Canadian Centre for Computational Genomics

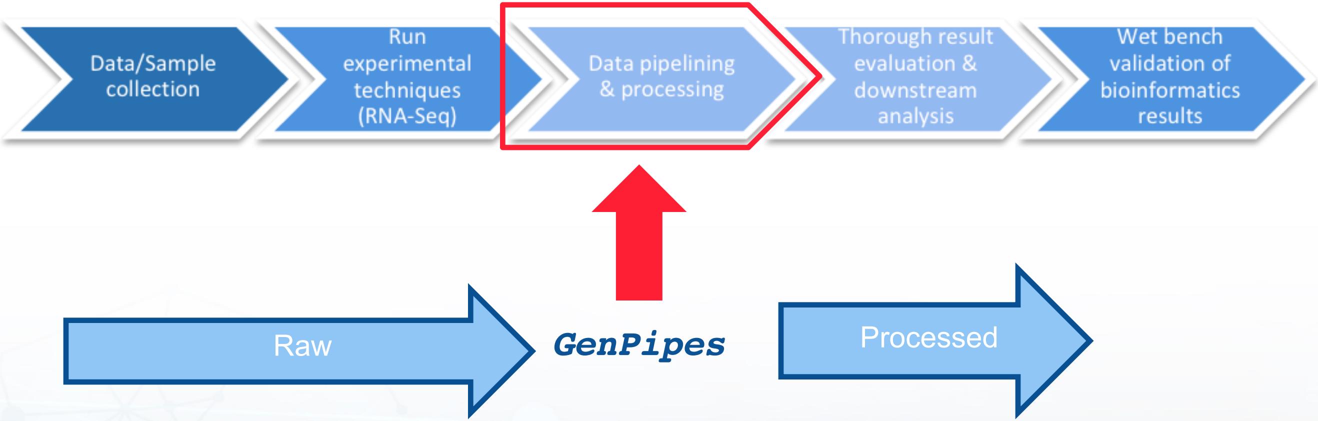
C3G Analysis Workshop: RNA-Seq

GenPipes Files: Creation & Validation

22-23 January 2019

GenPipes Overview

Experimental Processing:



Data/Sample collection

Run experimental techniques (RNA-Seq)

Data pipelining & processing

Thorough result evaluation & downstream analysis

Wet bench validation of bioinformatics results

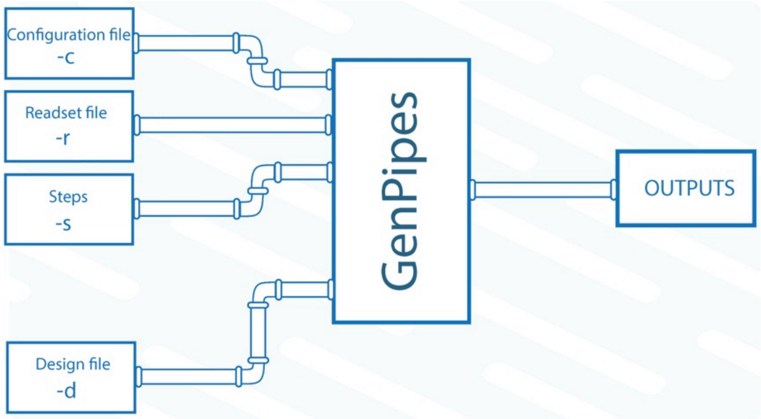
Raw

GenPipes

Processed

GenPipes Overview

GenPipes Requirements:



GenPipes requires setting up 3 files:

- Configuration “ini” file: contains parameters used by tools in the pipeline
- Readset file: contains details about your samples
- Design file: contains details about sample comparisons. NOT needed by all pipelines.

What are ‘configuration/ini’ files and why do we need them?

- A “configuration” file, also know as an “ini” file due to its file extention, is a file that stores all the parameters needed by tools in the pipelines.
- A pipeline is a multi-step framework. At every step, tools require several parameters to be set. Having to type all the parameters on the command line as you call the tool would be complicated and messy.
- As a solution, parameters are stored in these ini files.
- To open an ini file and check out its content try:

```
cat $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.base.ini
```

- ini content:

```
[DEFAULT]
assembly_dir =
$MUGQIC_INSTALL_HOME/genomes/C3G_workshop/%(scientific_name)s.%
(assembly)s
scientific_name=Homo_sapiens
assembly=GRCh38.chr19
assembly_synonyms=hg38

module_R=mugqic/R_Bioconductor/3.5.0_3.7

[picard_sam_to_fastq]
cluster_cpu=-N 1 -n 4
cluster_queue=--mem=16G
```

Configurations Files:

- ini files come with each pipeline in:
  $MUGQIC_PIPELINES_HOME/pipelines/<pipeline_name>/<pipeline_name>.*.ini

```
## for the rnapseq pipeline:
ls $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.*.ini
```

- ini files for different species are available in: $MUGQIC_PIPELINES_HOME/resources/genomes/config/

```
## to check available species ini files:
ls $MUGQIC_PIPELINES_HOME/resources/genomes/config/
```

Creating your own configuration/ini file

- If you are interested in editing parameters in the ini file, you can create your own ini file that over-writes the set parameters.

- For example, to edit the number of threads in the star_align step in the rnaseq pipeline from 20 to 5, you can create a file, myini.ini, and in it you can change the parameters. You need to indicate the [section].

```
[star_align]
threads=5
```

How to use the configuration/ini file

- Configuration files are provided to the pipelines through the flag -c or --config, as follows:

```
rnaseq.py -c $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.base.ini
```

- For every pipeline, there are several provided ini files. The <pipeline>.base.ini contains all the parameters but is optimized for our internal server.
- To run a pipeline on cedar, you need to add the <pipeline>.cedar.ini:

```
rnaseq.py -c $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.base.ini
$MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.cedar.ini
```

- To use your own ini, you add it at the end:

```
rnaseq.py -c $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.base.ini
$MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.cedar.ini myini.ini
```

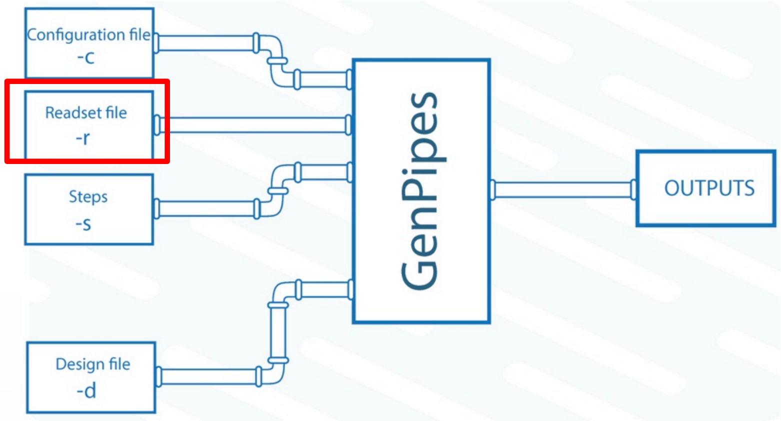
- To align to mouse mm9, instead of human, you need to add the mouse specific ini:

```
rnaseq.py -c $MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.base.ini
$MUGQIC_PIPELINES_HOME/pipelines/rnaseq/rnaseq.cedar.ini
$MUGQIC_PIPELINES_HOME/resources/genomes/config/Mus_musculus.mm9.ini
```

- The parameters in later ini files will over-write those in earlier ones.

GenPipes Overview

GenPipes Requirements:



GenPipes requires setting up 3 files:

- Configuration "ini" file: contains parameters used by tools in the pipeline
- Readset file: contains details about your samples
- Design file: contains details about sample comparisons. NOT needed by all pipelines.

What is the ‘Readset’ file?

Canadian Centre for Computational Genomics

- A Readset file is a file that contains information about your samples.
- It is a **tab-separated** file, that contains the following columns:

```
Sample Readset Library RunType Run Lane Adapter1 Adapter2 QualityOffset BED FASTQ1 FASTQ2 BAM
```

- Some fields are optional and others are mandatory (check here).

- Sample vs Readset:
  - Samples can be thought of as biological replicates while Readsets as technical replicates. If a library was divided across lanes, then each lane would be a “Readset” of the original “Sample”. In most pipelines, readsets belonging to the same sample are merged under a single sample name.

Creating and Using the Readset file

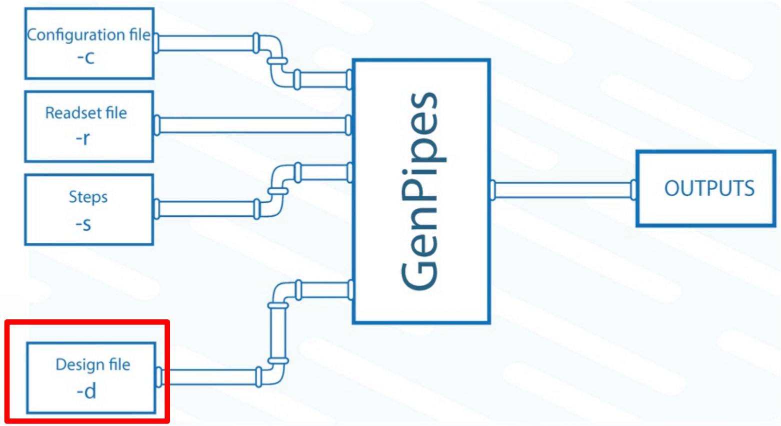
- How to create the Readset file:
  - There are several ways to create a readset file. The most basic way is to create it in Excel manually and copy it to the server.
  - Another way is to use the csvToReadset.R script if you have a csv file from Genome Quebec

- How to use a Readset file
  - a Readset file is provided to the pipeline using the (-r or --readsets)

```
rnaseq.py -r myreadset.txt
```

GenPipes Overview

GenPipes Requirements:



GenPipes requires setting up 3 files:

- Configuration “ini” file: contains parameters used by tools in the pipeline
- Readset file: contains details about your samples
- Design file: contains details about sample comparisons. NOT needed by all pipelines.

Design Files

Canadian Centre for Computational Genomics

- The Design file is needed by SOME of the pipelines.
- It contains data about sample comparisons. The format of the file depends on the pipeline. Consult the pipeline of interest for more information.
- In general, the design file is tab-separated, and has the following structure:
-

RNA-Seq

```
Sample  Contrast1  Contrast2
sampleA        1         1
sampleB        2         0
sampleC        0         2
```

ChIP-Seq

```
Sample  Contrast1,N  Contrast2,B
sampleA        1           1
sampleB        2           0
sampleC        0           2
```

- The first column is the Sample column. Every column after that is a contrast column:
  - '0' or '': the sample is not included in the analysis;
  - '1': the sample belongs to the control group;
  - '2': the sample belongs to the treatment test case group.
- For chipseq, N=NarrowPeaks and B=BroadPeaks
- Providing the pipeline with a design file (-d or --design):

```
rnaseq.py -d mydesign.tsv
```

Readset & Design File Validation

Canadian Centre for Computational Genomics

- Getting the Readset and Design files right is the most important task as a user.
- To avoid pipeline errors due to file issues, we have a file validator, mugqicValidator.py, that can let you know if your Readset or Design files have issues.

- To validate your Readset file:

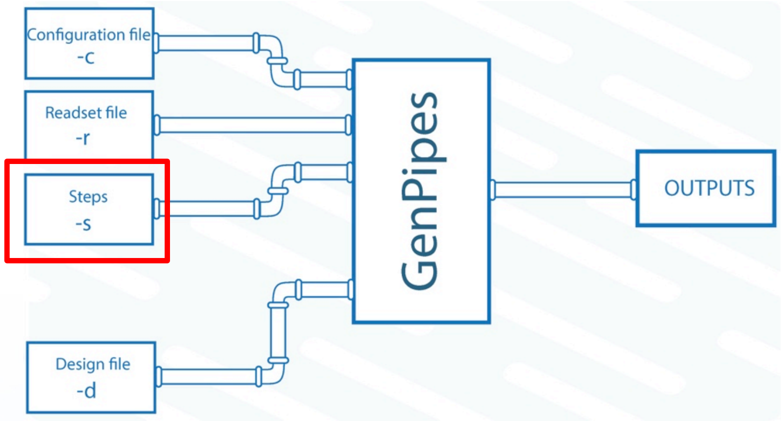
```
python $MUGQIC_PIPELINES_HOME/utils/mugqicValidator.py -r myreadsetfile
```

- To validate your Design file:

```
python $MUGQIC_PIPELINES_HOME/utils/mugqicValidator.py -d mydesignfile
```

GenPipes Overview

GenPipes Requirements:



GenPipes requires setting up 3 files:

- Configuration "ini" file: contains parameters used by tools in the pipeline
- Readset file: contains details about your samples
- Design file: contains details about sample comparisons. NOT needed by all pipelines.

Steps

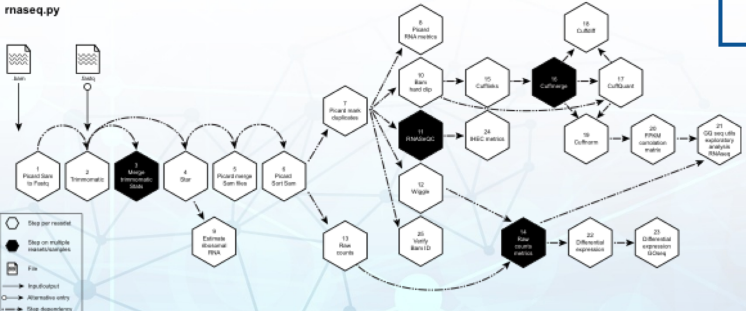
- The Steps parameter tells GenPipes which steps to run.
- To know what the steps are, you can use <pipeline>.py -h

```
rnaseq.py -h
```

- To specify the steps:

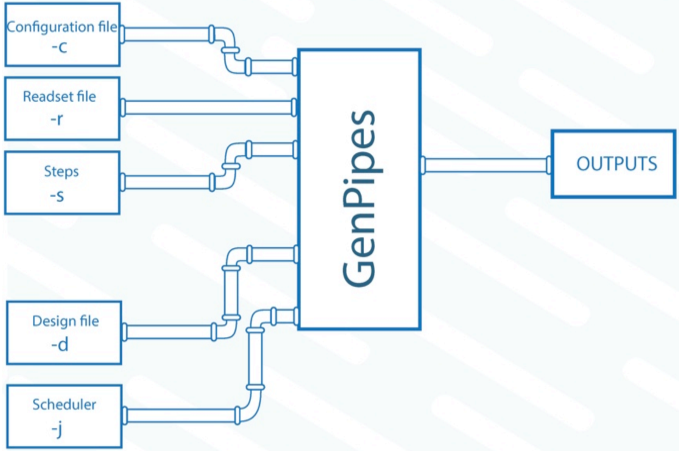
```
## full pipeline:
rnaseq.py -s 1-25

## specify steps 1 to 7 and 12
rnaseq.py -s 1-7,12
```

```
Steps:
------
1- picard_sam_to_fastq
2- trimmomatic
3- merge_trimmomatic_stats
4- star
5- picard_merge_sam_files
6- picard_sort_sam
7- picard_mark_duplicates
8- picard_rna_metrics
9- estimate_ribosomal_rna
10- bam_hard_clip
11- rnaseqc
12- wiggle
13- raw_counts
14- raw_counts_metrics
15- cufflinks
16- cuffmerge
17- cuffquant
18- cuffdiff
19- cuffnorm
20- fpkm_correlation_matrix
21- gq_seq_utils_exploratory_analysis_rnaseq
22- differential_expression
23- differential_expression_goseq
24- ihec_metrics
25- verify_bam_id
```



Running GenPipes:

Putting it all together



```
## create the command script
rnaseq.py     -c rnaseq.base.ini      -r readset.txt     -d designfile.txt      -s 1-15     -j slurm   >    commands.txt
```

pipeline   Ini file   readset file   design file   steps   scheduler   script

```
## run the script
bash commands.txt

## to check commands in queue:
squeue -u $USER
```

Running GenPipes:

Output Files, Logs and Errors

- As jobs finish, log files will be created in the "**job_output**" folder.

- If everything is Ok, you will get **MUGQICexitStatus:0** or **Exit_status=0**

- If you get an error that you cannot figure out, you can send us an email.

- For every analysis step, a folder will be created with the appropriate output files. Look into all the folders to know what each pipeline produces.

- If you made a mistake and would like to cancel ALL jobs in your queue, use:

```
scancel -u $USER
```

Running GenPipes:

Interpreting the output

- Our pipelines are a framework that runs third party software that is used in particular fields. We do not write the software ourselves. We control the framework that runs the tools.

- For knowledge about the format of the output, you need to read the documentation of the tools that produced the output.

- You can contact us for information that you can not find.

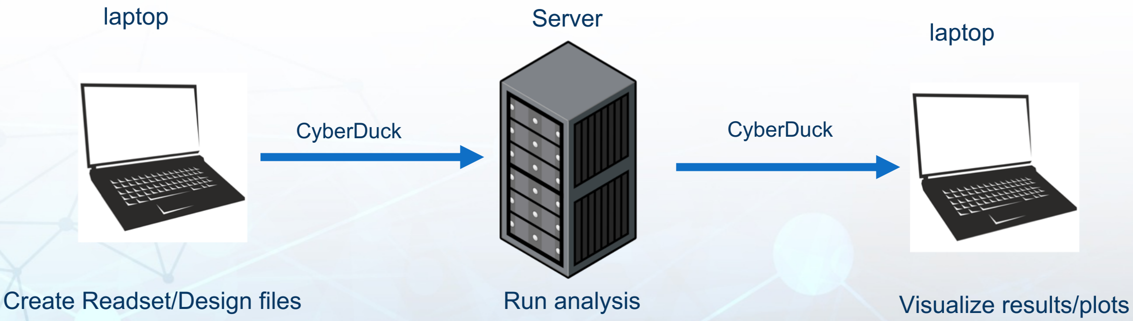
A note about your laptop and the server

Canadian Centre for Computational Genomics

- Your laptop has a graphic user interface and useful software like Excel but rarely has enough capacity to run a full bioinformatics analysis.

- The server has enough capacity to run analysis but has no graphic user interface and can be hard to create files, like the readset and design files, and to visualize plots.

- What usually happens is that we create and view files on our laptops but run larger computations on the server. So moving files between the server and your laptop is useful. In this workshop, we will use CyberDuck to transfer between the two systems.



A note about different Operating Systems

- If you transfer files (Readset or Design files) between Mac, Linux and Windows, sometimes you get errors related to differences in how files are handled by the different systems, specifically relating to end of file (EOF) characters.

- In order to fix file formats to suit the unix environment of most servers, try dos2unix:

```
dos2unix myfile.txt
```

- Another thing to try in vim is:

```
:%s/\r/\r/g
```

BioInformatics Resources

- Our website: http://www.computationalgenomics.ca/
- Our github pages: http://c3g.github.io
- Our email: support@computationalgenomics.ca
- C3G Open Door: http://www.computationalgenomics.ca/open-door/
- Guillimin Training: http://www.hpc.mcgill.ca/index.php/training
- Online Forums: BioStars, SeqAnswers, …
- Various Workshops/Resources: C3G, bioinfomatics.ca, Coursera courses, Software Carpentry …

Questions?

Fear is your worst enemy… Don’t be intimidated; experiment, test, fail and learn!

Backup First though!

GenomeQuébec

GenomeCanada

Ontario Genomics

McGill

Canadian Centre for Computational Genomics

C3G

SickKids®

MiCM McGill initiative in Computational Medicine

Thank you!

HPC4Health

bioinformatics.ca

IHEC International Human Epigenome Consortium

compute canada | calcul canada

Terry Fox PROFYLE

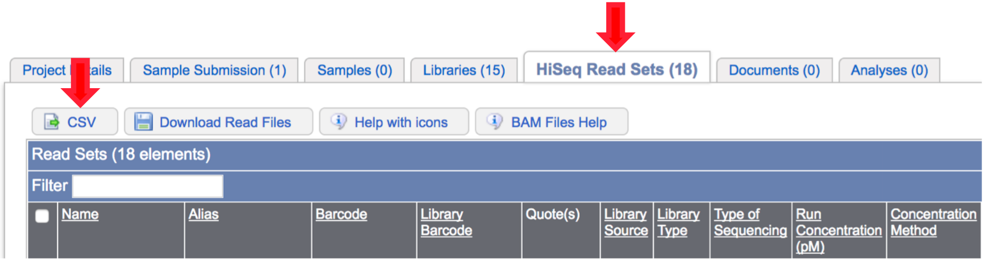
GenAP

Calcul Québec

canarie

Data from Nanuq

- csv file from Nanuq:



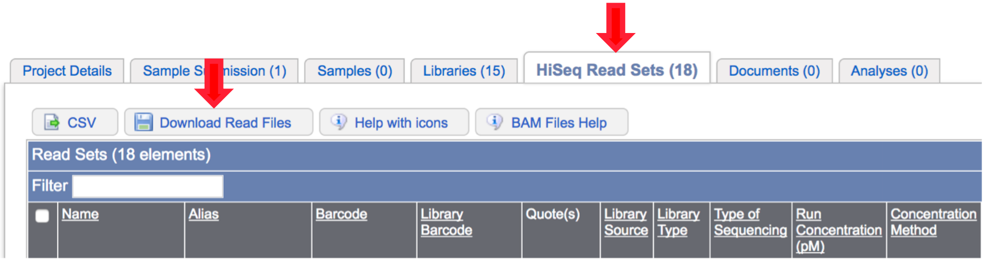
Read Sets (18 elements)

Filter

□	Name	Alias	Barcode	Library Barcode	Quote(s)	Library Source	Library Type	Type of Sequencing	Run Concentration (pM)	Concentration Method

- This file is very useful in helping you find your experimental data to prepare the readset file

Download your data from Nanuq

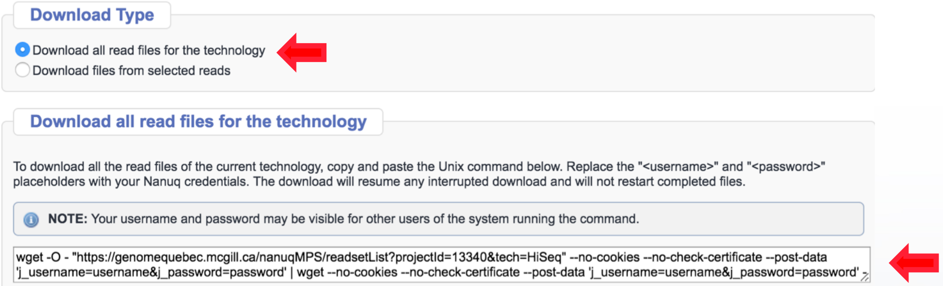


Read Sets (18 elements)

Filter

☐	Name	Alias	Barcode	Library Barcode	Quote(s)	Library Source	Library Type	Type of Sequencing	Run Concentration (pM)	Concentration Method

- if downloading all readsets in the project, copy and paste download command to the server



Download Type

● Download all read files for the technology
○ Download files from selected reads

Download all read files for the technology

To download all the read files of the current technology, copy and paste the Unix command below. Replace the "<username>" and "<password>" placeholders with your Nanuq credentials. The download will resume any interrupted download and will not restart completed files.

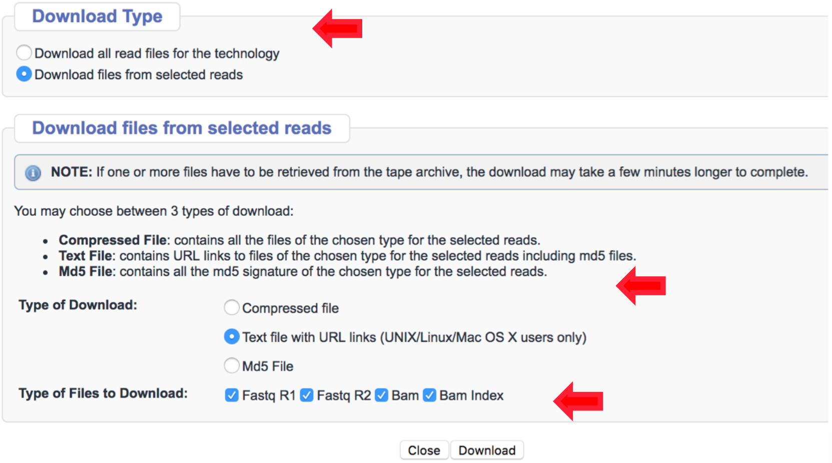
NOTE: Your username and password may be visible for other users of the system running the command.

```
wget -O - "https://genomequebec.mcgill.ca/nanuqMPS/readsetList?projectId=13340&tech=HiSeq" --no-cookies --no-check-certificate --post-data
'j_username=username&j_password=password' | wget --no-cookies --no-check-certificate --post-data 'j_username=username&j_password=password' -
```

Download your data from Nanuq

Canadian Centre for Computational Genomics

- if downloading selected readsets in the project:



Download Type

○ Download all read files for the technology
● Download files from selected reads

Download files from selected reads

NOTE: If one or more files have to be retrieved from the tape archive, the download may take a few minutes longer to complete.

You may choose between 3 types of download:

- **Compressed File**: contains all the files of the chosen type for the selected reads.
- **Text File**: contains URL links to files of the chosen type for the selected reads including md5 files.
- **Md5 File**: contains all the md5 signature of the chosen type for the selected reads.

Type of Download: ○ Compressed file
● Text file with URL links (UNIX/Linux/Mac OS X users only)
○ Md5 File

Type of Files to Download: ☑ Fastq R1 ☑ Fastq R2 ☑ Bam ☑ Bam Index

- Download the file, and follow the Readme.txt in order to download your files.