NCMBM COURSE MULTIOMICS SEQUENCING HANDS-ON

17-11-25

Samuele Cancellieri

samuele.cancellieri@ncmbm.uio.no

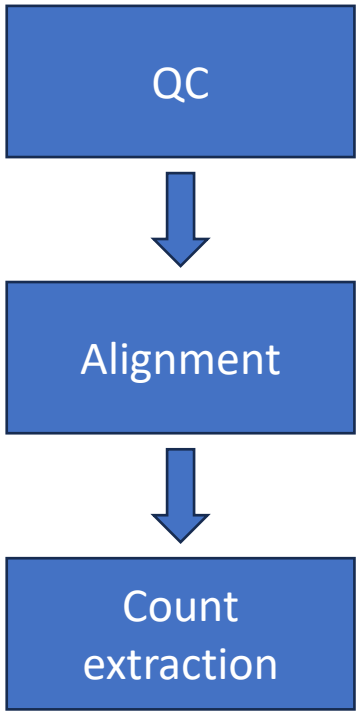
Download the git repo from

samuelecancellieri/ncmbm_practical_multiomics_course

Using the command:

git clone https://github.com/samuelecancellieri/ncmbm_practical_multiomics_course.git

RNA-seq pipeline



fastQC

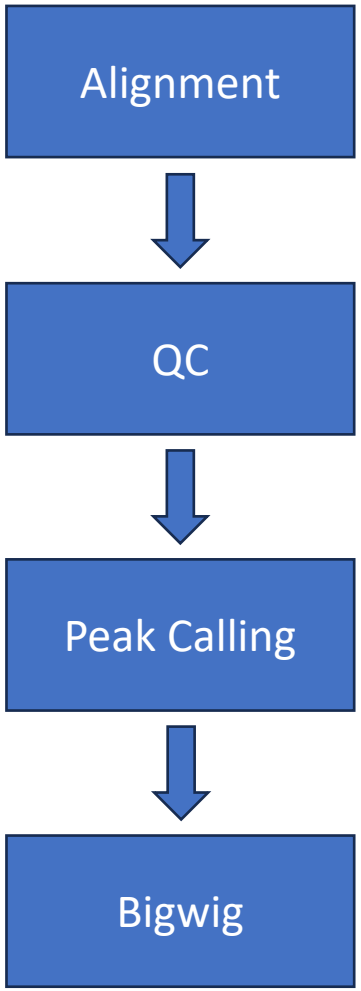
STAR

HT-seq

The pipeline requires as input a set of files to start.
The files are the paired reads (1 & 2), the name we want to give to our files, the output folder and the index plus the gtf to create a count matrix on the gene (we are aligning against the genome but using transcripts information).

This count matrix can be used for different purpose, the most common one is to find differentially expressed genes (i.e., genes that have a different amount of reads mapping the transcriptome) in two conditions (e.g., cancer vs healthy)

ATAC-seq pipeline



bowtie2

RemoveMithocondrial.py
Picard MarkDuplicates
PBC1,PBC2,NRF
Samtools filter low-quality

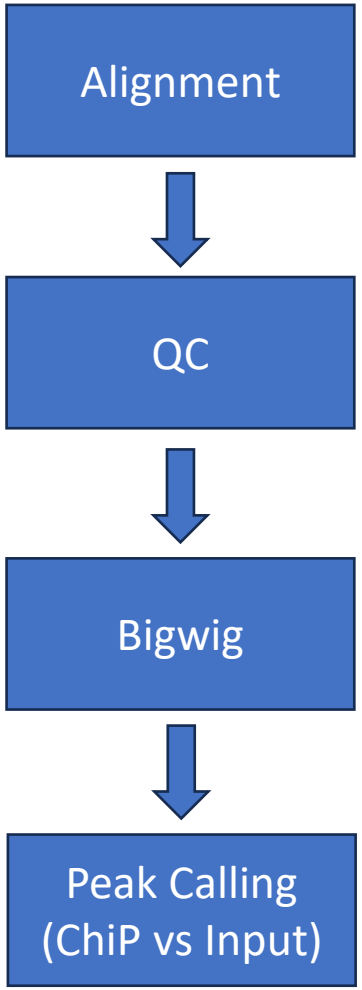
MACS2

deepTools
IGV

The pipeline requires as input a set of files to start.
The files are the paired reads (1 & 2), the name we want to give to our files, the output folder and the genome index plus the fasta file of the genome.

The resulting files are bigwig and peak files, these can be used to determine differentially open chromatin regions and find activity at the TSS (transcription start site), plus determine a footprint (possible occupancy of the OCR by a TF protein).
The visualization of the results is done in IGV to manually inspect if our data are good.

ChiP-seq pipeline



bowtie2

Picard MarkDuplicates
Samtools filter low-quality

deepTools

MACS2

This pipeline is split in two scripts.
The first one aligns the reads with bowtie and produced a bigwig file to manually inspect the alignment.
The first pipeline requires as input a read file (single end or first file of a paired end sequencing, paired end sequencing does not improve much the results), the name of the files, and output folder and the genome index. This process has to be repeated for both the control and the antibody enriched sample (e.g., CTCF v control).
The second pipeline takes as input the two bam produced before (antibody-enriched_experiment.bam and the control.bam) plus basename and output folder. This pipeline will use MACS to find significantly enriched regions (and therefore peaks). This means that these peaks are statistically robust when we compare the number of reads mapping into them between the antibody sample and the control.

We can use these peaks files to perform a differential peak analysis and find peaks that are not appearing in different sample (e.g., cancer cell line vs normal cell line)