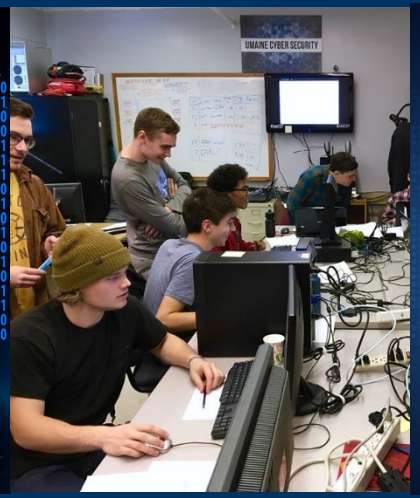
# Part 2- Introduction to HPC

**March 31, 2023**



*Laura Jackson, Ph.D. - Integrative Data Scientist, ARCSIM*

*Kate Beard, Ph.D. - Professor of Spatial Computing, School of Computing and Information Science*


THE UNIVERSITY OF MAINE

ARCSIM MAINE
Advanced Research Computing, Security & Information Management

MAINE eDNA

THE UNIVERSITY OF MAINE

# *Part 2:* Introduction to HPC on OSC

**March 31 12:00pm– 1:00pm**

Package managers for containerized environments

- Installing new software on HPC
- Use different software versions

Shell scripts and submitting jobs to a schedular

Running an interactive job using command line

# Accessing Terminal on your local computer

- If you are on a Mac, you have a Bash terminal pre-installed on your computer
- If you are on a PC, you can install terminal emulators free online
  - PuTTY – Most popular ([https://www.putty.org](https://www.putty.org))

➢ To connect using your terminal, use the shell command with your OSC username and password

```
ssh username@owens.osc.edu
```

Enter password at prompt and press enter

```
~ ssh lmjack@owens.osc.edu
*******************************************************************************

   This system is for the use of authorized users only.  Individuals using
   this computer system without authority, or in excess of their authority,
   are subject to having all of their activities on this system monitored
   and recorded by system personnel.  In the course of monitoring individuals
   improperly using this system, or in the course of system maintenance,
   the activities of authorized users may also be monitored.  Anyone using
   this system expressly consents to such monitoring and is advised that if
   such monitoring reveals possible evidence of criminal activity, system
   personnel may provide the evidence of such monitoring to law enforcement
   officials.

*******************************************************************************
lmjack@owens.osc.edu's password:
```

THE UNIVERSITY OF MAINE

https://ondemand.osc.edu/

# Ohio Supercomputer Center

An OH·TECH Consortium Member

Log in with your OSC username and password.

This system is for the use of authorized users only. Individuals using this computer system without authority, or in excess of their authority, are subject to having all of their activities on this system monitored and recorded by system personnel. In the course of monitoring individuals improperly using this system, or in the course of system maintenance, the activities of authorized users may also be monitored. Anyone using this system expressly consents to such monitoring and is advised that if such monitoring reveals possible evidence of criminal activity, system personnel may provide the evidence of such monitoring to law enforcement officials.

If making changes to your OSC user account, please allow up to 20 minutes for the changes to take effect. This includes password changes, group membership changes, and new users having login ability to OSC systems using SSH or OnDemand.

**Username**

**Password**

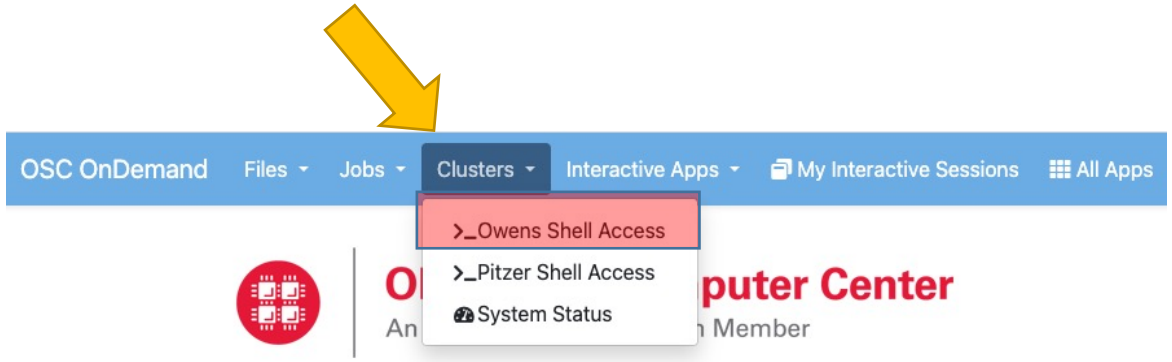
Remember me

Log in with your OSC account

Log in with third party through CILogon

**Step 1. Choose your identity provider**

CILogon provides access to identity providers from many academic institutions across the state.

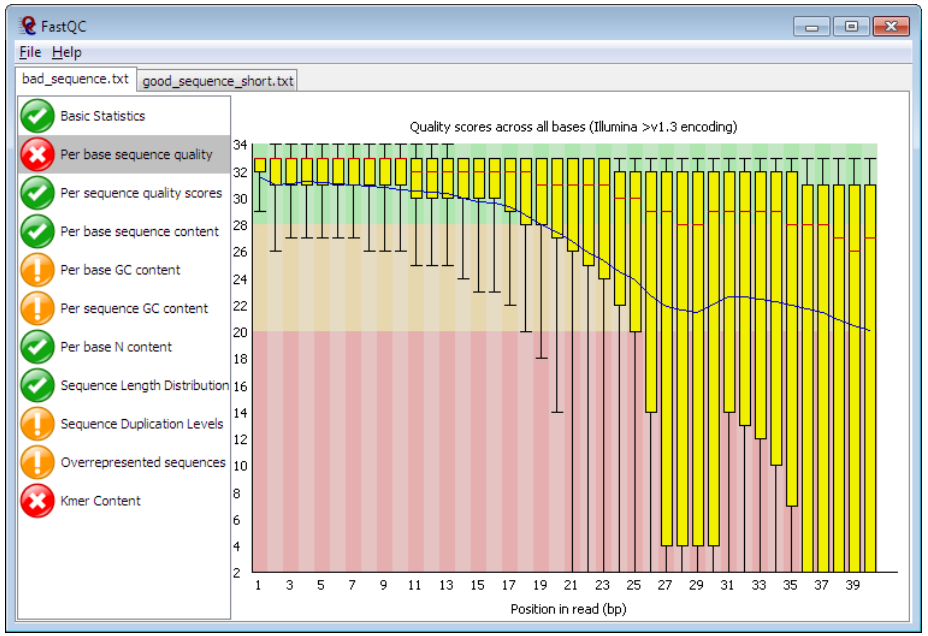


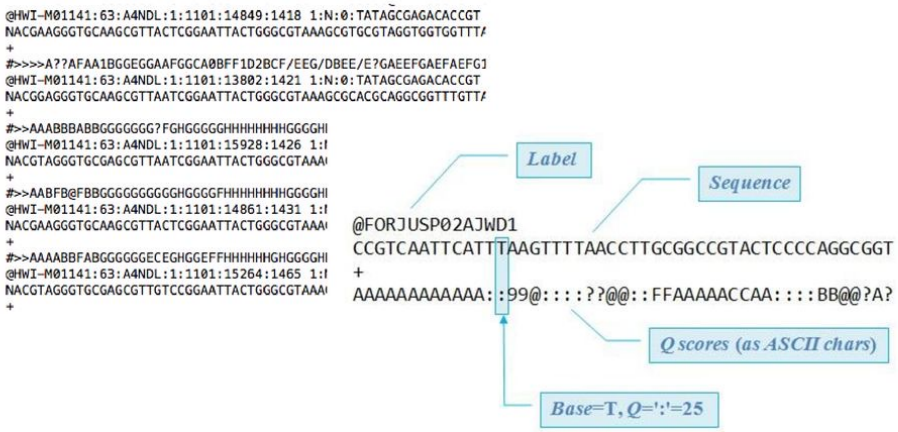

```
   officials.

*******************************************************************************
Last login: Wed Mar 22 12:01:29 2023 from web02.hpc.osc.edu
*******************************************************************************
Welcome to the Ohio Supercomputer Center!
PLEASE SEE THE OSC SUPERCOMPUTING DOCUMENTATION
FOR UP-TO-DATE INFORMATION ON OUR RESOURCES:
https://www.osc.edu/services/cluster_computing
For questions or assistance, contact oschelp@osc.edu

To check for currently known issues, please visit
https://www.osc.edu/resources/technical_support/known_issues
and follow @HPCNotices on Twitter.
*******************************************************************************
2020/03/16
--- OSC Classroom Support

CLASSROOM RESOURCES FOR DISTANCE LEARNING

If your class has lost or limited access to computer labs, the Ohio
Supercomputer Center might be able to help by providing no-cost access
to cloud computing resources. Classes and workloads of any size can
gain access. OSC's web-browser interface to its substantial Linux
computer systems provides novice users with virtual desktops preloaded
with applications, such as MATLAB, RStudio, or Jupyter Notebook.

As an example, an OSU undergrad statistics class recently used iPads
to remotely access RStudio on OSC systems. We can provide online
demonstrations or evaluations and potentially add additional software
packages.

Please contact OSCHelp@osc.edu to talk to OSC about distance-learning
support options available to you.

*******************************************************************************


Your primary project is the only group quota reported here.
See http://osc.edu/check-quotas to learn how to check the group storage quotas for your other projects.

As of 2023-03-22T14:40:01.000000 userid lmjack on /fs/ess/PUOM0014 used 13 GiB of quota 0 GiB and 419 f
As of 2023-03-22T14:40:01.000000 userid lmjack on /fs/ess/PUOM0015 used 0 GiB of quota 0 GiB and 1 file
As of 2023-03-22T14:40:01.000000 userid lmjack on /fs/ess/PUOM0012 used 0 GiB of quota 0 GiB and 1 file
As of 2023-03-22T14:42:10.000000 userid lmjack on /users/PYS1047 used 22.38 GiB of quota 500 GiB and 36

-bash-4.2$
```

# FastQC and MultiQC

**FastQC** - A quality control tool for high throughput sequence data

-Input raw sequence data (FastQ, SAM, BAM)
-Generates quality plots (.html) for each sample



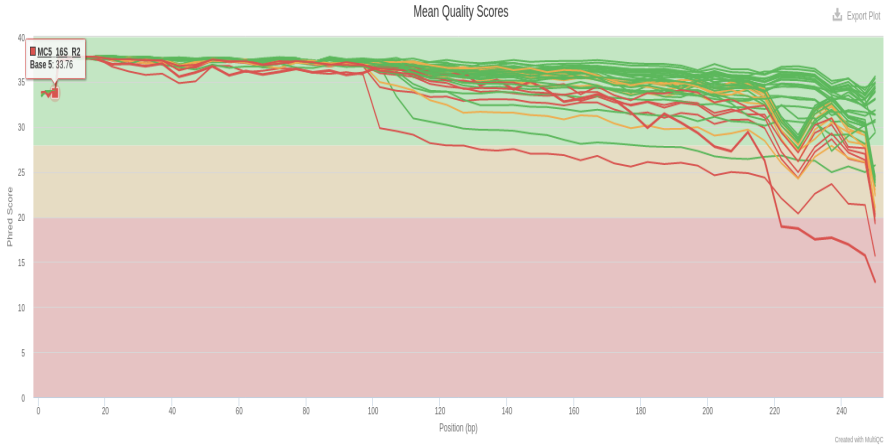



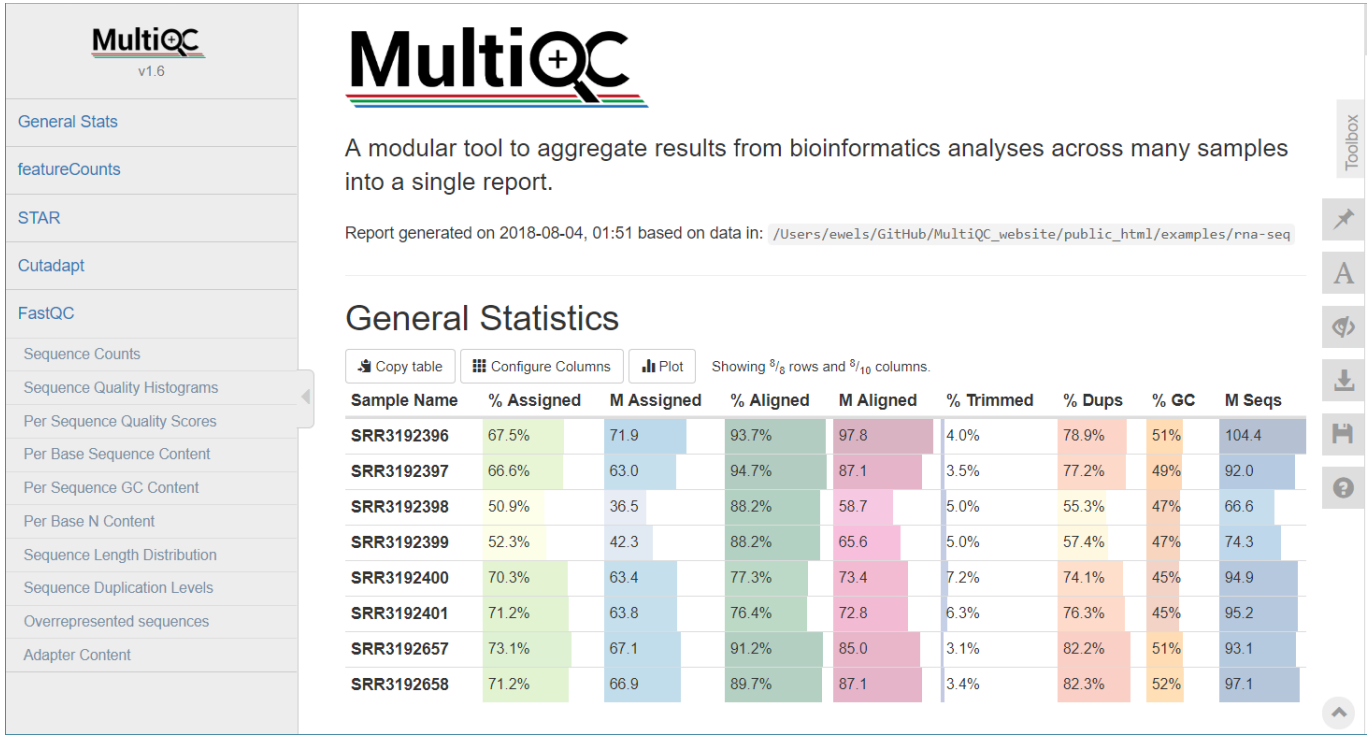

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

# FastQC and MultiQC

- Uses the output of FastQC to aggregate the quality metrics of the different reads together, to allow inter-read set comparisons

- Not available as a module,
  so we will need to install;
  will discuss in a minute…..




| Sample Name | % Assigned | M Assigned | % Aligned | M Aligned | % Trimmed | % Dups | % GC | M Seqs |
|---|---|---|---|---|---|---|---|---|
| SRR3192396 | 67.5% | 71.9 | 93.7% | 97.8 | 4.0% | 78.9% | 51% | 104.4 |
| SRR3192397 | 66.6% | 63.0 | 94.7% | 87.1 | 3.5% | 77.2% | 49% | 92.0 |
| SRR3192398 | 50.9% | 36.5 | 88.2% | 58.7 | 5.0% | 55.3% | 47% | 66.6 |
| SRR3192399 | 52.3% | 42.3 | 88.2% | 65.6 | 5.0% | 57.4% | 47% | 74.3 |
| SRR3192400 | 70.3% | 63.4 | 77.3% | 73.4 | 7.2% | 74.1% | 45% | 94.9 |
| SRR3192401 | 71.2% | 63.8 | 76.4% | 72.8 | 6.3% | 76.3% | 45% | 95.2 |
| SRR3192657 | 73.1% | 67.1 | 91.2% | 85.0 | 3.1% | 82.2% | 51% | 93.1 |
| SRR3192658 | 71.2% | 66.9 | 89.7% | 87.1 | 3.4% | 82.3% | 52% | 97.1 |

# HPC Modules (Pre-installed software)

-You can check all available modules (and versions) by using `module avail`
-To check a specific software, use the `module spider {software}` command

Example: Check all versions of FastQC available to use `module spider fastqc`
-Version 0.11.7 is the default version that is used

To use a module, you first need to load it using `module load fastqc/0.11.7`

When loaded, you can run the help command to see all functions `fastqc --help`

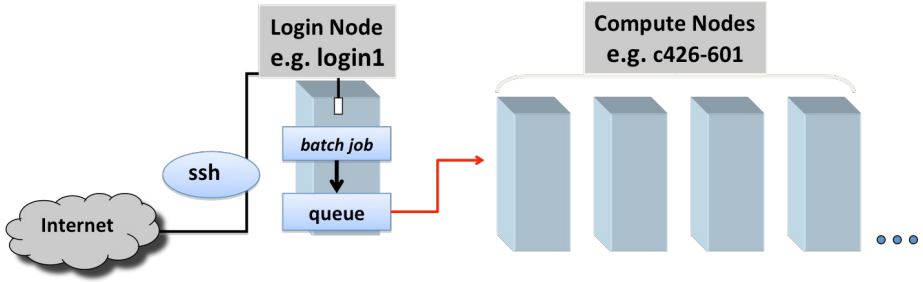
| Module Command | Shortcut | Description |
|---|---|---|
| module avail | ml avail | List available modules |
| module list | ml | Show modules currently loaded |
| module load *package*<br>module add *package* | ml load *package*<br>ml add *package* | Load a selected module |
| module unload *package*<br>module rm *package* | ml unload *package*<br>ml rm *package* | Unload a previously loaded module |
| module swap *package1 package2* | ml swap *package1 package2* | Unload package1 and load package2 |
| module purge | ml purge | Unload all loaded modules |

# Run Software Interactively

- Now that we have FastQC software module loaded, we can use it **interactively** or by **Slurm job submission**

- First, we need to navigate to our folder that contains our raw sequence files (fastq)
  `cd /fs/ess/PUOM0012/linux` now check that the fastq files are there by using `ls *.fastq`

- We are currently on a login node, we need to get access to an interactive/compute node to run jobs
  `sinteractive -A <project-code>` ***Our project is PUOM0012**
  -Once you are done using compute, you type `exit` to return to login node




- Now we are ready to run FastQC on our files to produce the quality plots
  `fastqc *.fastq`
  -Type `ls` to see our new FastQC output files (.zip and .html)

- To transfer files to your computer, you can use sftp command or GUI software like CyberDuck (https://cyberduck.io/)

# New Directory for FASTQC Outputs

- First, we need to make a new directory (*check if you are in /fs/ess/PUOM0012/linux by running **pwd**)

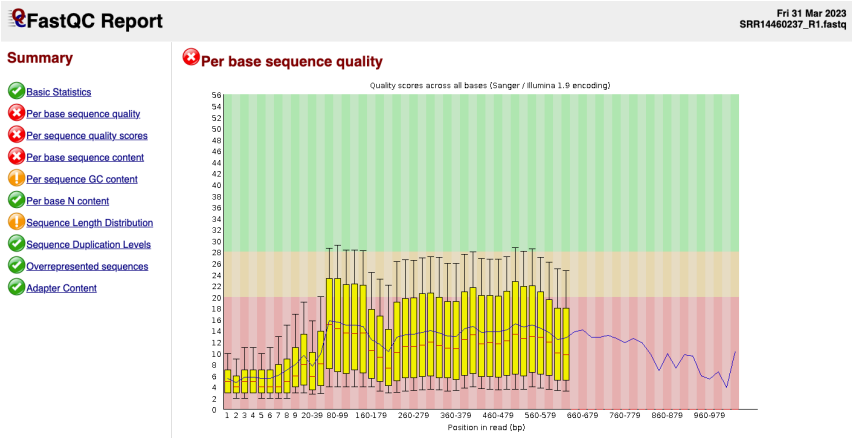
```
mkdir fastqc_interactive
```

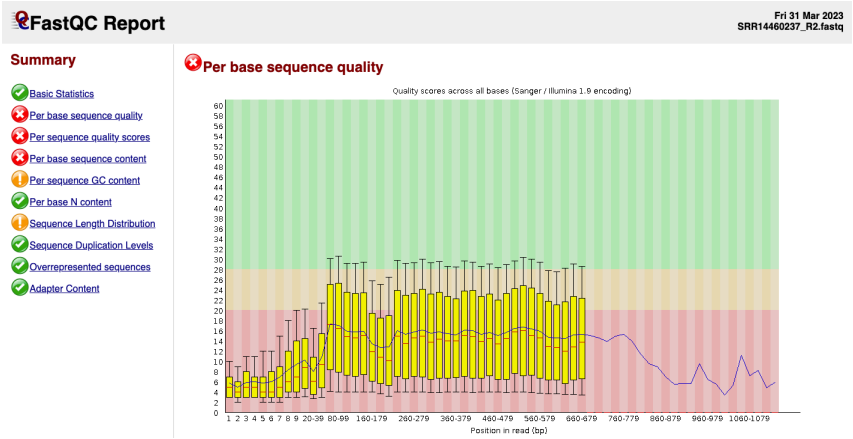
- Next, we will move FastQC outputs (.zip and .html) to the new folder

```
mv *.zip fastqc_interactive
mv *.html fastqc_interactive
```

-Now you can see if the new files moved correctly by typing ls and the path,

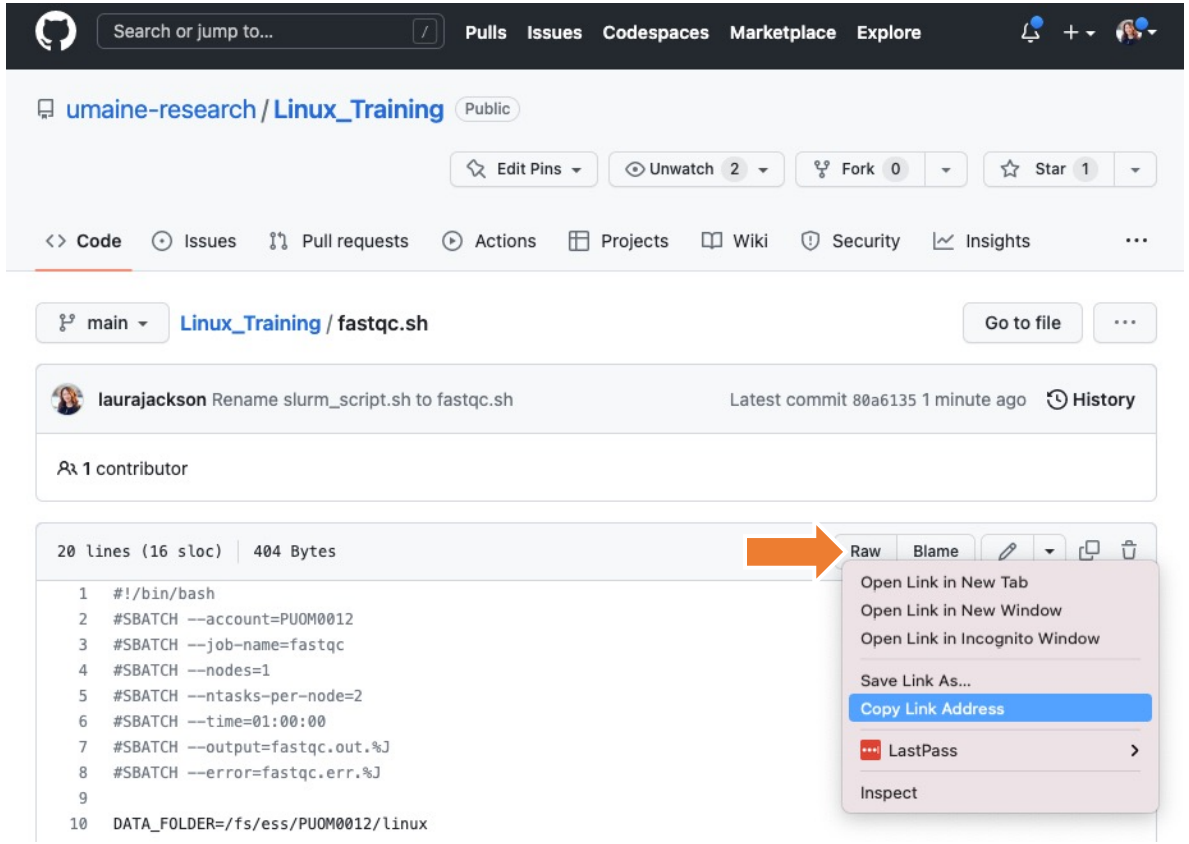
```
ls fastqc_interactive/
```



# Downloading files from Github

- Using the `wget` command
  - Right click on the Raw button for a file of interest
  - Copy link address to use with the `wget` command



# Run Software as a Job Submission

To run batch jobs, sent to a scheduling queue, you need to create a script

-Contains:

1- slurm syntax

2- path to files

3- some commands to run

```
#!/bin/bash                    # first line must identify interpreter
#SBATCH -J job-name            # create a short name for your job
#SBATCH -o output.o%j          # create a name for output file log
#SBATCH -e error.e%j           # create a name for error file log
#SBATCH -p normal              # submit job to designated queue
#SBATCH -N 1                   # node count
#SBATCH -n 34                  # total number of tasks across all nodes
#SBATCH -t 00:01:00            # total run time allocated (HH:MM:SS)
#SBATCH --mail-type=all        # send email on start, end and fault
#SBATCH --mail-user=<YourEmail>@maine.edu
#SBATCH -A projectNumber       # charge job to project allocation

date                            #prints date/time of run
```

- First, we will pull a demo script from Github using wget

  **wget "https://github.com/umaine-research/Linux_Training/raw/main/fastqc.sh"**

- We will open it using nano to edit the file

  **nano fastqc.sh**

- We need to make sure the script has execute permissions set

  **chmod u+x fastqc.sh**

- Now, we can submit the job as a batch script (*will run even if not logged in)

  **sbatch fastqc.sh**

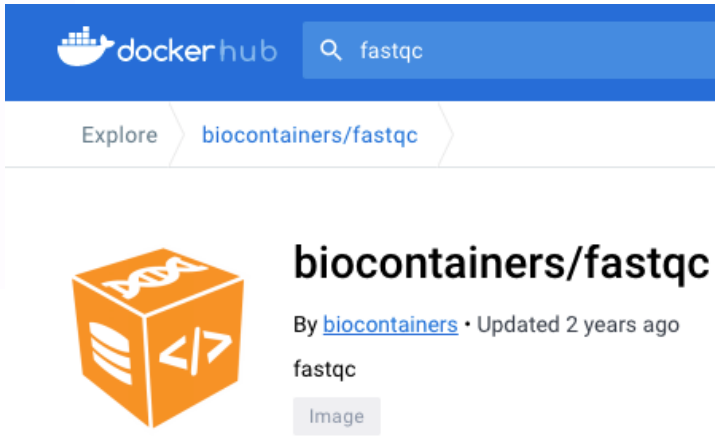
# Monitoring Job Submission

| Use | Command |
|---|---|
| Submit a batch job | sbatch <script_name> |
| Submit an interactive job | sinteractive [options] or salloc [options] |
| Check job status | squeue –u <username> |
| Delete a job | scancel <jobid> |
| Hold a job | scontrol hold <jobid> |
| Release a job from hold and send to queue | scontrol release <jobid> |

# Containerized Environments / Software Install

## CONTAINERS: PRIMARY USE CASES

- Reproducibility!
- Extreme mobility of compute (portability)
- Provides the users, developers and engineers FREEDOM!
  - Can easily leverage other people's work
  - Changes the software distribution and dependency paradigm
  - Depending on application and use-case, simple extreme scalability
  - Containers are in many ways the next logical progression from virtual machines



# Anaconda and Installing new software

- Over 8,000 open-source data science and machine learning packages
- **Anaconda**-built and compiled for all major OS
- Simplifies package management and deployment
- Miniconda3 module pre-installed at OSC

```
module load miniconda3
```

➢ You need to create new environments and install software into that environment
-Can have many environments based on different projects

➢ First, we will create a new environment and load that environment
conda create –n {environment_name}
**`conda create -n multiqc`**
proceed ([y]/[n])? **`type y+enter`**

➢ Then you need to activate that environment to use it
**`source activate multiqc`**

```
-bash-4.2$ source activate multiqc
(multiqc) -bash-4.2$ conda deactivate
```

## Installing on Linux

1. Download the installer:
   - Miniconda installer for Linux.
   - Anaconda installer for Linux.
2. Verify your installer hashes.
3. In your terminal window, run:
   - Miniconda:

```
bash Miniconda3-latest-Linux-x86_64.sh
```

https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html

# Running jobs with conda environment software

- Now that we have a conda environment called ‘multiqc’, we are now ready to install software

  `conda install -c bioconda multiqc=1.14`

  proceed ([y]/[n])? `type y+enter`

- If loaded correctly, using the `conda list` command will output all software and versions in the active environment

  -You should now see multiqc v1.14 listed

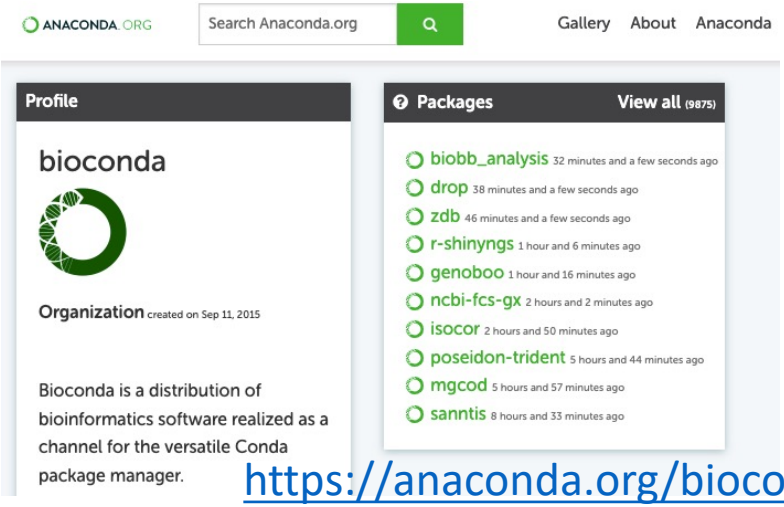
- To close an environment **use** `conda deactivate`

CONDA

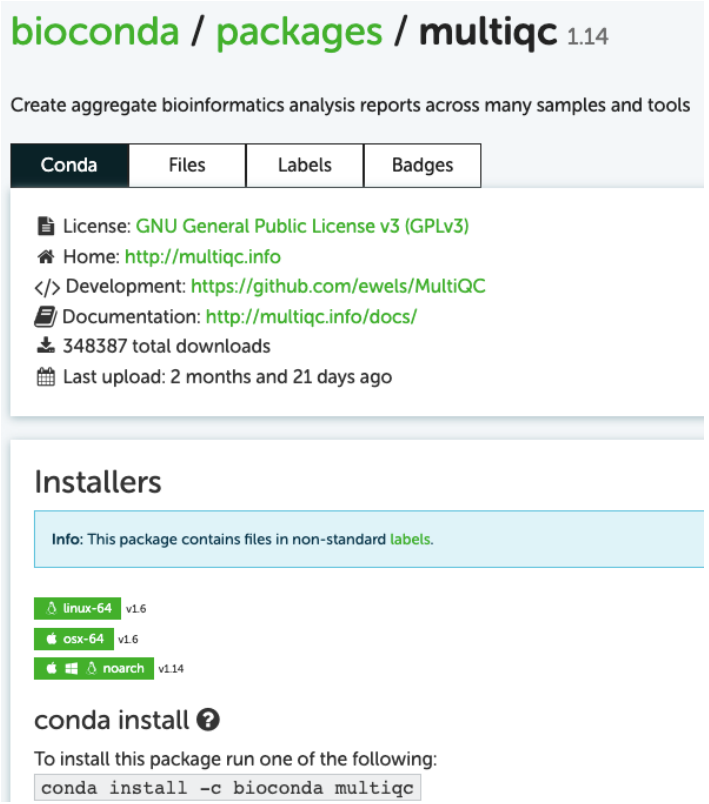
CONDA CHEAT SHEET

Command line package and environment manager

CHEAT SHEET Link




https://anaconda.org/bioconda




https://anaconda.org/bioconda/multiqc

# Running MultiQC Interactively

- To run on an **interactive node**

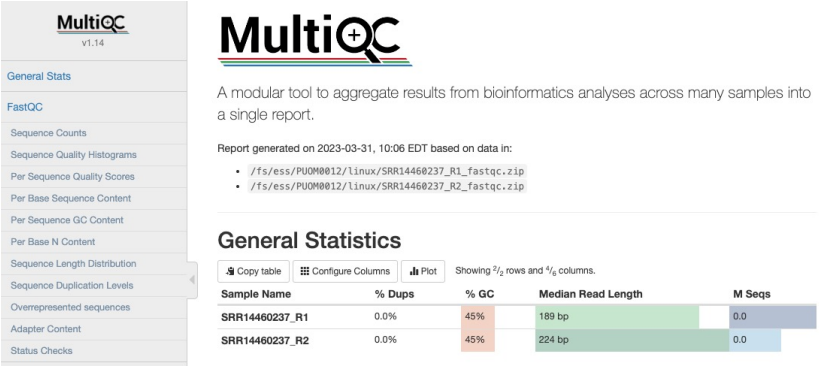
```
sinteractive -A <project-code>
module load miniconda3
source activate multiqc
```

```
-bash-4.2$ source activate multiqc
(multiqc) -bash-4.2$
```

- Now we can navigate to our files to run multiqc

```
cd /fs/ess/PUOM0012/linux
multiqc *fastqc.zip
```

```
(multiqc) -bash-4.2$ multiqc *fastqc.zip

 /// MultiQC 🔍 | v1.14

|           multiqc | Search path : /fs/ess/PUOM0012/linux/SRR14460237_R1_fastqc.zip
|           multiqc | Search path : /fs/ess/PUOM0012/linux/SRR14460237_R2_fastqc.zip
|          searching | ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 100% 2/2
|            fastqc | Found 2 reports
|           multiqc | Compressing plot data
|           multiqc | Report      : multiqc_report.html
|           multiqc | Data        : multiqc_data
|           multiqc | MultiQC complete
```




To run as a **batch job** include in script

```
module load miniconda3
source activate multiqc
```

## Slurm Script to Run as Batch Job

```
#!/bin/bash
#SBATCH --account=PUOM0012
#SBATCH --job-name=multiqc
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=2
#SBATCH --time=01:00:00
#SBATCH --output=multiqc.out.%J
#SBATCH --error=multiqc.err.%J

DATA_FOLDER=/fs/ess/PUOM0012/linux

#Load Required Module
module load miniconda3

#Load Conda Environment
source activate multiqc

#Move into data folder
cd $DATA_FOLDER

#Run MultiQC
#The files may end in .fastq.zip or .fastq.qz.zip
multiqc *fastq.zip
```

# Summary

- To run jobs interactively using a module

```
module load {software_name}
```

- To run jobs interactively using a conda environment

-Create a conda environment and install your software

```
module load miniconda3
source activate {environment_name}
```

- To submit batch jobs using a module, include in script

```
module load {software_name}
```

- To submit batch jobs using an environment, include in script

-Create a conda environment and install your software

```
module load miniconda3
source activate {environment_name}
```

```
-bash-4.2$ sinteractive -A PUOM0012
salloc: Pending job allocation 24086829
salloc: job 24086829 queued and waiting for resources
salloc: job 24086829 has been allocated resources
salloc: Granted job allocation 24086829
salloc: Waiting for resource configuration
salloc: Nodes o0647 are ready for job
(base) bash-4.2$ module load fastqc
(base) bash-4.2$ module load miniconda3
(base) bash-4.2$ source activate multiqc
(multiqc) bash-4.2$
```

```
#!/bin/bash
#SBATCH --account=PUOM0012
#SBATCH --job-name=multiqc
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=2
#SBATCH --time=01:00:00
#SBATCH --output=multiqc.out.%J
#SBATCH --error=multiqc.err.%J

DATA_FOLDER=/fs/ess/PUOM0012/linux

#Load Required Module
module load miniconda3

#Load Conda Environment
source activate multiqc

#Move into data folder
cd $DATA_FOLDER

#Run MultiQC
#The files may end in .fastq.zip or .fastq.qz.zip
multiqc *fastq.zip
```

# Any Questions

https://forms.gle/uZEgpUf7CrmqtjMCA

## Post-Linux/HPC Workshop Survey

We use your feedback to improve future programming and bioinformatics workshops and greatly appreciate your input on any aspect of your training experience.

This survey will be anonymous and all responses will not be linked to individual participants.