## Guide. Chromatin organization.

**# Alignment and filtering**

Start processing the HiC samples on HiC-Bench. You can follow the steps here:
https://github.com/NYU-BFX/hic-bench/wiki/Setting-a-HiC-Bench-run-on-BigPurple-(NYU)
HiC fastq files: /gpfs/data/courses/bmscga2604/2020/chromatin_organization/HiC

Make sure the data in the sample sheet is correct:

- The 'enzyme' column must be set to Arima. This was the enzyme cocktail used in the library preparation.
- The samples are mouse embryonic stem cells. Set the ‘genome’ column to mm10 (latest mouse genome-build).
- These 2 samples are replicates so the group column should be the same for both samples. This is important because in many steps hic-bench will generate results ‘by sample’ and ‘by group’. In the ‘by sample’ subdirectory you will find the results per sample while in the ‘by group’ subdirectory the results are obtained by merging the information of all the replicates in the same group.

Once the alignment and filtering steps are completed run the '02b-filter-stats' step. This step generates two plots useful for quality control (‘counts.pdf’ and ‘percentage.pdf’).

Tips: How many total intra-chromosomal valid pairs were obtained per sample? What is the percentage of intra-chromosomal valid pairs?

**# Contact matrix visualization**

Run the '03a-tracks' step. It will use ‘Juicer’ to  generate contact matrices in binary format (‘filtered.hic' file). These files can be visualized on Web Juicebox:  https://aidenlab.org/juicebox/

Note: Select the 'balanced' normalization (this is similar to IC normalization).
Juicer documentation: http://aidenlab.org/documentation.html

**# Compartments**

Run the '04a-compartments' step. This step uses 'Homer' to identify A/B compartments.

Relevant output files:
- compartments.scores.bedGraph: Useful for visualization. Positive values denote A-like regions, negative values B-like regions.
- A_compartments.bed: Bed file with the coordinates of all the identified A compartments.
- B_compartments.bed: Bed file with the coordinates of all the identified B compartments.

Tips: You can load and visualize the 3 files on IGV or Juicebox. I suggest you load the bedgraph and bed files on Juicebox so you can explore the contact matrix together with the compartment information. Select a region where you can see a nice TAD. Is this TAD in an active or inactive compartment? What gene/s are in this TAD?

Note: To identify A/B compartments this tool usually requires active mark information (e.g. H3K27ac peak coordinates) from the same cell-type. By default HiC-Bench uses the coordinates of house-keeping (HK) genes from a database. HK genes are expected to be found in active genomic regions and they have the advantage that their gene activity is usually conserved across cell-types. This can be specially useful when you don't have any active mark data for your samples. It was already tested on the ES samples so It should work well, you don’t need to set the ES H3K27ac peaks in this case.

Homer documentation: http://homer.ucsd.edu/homer/interactions2/HiCpca.html

**# Contact matrix and normalization**

Run steps 05a-matrix-filtered and 06a-matrix-ic.

Note: That's all you have to do here but just FYI, the normalized matrices could be next used to identify TADs, compute TAD boundary-scores, make contact heatmaps, and to perform other downstream analyses (e.g. intraTAD activity).

**# Downstream analysis**

Integrate the compartments information with gene density, H3K27ac peak density and house-keeping gene density.

<u>Data</u>:

House-keeping genes: /gpfs/data/tsirigoslab/hicbench-repository/data-repo/genomes/mm10/HK_genes.bed
Transcription start sites: /gpfs/data/tsirigoslab/hicbench-repository/data-repo/genomes/mm10/tss.bed
H3K27ac peaks: /gpfs/data/courses/bmscga2604/2020/chromatin_organization/H3K27ac_ChIP_Seq/peaks_ES.bed
A_compartments.bed and B_compartments.bed files that you generated in 04a-compartments.

<u>Tips</u>:

- How many transcription start sites fall in average in A compartments? And in B compartments? <gene density>
- Same for H3k27ac peaks and house-keeping genes.
- Which of these 3 features do you think will be the most accurate for predicting A compartments? Why?

You can use ‘bedtools' on bigpurple to find/count overlaps between 2 bed files:
https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html

Example:

```
$ module load bedtools
$ bedtools intersect -c -a A_compartments.bed -b tss.bed > tss_in_A_compartments.tsv
$ head tss_in_A_compartments.tsv
chr7    3000000    3800000    36
chr7    3900000    5100000    68
chr7    6100000    6400000    11
```

The command does the following: for each entry in '-a', reports the number of hits in ‘-b’. So in the previous example the 4th column shows the total number of tss sites mapping each compartment A. You can next load the ‘tss_in_A_compartments.tsv’ file as a data frame in R, python or Excel and do more analysis, make plots, etc.