**ChIP-seq data analysis protocol**

By Sevinc Ercan 06/06/2018

e mail: [se71@nyu.edu](mailto:annalena.kranz@gmail.com)

updated by Lena & Diogo ???

For short version of the protocol see ErcanLab\_ChIP-seq\_analysis\_v5\_slurm.docx

Located at https://drive.google.com/drive/folders/0B86yNkEPp\_kmcWk5UHd5ZmVIUFk

Note on directory and file naming:

Directories for each data set should be named as the sequence IDs (e.g. SE100).

Directories of average data should have names of all replicates (e.g. SE100\_SE103\_SE150)

File names should have in the same order descriptive name, extract, Sequence ID, and input ID.

E.g. DPY27\_N2\_Emb\_ext40\_SE100\_input\_SE101

Then you can have the appended \_ratio or \_subt etc.

For averaged files, you would list all the sequence IDs of the ChIPs and Inputs with “-“ separator.

Note on files to be stored:

All bam files in one folder. All median coverage output in one folder.

Input subt coverage, peak calls of each rep in its own folder. Do not store initial coverage file from Deeptools to calculate median coverage, and do not store any coverage (e.g. bedgraph) output coming from MACS2.

For average folders store subtract input and log2 ratio coverage, MACS2 output of combined bams (not coverage file), final filtered peak file, final filtered summits file.

Note on old ChIP analysis files:

Move all folders to the archive.

Note on old ChIP data:

Combine bams to get average tracks for two bigwigs with input subtract and log2 ratio.

Note on track hubs:

Put all the bigwigs of individual data into multiple hubs organized 50 at a time by sequence ID.

Put all the average tracks ran on old data into multiple hubs organized up to 50 at a time by target.

**FOR INDIVIDUAL REPLICATES**

**Sequence alignment using Bowtie**

Run new bowtie version (note on protocol) with default parameters.

Get output, name with sequencing ID, and store as a sorted bam file.

**Peak calling using MACS2**

Run MACS2 with default parameters for regular peak calling

macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g ce -n SequenceID -q 0.05

**Get median coverage**

Get the median coverage of each data set (to normalize to median coverage).

Run deeptools with “bamcoverage” without normalization for 100 bp windows.

Get coverage as bedgraph and a tool that will quickly tell you the median coverage in the bedgraph. Delete the bedgraph file afterwards. Use ID as the name of the text file reporting median coverage and all of these in one folder named (median coverage).

bamCoverage --bam a.bam -o a.SeqDepthNorm.bdg --binSize 100 --outFileFormat bedgraph

**Input Comparison**

Run deeptools with “bamcompare”

Get coverage as bigwig.

Normalize based on the ratio of the median coverage ChIP/ median coverage Input (obtained from median coverage folder by matching sequencing ID)

bamCompare -b1 ChIP.bam -b2 Input.bam –o NAME\_inputsubt.bw

--scaleFactors (ChIP median coverage):(input median coverage)

--operation subtract --binSize 10 --extendReads 200

**FOR COMBINING REPLICATES**

**Combining replicates for producing combined coverage**

Combine all bam files for all replicates for all samples and inputs, respectively.

Get median average as above. Run the program twice for outputting input subtracted signal as above, and log2ratio signal. For log2 ratio signal run:

bamCompare -b1 ChIP.bam -b2 Input.bam –o NAME\_log2rat.bw

--scaleFactors (ChIP median coverage):(input median coverage)

--binSize 10 --extendReads 200

**Average peak calling using MACS2**

Run MACS2 on the combined ChIp and Input bam files.

macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g ce -n SequenceID -q 0.01 --call-summits

**Peak filtering based on majority rules**

Take the averaged peak file from the combined replicate MACS2 run. Compare against the individual replicate peak files, and filter peaks that overlap in majority (1 out of 2, 2 out of 3, 2 out of 4, 3 out of 5, 3 out of 6 or better) of the replicates into a new file named \_finalpeaks. Take corresponding summits into a \_finalsummits.

Two parameters that I am not sure of:

The q value of 0.01 for the combined MACS2 run.

How the scaling factor of the median coverage will affect the y axis scale between replicates.

To settle these Test run the new pipeline on the replicates and average:

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Descriptive name | Matching Input | Mapped reads |
| CJ39 | DPY27\_JL001\_N2\_Emb\_ext19 | CJ43 | 10,104,944 |
| SE30 | DPY27\_JL001\_N2\_Emb\_extEE6 | EE6 input | 24437615 |
| CJ132 | DPY27\_JL001\_N2\_Emb\_ext10 | CJ19 | 7,655,704 |
| SE172 | DPY27\_ Q3995\_N2\_Emb\_ext 75 | SE176 | 16,484,861 |

Report median coverage values per reps and combined runs. Report peak numbers per reps and combined runs. Put individual and combined tracks and averaged and final peak calls into a track hubnamed “test\_chipanalysis\_vs5”