

Présentation du cluster de calculs intensifs de Genotoul : **GENOBIOINFO** (ex genologin)

Utilisation de quelques commandes basiques

File system

Les outils de base (find, sort, cut, grep ...)

Transférer et compresser des fichiers

Présentation de la capacité de calcul du cluster

Connexion au serveur bioinfogenotoul

Soumettre un job et le gérer

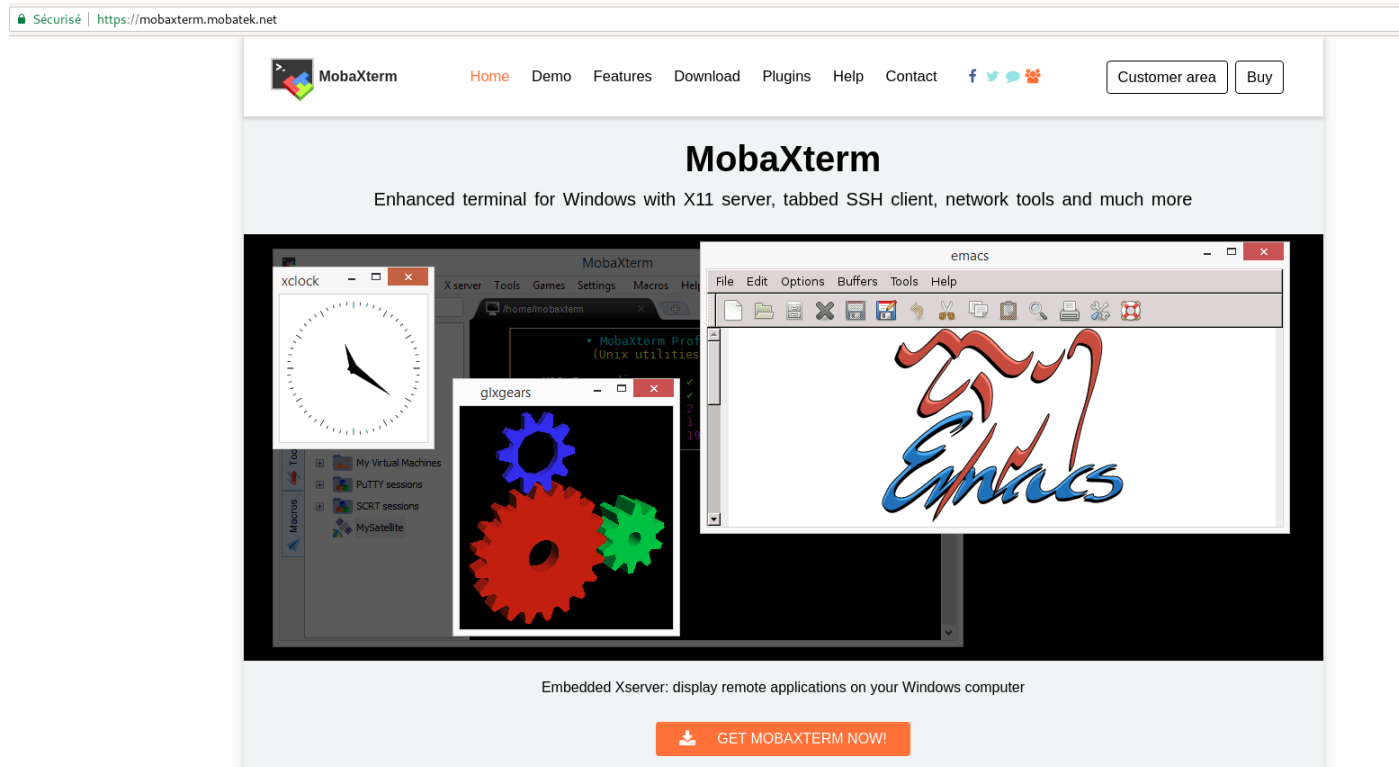


Linux





- Xming (Windows graphic)
- Putty (connection)
- <https://mobaxterm.mobatek.net/>



Source: <http://bioinfo.genotoul.fr/>

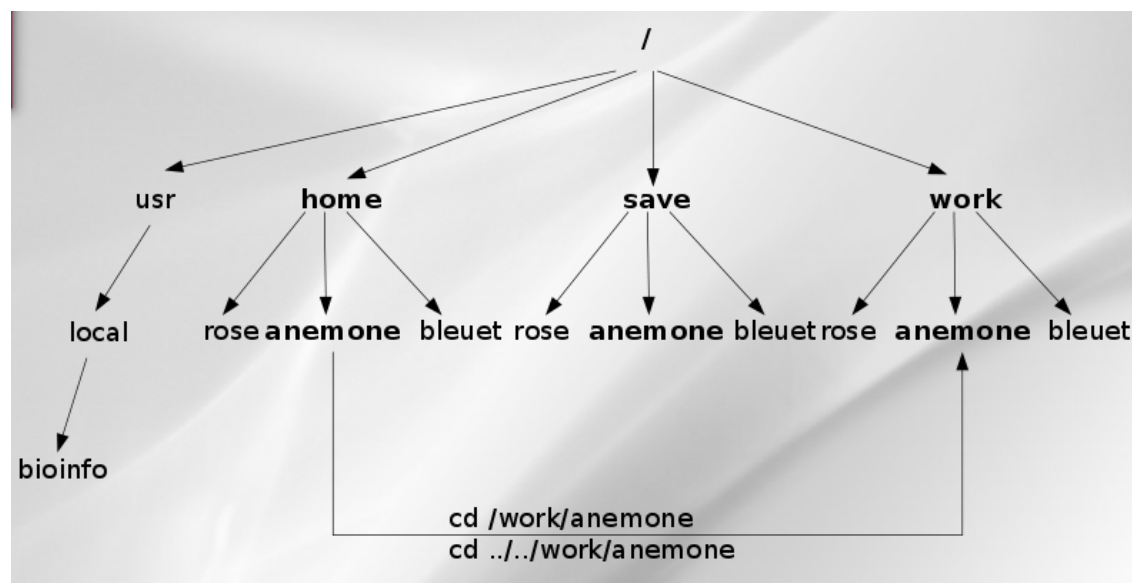
`cd` : change directory

`pwd` : print working directory

`ls [nom_répertoire]` : list directory contents

`passwd` : update user's authentication token

`history` : display the commands history



```
mkdir directory_name  
touch new_file_name  
cp file_source file_destination  
cp -r dir_source dir_destination  
mv file dir/.  
mv old_file_name new_file_name  
ls *.txt  
find . -type f *.txt  
cat file - more file - less file  
geany - gedit - vi
```



Comprendre le résultat d'un ls -ltrah

```
lrwxrwxrwx 1 cnoirot BIOINFO 3683591 Jun  9 11:56 Diapo_F10a.odp
drwxr-xr-x 3 cnoirot BIOINFO 4096 Jul  8 14:56 igv
-rwxr-xr-x 1 cnoirot BIOINFO 20 Apr 16 11:21 monscript.sh
-rw-r--r-- 1 cnoirot BIOINFO 954415 Oct  3 2009 Presentation_pyrocleaner.odp
lrwxrwxrwx 1 cnoirot BIOINFO 13 Mar 15 2009 save -> /save/cnoirot
lrwxrwxrwx 1 cnoirot BIOINFO 13 Mar 18 2009 work -> /work/cnoirot
```

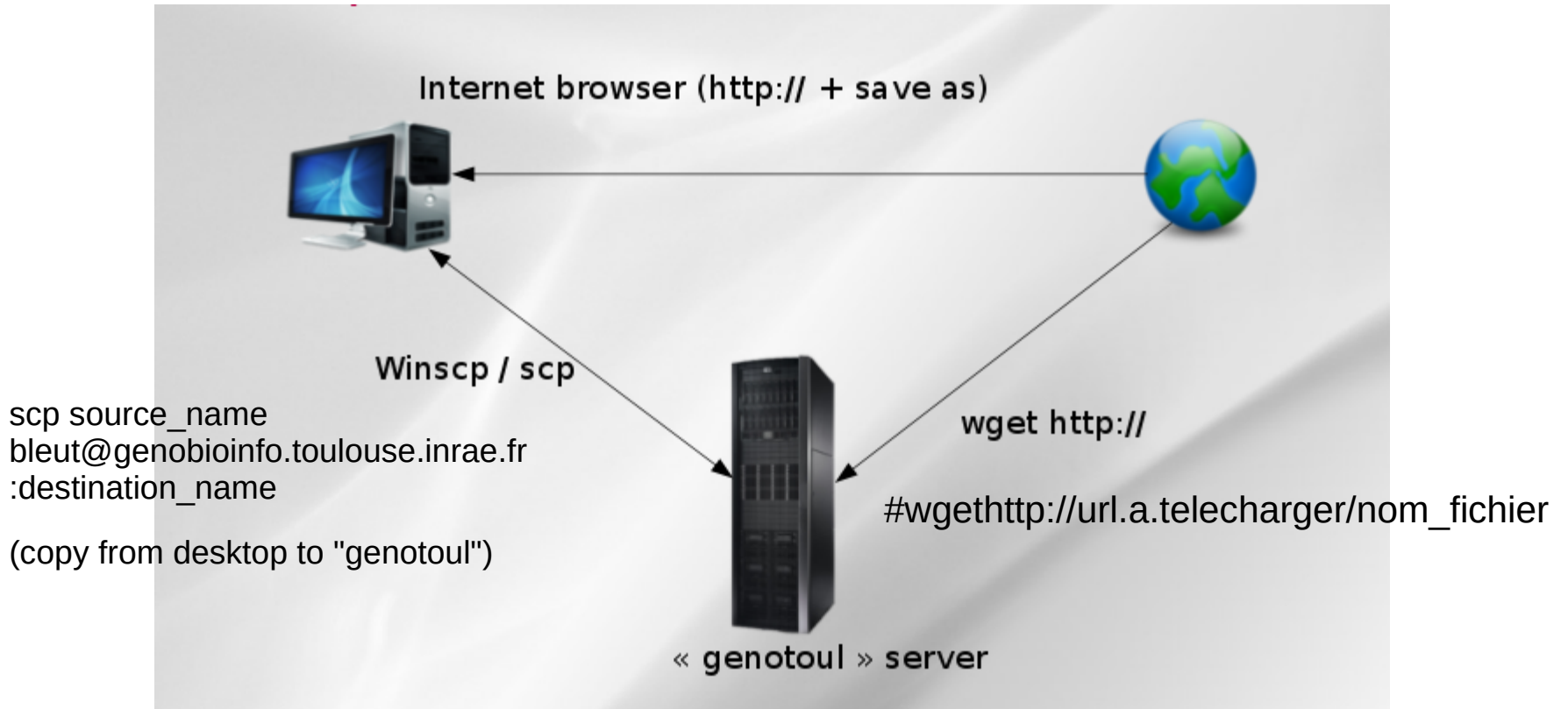
Changer les permissions avec chmod

chmod [options] filename

modifies the permissions of a file

→ **u** : user, **g** : group, **o** : other, **a** : all

→ **r** : read, **w** : write, **x** : execute



Ou **FILEZILLA**

Exercice en autonomie : Jouer avec vos fichiers

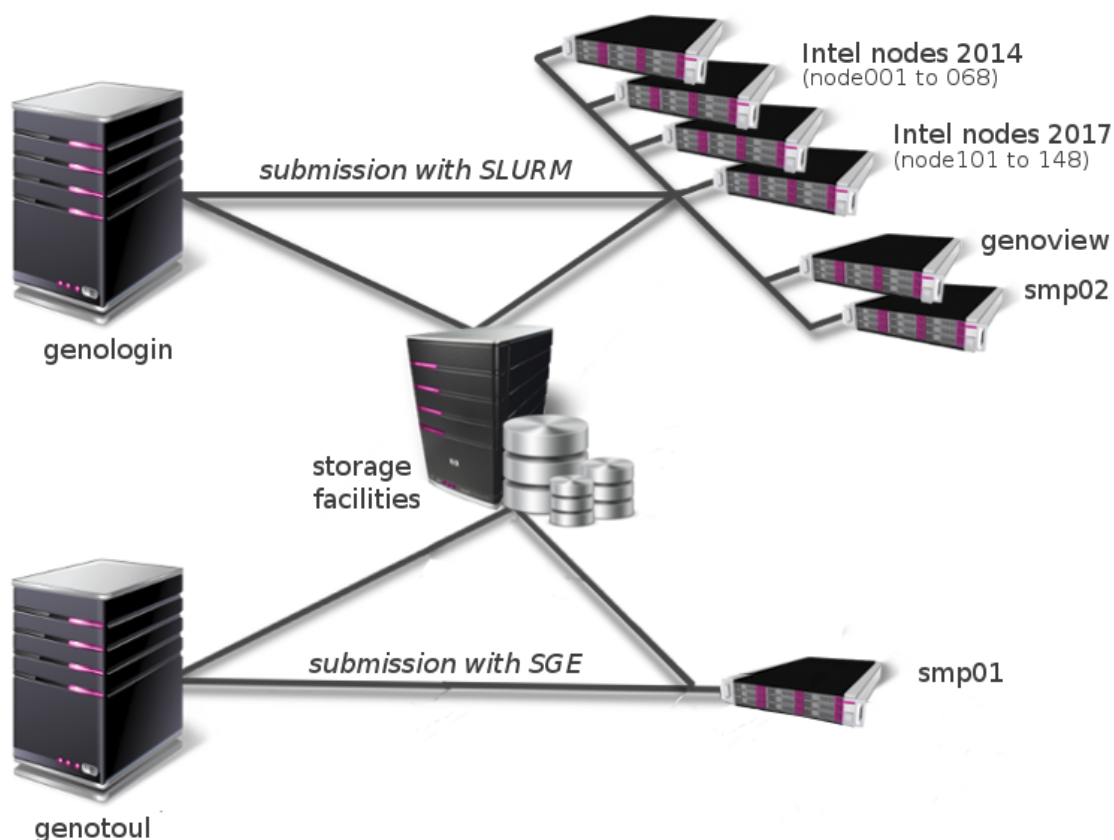
- 1 – Créer un fichier fasta nommé VosInitiales.fasta (rechercher un exemple sur Google ou inventez)
- 2 – Compter le nombre de lignes de VosInitiales.fasta et écrire le résultat dans un fichier VosInitiales_nbLines.txt, à l'aide de la commande >
- 3 – Récupérer les 20 premières lignes de votre fasta dans un fichier VosInitiales_20.fasta
- 4 – Partagez moi vos fichiers par mail (sarah.maman@inrae.fr) pour validation.

Se connecter au cluster Genologin

Connection en ssh sur les noeuds de Genologin .

Si besoin, demande de compte : <http://bioinfo.genotoul.fr/>

Vous avez ainsi accès à des ressources de stockage et de calculs.



Source: <http://bioinfo.genotoul.fr/>



```
smaman@smaman:~$ ssh -XY smaman@genobioinfo.toulouse.inrae.fr
smaman@genobioinfo.toulouse.inrae.fr's password:
Last login: Thu Aug 31 10:08:44 2023 from 92.88.171.166
```

```
=====
Bienvenue sur les serveurs d'accès de la plate-forme bioinformatique
=====
```

```
=====
Bienvenue sur les serveurs d'accès de la plate-forme bioinformatique
=====
```

Les serveurs "genologin" ne sont PAS destinés à lancer des traitements de données mais **uniquement à soumettre des jobs sur le cluster de calcul.**

Pour cela, utiliser les commandes "sbatch" (soumission de scripts) ou "srun" (interactif).

```
=====
Information concernant les espaces disques
=====
```

La partition **/home** est réservée aux **fichiers de configuration** (quota utilisateur **100Mo**)

La partition **/save** est réservée pour les **données à sauvegarder** (quota utilisateur **250Go**)

La partition **/work** est réservée pour les **fichiers temporaires de calcul** (quota utilisateur **1To**)

=> ATTENTION CE VOLUME N'EST PAS SAUVEGARDE ET PEUT ÊTRE PURGÉ SI BESOIN

=> IL EST DE **VOTRE RESPONSABILITÉ DE GÉRER VOS DONNÉES** (organisation, volumétrie, pertinence, ancienneté)

Pour connaître votre consommation d'espace disque, utilisez la commande suivante :

```
# du -csh /DIR_NAME/USER_NAME/*
```



=====

Informations concernant l'utilisation de la memoire

=====

Il existe une **limitation par défaut de 4Go de RAM par process utilisateur** (sur le cluster).

Pour obtenir plus de memoire, veuillez consulter la FAQ de notre site web (cf + bas).

=====

Informations concernant le quota de temps de calcul

=====

Il existe un **quota de temps de calcul annuel de 100.000H pour les academiques (500H pour les entreprises privees)**.

Au dela il faut renseigner le formulaire de demande de ressources exceptionnelles sur notre site web.

Vous pouvez verifier votre quota de calcul avec la commande: `squota_cpu`

=====

Support

=====

Pour plus d'informations, consulter le site web :

<http://bioinfo.genotoul.fr/>

Pour toute demande de support, adressez-vous a :

support.bioinfo.genotoul@inra.fr

=====

Informations sur le compte smaman (23-09-2019 00:09)

=====

Home: 38.21M (3%)

Work: 656G (64%)

Save: 175.947G (70%)

CPU : 1014h (0%)

Expiration du compte: 08-03-2029

Expiration du mot de passe: 28-01-2020

Se connecter sur les noeuds de calculs intensifs

Pas d'accès direct en ssh sur les noeuds de calculs

Car il y a une file d'attente pour la gestion des priorités/ressources par GRID SCHEDULER.

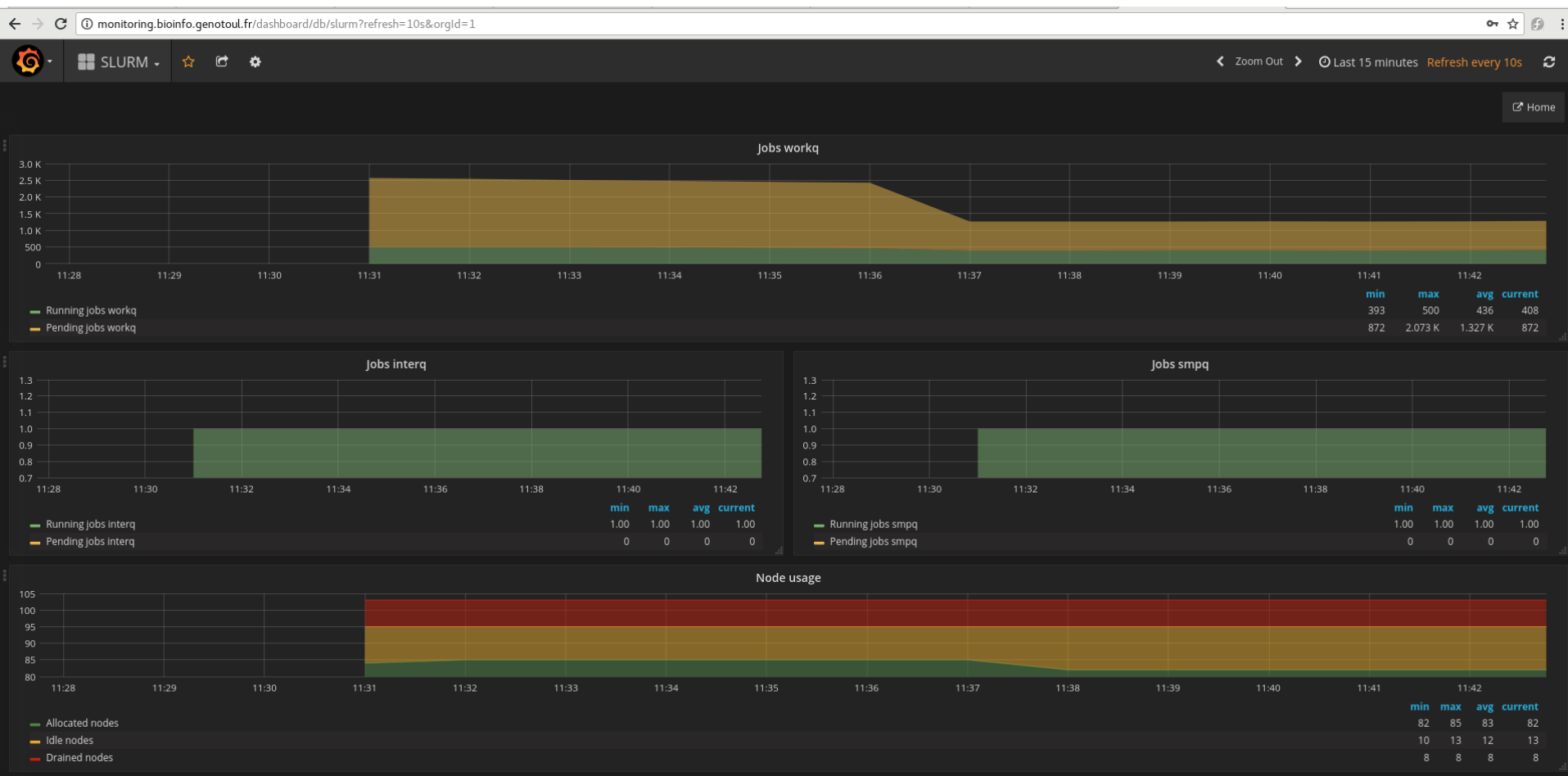


Donc, d'abord se connecter en ssh pour un accès au frontal
Puis lancer des calculs/jobs en sbatch UNIQUEMENT SUR VOTRE /WORK et enfin
exit pour sortir du cluster

Disk spaces

/usr /local/bioinfo/	Bioinformatics Software
/bank/	International genomics Databanks
/home/	User configuration files (ONLY) (100 MB user quota)
/save/	User disk space (with BACKUP) (250 GB user quota)
/work/	HPC TEMPORARY disk space (1 TB user quota)

Monitoring des ressources SLURM : GRAPHANA (exemple)



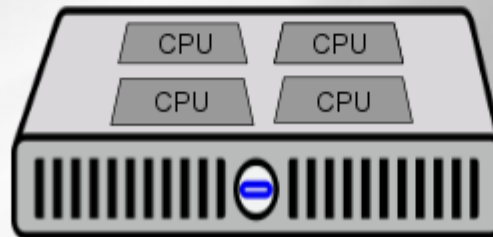
A montrer directement sur bioinfo Genotoul :
<http://bioinfo.genotoul.fr/index.php/resources-2/monitoring/>

Source: <http://bioinfo.genotoul.fr/>

- Cluster : set of nodes



- Node : Huge computer (with several CPUs)

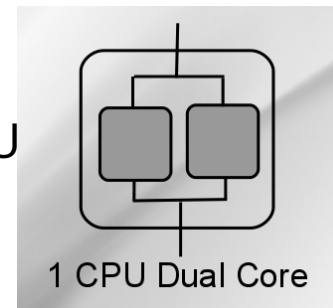


08/03/16

1 cluster = une série de noeuds

1 noeud = un énorme ordinateur de plusieurs CPU

1 CPU = Central Processing Unit = 2 coeurs



Plusieurs queues de traitement sont disponibles selon les ressources demandées

Queue	Access	Priority	Max time	Max slots
Workq (default)	everyone	300	96H	4120
unlimitq	everyone	100	unlimited	680
smpq	on demand	0	unlimited	240
hypermembq	On demand	0	unlimited	96
Interq (qlogin)	everyone	100	48H	40

Les ressources sont calculées et limitées selon le groupe linux auquel appartient l'utilisateur : INRA, Région, contributeur, autre.

Defaults parameters

- Workq
- 1 core
- 8 GB memory maximum
- Write only /work directory (temporary disk space)
- 1 TB quota disk per user (on /work directory)
- 120 days files without access automatic purged
- 100 000H annually computing time (more on demand)

Ouvrir un terminal et se connecter au cluster, dans son espace de travail

Ouvrir un terminal avec un compte de formation (fleur)

```
smaman@smaman:~$ ssh -XY smaman@genobioinfo.toulouse.inrae.fr
smaman@genobioinfo.toulouse.inrae.fr's password:
Last login: Thu Aug 31 10:08:44 2023 from 92.88.171.166

=====
Bienvenue sur les serveurs d'accès de la plate-forme bioinformatique
=====
```

Se positionner dans son répertoire de travail /work/user/username/

```
[smaman@genobioinfo1 TOTO]$ cd /work/user/smaman/
[smaman@genobioinfo1 smaman]$ pwd
/work/user/smaman
[smaman@genobioinfo1 smaman]$
```

1- Ouvrir un terminal en local sur votre ordinateur et se connecte en ssh sur le serveur frontal genologin avec les comptes de formation.

2- Se positionner dans /work/user/fleur . Exemple: cd /work/user/anemone/

3- Ecrire un script : myscript.sh

```
#!/bin/bash          <--- shebang bash
more tata;           <--- Ligne de commande
```

4- Créer un fichier « tata » et écrivez quelques lignes à l'intérieur de votre fichier.

4 – Lancer votre script

```
sbatch myscript.sh
```

```
sbatch myscript.sh
```

```
Submitted batch job 7897161
```

5- Visualiser le stdout dans le fichier slurm:

```
$ ls -ltrah
```

```
-rw-r--r--  1 smaman SIGENAE  24 23 sept. 11:59 myscript.sh
-rw-r--r--  1 smaman SIGENAE  66 23 sept. 12:00 slurm-7897161.out
```

```
$ more slurm-7897161.out
```

```
.....
```

```
tata
```

```
.....
```

```
test pour le master II bioinfo
```

Job submission

[BATCH]

- **sbatch** (qsub) : submit a batch job to slurm (default workq partition())
- **sarray** (qarray) : submit a batch job-array to slurm
- **scancel** (qdel) : kill the specified job

Job monitoring

- **sinfo** (qhost) : display nodes, partitions, reservations
- **squeue** (qstat) : display jobs and state
- **scontrol show** : get informations on jobs, nodes, partitions
- **sstat** (qstat -j) : show status of running jobs
- **sview** (qmon) : graphical user interface
- **sacct** (qacct) : display accounting data



SLURM Sample sbatch script

```
# !/bin/bash

#SBATCH --time=00:10:00 #job time limit

#SBATCH -J testjob      #job name

#SBATCH -o output.out  #output file name

#SBATCH -e error.out    #error file name

#SBATCH --mem=8G        #memory reservation

#SBATCH --cpus-per-task=4      #ncpu on the same node

#SBATCH --mail-type=BEGIN,END,FAIL (email address is LDAP account's)

#Purge any previous modules

module purge

#Load the application

module load bioinfo/ncbi-blast-2.2.29+

# My command lines I want to run on the cluster

blastall ...
```

15



New environment Search/Find a soft (Web)

Website (Resources/Software): <http://bioinfo.genotoul.fr/index.php/resources-2/software/>

Select a category:

All software

Search a software:

Search

Go

Search Results for "Admixture"

Application	Description	Availability/Use (SLURM Cluster available on 16/03/2018)
Admixtools	ADMIXTOOLS (Patterson et al. 2012) is a software package that supports formal tests of whether admixture occurred, and makes it possible to infer admixture proportions and dates.	Slurm Cluster: Ask for Install SGE Cluster: /usr/local/bioinfo/src/ADMIXtools
Admixture	ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets. It uses the same statistical model as STRUCTURE but calculates estimates much more rapidly using a fast numerical optimization algorithm.	SLURM Cluster: How to use SGE Cluster: How to use

Not installed on
SLURM Cluster-
link to ask for

Link to soft
website

Installed on
SLURM Cluster-
link to help

25



Software documentation

- official software documentation in the installation folder `/usr/local/bioinfo/src/<soft_name>/<soft_version>`
- our website Software page: link to software website.

Use on SLURM cluster

- "How_to_use_SLURM_<soft_name>" file:
 - software installation directory `/usr/local/bioinfo/src/<soft_name>`
 - our website Software page (Availability/Use column, click on SLURM cluster link).
- a basic « example_on_cluster » directory in the software installation directory
`/usr/local/bioinfo/src/<soft_name>/example_on_cluster`

HOW TO USE ON SLURM CLUSTER

```
SOFT : samtools
-----
Site du soft: http://samtools.sourceforge.net
-----

LICENSE:
-----

The MIT/Expat License

See software documentation for more informations.

Location: /usr/local/bioinfo/src/samtools
-----
```

Software
informations

```
Load binaries and environment:
-----

-> Version v0.1.19
module load bioinfo/samtools-0.1.19

-> Version 1.3.1
module load bioinfo/samtools-1.3.1

-> Version v1.4
module load bioinfo/samtools-1.4

or use absolute path
```

Usage and
versions

```
Example directory for use on cluster:
-----

/usr/local/bioinfo/src/samtools/example_on_cluster

To submit:

sbatch test v1.4.sh
```

Example

Ouvrir un terminal en local sur votre ordinateur.

ssh sur le serveur frontal genologin avec les comptes de formation.

Se positionner dans /work/user/username/

Chercher le « How to use » sur <http://bioinfo.genotoul.fr/> pour utiliser l'outil FastQC_v0.11.5

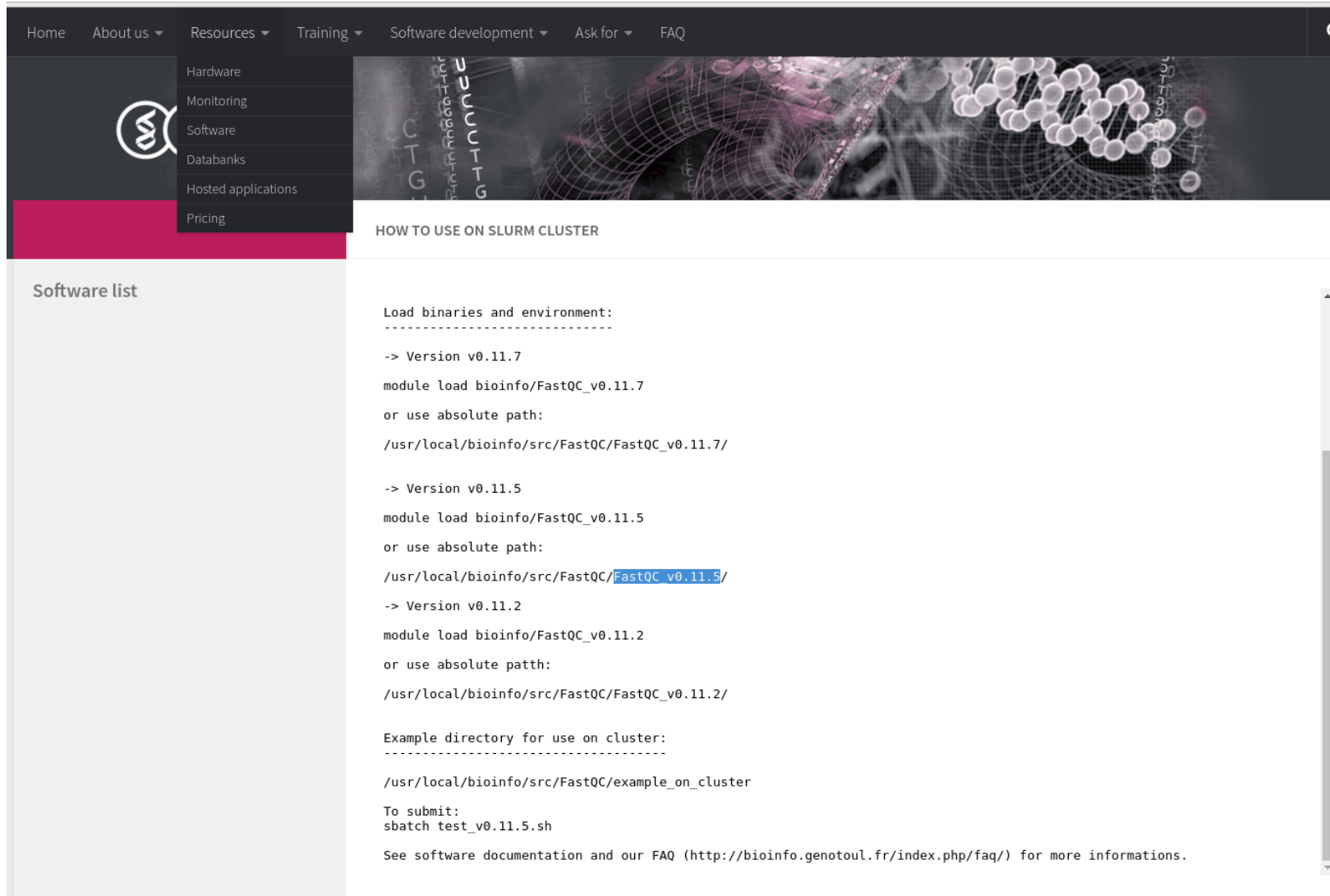
Récupérer un fichier de séquences à l'aide d'un wget :

http://genoweb.toulouse.inra.fr/~sigenae/Galaxy_Formation/test-data/2.fastq

Écrire le script quality.sh pour lancer l'outil FastQC report sur le fichier FASTQ ci-dessous.

Envoyez moi votre script pour validation : sarah.maman@inrae.fr

Chercher le « How to use » sur <http://bioinfo.genotoul.fr/> pour utiliser l'outil FastQC_v0.11.5



The screenshot shows the bioinfo.genotoul.fr website. The top navigation bar includes links for Home, About us, Resources, Training, Software development, Ask for, and FAQ. A left sidebar contains a search icon and a list of categories: Hardware, Monitoring, Software, Databanks, Hosted applications, and Pricing. The main content area is titled 'HOW TO USE ON SLURM CLUSTER' and contains instructions for loading FastQC binaries and environment modules. The instructions are organized by version: v0.11.7, v0.11.5, and v0.11.2. For each version, it shows the module load command and the absolute path to the FastQC directory. The v0.11.5 version is highlighted in blue. Below the module load instructions, there is an example directory path and a command to submit a job using sbatch. The page concludes with a reference to software documentation and the FAQ.

Home About us Resources Training Software development Ask for FAQ

Hardware
Monitoring
Software
Databanks
Hosted applications
Pricing

HOW TO USE ON SLURM CLUSTER

Software list

```
Load binaries and environment:
-----

-> Version v0.11.7
module load bioinfo/FastQC_v0.11.7
or use absolute path:
/usr/local/bioinfo/src/FastQC/FastQC_v0.11.7/

-> Version v0.11.5
module load bioinfo/FastQC_v0.11.5
or use absolute path:
/usr/local/bioinfo/src/FastQC/FastQC_v0.11.5/

-> Version v0.11.2
module load bioinfo/FastQC_v0.11.2
or use absolute path:
/usr/local/bioinfo/src/FastQC/FastQC_v0.11.2/

Example directory for use on cluster:
-----

/usr/local/bioinfo/src/FastQC/example_on_cluster

To submit:
sbatch test_v0.11.5.sh

See software documentation and our FAQ (http://bioinfo.genotoul.fr/index.php/faq/) for more informations.
```



Récupérer un fichier de séquences à l'aide d'un wget :

http://genoweb.toulouse.inra.fr/~sigenae/Galaxy_Formation/test-data/2.fastq

```
smaman@genologin1 /work/smaman $ wget http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/test-data/2.fastq
--2019-09-23 14:02:11-- http://snp.toulouse.inra.fr/~sigenae/Galaxy_Formation/test-data/2.fastq
Résolution de snp.toulouse.inra.fr (snp.toulouse.inra.fr)... 147.99.108.69
Connexion vers snp.toulouse.inra.fr (snp.toulouse.inra.fr)|147.99.108.69|:80...connecté.
requête HTTP transmise, en attente de la réponse...200 OK
Longueur: 255 [text/plain]
Sauvegarde en : «2.fastq»

100%[=====>] 255 --.-K/s ds 0s

2019-09-23 14:02:11 (34,7 MB/s) - «2.fastq» sauvegardé [255/255]
```

Idem pour 1.fastq :

wget http://genoweb.toulouse.inra.fr/~sigenae/Galaxy_Formation/test-data/1.fastq

Ecrire le script quality.sh pour lancer l'outil FastQC report sur le fichier fastq ci-dessous.

```
smaman@genologin1 /work/smaman $ ls /usr/local/bioinfo/src/FastQC/example_on_cluster
slurm-7168.out test_v0.11.5.sh xxxx_R1_sample.fastq xxxx_R1_sample_fastqc.html xxxx_R1_sample_fastqc.zip xxxx_R2_sample.fastq
smaman@genologin1 /work/smaman $ more /usr/local/bioinfo/src/FastQC/example_on_cluster/test_v0.11.5.sh
#!/bin/bash
#SBATCH -p workq

#Load binaries
module load bioinfo/FastQC_v0.11.5

#Command

fastqc -o /usr/local/bioinfo/src/FastQC/example_on_cluster/ /usr/local/bioinfo/src/FastQC/example_on_cluster/xxxx_R1_sample.fastq /
smaman@genologin1 /work/smaman $ vi quality.sh
```



Script sbatch : Exercice

```
smaman@genologin1 /work/smaman $ more quality.sh
#!/bin/bash
#SBATCH -p workq

#Load binaries
module load bioinfo/FastQC_v0.11.5

#Command

fastqc -o /work/smaman/ /work/smaman/1.fastq /work/smaman/2.fastq
smaman@genologin1 /work/smaman $ sbatch quality.sh
Submitted batch job 7900189
```

Donner les droits de lecture de votre script afin que votre formateur puisse valider votre script :

```
chmod a+r quality.sh
```



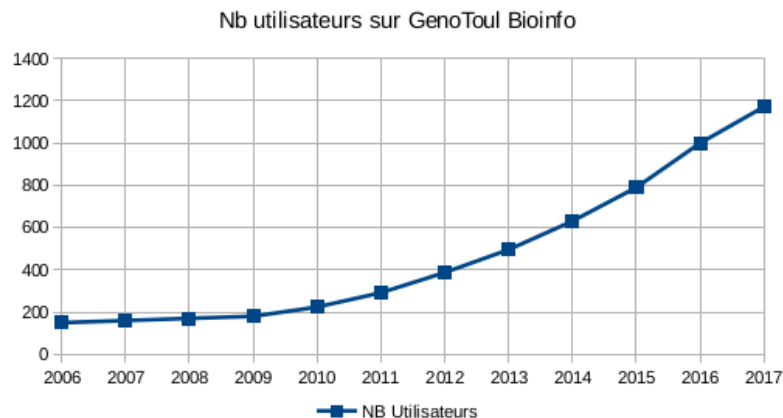
Penser à vérifier les résultats de votre traitement :

```
drwxr-xr-x 11 smaman SIGENAE 876K 23 sept. 14:02 nextflow_star_13m_cdr1links_cdrmerge_recirc
-rw-r--r-- 1 smaman SIGENAE 159 23 sept. 14:07 quality.sh
-rw-r--r-- 1 smaman SIGENAE 6,3K 23 sept. 14:07 1_fastqc.zip
-rw-r--r-- 1 smaman SIGENAE 930 23 sept. 14:07 slurm-7900189.out
-rw-r--r-- 1 smaman SIGENAE 277K 23 sept. 14:07 2_fastqc.zip
drwxrwxrwx 14 smaman BIOINFO 16K 23 sept. 14:07 .
-rw-r--r-- 1 smaman SIGENAE 257K 23 sept. 14:07 2_fastqc.html
smaman@genologin1 /work/smaman $ more slurm-7900189.out
Started analysis of 1.fastq
Analysis complete for 1.fastq
Failed to process file 1.fastq
java.lang.IllegalArgumentException: No known encodings with chars < 33 (Yours was )
    at uk.ac.babraham.FastQC.Sequence.QualityEncoding.PhredEncoding.getFastQEncodingOffset(PhredEncoding.java:32)
    at uk.ac.babraham.FastQC.Modules.PerBaseQualityScores.getPercentages(PerBaseQualityScores.java:71)
    at uk.ac.babraham.FastQC.Modules.PerBaseQualityScores.raisesError(PerBaseQualityScores.java:166)
    at uk.ac.babraham.FastQC.Report.HTMLReportArchive.startDocument(HTMLReportArchive.java:336)
    at uk.ac.babraham.FastQC.Report.HTMLReportArchive.<init>(HTMLReportArchive.java:84)
    at uk.ac.babraham.FastQC.Analysis.OfflineRunner.analysisComplete(OfflineRunner.java:155)
    at uk.ac.babraham.FastQC.Analysis.AnalysisRunner.run(AnalysisRunner.java:110)
    at java.lang.Thread.run(Thread.java:745)
Started analysis of 2.fastq
Analysis complete for 2.fastq
```



Context users, contributions

- 1200 user accounts
(50% INRA, 50 % OCCITANIE)
- 845 cluster users
(galaxy users included)
- 1800 annual support tickets





New services

Account information and password change

Self Service → <http://selfservice.bioinfo.genotoul.fr>

Change your password
(every year)

Genotoul
Bioinfo

Welcome to GenoToul Bioinfo LDAP self service.

Please login with your cluster's credentials to access your account's details.

User name

Password

Language

Login

Forgot password?



Best practices Security

One user = one account

You are responsible of the damage caused by your login.

Default permissions directories

- **home:** drwxr-x—x : **R**ead, **W**rite, **eX**ecution for the owner, **R**ead and **eX**ecution for the group members, **eX**ecution for all.
- **save and work:** drwxr-x--- : **R**ead, **W**rite, **eX**ecution for user, **R**ead and **E**xecution for your group members, no permissions for all.

To change permissions: **chmod** command

Merci pour votre écoute