Rep1_peaks_filter.narrowPeak

bedtools intersect -v -a CS_Rep2_peaks.narrowPeak -b sacCer3_blacklist.bed >
CS_Rep2_peaks_filter.narrowPeak
```

   Compare the difference after filtering out the blacklist regions

```
# diff command compares the differences between two files
diff CS_Rep1_peaks.narrowPeak CS_Rep1_peaks_filter.narrowPeak
```

3.4 Download peak files to your laptop

Use Filezilla to download:

CS_Rep1_peaks_filter.narrowPeak

CS_Rep2_peaks_filter.narrowPeak

CS_Rep1_treat_pileup.bdg

CS_Rep2_treat_pileup.bdg

3.5 Visualize in IGV

Open *narrowPeak and *bdg (bedGraph) files in IGV. Check those 3 genomic coordinate again and confirm peaks were found there (or not!). Right-click on the file name in the IGV interface after loading ithem and make sure 'Autoscale' is selected (should be 5th option from the bottom of the right-click menu).

4. **Call ChIP-seq peaks with GEM**

4.1 How to call peaks using GEM

```
cd /workdir/$USER/exercise2

java -Xmx8G -jar gem.jar --t 4 --q 0.05 --d Read_Distribution_default.txt --g
sacCer3.chrom.sizes --genome . --exptCS ChIP-seq_SRR5179211_Rep1.bam --exptCS
ChIP-seq_SRR5179212_Rep2.bam --f BAM --out PeakOutput --k_min 6 --k_max 13

ls -l
```

4.2 Download data files to your laptop.

Using FileZilla, download the 'PeakOutput' folder to your local computer.

4.3 Examine peak results

Open 'PeakOutput.results.htm' in a browser of your choice. The ChIP-seq dataset is from the Reb1 yeast protein with sequence specificity for the 'TTACCCK' motif. Did GEM correctly identify the motif at the peak?

5. **Call ATAC-seq peaks with MACS2**

5.1 How to call ATAC-seq peaks using MACS2

```
cd /workdir/$USER/exercise2

macs2 callpeak -t ATAC-seq_Rep1_SRR7696734.bam -f BAM -g 1.2e7 -n ATAC_Rep1 -B -q
0.05 --shift -75 --extsize 150 --nomodel --SPMR --keep-dup all --call-summits

macs2 callpeak -t ATAC-seq_Rep2_SRR7696734.bam -f BAM -g 1.2e7 -n ATAC_Rep2 -B -q
0.05 --shift -75 --extsize 150 --nomodel --SPMR --keep-dup all --call-summits

ls -l
```

5.2 Filter peaks with blacklist file

```
cd /workdir/$USER/exercise2

bedtools intersect -v -a ATAC_Rep1_peaks.narrowPeak -b sacCer3_blacklist.bed >
ATAC_Rep1_peaks_filter.narrowPeak

bedtools intersect -v -a ATAC_Rep2_peaks.narrowPeak -b sacCer3_blacklist.bed >
ATAC_Rep2_peaks_filter.narrowPeak
```

5.3 Download peak files to your laptop

Use Filezilla to download:

ATAC_Rep1_peaks_filter.narrowPeak

ATAC_Rep2_peaks_filter.narrowPeak

ATAC_Rep1_treat_pileup.bdg

ATAC_Rep2_treat_pileup.bdg

5.4 Visualize in IGV

Open *narrowPeak and *bdg (bedGraph) files in IGV. Check those 3 genomic coordinate again and confirm peaks were found there (or not!). Right-click on the file name in the IGV interface after loading ithem and make sure 'Autoscale' is selected (should be 5th option from the bottom of the right-click menu).

Appendix

A. **Additional reading**

NIH ENCODE Consortium's recommended scripts, parameters, and workflows for ChIP-seq and ATAC-seq data analysis:

https://github.com/ENCODE-DCC/chip-seq-pipeline2

https://github.com/ENCODE-DCC/atac-seq-pipeline

B. **Performing replicate comparison with IDR**

There are plenty of additional tutorials on running IDR with ChIP-seq data (ATAC-seq would be very similar).

https://hbctraining.github.io/Intro-to-ChIPseq/lessons/07_handling-replicates-idr.html

```
# Set environment
export PYTHONPATH=/programs/idr-2.0.3/lib64/python3.6/site-packages
export PATH=/programs/idr-2.0.3/bin:$PATH

# Sample code (requires IDR to be installed and ChIP/ATAC-seq files to be
processed additionally)
idr --samples CS_Rep1_peaks.narrowPeak CS_Rep2_peaks.narrowPeak --input-file-type
narrowPeak --output-file CS_IDRoutput --rank p.value --soft-idr-threshold 0.05 --
plot --use-best-multisummit-IDR
```