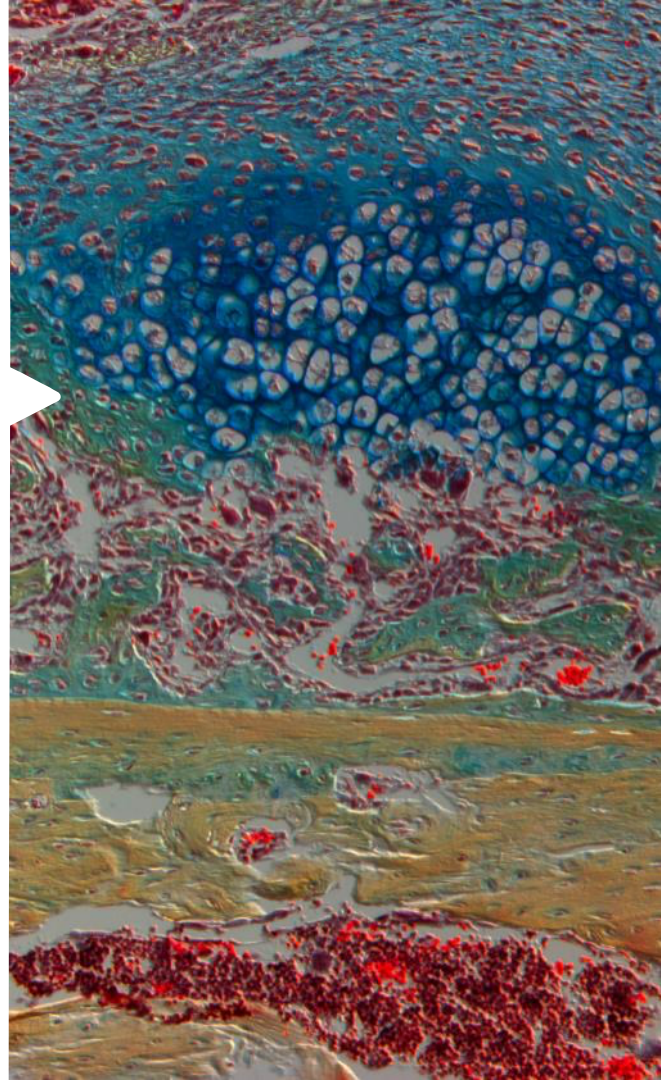




Working on Big Purple

Mark Grivainis

08 October 2020

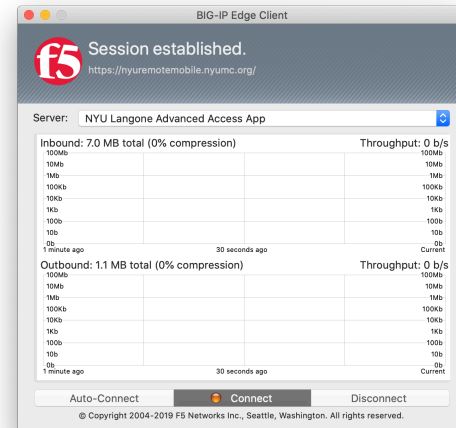
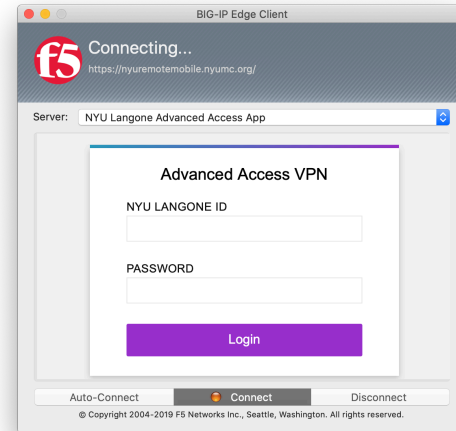


Content

- Connecting to the VPN
- Connecting to BP
- Loading modules
- Running interactive jobs
 - Running a script in interactive mode
- Installing Python Packages
- Shell Scripts
- Running a Batch job

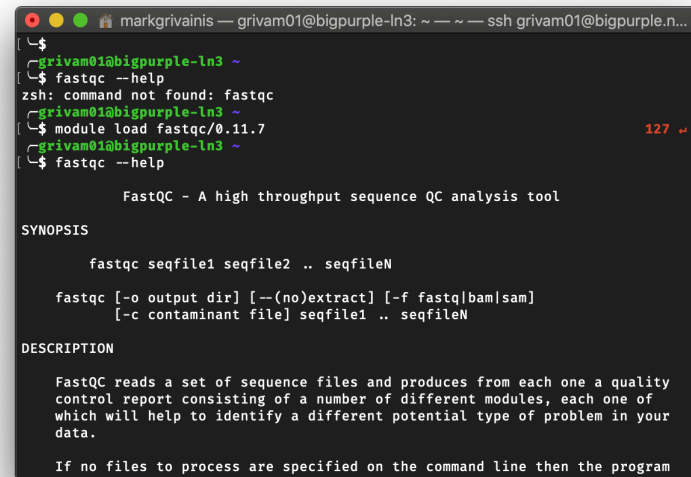
Connecting to the VPN

- Open the F5 App
- Enter your Kerberos ID and Password
- Complete the 2 – Factor authentication
- You are now logged in



Loading Modules

- If you do not know exactly what module you need use:
 - module avail
- If a module is not loaded you will get an error:
 - Command not found: <module_name>
- Load a module using:
 - module load <module_name>/version/...



```
markgrivainis — grivam01@bigpurple-ln3: ~ — ssh grivam01@bigpurple.n...
$
grivam01@bigpurple-ln3 ~
$ fastqc --help
zsh: command not found: fastqc
grivam01@bigpurple-ln3 ~
$ module load fastqc/0.11.7
grivam01@bigpurple-ln3 ~
$ fastqc --help

FastQC - A high throughput sequence QC analysis tool

SYNOPSIS

fastqc seqfile1 seqfile2 .. seqfileN

fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
        [-c contaminant file] seqfile1 .. seqfileN

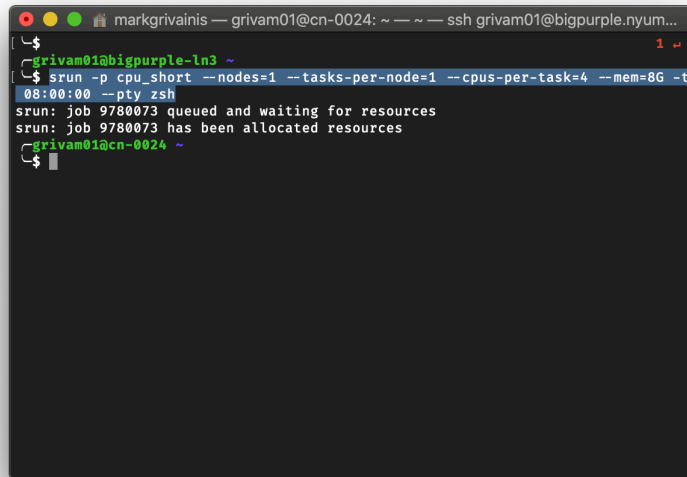
DESCRIPTION

FastQC reads a set of sequence files and produces from each one a quality
control report consisting of a number of different modules, each one of
which will help to identify a different potential type of problem in your
data.

If no files to process are specified on the command line then the program
```

Running an interactive job

- Use the 'srun' slurm command
 - `srun -p cpu_short --nodes=1 --tasks-per-node=1 --cpus-per-task=4 --mem=8G -t 08:00:00 --pty bash`
- -p: the type of node
- --nodes: How many nodes (use 1)
- --tasks-per-node: How many tasks (use 1)
- --cpus-per-task: How many CPUs to assign
 - This varies depending on what you are doing
 - For alignment use 8 or 16
 - For trimming / fastqc use 4 or 8 (remember to set the flag when calling the program)



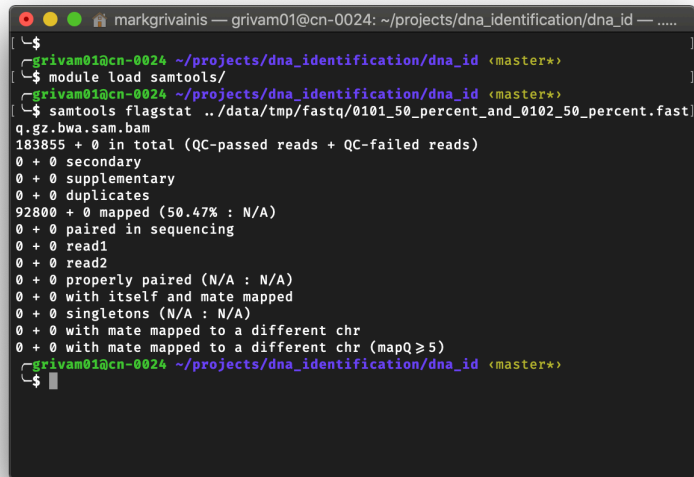
```
markgrivainis — grivam01@cn-0024: ~ — — ssh grivam01@bigpurple.nyum...
$
grivam01@bigpurple-ln3 ~
$ srun -p cpu_short --nodes=1 --tasks-per-node=1 --cpus-per-task=4 --mem=8G -t 08:00:00 --pty zsh
srun: job 9780073 queued and waiting for resources
srun: job 9780073 has been allocated resources
grivam01@cn-0024 ~
$
```

BBMap CPU/Ntask comparisons

- 1 node; 1 task; 8 CPUs per task
 - 21978.373 seconds
- 1 node; 1 task; 16 CPUs per task
 - 10641.761 seconds.
- 1 node; 1 task; 1 CPUs per task
 - > 12 hours
- 1 node; 8 tasks; 1 CPUs per task
 - 24770.645 seconds.

Running a Module in Interactive Mode

- Now that you are no longer on the head node you can run any command line applications you load
- Remember to read the documentation
 - Do this by adding the --help flag
- Once you know what the command is and what flags you need to set you can run it.
- The output that is sent to Stdout or Stderr will be printed in the terminal

A terminal window with a dark background and light-colored text. The window title is 'markgrivainis — grivam01@cn-0024: ~/projects/dna_identification/dna_id —'. The prompt is '\$'. The user enters 'module load samtools/' and the prompt changes to 'grivam01@cn-0024 ~/projects/dna_identification/dna_id <master*>'. The user then enters 'samtools flagstat ../data/tmp/fastq/0101_50_percent_and_0102_50_percent.fastq.gz.bwa.sam.bam'. The output shows statistics for the BAM file, including total reads, mapped reads, and various quality metrics.

```
$ module load samtools/
grivam01@cn-0024 ~/projects/dna_identification/dna_id <master*>
$ samtools flagstat ../data/tmp/fastq/0101_50_percent_and_0102_50_percent.fastq.gz.bwa.sam.bam
183855 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
92800 + 0 mapped (50.47% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ≥5)
grivam01@cn-0024 ~/projects/dna_identification/dna_id <master*>
$
```


Installing Python Packages

- You can either use conda (anaconda / miniconda) or pip
- I prefer using pip as I have encountered less issues
- My go-to versions of python is:
 - `module load python/cpu/3.7.2`
 - `module load python/gcc/3.6.5`
- Pip should be installed with this version of Python, it is not with all the versions on BP
- If it is not installed, this command will install it after loading the python module
 - `python -m pip install --user pip`
- Pip will normally try to install packages into a folder that you do not have access to on BP
- To avoid any issues, install packages into your home directory using the `--user` flag
 - `pip install --user cutadapt`
- Remember to load the same version of python that was used to install the packages every time.

Shell Scripts

- Shell scripts can range from a few lines to thousands of lines of code
- For this course, shell scripts will act as wrapper scripts when running sbatch.
 - A wrapper script will configure the environment for the batch job run by Slurm
 - The main components in the script are:
 - Shebang
 - Setting up the Slurm parameters
 - Purging and loading the correct modules
 - Loading a virtual environment / conda environment
 - Setting up any folders for output.
 - Running the program required for the analysis

```
driver.sh
1 #!/bin/bash shebang
2 #SBATCH --job-name=optimization_driver # Job name
3 #SBATCH --ntasks=1 # Run on a single CPU
4 #SBATCH --cpus-per-task=1
5 #SBATCH --mem=4gb # Job memory request
6 #SBATCH --time=12:00:00 # Time limit hrs:min:sec
7 #SBATCH -p=cpu_short
8
9
10 # Load all the packages you need
11 module purge
12 module load anaconda3
13 #conda activate py36
14 module load slurm/current
15
16 echo "Running Driver"
17 # run the python driver script
18 python driver.py
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

Slurm Parameters

Set up required modules

Run the script

NORMAL driver.sh sh 77% 14/18 h:25

My Workflow

1. Start with an interactive job
 1. Load the modules you require (Keep track of the modules you use for writing .sh script)
 2. Use the command line application to run your analysis
 1. If your dataset is large, try to make a small version of it for testing
 1. For example, if you are working with FASTQ files, just include the first 100 reads from both files
 2. Read the documentation, focus on what flags are available to you, and what flags you require
 3. Once the script is running build your shell script using the commands you ran in the interactive session
2. Run the shell script using sbatch and the subset of the data, does it match the ijob
3. Run the same shell script using the entire dataset

Adding Command Line Arguments to Shell

- Positional command line arguments are represented as variables in a shell script
- Variables start with a '\$' symbol
 - \$0 is the name of the script
 - Any subsequent numbers will be the argument in those positions
- Use these variables to pass on command line arguments to the wrapped script
 - In qc.sh
 - ...
 - fastqc <flags> \$1 \$2
 - sbatch qc.sh test.R1 test.R2

```
# shell script 'cmd_args.sh'
echo $1
echo $2
# ---- end of script ----
```

```
$ sh cmd_args.sh argument1 argument2
argument1
argument2
```

Some Useful Commands

- `head <file>`
 - Output the first 10 lines of the file
 - “-n 100” flag will output the first 100 lines
- Piping
 - `<cmd> | <cmd>`
 - will feed the output of one command to the next
 - `cat <filename> | head`
 - `cmd > file`
 - output the result of the command to specified file
- `cmd >> file`
 - Append the result of the command to the file
- `grep “pattern” file`
 - Search for the specified pattern in the file
- `cat <file>, less <file>, more <file>`
 - Will all output the file to the terminal, all have different features.
 - `cat` can also be used to join files together

ssh config file

- On mac / unix the config file is here:
 - ~/.ssh/config
- This might be the same on Windows when using powershell
- Add the contents in the right text box to your config file,

```
Host bp
```

```
HostName bigpurple.nyumc.org
```

```
User <Kerberos id>
```

Sbatch

- Sbatch will run a shell script in batch mode
 - This runs the script in the background, so you can carry on working on BP while it is running
 - When the job completes any output that would have been written to the terminal is written to:
 - `slurm-<job_id>.out`
 - Use `cat` / `more` / `less` to view the contents
 - If there are any errors they will be shown here
- If you have set your shell script up to use arguments you would pass them to the bash script
 - `sbatch <script_name.sh> <arg1> <arg2>`
- Check the progress of your job using:
 - `squeue -u <Kerberos Id>`
- The 'ST' column will tell you the status of the job. R means it is running, PD means the job is queued

Issues

- If you get any write issues when running apps like ‘sra-toolkit’ it might mean that you have gone over your file limit quota
 - When you go over the limit your home directory is set to be read only
 - Always use your scratch space for large files / directories with lots of folders

VIM

- This article covers the basics
 - <https://opensource.com/article/19/3/getting-started-vim>
- I have added my vim dot file to the Brightspace page for this lecture.
 - I recommend building your own dotfile using mine as a template
 - Dotfile basics tutorial:
 - https://linuxhint.com/configure_vim_vimrc/
- If you have any questions, send me a slack message

SCP

- Secure copy
- Use scp to copy files to and from big purple
 - `scp <path_to_original> <path_to_copy>`
- Copying to BP
 - `scp <path_to_file> <Kerberos>@bigpurple.nyumc.org:<path_to_save_to>`
- Copying from BP
 - `scp <Kerberos>@bigpurple.nyumc.org:<path_to_file> <path_to_save_to>`
- You can add the ‘-r’ flag to copy recursively (use this to copy directories)