Analysis protocol for GENE-SWitCH WP5 ATAC-seq data (Influence of maternal diet on fetus and piglet transcriptome and epigenome)
Version of March 2022

Table of content

GENE-SWitCH WP5 pig ATAC-seq data were processed using the nf-core atacseq pipeline (https://nf-co.re/atacseq) version 1.2.1.

Pipeline Ressources

In this project 320 samples were assayed by ATAC-seq (2 technical replicates, also called libraries in the nf-core pipeline, by sample), at a sequencing depth of about 50M PE reads per sample. To generate peaks from these 320 samples, representing 12 biological conditions (one for each combination of tissue, developmental stage and mother diet), we used mapped reads from about 20-28 samples for each condition, thus requiring very large computing resources.

For this reason we had to change the following files from version 1.2.1 of the nf-core atacseq pipeline before running it (note: a better option is to only change the custom configuration file, but for the sake of reproducibility we report here exactly what was done):

- In `conf/base.config`:
  - `72.GB` was changed to `200.GB`
  - `16.h` was changed to `96.h`
- In `nextflow.config`:
  - `128.GB` was changed to `250.GB`
- In `main.nf`:
  - `label 'process_medium'` was changed to `label 'process_high'` for the following processes:

```
    - MERGED_LIB_BAM
    - MERGED_LIB_BAM_FILTER
    - MERGED_LIB_CONSENSUS_COUNTS
    - MERGED_LIB_CONSENSUS_DESEQ2
    - MERGED_REP_BAM
    - MERGED_REP_BIGWIG
    - MERGED_REP_MACS2
    - MERGED_REP_CONSENSUS_COUNTS
    - MERGED_REP_CONSENSUS_DESEQ2
```

- `label 'process_low'` was changed to `label 'process_high'` for the `MERGED_LIB_PRESEQ` process
- `label 'process_high'` was added to the `MULTIQC` process

Also, since running the differential analysis took more than 4 days, we decided to skip it using the `--skip_diff_analysis` option (see below)

Input files

Nextflow-run script

We used a specific in-house `nextflow-run` script that allows more flexible handling of nextflow parameters on the command line.

We downloaded it from its repository in the `<resdir>` folder and gave it execution permission using the following commands:

```
cd <resdir>
wget
https://gist.githubusercontent.com/chbk/2f9122538c5db222a822cfade05f81f4/raw/
63e0442914767a255f95b7d70ba2efa6afbadce0/nextflow-run
chmod +x nextflow-run
```

Custom configuration file

In order to properly use singularity and to let the pipeline know it is being run on a slurm cluster, we made this custom configuration `nextflow.config` file (passed on to the pipeline via the command line, see below) in the `<resdir>` directory:

```
singularity {
  enabled = true
  autoMounts = true
  runOptions = '-B /bank -B /work2 -B /work -B /save -B /home'
}
```

```
process {
  executor = 'slurm'
}
```

Sample sheet file

The sample sheet file we used as input to the pipeline can be found here: https://api.faang.org/files/nextflow_files/nextflow_spreadsheet/SSC_INRAE_GS_WP5_ATACseq_analysis_samplesheet.csv

It was stored in the <resdir> result directory.
It is a csv file including 641 rows (a header and 640 rows each representing pairs of fastq files from one of the two technical replicates, also called libraries in the nf-core pipeline, of the 320 samples assayed by ATAC-seq experiments) whose 3 first rows look like this:

```
group,replicate,fastq_1,fastq_2
liver_fetus_control,1,/work/project/geneswitch/data/reads/atacseq/sus_scrofa/wp5/complete/Liv-Fetus-109-1_CCAAGTCT-TCATCCTT-AHJK7WDSX2_L004_R1.fastq.gz,/work/project/geneswitch/data/reads/atacseq/sus_scrofa/wp5/complete/Liv-Fetus-109-1_CCAAGTCT-TCATCCTT-AHJK7WDSX2_L004_R2.fastq.gz
liver_fetus_control,1,/work/project/geneswitch/data/reads/atacseq/sus_scrofa/wp5/complete/Liv-Fetus-109-1_CCAAGTCT-TCATCCTT-BHMYNVDSX2_L004_R1.fastq.gz,/work/project/
geneswitch/data/reads/atacseq/sus_scrofa/wp5/complete/Liv-Fetus-109-1_CCAAGTCT-TCATCCTT-BHMYNVDSX2_L004_R2.fastq.gz
641 (1 fields)
```

For each row, the group represents the condition of the sample (<tissue_dvtstage_motherdiet>) and the replicate represents the sample number within this condition (from 1 to 28 for all conditions except for the two low fibre diet conditions, where it goes from 1 to 20). Note that this number is the same for the two technical replicates (or libraries) of a given sample.

Genome and annotation files

We used the 11.1 version of the pig genome and the 102th version of the ensembl annotation of this genome, both available on the ensembl ftp site. These two files were stored in our <datadir> data directory.

Running the pipeline

The content of the `sbatch.sh` bash script run on our slurm cluster is as follows, given a pipeline directory `<pipdir>`, a singularity directory `<singdir>`, a data directory `<datadir>` and a result directory `<resdir>`:

```
#!/bin/sh
#SBATCH -o <resdir>/output.out
#SBATCH -e <resdir>/error.out
#SBATCH -D <resdir>
#SBATCH --mem 8G

cd <resdir>
module load bioinfo/Nextflow-v20.10.0
export NXF_SINGULARITY_CACHEDIR=<singdir>
module load system/singularity-3.6.4
export SINGULARITY_PULLFOLDER=<singdir>
export SINGULARITY_CACHEDIR=<singdir>
export SINGULARITY_TMPDIR=<singdir>
./nextflow-run \
<pipdir>/main.nf \
--config nextflow.config \
--outdir <resdir> \
--input SSC_INRAE_GS_WP5_ATACseq_analysis_samplesheet.csv \
--fasta <datadir>/sus_scrofa.fa \
--gtf <datadir>/sus_scrofa.gtf \
--mito_name MT \
--macs_gsize 1341049888 \
--save_reference \
--min_reps_consensus 2 \
--skip_diff_analysis \
--resume
```

This bash script was then sent to our slurm cluster using the following command:

```
cd <resdir>
sbatch --mem=8G --cpus-per-task=1 -J wp5 --mail-user=<my.email>
--mail-type=END,FAIL --export=ALL -p <queue.name> sbatch.sh
```

Analysis files submitted to the FAANG DCC

Once the pipeline finished running, the following files were selected for submission to the FAANG DCC:

1. Mapped reads in each sample:

    a. 320 bam files located in `<resdir>/bwa/mergedLibrary/*bam`
2. Chromatin accessibility peaks called in each of the 12 conditions (a condition is a combination of tissue, developmental stage and mother diet, so a total of 2*2*3=12 conditions):
    a. 12 bed files located in `<resdir>/bwa/mergedReplicate/macs/broadPeak/*broadPeak`
3. Chromatin accessibility consensus peaks across all samples and conditions:
    a. 1 bed file located in `<resdir>/bwa/mergedReplicate/macs/broadPeak/consensus/consensus_peaks.mRp.clN.bed`