Genetic data and analysis on a cluster

FEB 2021

Overview

I. Data

   1. Format

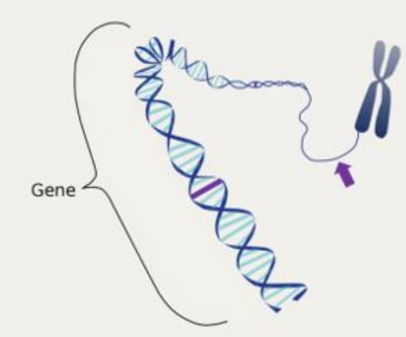
   2. IBD project

II. Cluster

   1. Uliege cluster

   2. Run an analysis

I. Data
1. Format



PLINK: free open source command-line program for genomic analysis

Download: https://www.cog-genomics.org/plink/

Plink formats: https://www.cog-genomics.org/plink/1.9/formats#bed

- Bed: representation of genotype calls at biallelic variants, so the markers/SNPs. The file can't be open.
- Bim: variant information file (Chromosome code, variant identifier, alleles…)
- Fam: Sample information file (family ID, sex code, phenotype value…)

-> go together and represent the entire dataset.

I. Data
1. Format

2. Obtain via –recode, loaded via -file

- Ped: The first six fields are the same as .fam. Then, variant information.
- Map: variant information file accompanying a .ped (chromosome code, variant identifier, …)

-> go together and represent the entire dataset.

I. Data
1. Format

Basic plink functions for input filtering:

- removes all unlisted samples: --keep
- Remove all listed samples: --remove
- Extract a subset of SNP based on chromosomes: --chr
- removes all unlisted variants: --extract
- removes all listed variants: --exclude
- Linkage disequilibrium: --indep-pairwise
- Minimum allele frequency= --maf
- …

I. Data
1. Format

Work with R from PLINK files:

- ❑ Change the format of the files using PLINK software so R can import them:

  from bed, bim, fam to map and ped using option –recode

  raw from –recodeA

  -> can be read in R but will be huge

- ❑ Use specific R functions, for example read.bed()

I. Data
1. IBD

Projects: Detect epistasis with multiple tools and same dataset

IBD: Inflammatory Bowel Disease.

Two main Datasets:

Same 66,280 individuals (~50% cases, +50% controls)

Same initial quality controls (LD, MAF, HWE…)

- Unfiltered
- Functional: biological filters

I. Data
1. IBD

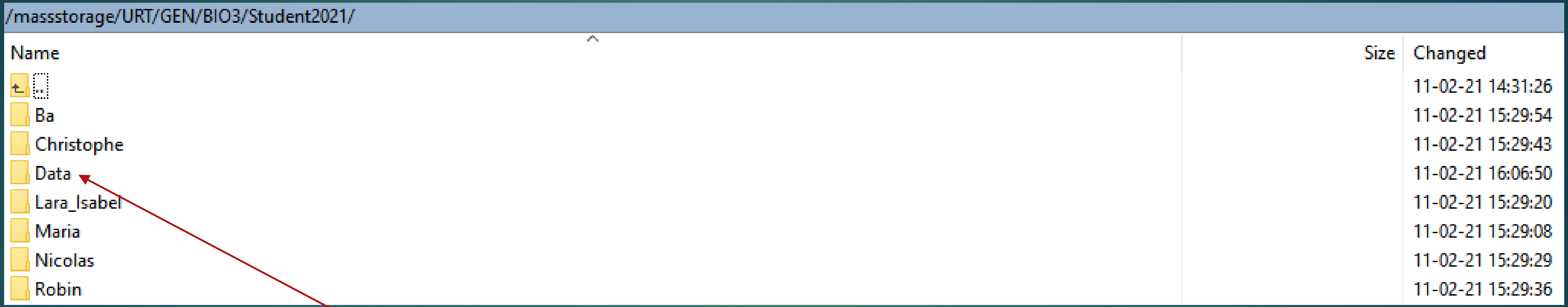
For the 2 datasets, multiple variations: specific requirement of analysis

- More SNP filters (relief and epiblaster) for analysis that can’t handle large amount of SNPs
- Imputation (knn) for analysis that can’t handle missing values
- Phenotypes:
  - continuous
  - binary

I. Data

1. IBD

Folder structure



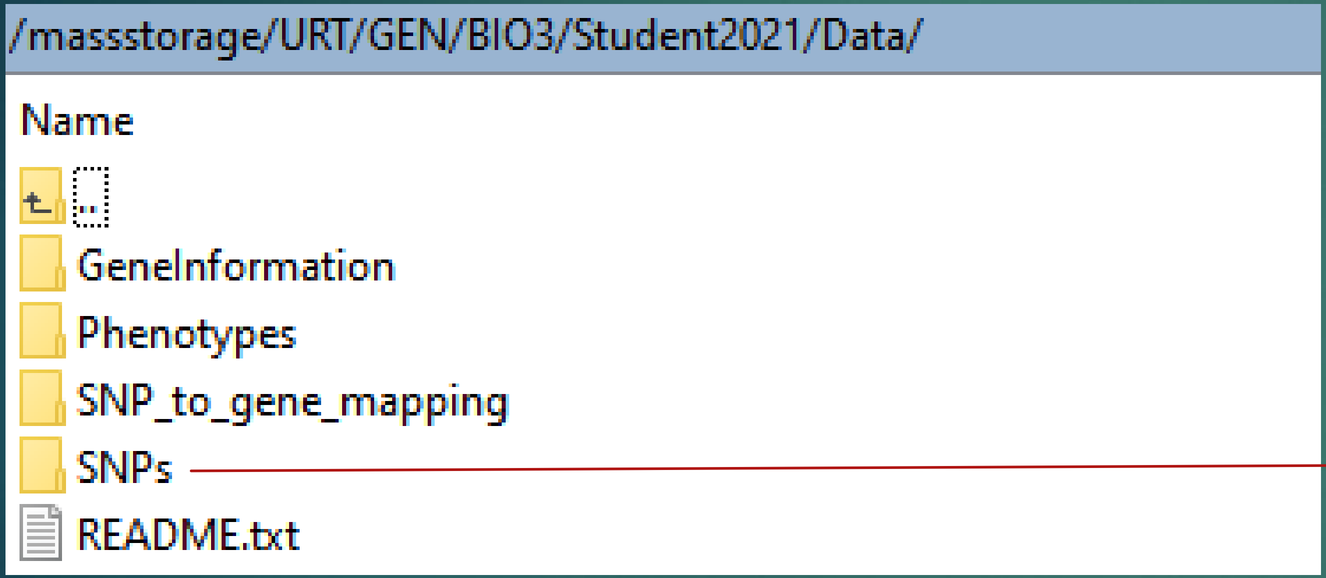
/massstorage/URT/GEN/BIO3/Student2021/

Name	Size	Changed
..		11-02-21 14:31:26
Ba		11-02-21 15:29:54
Christophe		11-02-21 15:29:43
Data		11-02-21 16:06:50
Lara_Isabel		11-02-21 15:29:20
Maria		11-02-21 15:29:08
Nicolas		11-02-21 15:29:29
Robin		11-02-21 15:29:36

Input

I. Data
1. IBD

Folder structure



/massstorage/URT/GEN/BIO3/Student2021/Data/

Name

..

GeneInformation

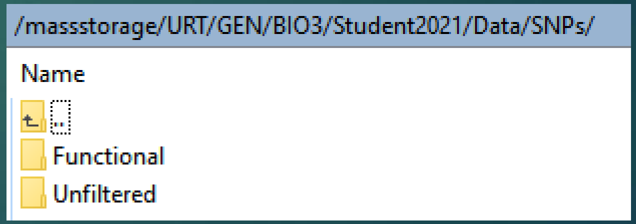
Phenotypes

SNP_to_gene_mapping

SNPs

README.txt

Two main SNP sets



/massstorage/URT/GEN/BIO3/Student2021/Data/SNPs/

Name

..

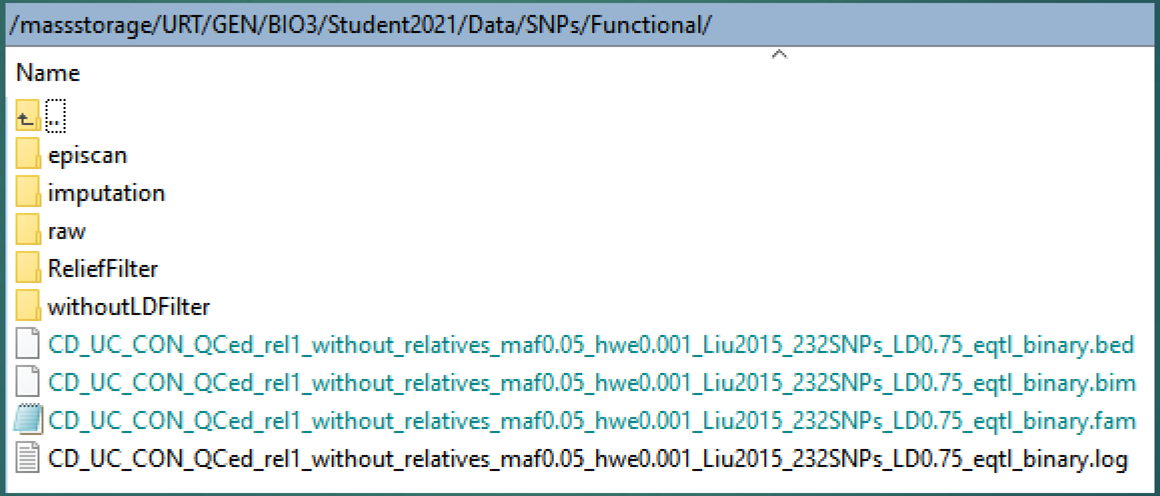
Functional

Unfiltered

I. Data
1. IBD

Folder structure

Example: Functional dataset



```
/massstorage/URT/GEN/BIO3/Student2021/Data/SNPs/Functional/
Name
..
episcan
imputation
raw
ReliefFilter
withoutLDFilter
CD_UC_CON_QCed_rel1_without_relatives_maf0.05_hwe0.001_Liu2015_232SNPs_LD0.75_eqtl_binary.bed
CD_UC_CON_QCed_rel1_without_relatives_maf0.05_hwe0.001_Liu2015_232SNPs_LD0.75_eqtl_binary.bim
CD_UC_CON_QCed_rel1_without_relatives_maf0.05_hwe0.001_Liu2015_232SNPs_LD0.75_eqtl_binary.fam
CD_UC_CON_QCed_rel1_without_relatives_maf0.05_hwe0.001_Liu2015_232SNPs_LD0.75_eqtl_binary.log
```

Available options for specific requirements:

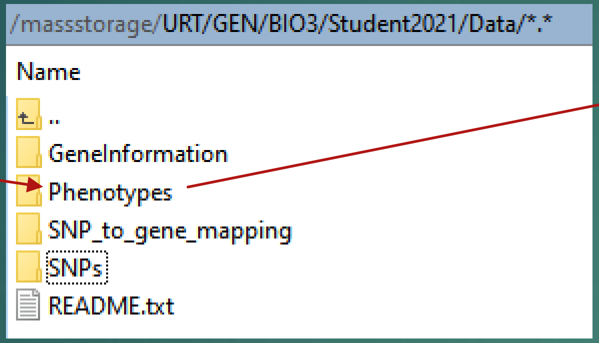
Imputation of 1. and 2.

Reduction of the dataset via episcan and relief

I. Data
1. IBD

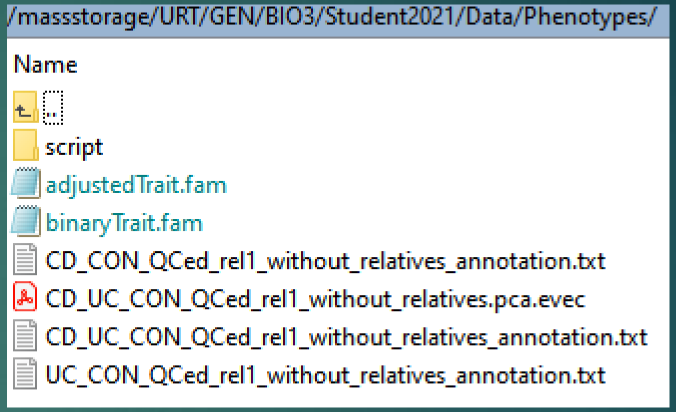
Folder structure

Phenotypes



/massstorage/URT/GEN/BIO3/Student2021/Data/*.*

Name
- ..
- GeneInformation
- Phenotypes
- SNP_to_gene_mapping
- SNPs
- README.txt



/massstorage/URT/GEN/BIO3/Student2021/Data/Phenotypes/

Name
- ..
- script
- adjustedTrait.fam
- binaryTrait.fam
- CD_CON_QCed_rel1_without_relatives_annotation.txt
- CD_UC_CON_QCed_rel1_without_relatives.pca.evec
- CD_UC_CON_QCed_rel1_without_relatives_annotation.txt
- UC_CON_QCed_rel1_without_relatives_annotation.txt

(Order is important)

1. If your tool can adjust for covariates: binary phenotypes and adjust for the first 7 PCs.
2. Else, if your tool can't include covariates but can handle continuous phenotypes: continuous phenotypes that are already adjusted for first 7 PCs
3. Else, if your tool can't include covariates and can't handle continuous phenotypes: binary phenotypes.

I. Cluster
1. Uliege cluster

What is a cluster?

Set of connected computers that work together.

Why are we using a cluster?

- Big dataset, big analysis -> improve performance and availability
- Legal agreement

Advice:

Create and try your code on a small dataset* on your own computer. Then, run the real analysis on the cluster once you made sure your code is ok.

Why: Easier and faster to find errors.

*a public dataset, not the IBD one which can't be downloaded.

I. Cluster

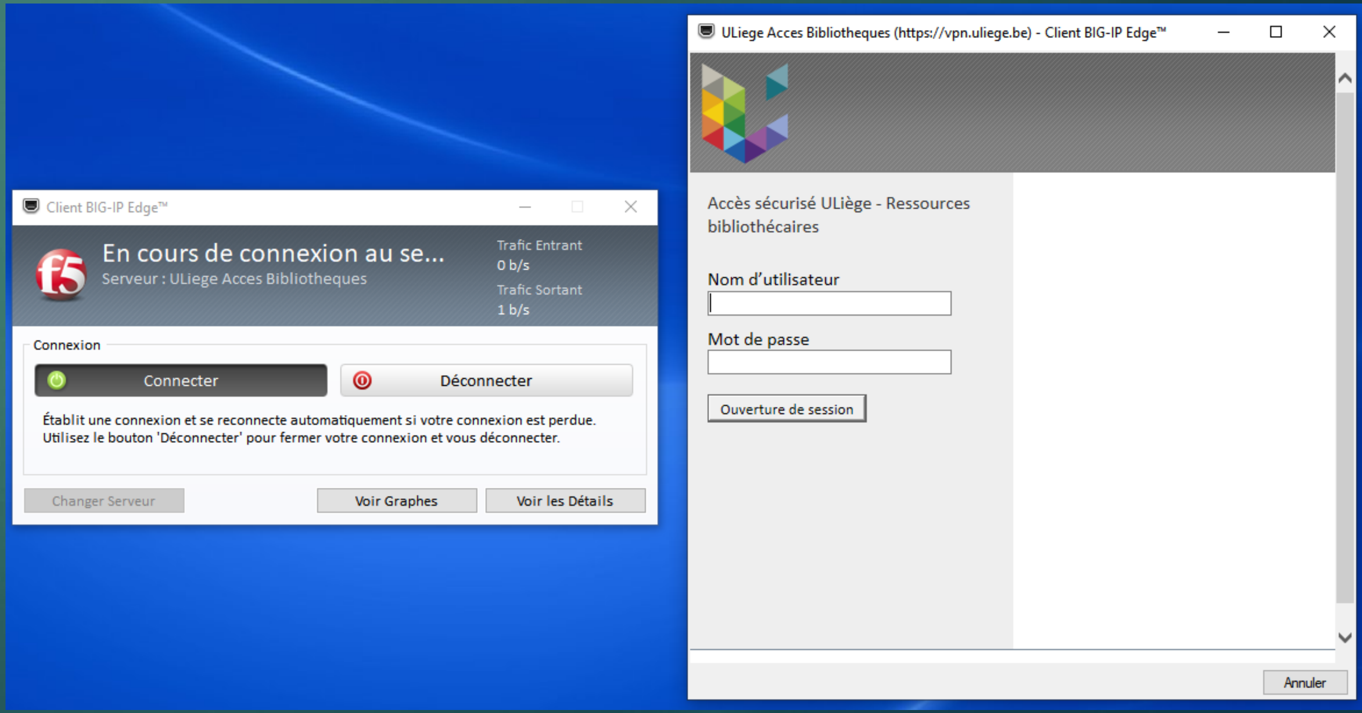
1. Uliege cluster

Connect to the cluster

If not onsite (wifi of university of Liège), download the VPN:

https://my.segi.uliege.be/cms/c_11650735/fr/mysegi-new-vpn

Enter your id and passwork to connect



Client BIG-IP Edge™

En cours de connexion au se...

Serveur : ULiege Acces Bibliotheques

Trafic Entrant 0 b/s

Trafic Sortant 1 b/s

Connexion

Établit une connexion et se reconnecte automatiquement si votre connexion est perdue. Utilisez le bouton 'Déconnecter' pour fermer votre connexion et vous déconnecter.

ULiege Acces Bibliotheques (https://vpn.uliege.be) - Client BIG-IP Edge™

Accès sécurisé ULiège - Ressources bibliothécaires

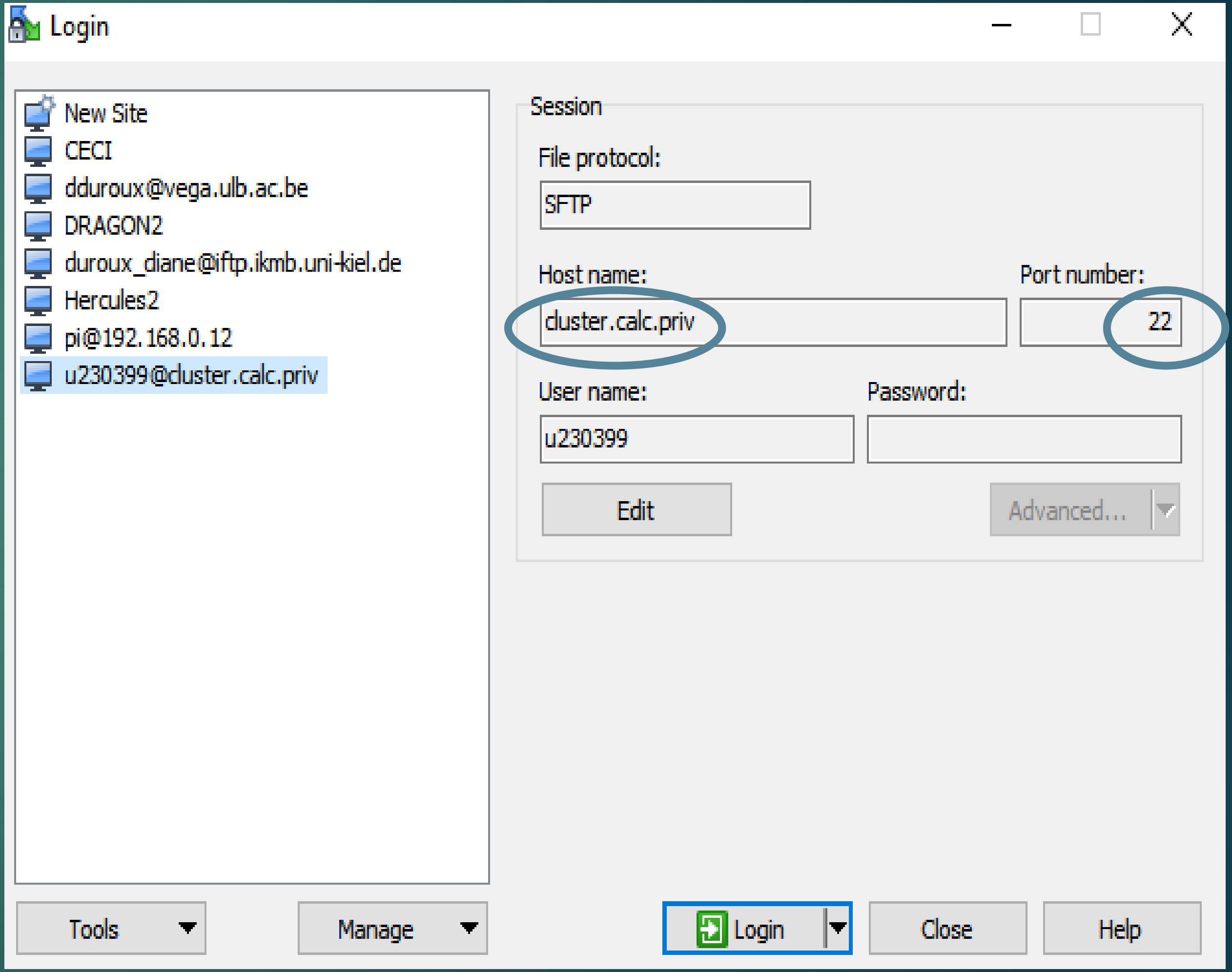
Nom d'utilisateur

Mot de passe

I. Cluster
1. Uliege cluster

Access and **visualize** your individual folder and the data

Windows:
https://winscp.net/eng/index.php



Login

New Site
CECI
dduroux@vega.ulb.ac.be
DRAGON2
duroux_diane@iftp.ikmb.uni-kiel.de
Hercules2
pi@192.168.0.12
u230399@cluster.calc.priv

Session

File protocol: SFTP

Host name: cluster.calc.priv

Port number: 22

User name: u230399

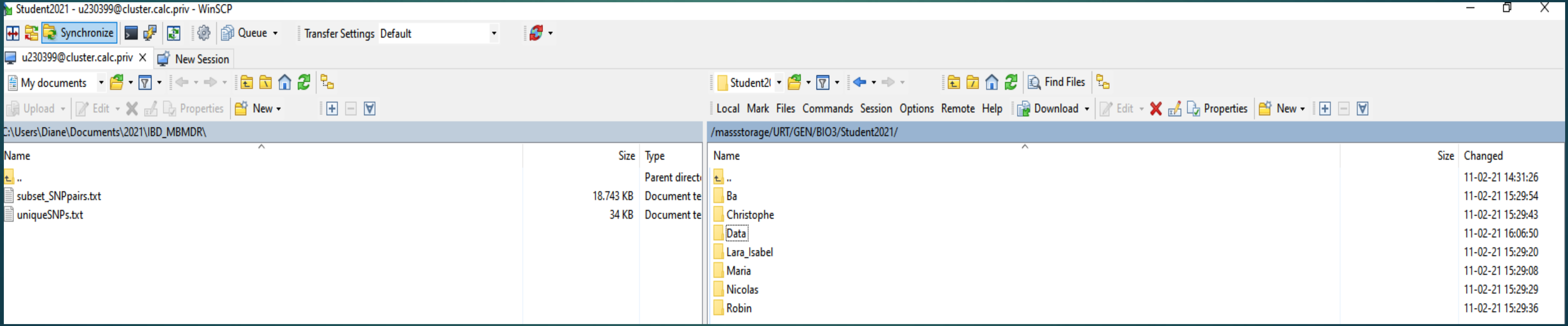
Password:

I. Cluster

1. Uliege cluster

Access and **visualize** your individual folder and the data

Windows:
https://winscp.net/eng/index.php



Student2021 - u230399@cluster.calc.priv - WinSCP

u230399@cluster.calc.priv

C:\Users\Diane\Documents\2021\IBD_MBMDR\

Name	Size	Type
..		Parent direct
subset_SNPpairs.txt	18.743 KB	Document te
uniqueSNPs.txt	34 KB	Document te

/massstorage/URT/GEN/BIO3/Student2021/

Name	Size	Changed
..		11-02-21 14:31:26
Ba		11-02-21 15:29:54
Christophe		11-02-21 15:29:43
Data		11-02-21 16:06:50
Lara_Isabel		11-02-21 15:29:20
Maria		11-02-21 15:29:08
Nicolas		11-02-21 15:29:29
Robin		11-02-21 15:29:36

I. Cluster

1. Uliege cluster

Access and **visualize** your individual folder and the data

Windows:
https://winscp.net/eng/index.php

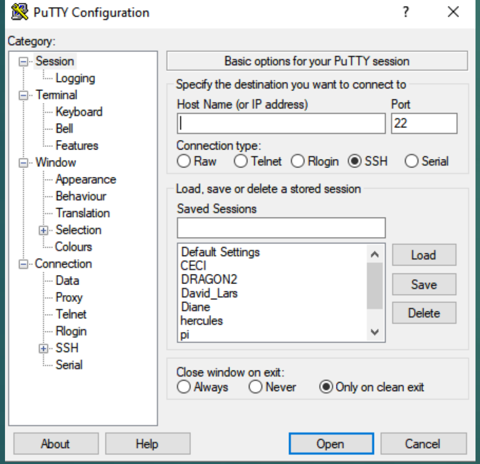
Linux: ssh command
https://docs.oracle.com/en/cloud/paas/big-data-cloud/csbdi/connect-cluster-node-secure-shell-ssh.html#GUID-E6F4421D-3D7F-415B-ABD6-D3CC0C870947

I. Cluster
2. Run an analysis

How to **communicate** with the cluster (sofware)

- Windows: puTTY software, open source SSH client

  https://www.putty.org/



PuTTY Configuration

Category:

- Session
  - Logging
- Terminal
  - Keyboard
  - Bell
  - Features
- Window
  - Appearance
  - Behaviour
  - Translation
  - Selection
  - Colours
- Connection
  - Data
  - Proxy
  - Telnet
  - Rlogin
  - SSH
  - Serial

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address):

Port: 22

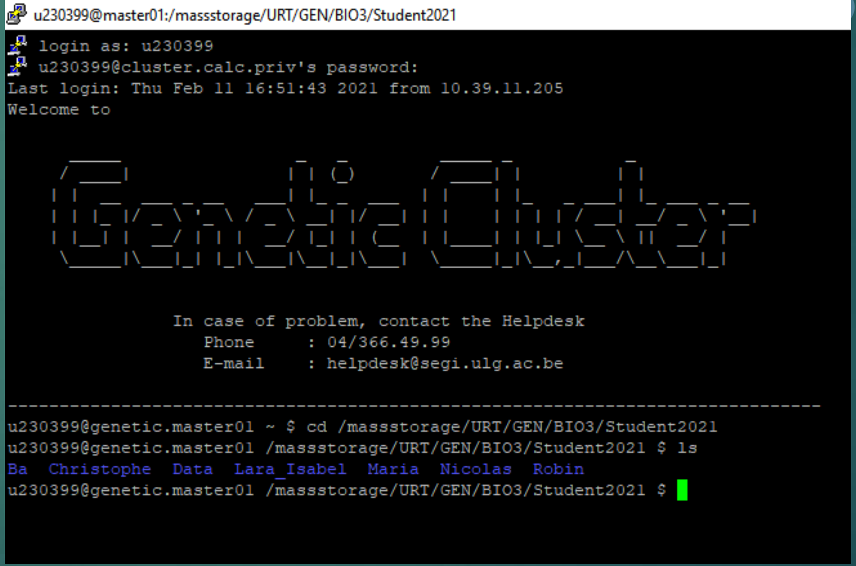
Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions

Default Settings
CECI
DRAGON2
David_Lars
Diane
hercules
pi

Close window on exit: ○ Always ○ Never ● Only on clean exit



u230399@master01:/massstorage/URT/GEN/BIO3/Student2021

```
login as: u230399
u230399@cluster.calc.priv's password:
Last login: Thu Feb 11 16:51:43 2021 from 10.39.11.205
Welcome to

     _____                 _   _         _____ _           _
    / ____|               | | (_)       / ____| |         | |
   | |  __  ___ _ __   ___| |_ _  ___  | |    | |_   _ ___| |_ ___ _ __
   | | |_ |/ _ \ '_ \ / _ \ __| |/ __| | |    | | | | / __| __/ _ \ '__|
   | |__| |  __/ | | |  __/ |_| | (__  | |____| | |_| \__ \ ||  __/ |
    \_____|\___|_| |_|\___|\__|_|\___|  \_____|_|\__,_|___/\__\___|_|


                In case of problem, contact the Helpdesk
                   Phone      : 04/366.49.99
                   E-mail     : helpdesk@segi.ulg.ac.be

-------------------------------------------------------------------------------
u230399@genetic.master01 ~ $ cd /massstorage/URT/GEN/BIO3/Student2021
u230399@genetic.master01 /massstorage/URT/GEN/BIO3/Student2021 $ ls
Ba  Christophe  Data  Lara_Isabel  Maria  Nicolas  Robin
u230399@genetic.master01 /massstorage/URT/GEN/BIO3/Student2021 $
```

- Linux: directly in terminal

I. Cluster
2. Run an analysis

How to **communicate** with the cluster (language)

Slurm: ressource manager / job scheduler

Goal: organize ressource sharing on a supercomputer

How: Users submit jobs, which are scheduled and allocated resources (CPU time, memory, etc.)

I. Cluster
2. Run an analysis

Basic commands to navigate in your folders and check your files (bash, shell):

- Cd folderName: change directory (go into another directory)
- Ls: display what’s in a directory
- Head fileName: See the top of the file
- Tail filename: see the end of the file
- wc –l: count the number of rows in a file
- du –sh folderName: check the size of a folder
- rm fileName: delete a file
- mkdir folderName: create a new folder
- …

More info: https://www.educative.io/blog/bash-shell-command-cheat-sheet

I. Cluster
2. Run an analysis

Run an analysis / a script / a job on the cluster:

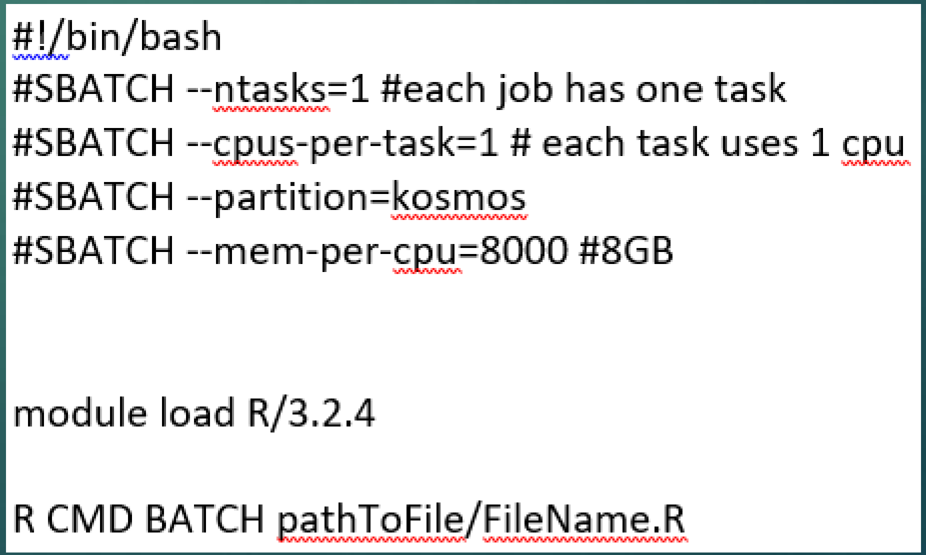
Create a .sh file (for example: run.sh).

This file has a specific structure so the cluster understands whit it needs to do

Header: must to start with #
Specify the ressource required

Load softwares needed

Analysis: here call an external R script



```
#!/bin/bash
#SBATCH --ntasks=1 #each job has one task
#SBATCH --cpus-per-task=1 # each task uses 1 cpu
#SBATCH --partition=kosmos
#SBATCH --mem-per-cpu=8000 #8GB


module load R/3.2.4

R CMD BATCH pathToFile/FileName.R
```

I. Cluster
2. Run an analysis

Header is very important:

If too much ressources asked: will never start

If not enough: job will stop before the end

Need to investigate the ressources needed: time, nb of CPUs…

Example:

```
#!/bin/bash
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=6
#SBATCH --partition=kosmos
#SBATCH --mem-per-cpu=8000
--time=01:00:00
```

Always required

Number of core per task

Each task uses 6 cpus

Select a partition

8GB required

Time limit for the job.

More info: https://ubccr.freshdesk.com/support/solutions/articles/5000688140-submitting-a-slurm-job-script

I. Cluster
2. Run an analysis

Some basic slurm commands:

- Submit/start a job: sbatch pathToFile/FileName.sh.

  What does it do: You ask permission to run a job on the cluster.

  If resources are available, it will start.

  If not, it will wait in the queue until enough resources are available.

- Ask if a program is running or pending: squeue -u yourUsername
- Get more info about the cluster: sinfo
- Stop a job: scancel jobNumber

More info: https://support.ceci-hpc.be/doc/_contents/QuickStart/SubmittingJobs/SlurmTutorial.html

I. Cluster
2. Run an analysis

Tips:

- Tests or debugging:

  Slurm jobs are normally batch jobs: they are run unattended.

  If you want to have a direct view on your job, run: srun –pty bash

- NEVER work on the master node of the cluser (ie without srun or sbatch)

- Always google your problem (stackOverflow, mathOverflow, …)

Thank you