

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>

FTP Download

You can download via a browser from our [FTP site](#), use a script, or even use [rsync](#) from the command line.

Globus

For rapid bulk download of files, the Ensembl FTP site is available as an end point in the [Globus Online system](#). In order to access the data you need to sign up for an account with Globus, install the Globus Connect Personal software and setup a personal endpoint to download the data. The Ensembl data is hosted at the EMBL-EBI end point called "Shared EMBL-EBI public endpoint". Data from the Ensembl FTP site can then be found under the "gridftp/ensemblorg/pub" directory within the EMBL-EBI public end point.

API Code

If you do not have access to git, you can obtain our latest API code as a zipped tarball:
[Download complete API for this release](#)

Note: the API version needs to be the same as the databases you are accessing, so please use git to obtain a previous version if querying older databases.

Database dumps

Entire databases can be downloaded from our FTP site in a variety of formats. Please be aware that some of these files can run to many gigabytes of data.

Looking for [MySQL dumps](#) to install databases locally? See our [web installation instructions](#) for full details.

Each directory on <http://ftp.ensembl.org> contains a [README](#) file, explaining the directory structure.

Multi-species data

Database	MySQL	EMF	MAE	BED	XML	Ancestral Alleles
Comparative genomics	MySQL	-	-	-	-	-
BioMart	MySQL	-	-	-	-	-
Stable ids	MySQL	-	-	-	-	-

Single species data

Popular species are listed first. You can customise this list via our [home page](#).

Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAMBigWig
Human <i>Homo sapiens</i>	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF SS20	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAMBigWig

- Pour la formation, ce TP et les données de formation sont disponibles via un wget sur
<http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/>

Génome :

http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/ITAG2.3_genomic_Ch6.fasta

Transcriptome:

http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/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

Exemple:

```
$ more part1_sample_sheet_V2.csv  
  
group,replicate,fastq_1,fastq_2,strandedness,sample_code_barre,animal,tissue,sexe,maturity,T  
G,dg  
  
E_L1_90_LL_F_M-,1,/path/to/FASTQ/22_R1.fastq.gz,/path/to/FASTQ/  
22_R2.fastq.gz,reverse,0233078348,Foetus896,endometrium,F,M-,LW,90j
```

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

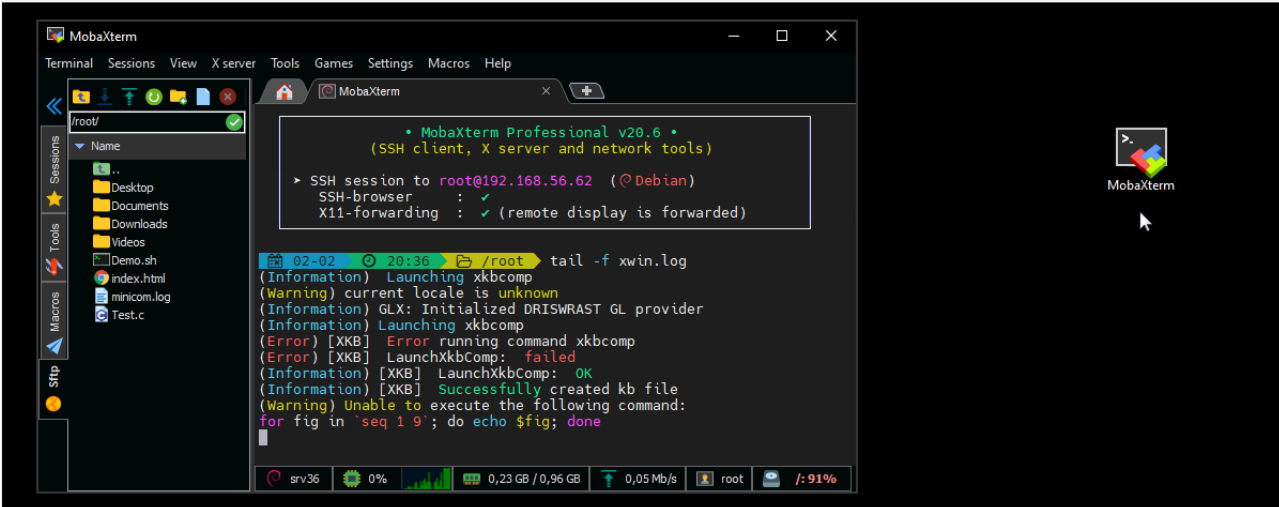
Traitements bioinformatiques

Préparation de l'espace de travail

1. Ouverture de votre terminal ou téléchargement de [MobaXterm](https://mobaxterm.mobatek.net/)
<https://mobaxterm.mobatek.net/>

MobaXterm

Enhanced terminal for Windows with X11 server, tabbed SSH client, network tools and much more



Dark mode: helps to reduce eye strain

[GET MOBAXTERM NOW!](#)

2. Principales commandes Linux:

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

3. Connection aux comptes de formation:

Les comptes suivants:

anemone arome aster bleuete camelia capucine chardon clematite cobee coquelicot cosmos

cyclamen dahlia digitale geranium gerbera glaieul hortensia iris jacinthe
ont été réservés pour la formation UPS du 10 au 17 Septembre 2021.

Exemple d'host pour MobaXterm: aster@genologin.toulouse.inra.fr

```
(base) [smaman@localhost ~]$ ssh -XY aster@genologin.toulouse.inra.fr
aster@genologin.toulouse.inra.fr's password:
Last login: Tue Jul 27 14:08:06 2021 from 147.100.120.100

aster@genologin1 ~ $
aster@genologin1 ~ $ cd /work/aster/
aster@genologin1 /work/aster $
```

4. Récupération du génome et de l'annotation:

```
cd /path/to/NEXTFLOW/
mkdir /path/to/NEXTFLOW/genome/ ; cd /path/to/NEXTFLOW/genome/
wget
http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/ITAG2.3\_genomic\_Ch6.fasta

mkdir /path/to/NEXTFLOW/annotation/ ; cd /path/to/NEXTFLOW/annotation/
wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/ITAG2.3\_genomic\_Ch6.gtf
```

5. Récupération des séquences:

```
mkdir /path/to/NEXTFLOW/FASTQ/ ; cd /path/to/NEXTFLOW/FASTQ/
wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/MT\_rep1\_1\_Ch6.fastq.gz
wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/MT\_rep1\_2\_Ch6.fastq.gz
wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/WT\_rep1\_1\_Ch6.fastq.gz
wget http://genoweb.toulouse.inra.fr/~sigenae/sarah/UPS/NEXTFLOW/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

```
$ 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

```
aster@genologin1 /work/aster $ more run_pipeline.sh
#!/bin/bash
#SBATCH -J nfcorernaseq
#SBATCH -p unlimitq
#SBATCH --mem=6G

module purge
module load bioinfo/nfcore-Nextflow-v20.11.0-edge/

input=/path/to/inputs.csv
gtf=/path/to//annotation/GCF_013265735.2_USDA_Omyka_1.1_genomic.gff
fasta=/pathe/to/genome/GCF_013265735.2_USDA_Omyka_1.1_genomic.fna
config=/path/to/sm_config.cfg

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

Lancement du pipeline

```
aster@genologin1 /work/aster $ sbatch run_pipeline.sh
Submitted batch job 27335211
aster@genologin1 /work/aster $ ls
annotation FASTQ genome inputs.csv inputs_test.csv README run_pipeline.sh
sm_config.cfg
```

Pour suivre l'état du job :

```
$ seff 27600449

Job ID: 27600449
Cluster: genobull
User/Group: galaxy-prod/wbioinfo
State: RUNNING
Nodes: 2
Cores per node: 3
CPU Utilized: 00:00:00
CPU Efficiency: 0.00% of 5-02:28:36 core-walltime
Job Wall-clock time: 20:24:46
Memory Utilized: 0.00 MB (estimated maximum)
Memory Efficiency: 0.00% of 205.08 GB (34.18 GB/core)
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" slurm*
```

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

← → ↻ ⚠ Non sécurisé | genoweb.toulouse.inra.fr/~sigenae/sarah/CO-LOCATION/part1_0907202



General Stats

STAR_RSEM DESeq2 sample similarity

STAR_RSEM DESeq2 PCA plot

Biotype Counts

DupRadar

Picard

Preseq

QualiMap

Genomic origin of reads

Gene Coverage Profile

Rsem

Mapped Reads

Multimapping rates

RSeQC

Read Distribution

Inner Distance

Read Duplication

E_L1_110_LL_M_M- R3	139.2	0
E_L1_110_LL_M_M- R3_1		
E_L1_110_LL_M_M- R3_2		
E_L1_110_LL_M_N R1	118.9	0

STAR_RSEM DESeq2 sc

is generated from clustering by Euclidean distances by

Sort by highlight

DESeq2: Heatmap of the sample-to

