Analysis protocol for GENE-SWitCH small RNA-seq data (pig and chicken)
Version of November 29th 2021

Pig and chicken GENE-SWitCH small RNA-seq data (*miRNA profiling by high throughput sequencing*) were processed using the nf-core smrnaseq pipeline (https://github.com/nf-core/smrnaseq) version 1.1.0 with a small modification made to the `bin/edgeR_miRBase.r` script (see below).

For pig data on a slurm cluster, here are the actions that we performed :

1) **Make a proper code for the 1.1.0 version of the nf-core smrnaseq pipeline**
    a) *Download the pipeline version 1.1.0 code*

```
cd <codedir>
wget
https://github.com/nf-core/smrnaseq/archive/refs/tags/smrnaseq-1.1.0.tar.gz
tar -xvzf smrnaseq-1.1.0.tar.gz
rm smrnaseq-1.1.0.tar.gz
```

    b) *Edit the script `<codedir>/smrnaseq-1.1.0/bin/edgeR_miRBase.r`*

replace

```
nr_keep <- as.numeric(nr_keep[names(nr_keep) == TRUE])
```

by

```
nr_keep <- sum(nr_keep == TRUE)
```

This allows a gene to be associated to at least one read in each sample.

2) **Download the singularity image corresponding to the 1.1.0 version of the nf-core smrnaseq pipeline**
    a) *make a script called `launch.sing_pull_srnaseq.sh` with the following content:*

```
#!/bin/sh
cd <singdir>
module load system/singularity-3.6.4
export SINGULARITY_PULLFOLDER=<singdir>
export SINGULARITY_CACHEDIR=<singdir>
export SINGULARITY_TMPDIR=<singdir>
singularity pull docker://nfcore/smrnaseq:1.1.0
```

    b) *launch it on the slurm cluster with the following command:*

```
sbatch --mem=8G --cpus-per-task=1 -J sing.pull --workdir=$PWD --export=ALL -p
<queuename> launch.sing_pull_srnaseq.sh
```

3) **Launch the nf-core smrnaseq pipeline on the 84 pig GENE-SWitCH small rnaseq experiments (only read 1 were used)**
    a) *make a bash script called `launch.nf.gs.allsamples.sus_scrofa.sh` with the following content:*

```
#!/bin/sh
cd <resdir>
```

```
wget
https://gist.githubusercontent.com/chbk/2f9122538c5db222a822cfade05f81f4/raw/
63e0442914767a255f95b7d70ba2efa6afbadce0/nextflow-run
chmod +x nextflow-run
module load bioinfo/Nextflow-v21.04.1
module load system/singularity-3.7.3
export SINGULARITY_PULLFOLDER=<singdir>
export SINGULARITY_CACHEDIR=<singdir>
export SINGULARITY_TMPDIR=<singdir>
./nextflow-run <codedir>/smrnaseq-1.1.0/main.nf --input
'<readdir>/*_1.fastq.gz' --fasta <datadir>/Sscrofa11.1/sus_scrofa.fa
--mirtrace_species ssc -profile singularity --protocol illumina --outdir
<resdir> --max_memory '16.GB' --max_cpus 2 --email <email> -resume >
nextflow.out 2> nextflow.err
```

The input files are the following:

- `<readdir>/*_1.fastq.gz` are read 1 files of all samples (there are 84 of them)
- `<datadir>/Sscrofa11.1.102/sus_scrofa.fa` is the pig genome assembly 11.1

b) launch it on the slurm cluster with the following command:

```
sbatch --mem=16G --cpus-per-task=1 -J nf.gs.all.ssc --workdir=$PWD
--export=ALL -p <queuename> launch.nf.gs.allsamples.sus_scrofa.sh
```

4) Submitted analysis files

a) Read mapping to the `mirbase` mature sequences

Those are 84 bam files that were located in the `<resdir>/bowtie/miRBase_mature` directory

b) Read mapping to the `mirbase` hairpin sequences

Those are 84 bam file that were located in the `<resdir>/bowtie/miRBase_hairpin` directory

c) `Mirbase` mature normalized expression in all samples

This is a single tsv file with header that has mature sequences as rows and CPM normalized expression in each sample as columns and that was located in the `<resdir>/edgeR/miRBase_mature` directory

d) `Mirbase` hairpin normalized expression in all samples

This is a single tsv file with header that has hairpin sequences as rows and CPM normalized expression in each sample as columns and that was located in the `<resdir>/edgeR/miRBase_hairpin` directory

e) Known and novel miRNAs found by `mirdeep2`

This is a single bed file that includes all the known and novel miRNAs found by mirdeep2 in all samples and with an <<estimated probability that the miRNA is a true positive>> above 75% in a sample. This file was made out of the 84 files in `<resdir>/mirdeep2/mirdeep/*_collapsed.csv`