Assignment 2

2022-09-24

Description

The goal of this assignment is for you to run a "rudimentary" sequencing analysis pipeline and get familiar with the tools. As usual, you should submit a reproducible[1] report. For all the steps you should describe what you are doing and make comments about the data/results, eg if they look okay and what justifies your next steps[2]. The analysis is based on
this tutorial you can use it as a guide but **DO NOT** run it in Galaxy! Run the analysis locally or in BigPurple and produce a report. When asked to diverge do so.

Notes on Notation

- `signal` = H3K27m3 or H3K4m3 for this exercise.
- `log2FC(x, y)` = $\log_2$ fold-change = $\log_2 \frac{x}{y} = \log_2 x - \log_2 y$
- ROI = Region of interest. For this exercise, ROI = `chrX:151,340,000-152,862,000` (see Figure 2 of Wang et al)

FASTQ

1. Download the following fastq files and check their `md5sum` to make sure they are not corrupted.
2. Describe the data (eg length, pair/single, number of reads)
3. Perform QC analysis and comment on the results
4. If deemed necessary, take necessary steps to improve the quality of the data (trim/filter etc)

USE fastp: for steps 3 & 4.

Links:

- https://zenodo.org/record/1324070/files/wt_H3K4me3_read1.fastq.gz
- https://zenodo.org/record/1324070/files/wt_H3K4me3_read2.fastq.gz

MD5SUMS (see command `md5sum --check`):

```
5b6054c8467f98afccb48e6b21d5494c  wt_H3K4me3_read1.fastq.gz
a2b7ea4849aa4a137c96bf135e9bde9a  wt_H3K4me3_read2.fastq.gz
```

Map Reads

1. Align the reads to the `mm10` reference genome.
2. Using `samtools` manipulate the resulting alignments to:
   1. convert to BAM
   2. sort
   3. index
   4. mark duplicates
3. Perform QC analysis of the mapping using `samtools` statistics (`flagstat`, `idxstats`, `stats`) and comment on the results.

[1] more or less I will not run it…
[2] besides me asking you

Try 2 different aligners (eg, `bowtie2`, `bwa`, `Rsubread`) and compare the results (eg % mapped, overlap). Pick the result of 1 aligner for the rest.

4. Clean up the data from unmapped reads and low quality alignments.
5. Visualize your BAM in IGV and find an area of high coverage (peak) comment on the image (eg mismatches, mis-paired).

ChIP QC

1. Download the files from this link and check their md5sums (`md5sum --check md5sum.txt`)[3].
2. Report how these files were generated, for example:
   - which programs were used to create these files?
   - are they sorted?
   - do they cover the whole genome?
   - how many reads do they represent and what is the average fragment length?
3. Compute the coverage (in `RPKM`) at 1000bp resolution for chrX and visualize the track for the ROI.
4. Visualize the correlation matrix of the coverage of the samples and their "fingerprint"

ChIP Analysis

1. Merge the replicates by averaging them and compute the `log2FC(signal, input)` for every signal.
2. Using a threshold, find the `ko_H3K4me3` peaks visible in IGV for the ROI. Write the coordinates in a `bed` file and add it as an IGV track to validate them.
3. Find the genes these peaks overlap or if they don't the nearest gene in either strand.
4. Compute `log2FC(KO, WT)` for both signals and add it to the track as well. Which genes are most affected?

[3] if you work on BigPurple the data are at `/gpfs/data/tsirigoslab/public/teaching/bioinformatics/GSE99991`