Analysis protocol for GENE-SWitCH WP2 ATAC-seq chicken data

Chicken GENE-SWitCH ATAC-seq data were processed using the nf-core atacseq pipeline (https://github.com/nf-core/atacseq) version 1.2.1, on a Kubernetes cluster.

This infrastucture required modifications in the default configuration of the pipeline.

1. Execution environment

The pipeline is run on a docker pod running a Docker image based on `nextflow/nextflow:21.04.1`

The image is accessible here : https://gitlab.com/geneswitch/nextflow

2. Pull the pipeline, download the data and prepare the configuration file

2.1 Pull the pipeline

```
codeDir=/data/pipelines
cd $codeDir
export NXF_ASSETS=$codeDir
nextflow pull nf-core/atacseq -r 1.2.1
```

2.2 Download the data

- Reference genome is GRCg6a, and can be downloaded here. It is stored on the cluster at `/data/ref/gallus_gallus/gallus_gallus.fa`
- Reference annotation is Ensembl version 102 and can be downloaded here. It is stored at `/data/ref/gallus_gallus/gallus_gallus.gtf`
- Samplesheet metadata file for the WP2 samples can de downloaded here and is stored as `/data/metadata/atacseq_nf-core-design-chicken_4cols.csv`

Configuration file

We need to replace the default configuration file, because the check_max function does not work on kubernetes environments (see this issue : https://github.com/nf-core/methylseq/issues/158)

A suitable configuration file with this problem fixed can be downloaded here and saved as `/data/pipelines/k8s.config`

3. Create submission scripts and launch the pipeline

3.1 Nextflow script

On the cluster, prepare the bash script `/data/code/atacseq/run_atacseq_chicken.sh` for the pipeline :

```
#!/bin/bash

export NXF_ASSETS=/data/pipelines/

nextflow run nf-core/atacseq \
        -profile docker \
        -revision 1.2.1 \
        -c /data/pipelines/nf-core/atacseq/nextflow.config \
        -c /data/pipelines/k8s.config \
        --max-cpus 7 --max-memory 31GB \
        -resume \
        --fasta /data/ref/gallus_gallus/gallus_gallus.fa \
        --gtf /data/ref/gallus_gallus/gallus_gallus.gtf \
        --input  /data/metadata/atacseq_nf-core-design-chicken_4cols.csv \
        --mito_name MT --macs_gsize 897512367 \
        --save_reference \
        --keep-temp \
        --outdir /data/run/atacseq/gallus_gallus/
```

3.2 Kubernetes job file

Locally, prepare the following job file called `atacseq_chicken_job.yaml` :

```
kind: Job
apiVersion: batch/v1
metadata:
  name: atacseq-chicken
spec:
  backoffLimit: 1
  template:
    spec:
      restartPolicy: Never
      containers:
        - name: container
          image: registry.gitlab.com/geneswitch/nextflow
          command: ["/bin/bash"]
          workingDir: /data/run/atacseq/gallus_gallus/
          args: ["-c", "/data/code/atacseq/run_atacseq_chicken.sh"]
          env:
            - name: NXF_ASSETS
              value: /data/pipelines
          volumeMounts:
            - name: volume
              mountPath: /data
      volumes:
        - name: volume
          persistentVolumeClaim:
            claimName: data
```

3.3 Submit the job

```
kubectl create -f atacseq_chicken_job.yaml
```

4. Results submission

The following results in `/data/run/atacseq/gallus_gallus/` are submitted to ENA :

- bam files : `bwa/mergedLibrary/*.mLb.clN.sorted.bam`
- peaks : `bwa/mergedReplicate/macs/broadPeak/*.mRp.clN_peaks.gappedPeak`
- consensus peaks : `bwa/mergedReplicate/macs/broadPeak/consensus/consensus_peaks.mRp.clN.featureCounts.txt`