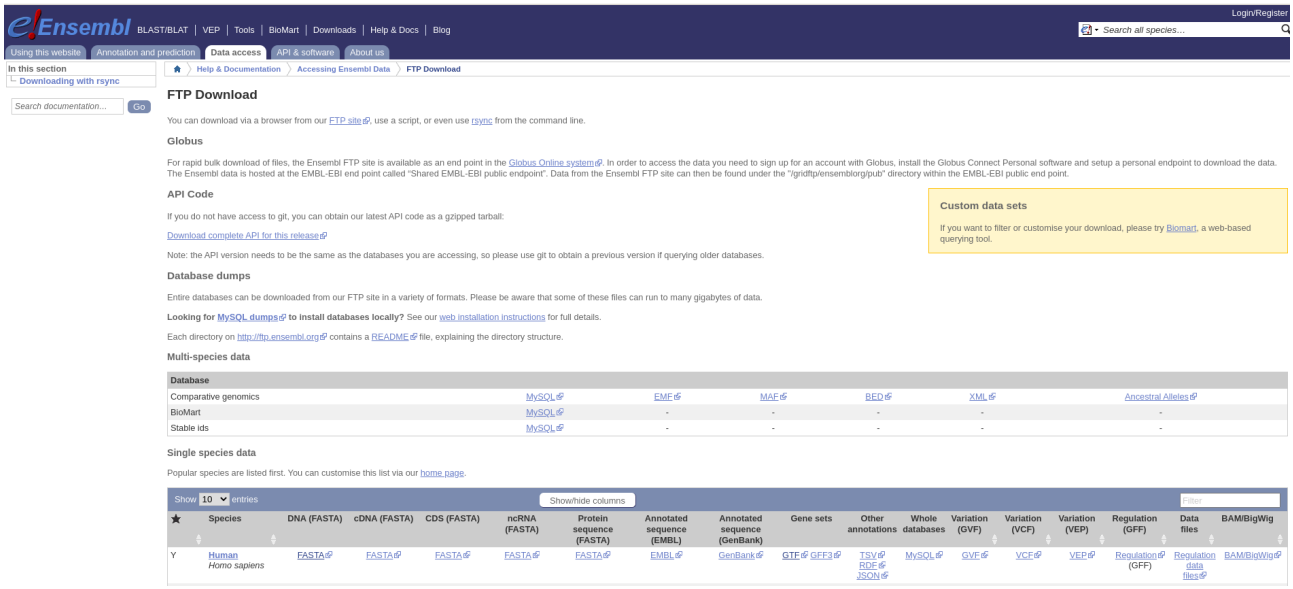
Lancement d’un pipeline Nextflow nf-core rnaseq

Contexte biologique

Références

- Génome et annotation (transcriptome GTF)
  http://www.ensembl.org/info/data/ftp/index.html



- Pour la formation, ce TP et les données de formation sont disponibles via un wget :

  Génome :

  http://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.fasta

  Transcriptome:

  http://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.gtf

Plan d'expérience : fichiers - réplicats - conditions

Descriptif des échantillons : nom du groupe, numéro du réplicat, path FASTQ R1, path FASTQ R2, foward/reverse/unstranded

NB : Remplacer *path/*to/ par le chemin d’accès à vos répertoires/fichiers de votre compte de formation.

Par exemple:

```
$ more part1_sample_sheet_V2.csv

group,replicate,fastq_1,fastq_2,strandedness,sample_code_barre,animal,tissue,sexe,maturity,TG,dg
```

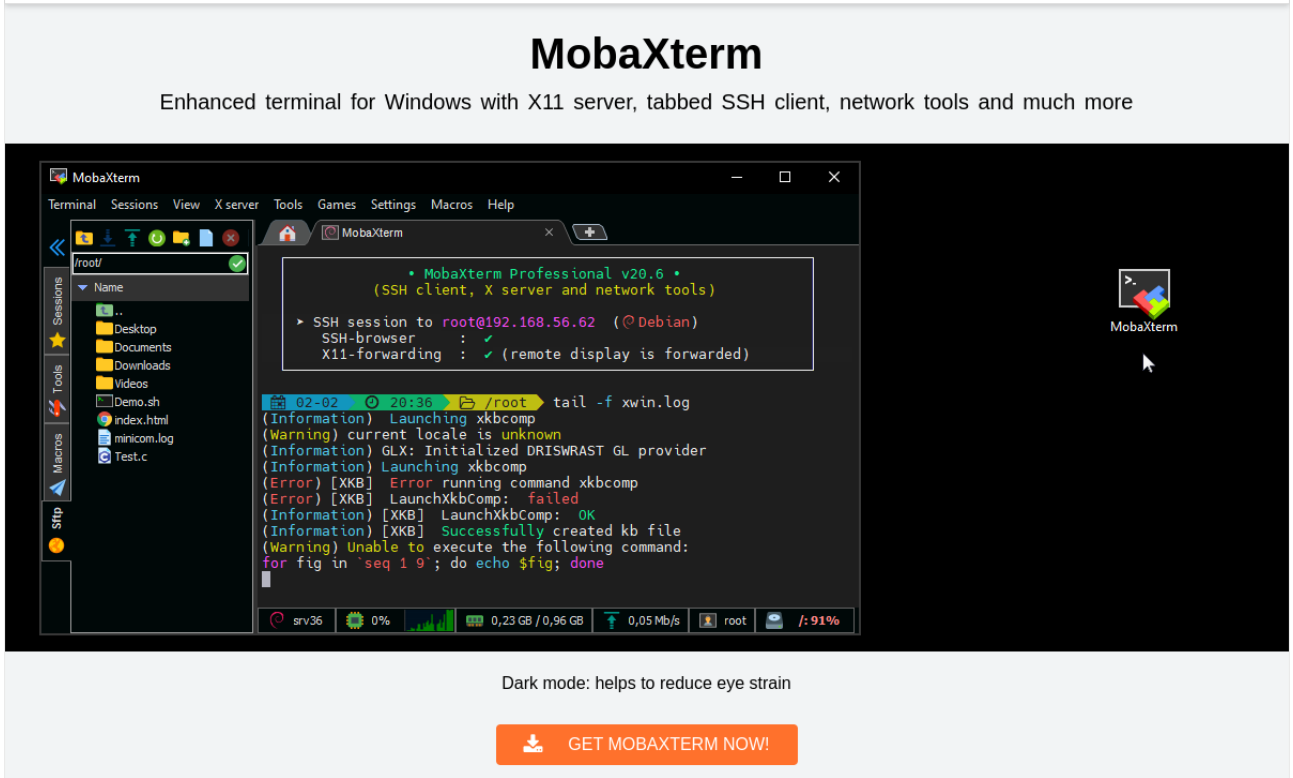
E_L1_90_LL_F_M-,1,/work/user/anemone/FASTQ/22_R1.fastq.gz,/work/user/anemone/FASTQ/22_R2.fastq.gz,reverse,0233078348,Foetus896,endometrium,F,M-,LW,90j

E_L1_90_LM_M_M+,1,/work/user/anemone/FASTQ/23_L001_R1.fastq.gz,/work/user/anemone/FASTQ/23_R2.fastq.gz,reverse,0233078321,Foetus964,endometrium,M,M+,LwxMS,90j

Traitements bioinformatiques

Préparation de l'espace de travail

1. Ouverture de votre terminal ou téléchargement de MobaXterm https://mobaxterm.mobatek.net/



2. Principales commandes Linux:

```
cd /work/user/aster
touch toto
rm -rf toto
ls
touch README
geany READMY &
more README
mkdir FASTQ
ls -ltrah
cd FASTQ
wget http://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/test-data-galaxy/1.fastq
more 1.fastq
mv 1.fasta reference.fasta
mkdir genome
mv /work/user/aster/FASTQ/reference.fasta genome/.
```

3. Connexion aux comptes de formation:

Exemple d'host pour MobaXterm: aster@genobioinfo.toulouse.inrae.fr

```
(base) [smaman@localhost ~]$ ssh -XY aster@genobioinfo.toulouse.inrae.fr
aster@genologin.toulouse.inra.fr's password:
aster@genologin1 ~ $
aster@genologin1 ~ $ cd /work/user/aster/
aster@genologin1 /work/user/aster $
```

4. Récupération du génome et de l'annotation:

```
cd  /path/to/NEXTFLOW/
mkdir /path/to/NEXTFLOW/genome/ ; cd /path/to/NEXTFLOW/genome/
wget https://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.fasta
mkdir /path/to/NEXTFLOW/annotation/ ; cd /path/to/NEXTFLOW/annotation/
wget https://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/ITAG2.3_genomic_Ch6.gtf
```

5. Récupération des séquences:

```
mkdir /path/to/NEXTFLOW/FASTQ/ ; cd /path/to/NEXTFLOW/FASTQ/
wget https://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep1_1_Ch6.fastq.gz
wget https://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/MT_rep1_2_Ch6.fastq.gz
wget https://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/WT_rep1_1_Ch6.fastq.gz
wget https://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/WT_rep1_2_Ch6.fastq.gz
```

Votre pipeline Nextflow

Pour lancer un pipeline sur le cluster de calcul BioInfo Genotoul, nous préparons 3 fichiers:

1/ Un fichier de lancement en sbatch

2/ Un fichier de configuration qui surcharge le fichier en local.

3/ Un fichier de description des échantillons

Fichier de configuration

```
$ touch sm_config.cfg
$ more sm_config.cfg
trace {
    enabled = true
    file = 'pipeline_trace.txt'
    fields = 'task_id,name,status,exit,realtime,%cpu,rss,script'
}
```

Le fait de rajouter ce module «trace» permet de récupérer les lignes de commande complètes lancées à chaque étape du pipeline. Voici un exemple:

```
$ more pipeline_trace.txt

task_id     name   status exit   realtime     %cpu   rss    script
3    RNASEQ:INPUT_CHECK:SAMPLESHEET_CHECK (part1_sample_sheet_V2.csv)  COMPLETED 0  1s
     24.4%  1 MB
5    RNASEQ:CAT_FASTQ (E_L1_110_LL_F_M+_R1)  COMPLETED    0     18ms   12.3%  0
ln -s 0233078320_CCGTGAAG-ATCCACTG-AHV5H7DSXY_L001_R1.fastq.gz
E_L1_110_LL_F_M+_R1_1.merged.fastq.gz
ln -s 0233078320_CCGTGAAG-ATCCACTG-AHV5H7DSXY_L001_R2.fastq.gz
E_L1_110_LL_F_M+_R1_2.merged.fastq.gz
```

Fichier de description des échantillons tests

```
$ more inputs.csv
group,replicate,fastq_1,fastq_2,strandedness
mutant,1,/path/to/data/MT_rep1_1_Ch6.fastq.gz,path/to/data/
MT_rep1_2_Ch6.fastq.gz,unstranded
wild,1,/path/to/data/WT_rep1_1_Ch6.fastq.gz,path/to/data/
WT_rep1_2_Ch6.fastq.gz,unstranded
```

Fichier de lancement du pipeline

```
$ more run_pipeline.sh
#!/bin/bash
#SBATCH -J nfcorernaseq
#SBATCH -p unlimitq
#SBATCH --mem=6G

module purge
module load  bioinfo/nfcore-Nextflow-v21.04.1

input=/path/to/inputs.csv
gtf=/path/to/annotation/nameToComplete.gff
fasta=/path/to/genome/nameToComplete.fna
config=/path/to/config/file

nextflow run nf-core/rnaseq -r 3.0 -profile genotoul --input $input --fasta $fasta --
gtf $gtf --aligner star_rsem  -c $config
```

PS : Peut-être que la révision sera à mettre à jour.

Attention, la version 22.12.0 pas compatible avec code Nextflow DSL1

Lancement du pipeline

```
$ sbatch run_pipeline.sh
Submitted batch job 27335211
$ ls
annotation  FASTQ  genome  inputs.csv  inputs_test.csv  README  run_pipeline.sh
sm_config.cfg
```

Pour suivre l'état du job :

```
$ seff 640143

Job ID: 640143

Cluster: genologin

User/Group: smaman/SIGENAE

State: RUNNING

Cores: 1

CPU Utilized: 00:00:00

CPU Efficiency: 0.00% of 00:36:46 core-walltime

Job Wall-clock time: 00:36:46

Memory Utilized: 0.00 MB (estimated maximum)

Memory Efficiency: 0.00% of 20.00 GB (20.00 GB/node)

WARNING: Efficiency statistics may be misleading for RUNNING jobs.
```

Pour vérifier s'il y a une erreur dans le log sbatch:

```
grep --color -i "error" slu*
```

Analyse des résultats

```
$ ls results/
fastqc/  genome/  multiqc/  pipeline_info/  star_rsem/  trimgalore/
```


fastqc/


multiqc/


pipeline_info/


pipeline_trace.txt


slurm-27488902.out


star_rsem/


trimgalore/

En détails:

```
results/fastqc:
1_R1_fastqc.html
1_R1_fastqc.zip

results/genome:
rsem/  ref.fa.fai  ref.fa  annotation_genes.gtf

results/multiqc:
star_rsem/

results/pipeline_info:
execution_report.html  execution_timeline.html  pipeline_report.html
pipeline_report.txt  samplesheet.valid.csv  software_versions.csv

results/star_rsem:
bigwig/
deseq2_qc/
dupradar/
featurecounts/
picard_metrics/
preseq/
qualimap/
rseqc/
samtools_stats/
stringtie/
rsem.merged.gene_counts.tsv        rsem.merged.gene_tpm.tsv
rsem.merged.transcript_counts.tsv  rsem.merged.transcript_tpm.tsv

results/trimgalore:
...._trimming_report.txt
```

Analyse de multiQC report

results/multiqc/star_rsem/multiqc_report.html