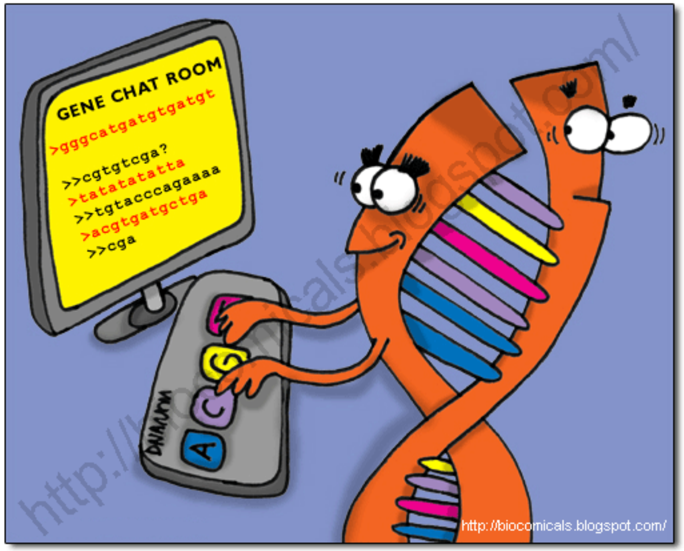
USING BIOINFORMATICS TOOLS ON RIVANNA

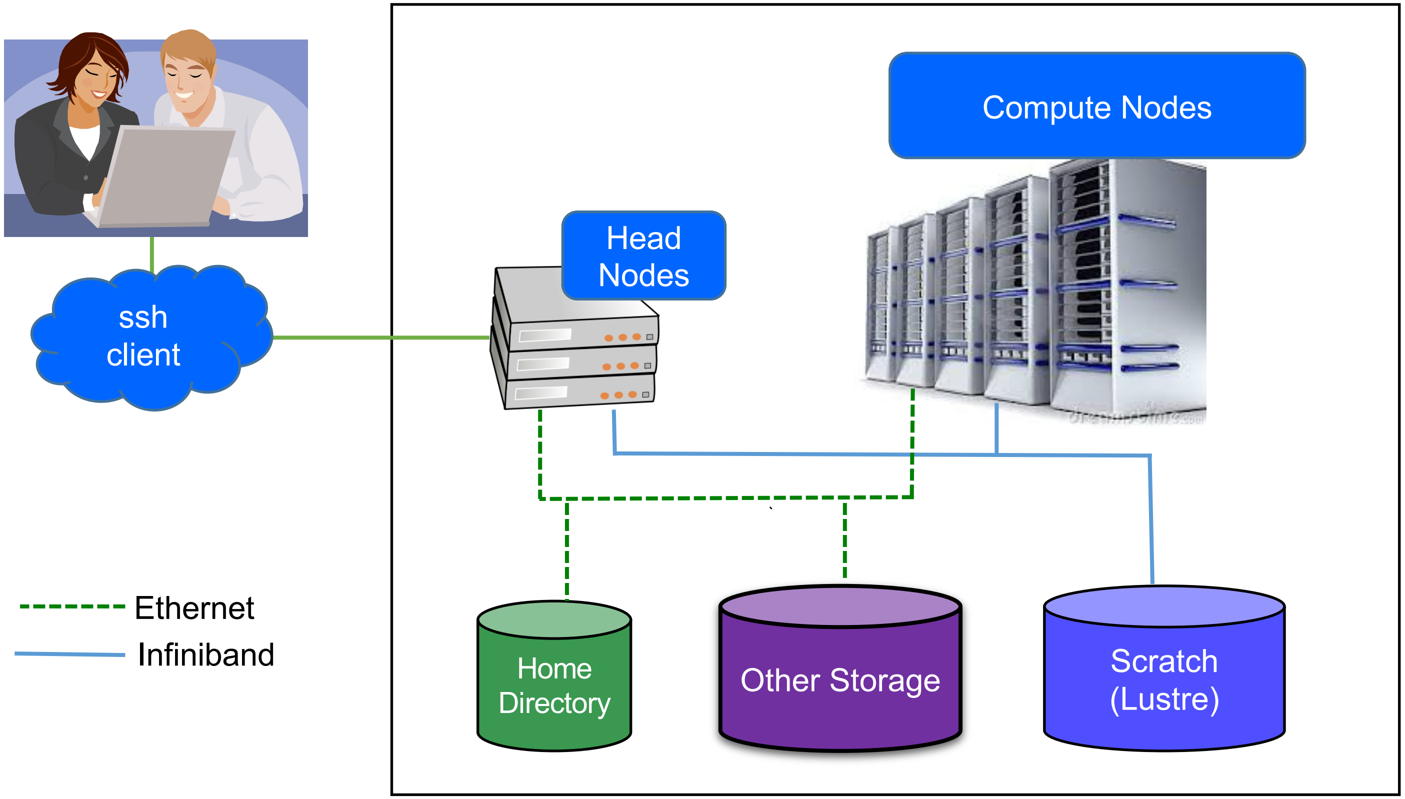
Gladys K Andino, PhD
Senior Computational Scientist
E gka6a@virginia.edu

OUTLINE

- Logging in
  - OOD
  - MobaXterm – PC (ssh, shell, SFTP)
  - Terminal/SSH - Mac
- Basic Unix commands
- Modules -How to load modules
- Practical
  - Fastqc
  - Trimmomatic
  - Bowtie2
  - Samtools
  - Qualimap



RIVANNA

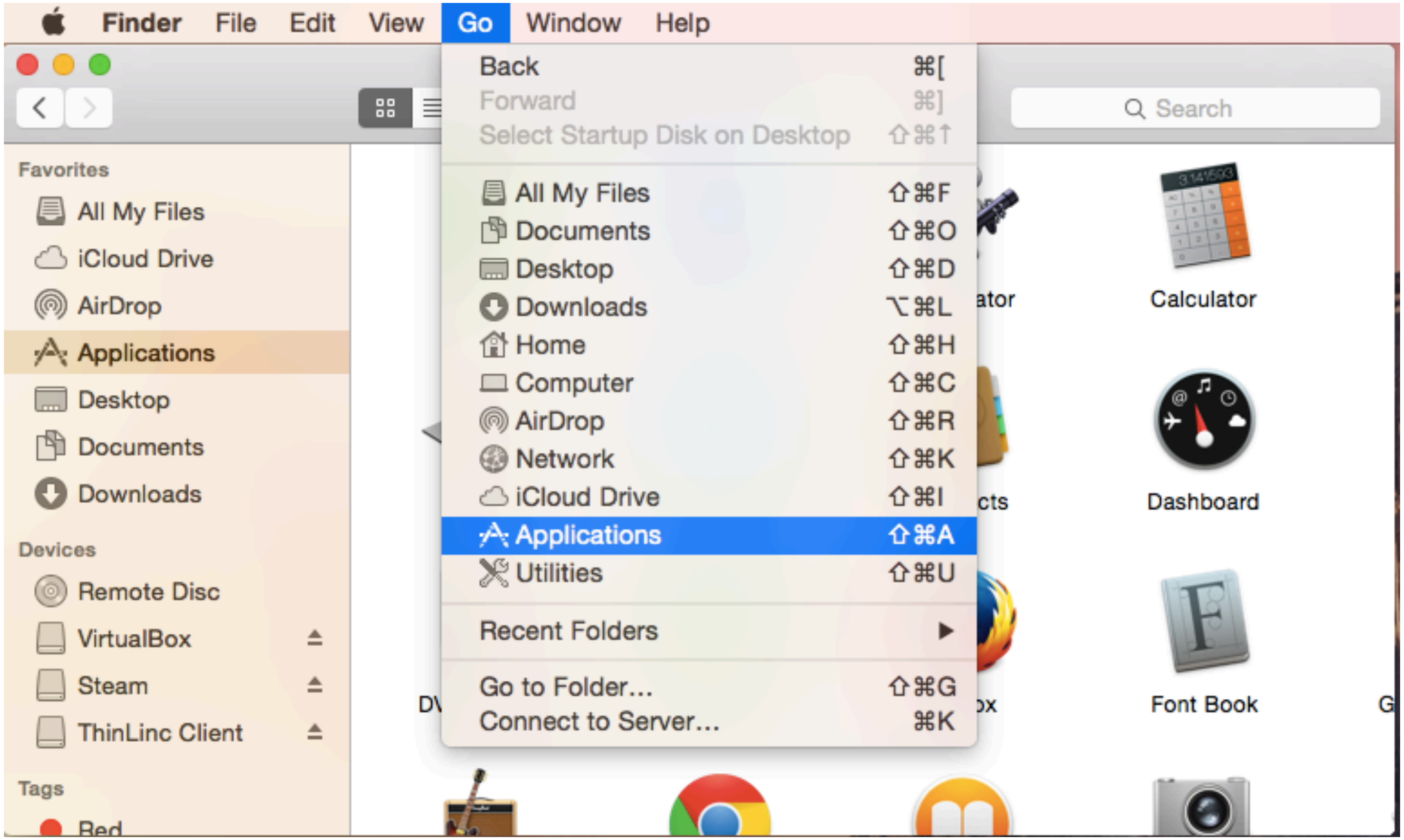

Compute Nodes
Head Nodes
ssh client
Ethernet
Infiniband
Home Directory
Other Storage
Scratch (Lustre)

LOGGING IN

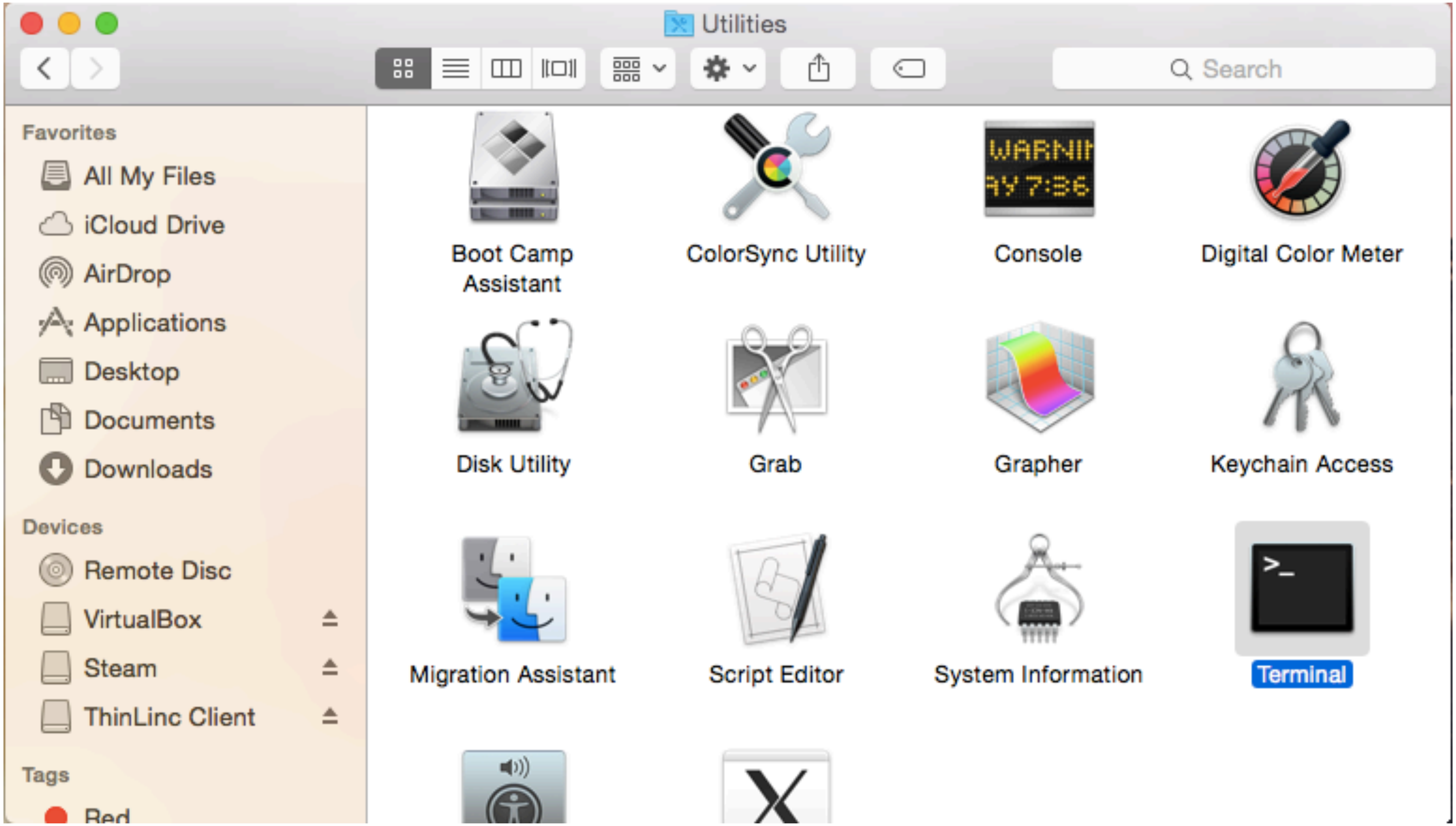
- Logging into a remote UNIX based system requires a client
- Based on the “SSH” or Secure Shell protocol
  - Encrypted
  - Used on most UNIX systems
  - Variety of clients for all platforms

LOGGING IN – using a MAC

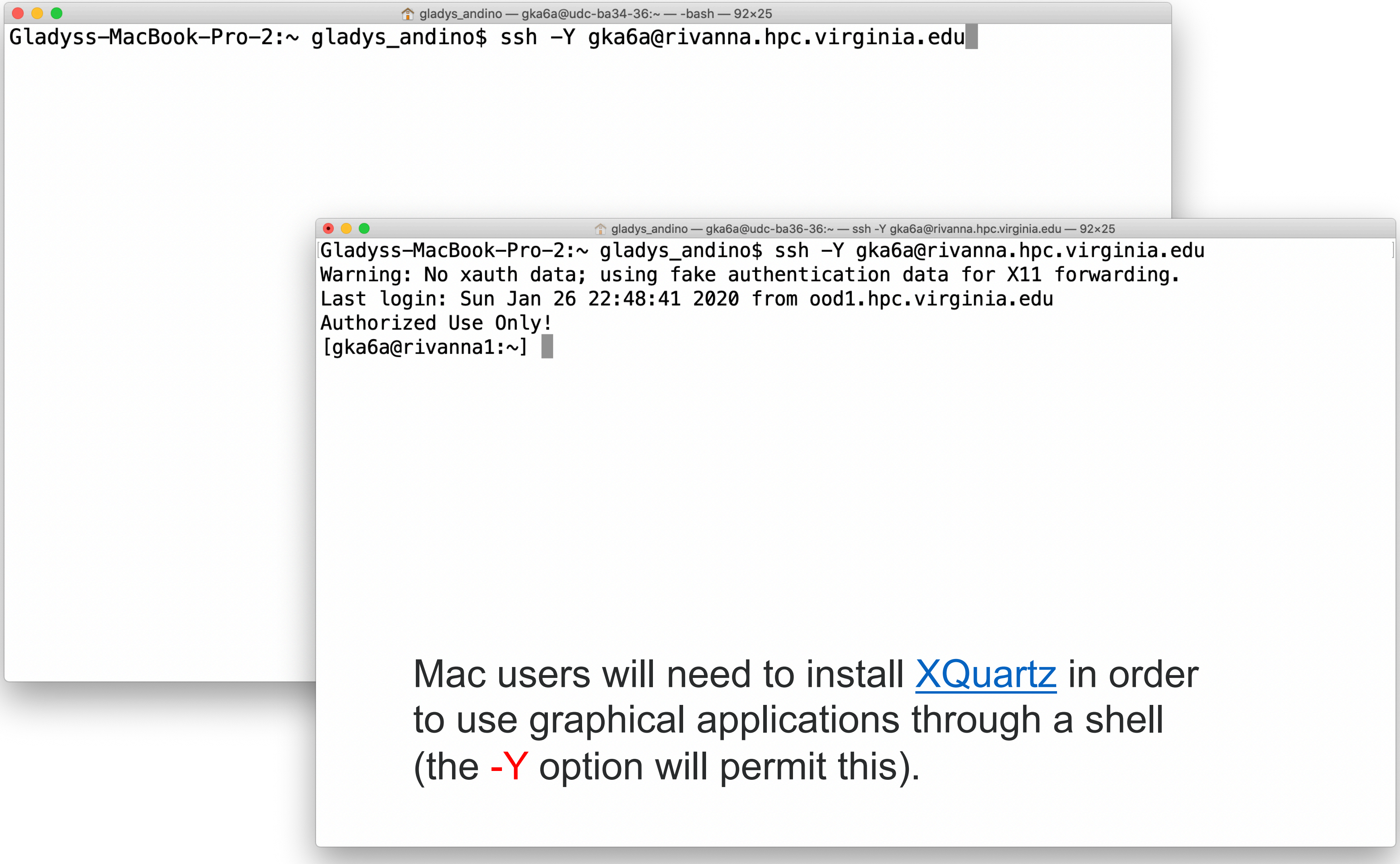
- Mac OS X has built in Terminal app that can use SSH
- Open Finder and Go to Applications



Utilities > Terminal app



Connect using ssh -Y username@rivanna.hpc.virginia.edu


gladys_andino — gka6a@udc-ba34-36:~ — -bash — 92×25
gladys_andino — gka6a@udc-ba36-36:~ — ssh -Y gka6a@rivanna.hpc.virginia.edu — 92×25

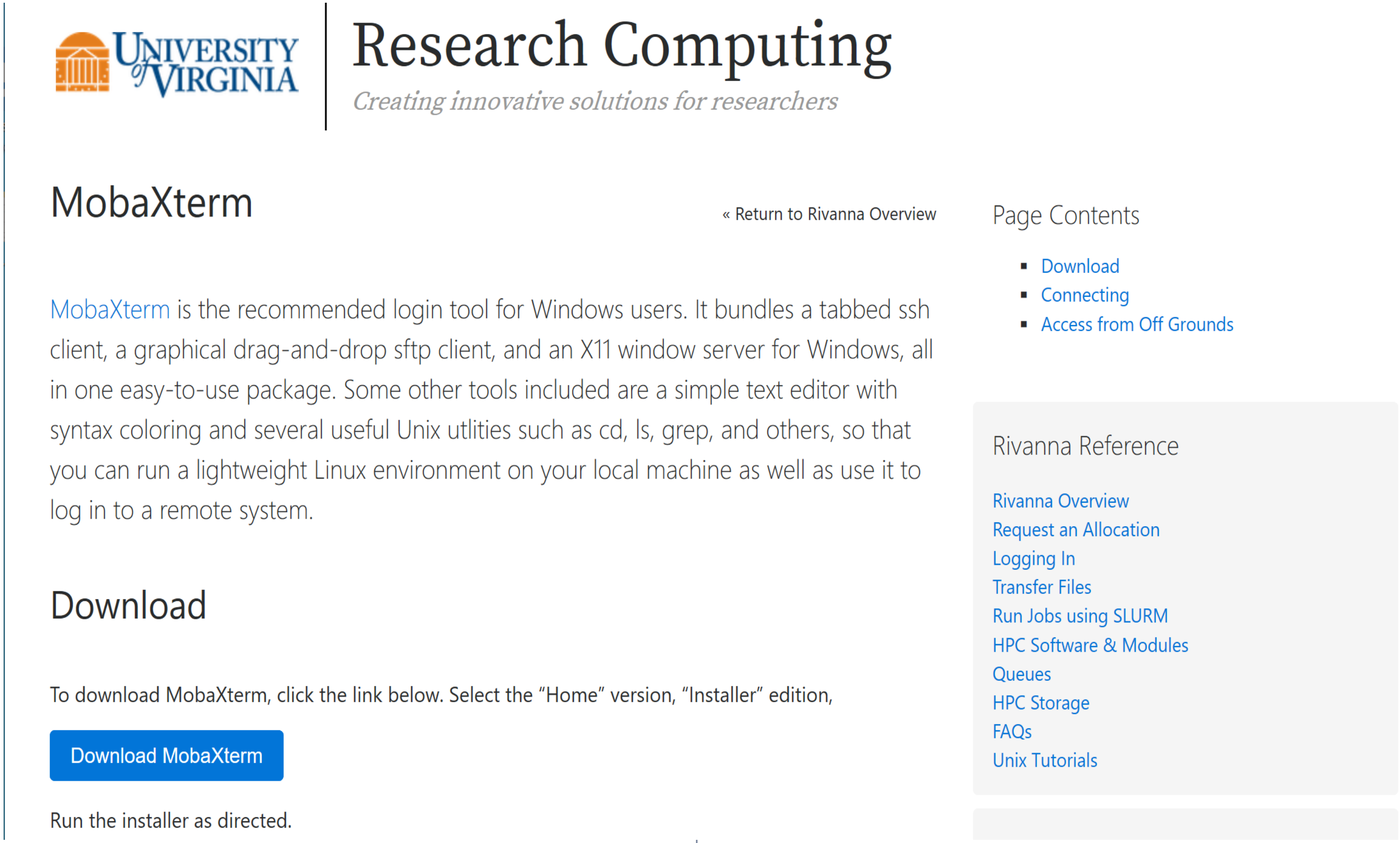
```
Gladyss-MacBook-Pro-2:~ gladys_andino$ ssh -Y gka6a@rivanna.hpc.virginia.edu
```

```
Gladyss-MacBook-Pro-2:~ gladys_andino$ ssh -Y gka6a@rivanna.hpc.virginia.edu
Warning: No xauth data; using fake authentication data for X11 forwarding.
Last login: Sun Jan 26 22:48:41 2020 from ood1.hpc.virginia.edu
Authorized Use Only!
[gka6a@rivanna1:~]
```

Mac users will need to install XQuartz in order to use graphical applications through a shell (the -Y option will permit this).

LOGGING IN – using MobaXterm

https://www.rc.virginia.edu/userinfo/rivanna/login/



Research Computing

Creating innovative solutions for researchers

MobaXterm

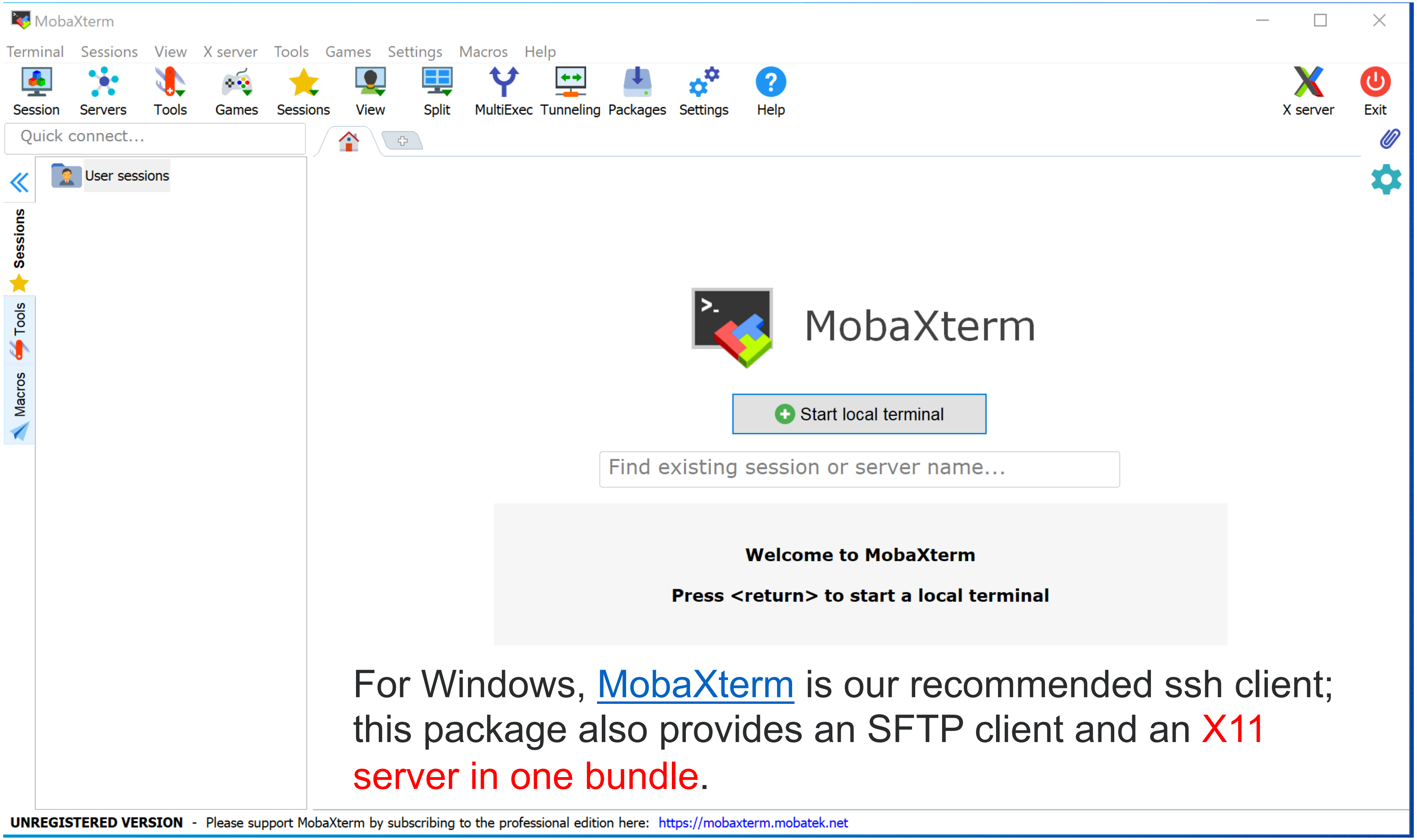
« Return to Rivanna Overview

MobaXterm is the recommended login tool for Windows users. It bundles a tabbed ssh client, a graphical drag-and-drop sftp client, and an X11 window server for Windows, all in one easy-to-use package. Some other tools included are a simple text editor with syntax coloring and several useful Unix utlities such as cd, ls, grep, and others, so that you can run a lightweight Linux environment on your local machine as well as use it to log in to a remote system.

Download

To download MobaXterm, click the link below. Select the “Home” version, “Installer” edition,

Run the installer as directed.



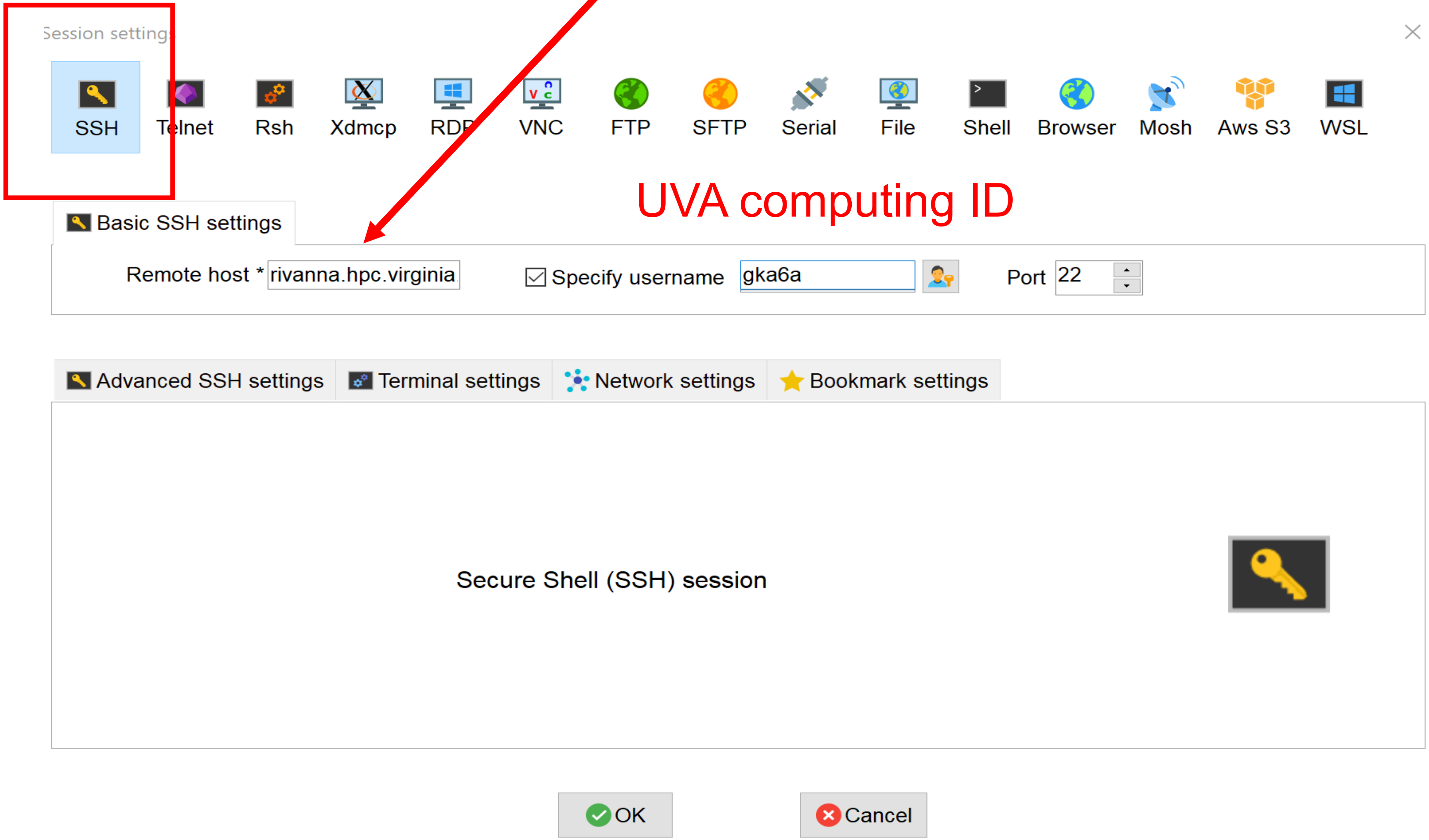
Welcome to MobaXterm

Press <return> to start a local terminal

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

For Windows, MobaXterm is our recommended ssh client; this package also provides an SFTP client and an X11 server in one bundle.

Remote host: rivanna.hpc.virginia.edu



UVA computing ID

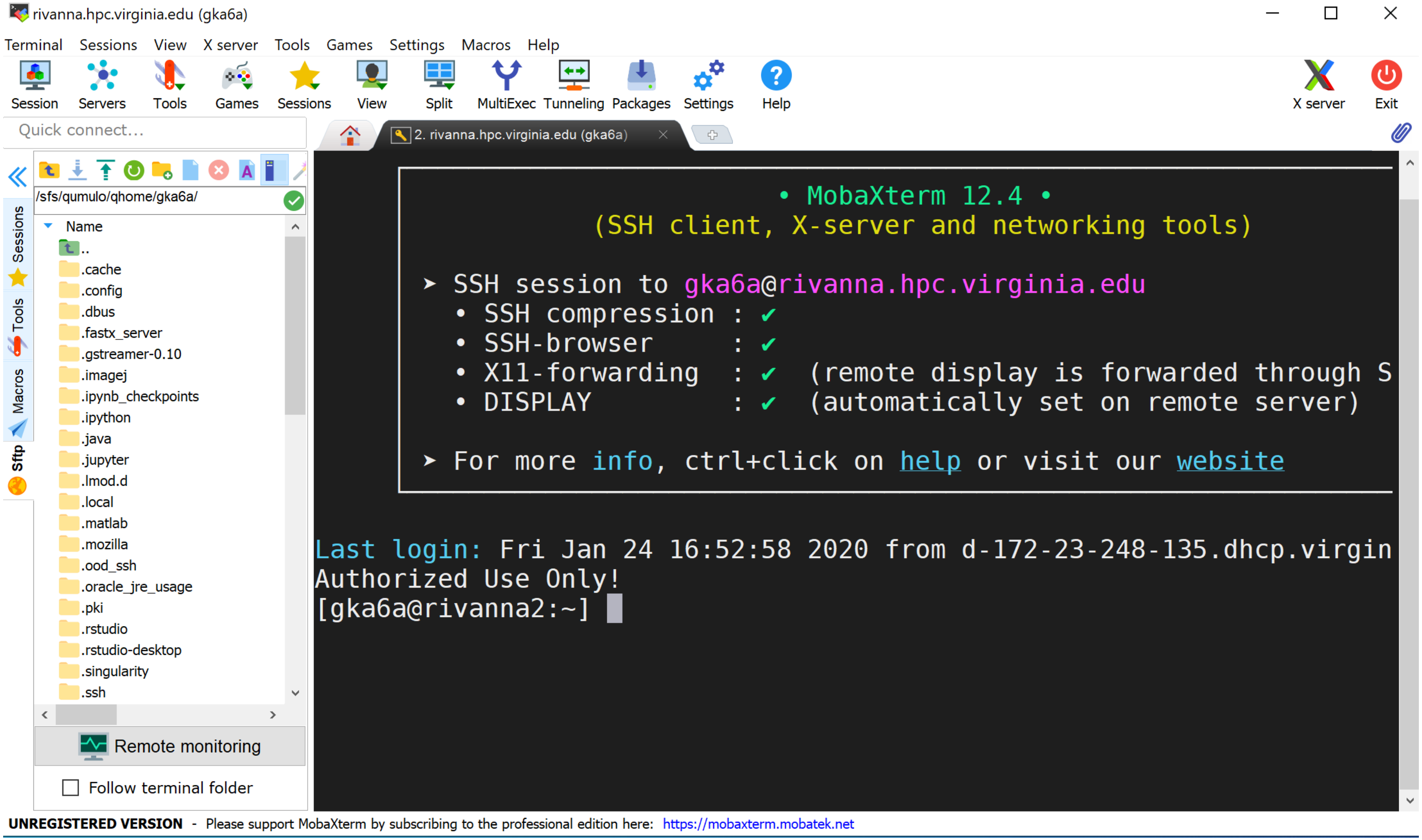
Remote host * rivanna.hpc.virginia

☑ Specify username gka6a

Port 22

Secure Shell (SSH) session

University of Virginia | Research Computing



/sfs/qumulo/qhome/gka6a/

- ..
- .cache
- .config
- .dbus
- .fastx_server
- .gstreamer-0.10
- .imagej
- .ipynb_checkpoints
- .ipython
- .java
- .jupyter
- .lmod.d
- .local
- .matlab
- .mozilla
- .ood_ssh
- .oracle_jre_usage
- .pki
- .rstudio
- .rstudio-desktop
- .singularity
- .ssh

```
                • MobaXterm 12.4 •
      (SSH client, X-server and networking tools)

 ➤ SSH session to gka6a@rivanna.hpc.virginia.edu
   • SSH compression : ✔
   • SSH-browser     : ✔
   • X11-forwarding  : ✔  (remote display is forwarded through S
   • DISPLAY         : ✔  (automatically set on remote server)

 ➤ For more info, ctrl+click on help or visit our website

Last login: Fri Jan 24 16:52:58 2020 from d-172-23-248-135.dhcp.virgin
Authorized Use Only!
[gka6a@rivanna2:~]
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

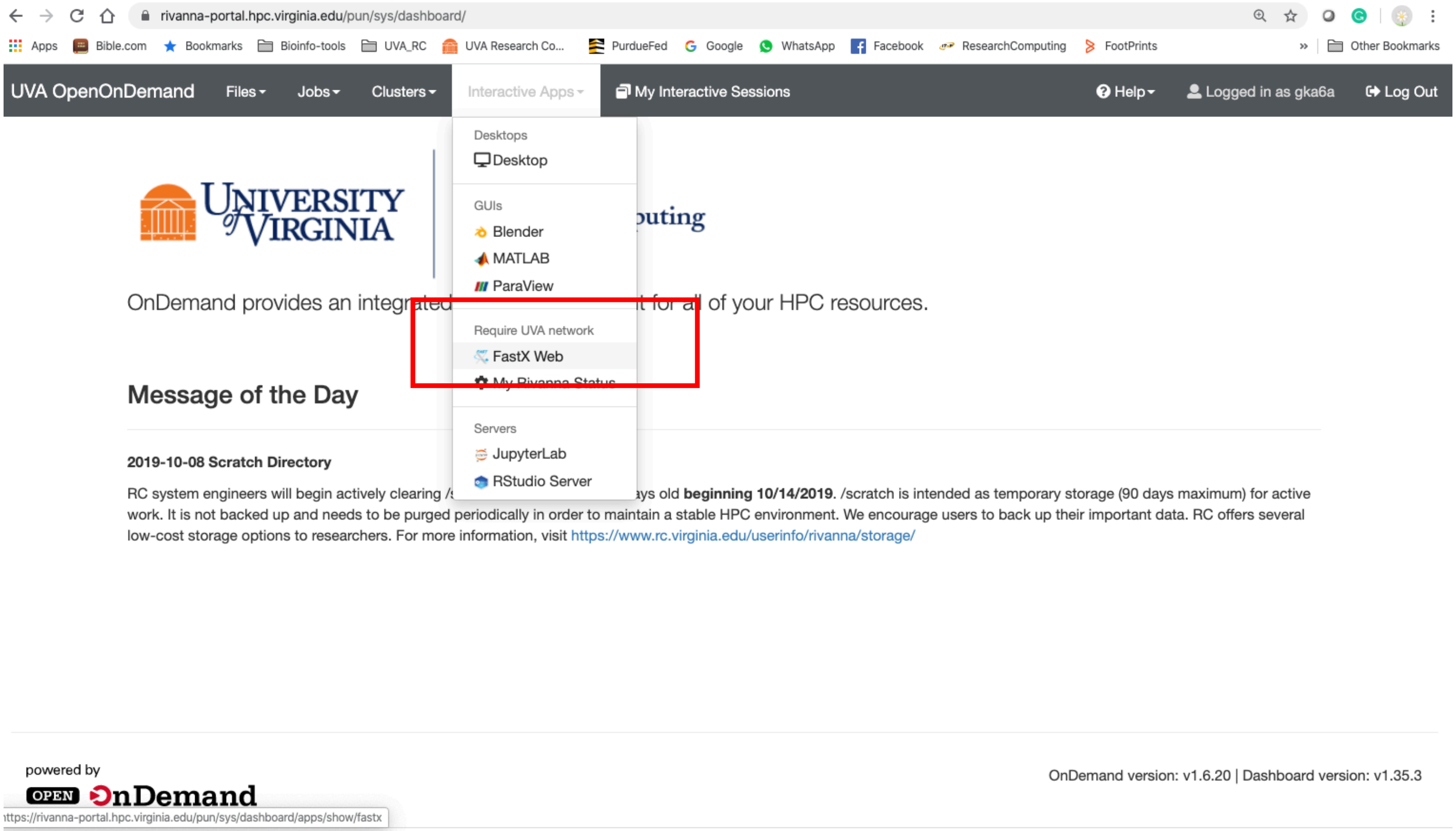
LOGGING IN - using (OOD)

https://www.rc.virginia.edu/userinfo/rivanna/ood/overview/

- OpenOnDemand is a graphical user interface that allows you to examine and manipulate files and submit jobs.
- It is very easy and intuitive but, is limited. It's a good way to get started.
- OOD also provides portals to applications such as Jupyterlab and R Studio Server.
- When you first log in (**through Netbadge**) you will see the Dashboard.

LOGGING IN - using FASTX WEB

From the Dashboard go to Interactive Apps > FastX web



OnDemand provides an integrated ... t for all of your HPC resources.

Message of the Day

2019-10-08 Scratch Directory

RC system engineers will begin actively clearing /s ... ys old **beginning 10/14/2019**. /scratch is intended as temporary storage (90 days maximum) for active work. It is not backed up and needs to be purged periodically in order to maintain a stable HPC environment. We encourage users to back up their important data. RC offers several low-cost storage options to researchers. For more information, visit https://www.rc.virginia.edu/userinfo/rivanna/storage/

OnDemand version: v1.6.20 | Dashboard version: v1.35.3

COMMAND LINE - REVIEW

Using Rivanna from the Command Line

https://learning.rc.virginia.edu/notes/rivanna-command-line/

COMMAND LINE - BASICS

- List a directory

```
ls -l {path}
ls -a {path}
ls {path} | more
```

- Change to directory

```
cd {dirname}
cd ~
cd ..
```

- Make a new directory

```
mkdir {dirname}
```

- Remove a directory

```
rmdir {dirname}
rm -r {dirname}
```

- Print working directory

```
pwd
```

- Copy a file or directory

```
cp {file1} {file2}
cp -r {dir1} {dir2}
cat {newfile} >> {oldfile}
```

- Move (or rename) a file

```
mv {oldfile} {newfile} # change
name
mv {oldname} {newname}
```

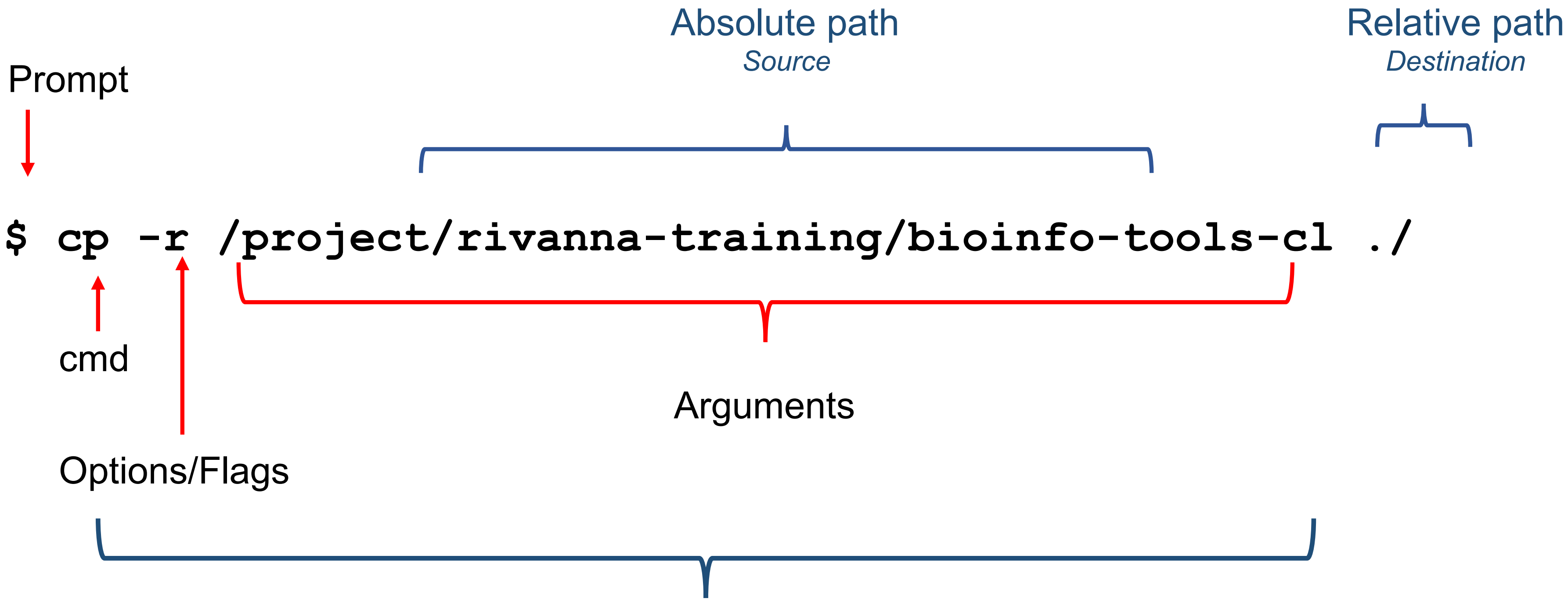
- Delete a file

```
rm {filename}
```

- View a text file

```
more {filename}
less {filename}
cat {filename}
```

LET’S GRAB SOME FILES



```
$ cp -r /project/rivanna-training/bioinfo-tools-cl ./
```

Spaces separate these parts!

YOUR DIRECTORIES – ON RIVANNA

- The default `/home` directory has 50GB of storage capacity.
  - The home directory is for personal use and is not shareable with other users.

- Secondary directory `/scratch` each user will have access to 10 TB of **temporary** storage.
  - It is located in a subdirectory under `/scratch`, and named with your userID
  - e.g., `/scratch/gka6a`
  - You are limited to 350,000 files in your scratch directory.
  - The `/scratch` directory is for personal use and is not shareable with other users

Important:
`/scratch` is **NOT permanent** storage and files that have not been accessed for more than **90 days** will be marked for deletion.

CHECKING YOUR STORAGE

- To see how much disk space you have used in your home directory, open a terminal window and type **`hdquota`** at the command-line prompt:

```
$ hdquota

Type    Location         Name                    Size Used Avail Use%
========================================================================
home       /home        gka6a                     51G   12G   39G  24%
Project    /project     slurmtests              2.0P  1.9P  144T  93%
Project    /project     arcs                     16T   12T  3.8T  75%
Project    /project     rivanna_software        1.1T  4.2M  1.0T   1%
Project    /project     ds5559                   51G  3.7G   47G   8%
Value      /nv          vol174                  5.5T  1.2T  4.4T  21%
...

Location        Age_Limit(Days) Disk_Limit(GB)  Use(GB) File_Limit      Use
==============================================================================
/scratch/gka6a         90        10240           541      350000        1273
```

STORAGE - DETAILS

https://www.rc.virginia.edu/userinfo/storage/

MODULES COMMANDS

- `module spider`
  - List all available packages (may be a lot!)
- `module spider <package>`
  - List all versions of <package>, if any
- `module spider <package>/<version>`
  - Describes how to load <package>/<version>. There may be prerequisite modules.
- `module list`
  - List modules loaded in current shell
- `module purge`
  - Remove all module modifications to the environment
- `module load <package>/[<version>]`
  - Load the module for (optionally) <version> of <package>
- `module unload <package>`
  - Delete the changes made by the <package> module
- `module swap <package>/<current> <package>/<newver>`
  - Exchange one version of a package for another

MODULES - DETAILS

- Any application software that you want to use will need to be loaded with the **`module load`** command.
- For example:
    - `module spider fastqc`
    - `module load fastqc/0.11.5`
    - `module list`

- You will need to load the module any time that you create a new shell
    - Every time that you log out and back in
    - Every time that you run a batch job on a compute node

MODULES - DETAILS

https://www.rc.virginia.edu/userinfo/rivanna/software/modules/

SLURM

```
$ qlist # Usage: qlist [-p] [-c] [-m]
$ hdquota
$ sbatch
$ squeue -u $USER
$ scontrol show job <jobid>
$ squeue --start -j <jobid> # to request an estimate when your pending job will run
```

QUEUES/PARTITIONS

SLURM refers to queues as **partitions**. We do not have a default partition; each job must request one explicitly.

Queue Name	Purpose	Job Time Limit	Memory / Node	Cores / Node
standard	**For jobs on a single compute node**	**7 days**	**256 GB** **384 GB**	**28** **40**
gpu	For jobs that can use general purpose graphical processing units (GPGPUs) (K80, P100 and V100)	3 days	256 GB	28
parallel	For large parallel jobs on up to 120 nodes (<= 2400 CPU cores)	3 days	128 GB	20
largemem	**For memory intensive jobs (<= 16 cores/node)**	**4 days**	**1 TB**	**16**
dev	To run jobs that are quick tests of code	1 hour	128 GB	4

QUEUES/PARTITIONS

SLURM refers to queues as **partitions**. We do not have a default partition; each job must request one explicitly.

```
$qlist
```

```
Queue            Total        Free        Jobs        Jobs        Time         SU
(partition)      Cores        Cores       Running     Pending     Limit        Charge
 ===============================================================================
bii              4600         2427        40          41          7-00:00:00   1
standard         3660         1020        766         121         7-00:00:00   1
dev              2820         2106        0           0             1:00:00    0
parallel         3900         2898        11          0           3-00:00:00   1
instructional    600          336         0           0           3-00:00:00   1
largemem         80           60          3           0           4-00:00:00   1
gpu              364          272         27          4           3-00:00:00   3
bii-gpu          320          316         1           0           3-00:00:00   1
knl              2048         1024        0           0           3-00:00:00   1
pcore            144          72          0           1246         infinite    1
```

CHECKING YOUR ALLOCATION

To see how many SUs you have available for running jobs, type at the command-line prompt: **`allocations`**

```
$ allocations

Allocations available to Gladys_Karina_Andino_Bautista (gka6a):

 * arcs_admin: less than 500 service-units remaining
 * ds5559: less than 25,000 service-units remaining
 * ga_bioinfo-test: less than 100,000 service-units remaining
 * hpc_build: less than 203,417 service-units remaining
 * rivanna-training: less than 20,000 service-units remaining


 for more information about a specific allocation, please run:
  'allocations -a <allocation name>'
```

SLURM - DETAILS

https://www.rc.virginia.edu/userinfo/rivanna/slurm/

SEQUENCING BASICS– FASTQ FORMAT

- Typically will have the suffix .fastq or .fq
  - may be compressed .fastq.gz or .fq.gz
  - some but not all programs can read the compressed version

- Four lines per sequence
  - line 1: @Sequence ID<space>optional description

    @ often occurs in quality lines so it is an unreliable way to identify this line
  - line 2: sequence
  - line 3: + optional description (NCBI repeats ID line)

    + often occurs in quality lines so it is an unreliable way to identify this line
  - line4: quality (one value per base, Phred encoded)

- Quality is the Probability that the reported base is incorrect
  - Quality values are converted to letters in the ASCII alphabet by adding 33 to the log transformed quality
  - ascii value = quality + 33
  - other offsets than 33 are sometimes used (rare)

SEQUENCING BASICS– FASTQ FORMAT

- Quality is the probability that the reported base is incorrect

- Usually reported as Q = -10 log10 P(incorrect)
  - quality = 10 is 10 % error
  - quality = 20 is 1% error
  - quality = 30 is 0.1% error

- Encoded as a single ASCII letter
  - value = quality + 33

- Other offsets than 33 are sometimes used (rare)

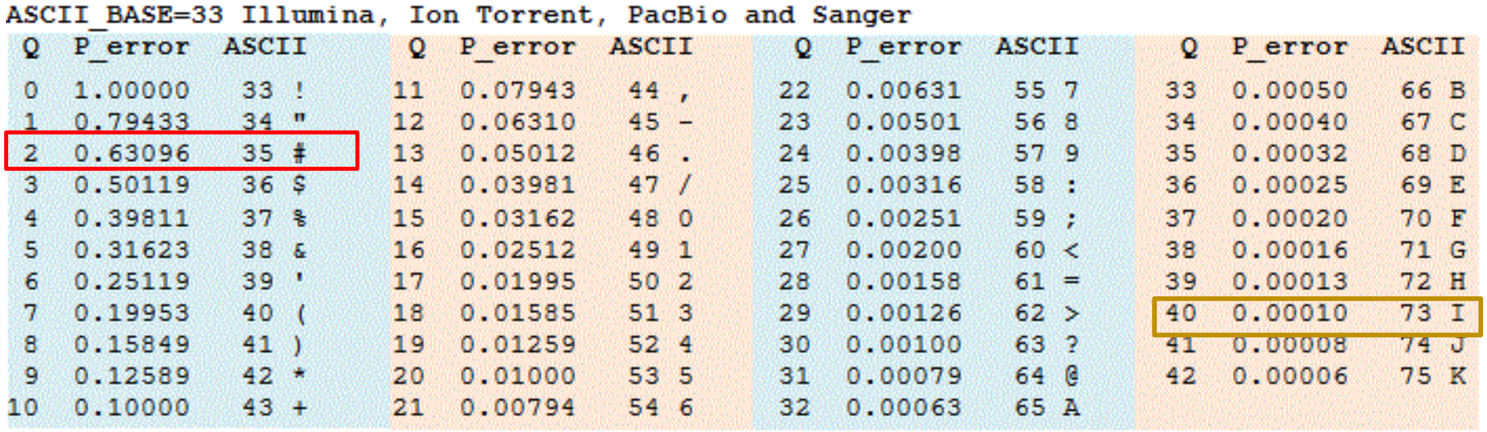
SEQUENCING BASICS– FASTQ FORMAT

```
instrument:run:flowcell:lane:tile:x:y                    pair:filtered:control:bar-code
@HISEQ02:319:C22FKACXX:2:1101:1699:1972     1:N:0:GTAGAG
GACCCATCCATTGTTGGACAGCTGAAGACGGGACGATCGTGCTCGTGTTTTGAATGCGAGAATCCCTGCAGAGGCTGCG
+
CCCFFFFFHHHHHJIJJJJGIJJJJJJJJJJJJIIJIJJJIIJIIHAFGIJJEHHHHFFFDCDDDDDDCDDDD###<<@B
```

# = ascii 35
Q = 35 – 33 = 2
$\varepsilon = 10^{-0.2}$ = 0.63 totally bogus

I = ascii 73
Quality = 73 – 33 = 40
Quality = $-10 \log_{10}\varepsilon$, $\varepsilon = 10^{-4}$

- Phred quality score 33 - program (Phil Green, UWa) ca. 1998
- where ε is the expected error rate (probability of calling an incorrect base)

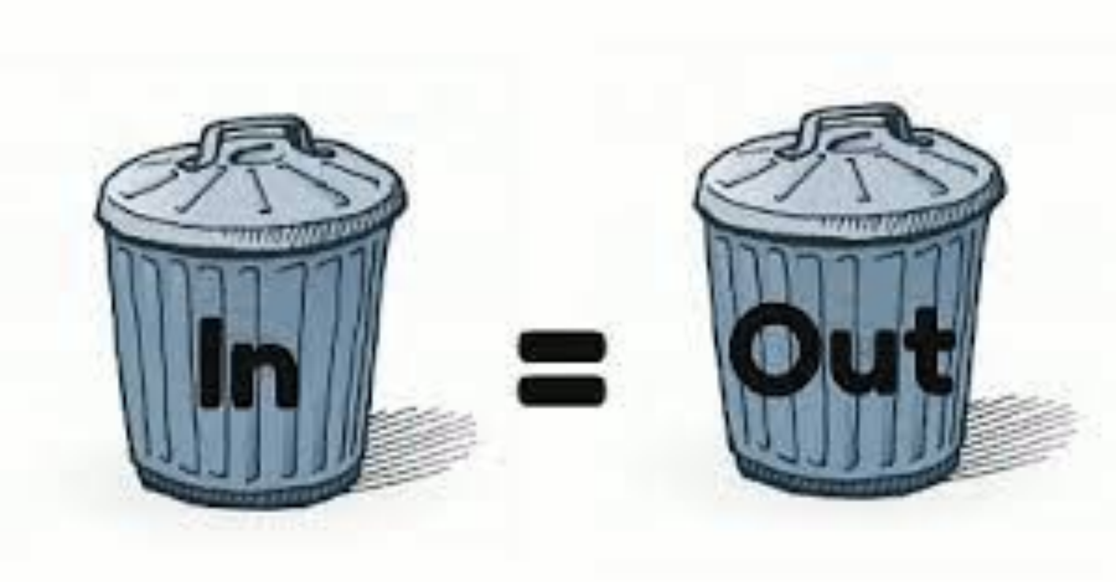


ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

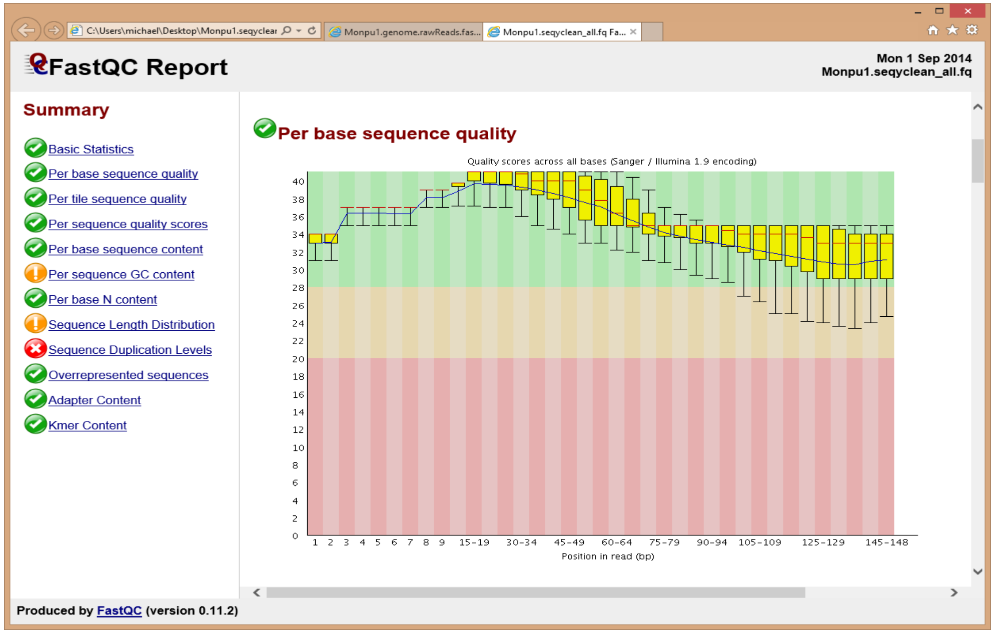
DATA PREPROCESSING = CLEANING

- What should we clean?
  - All big data projects begin with data cleaning



FASTQC

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.



C:\Users\michael\Desktop\Monpu1.seqyclear
Monpu1.genome.rawReads.fas...
Monpu1.seqyclean_all.fq Fa...

FastQC Report

Mon 1 Sep 2014
Monpu1.seqyclean_all.fq

Summary

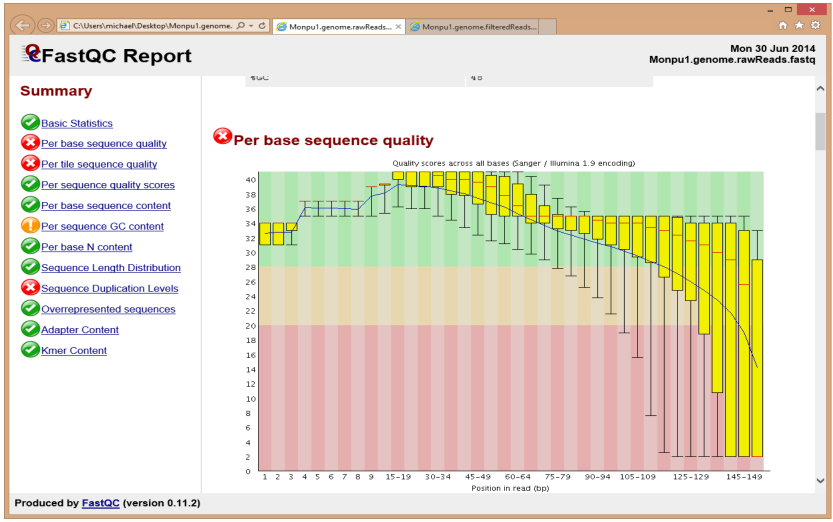
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✓ **Per base sequence quality**

Produced by FastQC (version 0.11.2)

Simon Andrews of Babraham Bioinformatics
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC: QC OF THE DATA



C:\Users\michael\Desktop\Monpu1.genome. | Monpu1.genome.rawReads... | Monpu1.genome.filteredReads...

FastQC Report

Mon 30 Jun 2014
Monpu1.genome.rawReads.fastq

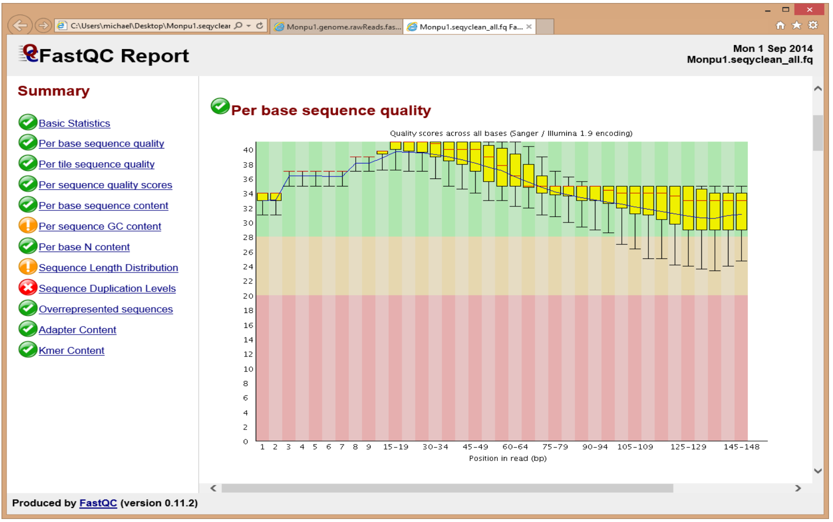
Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

%GC 48

✗ **Per base sequence quality**

Produced by FastQC (version 0.11.2)



C:\Users\michael\Desktop\Monpu1.seqyclear | Monpu1.genome.rawReads.fas... | Monpu1.seqyclean_all.fq Fa...

FastQC Report

Mon 1 Sep 2014
Monpu1.seqyclean_all.fq

Summary

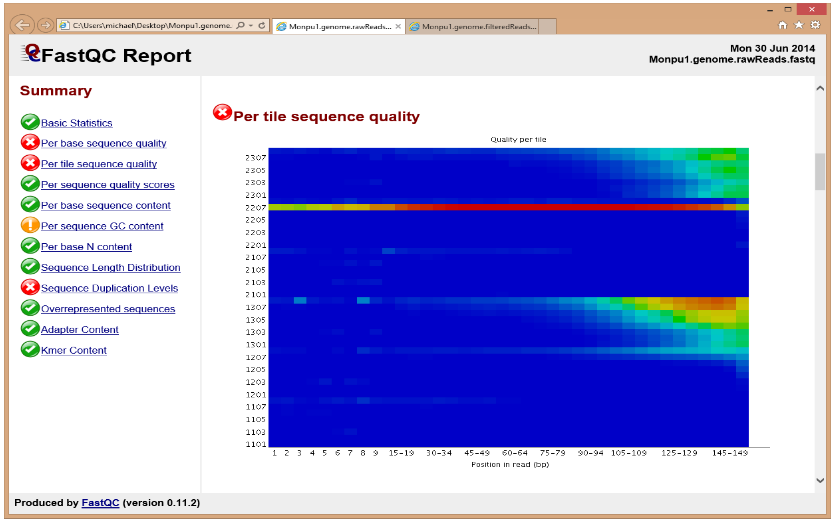
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✓ **Per base sequence quality**

Produced by FastQC (version 0.11.2)

Before

After



C:\Users\michael\Desktop\Monpu1.genome. | Monpu1.genome.rawReads... | Monpu1.genome.filteredReads...

FastQC Report

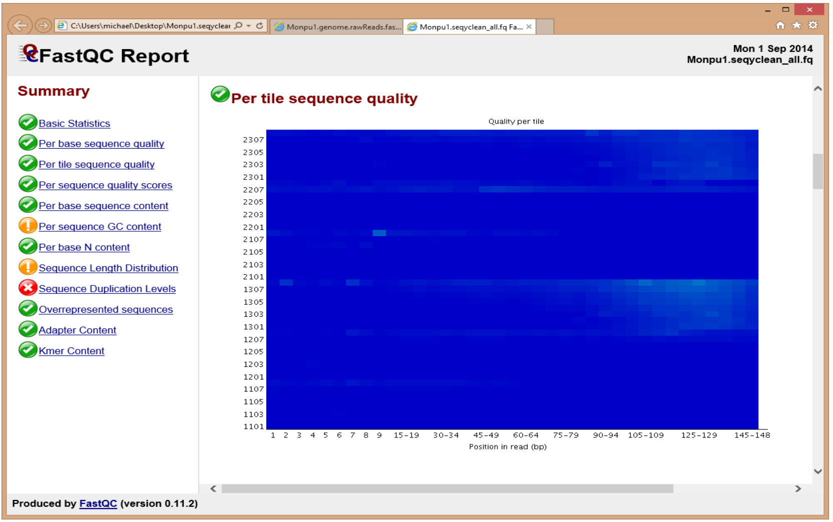
Mon 30 Jun 2014
Monpu1.genome.rawReads.fastq

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✗ **Per tile sequence quality**

Produced by FastQC (version 0.11.2)



C:\Users\michael\Desktop\Monpu1.seqyclear | Monpu1.genome.rawReads.fas... | Monpu1.seqyclean_all.fq Fa...

FastQC Report

Mon 1 Sep 2014
Monpu1.seqyclean_all.fq

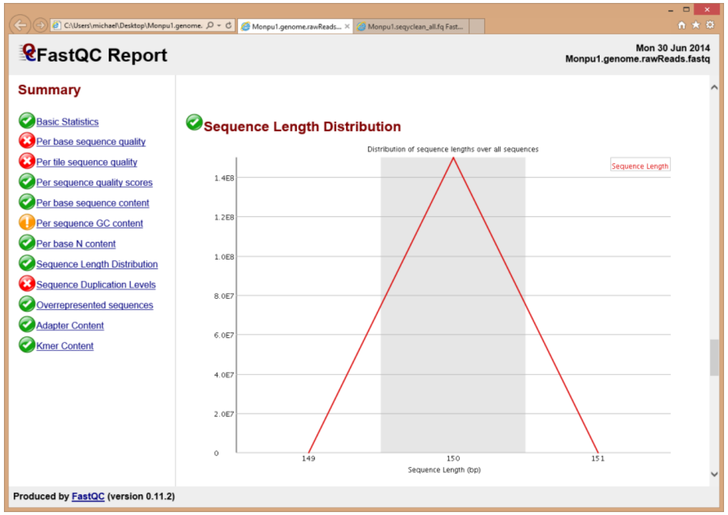
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✓ **Per tile sequence quality**

Produced by FastQC (version 0.11.2)

FASTQC: QC OF THE DATA


C:\Users\michael\Desktop\Monpu1.genome.
Monpu1.genome.rawReads...
Monpu1.seqyclean_all.fq Fast...
Distribution of sequence lengths over all sequences
Sequence Length
1.4E8
1.2E8
1.0E8
8.0E7
6.0E7
4.0E7
2.0E7
0
149
150
151
Sequence Length (bp)

FastQC Report

Mon 30 Jun 2014
Monpu1.genome.rawReads.fastq

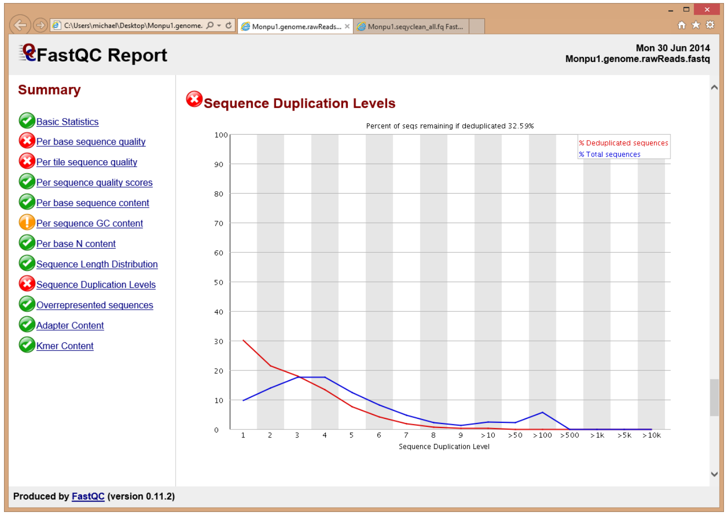
Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✓ **Sequence Length Distribution**

Produced by FastQC (version 0.11.2)

Before


C:\Users\michael\Desktop\Monpu1.genome.
Monpu1.genome.rawReads...
Monpu1.seqyclean_all.fq Fast...
Percent of seqs remaining if deduplicated 32.59%
% Deduplicated sequences
% Total sequences
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k
Sequence Duplication Level

FastQC Report

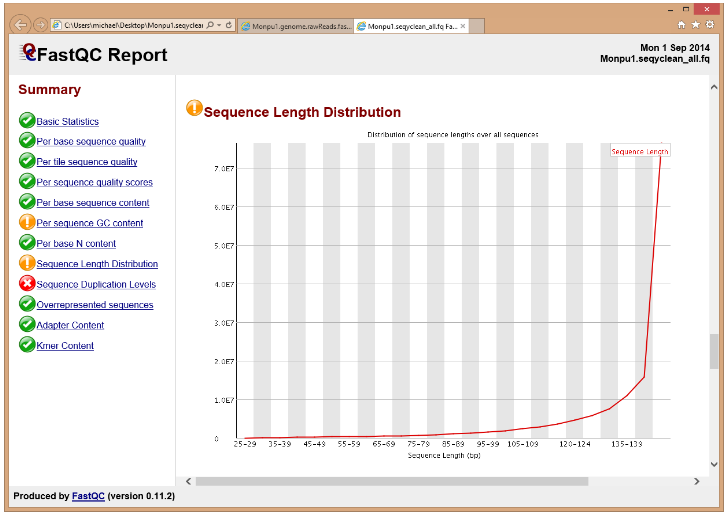
Mon 30 Jun 2014
Monpu1.genome.rawReads.fastq

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✗ **Sequence Duplication Levels**

Produced by FastQC (version 0.11.2)


C:\Users\michael\Desktop\Monpu1.seqyclear
Monpu1.genome.rawReads.fas...
Monpu1.seqyclean_all.fq Fa...
Distribution of sequence lengths over all sequences
Sequence Length
7.0E7
6.0E7
5.0E7
4.0E7
3.0E7
2.0E7
1.0E7
0
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
105-109
120-124
135-139
Sequence Length (bp)

FastQC Report

Mon 1 Sep 2014
Monpu1.seqyclean_all.fq

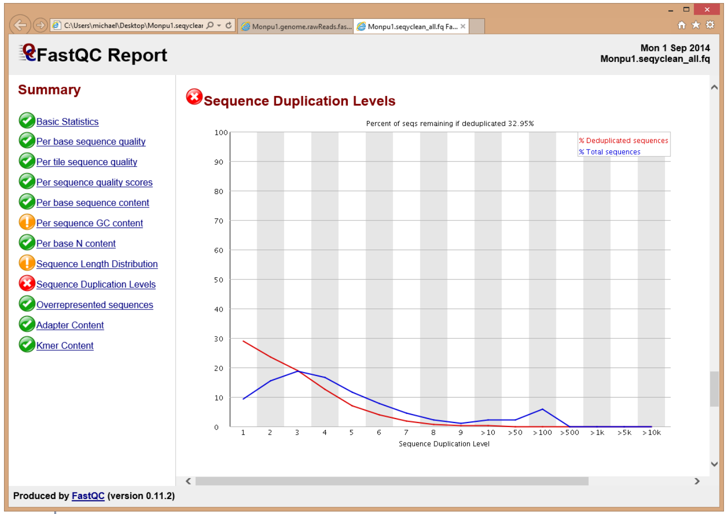
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

! **Sequence Length Distribution**

Produced by FastQC (version 0.11.2)

After


C:\Users\michael\Desktop\Monpu1.seqyclear
Monpu1.genome.rawReads.fas...
Monpu1.seqyclean_all.fq Fa...
Percent of seqs remaining if deduplicated 32.95%
% Deduplicated sequences
% Total sequences
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k
Sequence Duplication Level

FastQC Report

Mon 1 Sep 2014
Monpu1.seqyclean_all.fq

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✓ Kmer Content

✗ **Sequence Duplication Levels**

Produced by FastQC (version 0.11.2)

RUNNING FASTQC

- You can run FastQC in one of two modes, either as an interactive graphical application in which you can dynamically load FastQ files and view their results.
- Alternatively you can run FastQC in a non-interactive mode where you specify the files you want to process on the command line and FastQC will generate an HTML report for each file without launching a user interface. This would allow FastQC to be run as part of an analysis pipeline.

RUNNING FASTQC

```
$ module purge
$ module spider fastqc
--------------------------------------------------------------------------
fastqc: fastqc/0.11.5
--------------------------------------------------------------------------
Description
===========
FastQC is a Java application which takes a FastQ file and runs a series of
tests on it to generate a comprehensive QC report.
More information
================
- Homepage: https://www.bioinformatics.babraham.ac.uk/projects/fastqc

$ module load fastqc
$ ml # short for module list
Currently Loaded Modules:
  1) java/1.8.0   2) fastqc/0.11.5
```

RUNNING FASTQC

```
$ module show fastqc
--------------------------------------------------------------------
   /apps/modulefiles/standard/core/fastqc/0.11.5.lua:
---------------------------------------------------------------------
Description
===========
FastQC is a Java application which takes a FastQ file and runs a series
of tests on it to generate a comprehensive QC report.
More information
================
 - Homepage: https://www.bioinformatics.babraham.ac.uk/projects/fastqc
whatis("Description: FastQC is a Java application which takes a FastQ file
and runs a series of tests on it to generate a comprehensive QC report.")
whatis("Homepage:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc")
setenv("EBROOTFASTQC","/apps/software/standard/core/fastqc/0.11.5")
setenv("EBVERSIONFASTQC","0.11.5")
```

RUNNING FASTQC

…as an interactive graphical application in which you can dynamically load FastQ files and view their results.

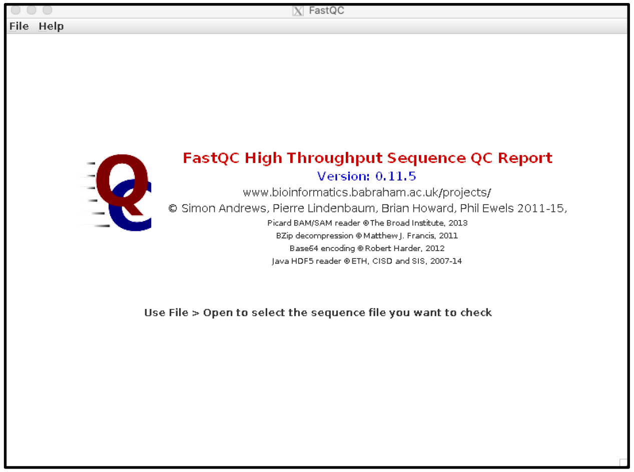
- Open FastX web

- Start an interactive job

```
ijob -N1 -c 1 --ntasks=1 -J fastqc-inte -p standard  -A rivanna-training
```

- Load the module

```
module load fastqc
module list
fastqc &
```

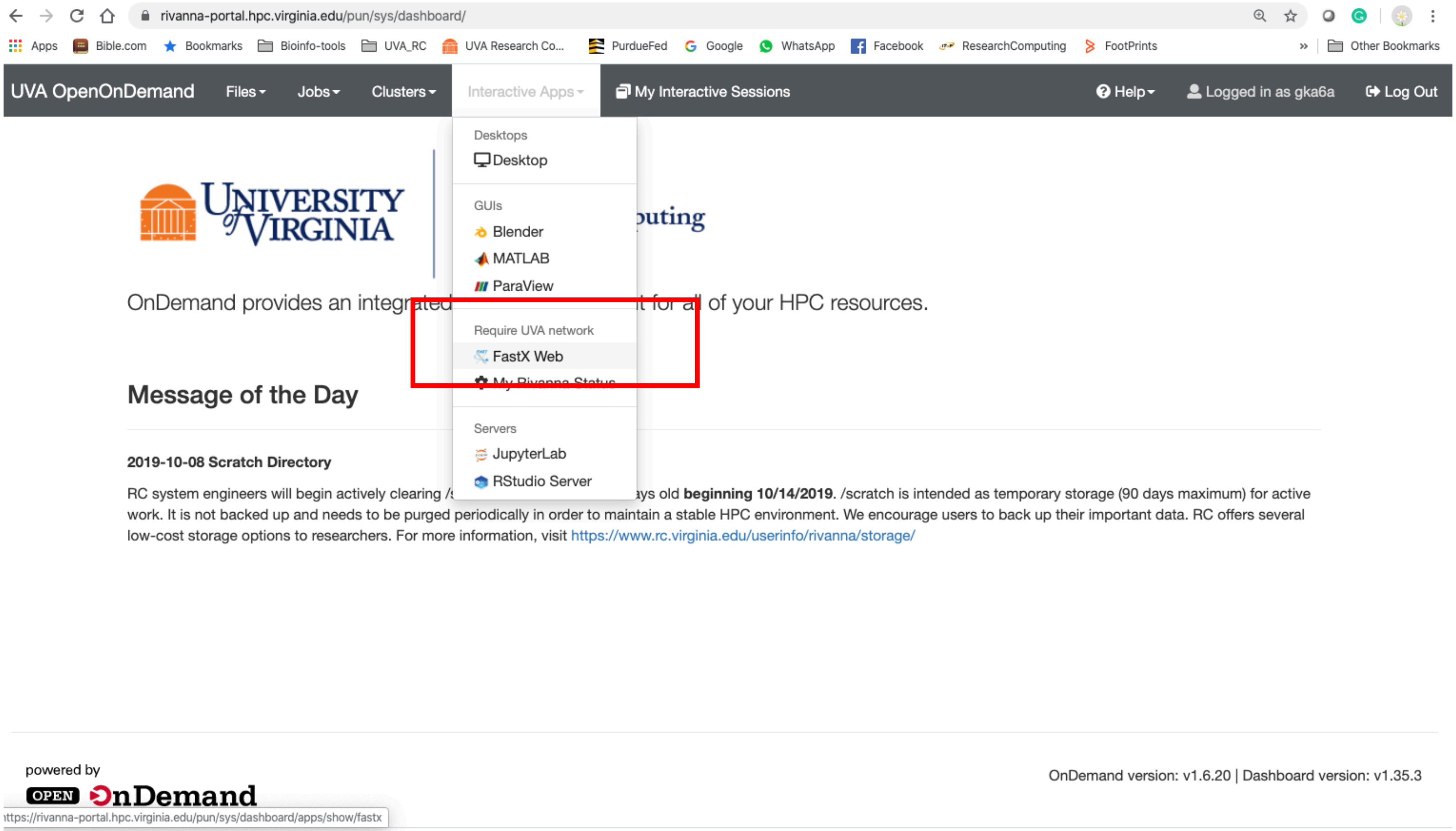


FastQC High Throughput Sequence QC Report
Version: 0.11.5
www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewels 2011-15,
Picard BAM/SAM reader ©The Broad Institute, 2013
BZip decompression © Matthew J. Francis, 2011
Base64 encoding © Robert Harder, 2012
Java HDF5 reader © ETH, CISD and SIS, 2007-14

Use File > Open to select the sequence file you want to check

RUNNING FASTQC

From the Dashboard go to Interactive Apps > FastX web



OnDemand provides an integrated ... t for all of your HPC resources.

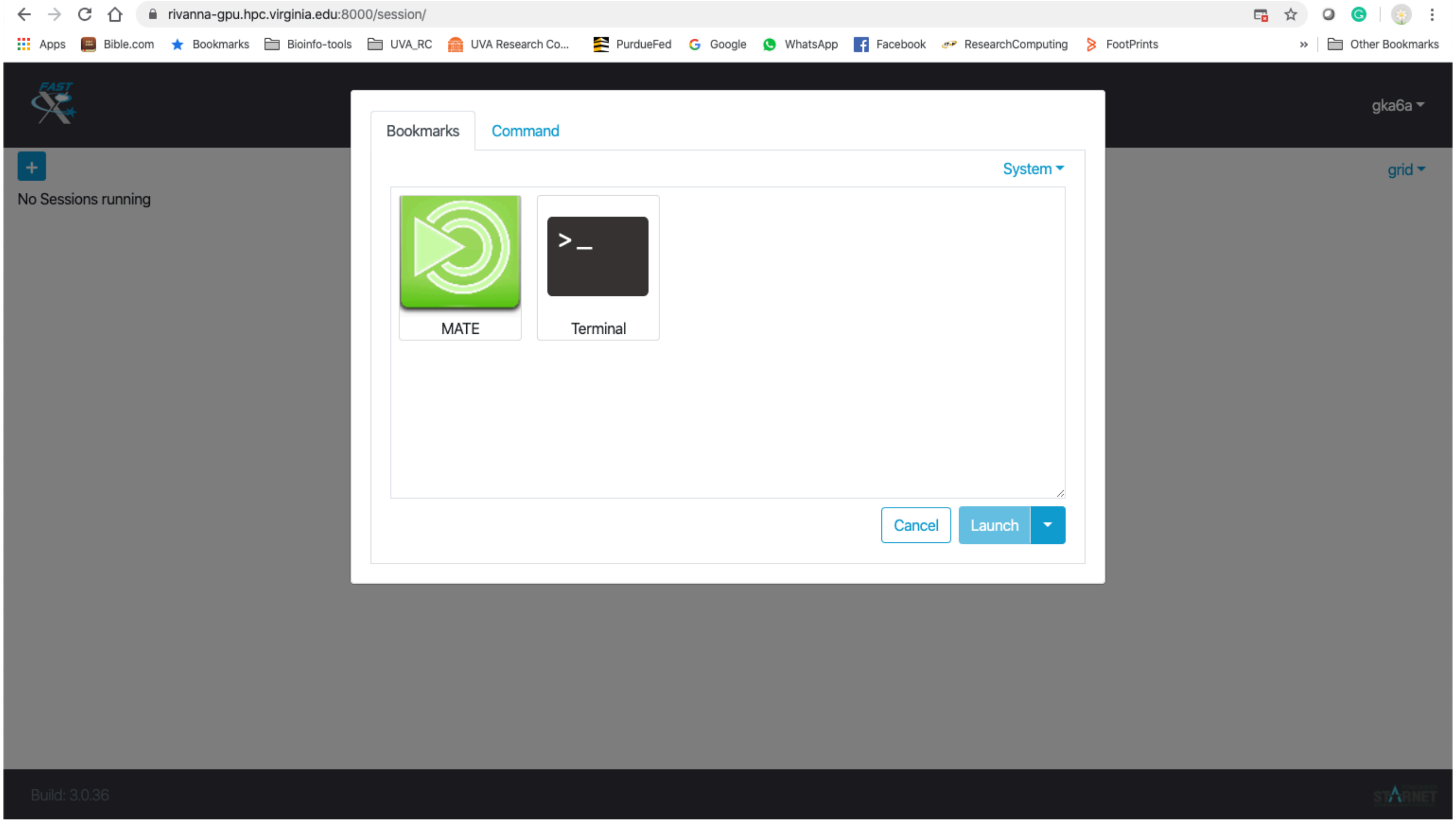
Message of the Day

2019-10-08 Scratch Directory

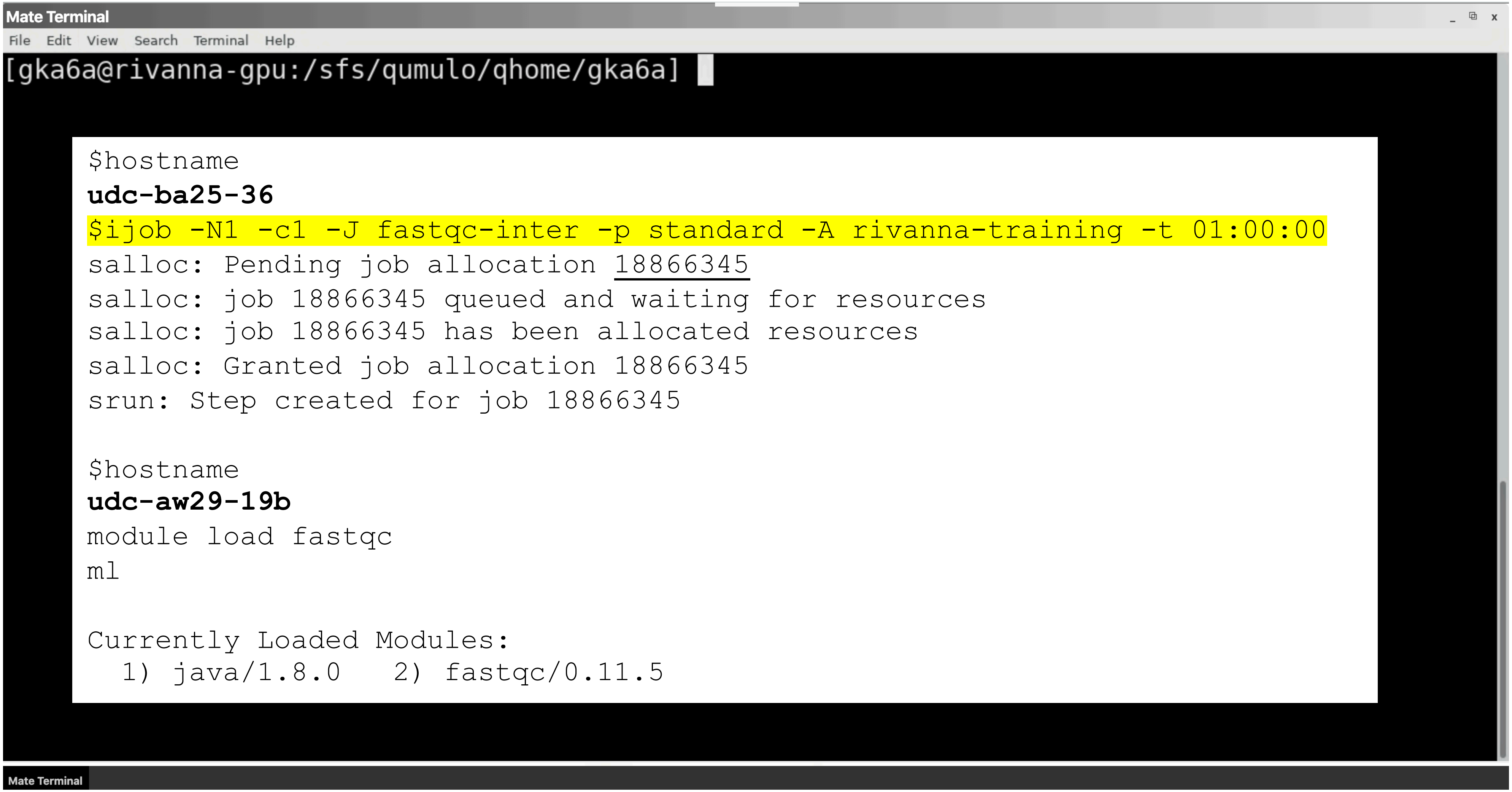
RC system engineers will begin actively clearing /s... ays old **beginning 10/14/2019**. /scratch is intended as temporary storage (90 days maximum) for active work. It is not backed up and needs to be purged periodically in order to maintain a stable HPC environment. We encourage users to back up their important data. RC offers several low-cost storage options to researchers. For more information, visit https://www.rc.virginia.edu/userinfo/rivanna/storage/

OnDemand version: v1.6.20 | Dashboard version: v1.35.3

FastX web > + > Terminal > Launch



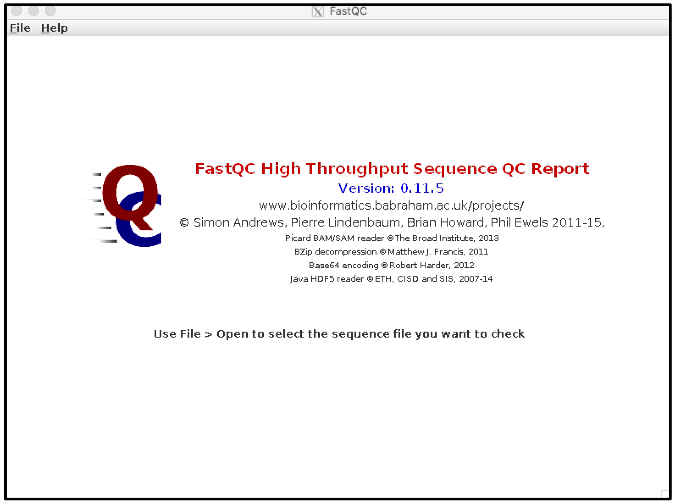
Terminal



```
[gka6a@rivanna-gpu:/sfs/qumulo/qhome/gka6a]
```

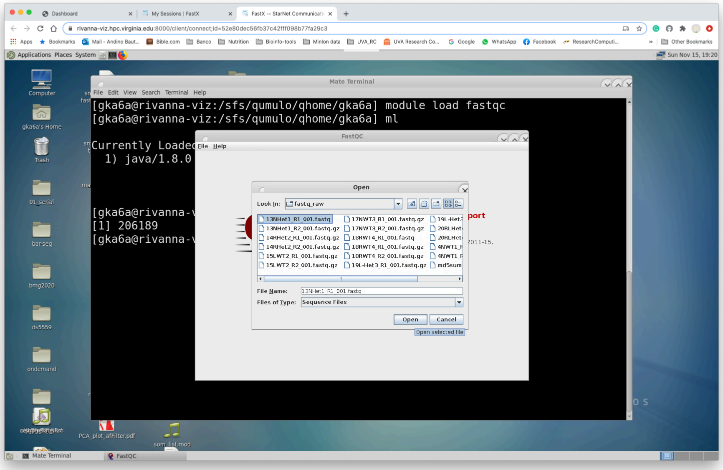
```
$hostname
udc-ba25-36
$ijob -N1 -c1 -J fastqc-inter -p standard -A rivanna-training -t 01:00:00
salloc: Pending job allocation 18866345
salloc: job 18866345 queued and waiting for resources
salloc: job 18866345 has been allocated resources
salloc: Granted job allocation 18866345
srun: Step created for job 18866345

$hostname
udc-aw29-19b
module load fastqc
ml

Currently Loaded Modules:
  1) java/1.8.0   2) fastqc/0.11.5
```

RUNNING FASTQC



FastQC High Throughput Sequence QC Report
Version: 0.11.5
www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewels 2011-15,
Picard BAM/SAM reader ©The Broad Institute, 2013
BZip decompression ©Matthew J. Francis, 2011
Base64 encoding ©Robert Harder, 2012
Java HDF5 reader ©ETH, CISD and SIS, 2007-14

Use File > Open to select the sequence file you want to check

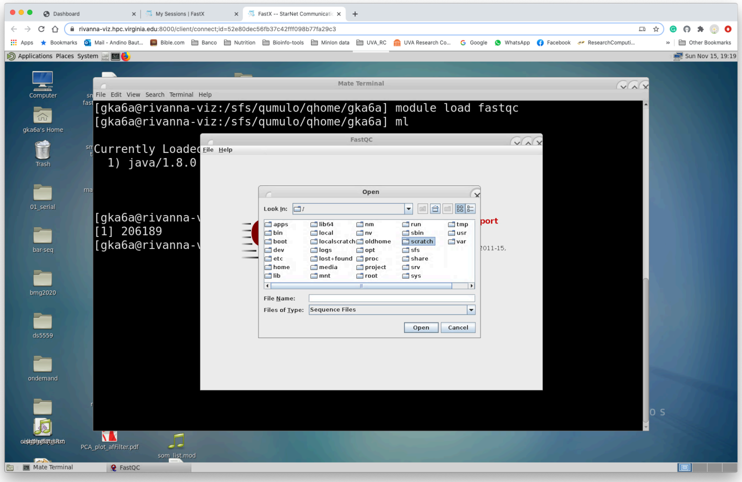

Dashboard
My Sessions | FastX
FastX -- StarNet Communicat
rivanna-viz.hpc.virginia.edu:8000/client/connect;id=52e80dec56fb37c42fff098b77fa29c3
Applications Places System
Sun Nov 15, 19:20
Mate Terminal
File Edit View Search Terminal Help
FastQC
File Help
Open
Open Cancel
Open selected file
Mate Terminal
FastQC

```
[gka6a@rivanna-viz:/sfs/qumulo/qhome/gka6a] module load fastqc
[gka6a@rivanna-viz:/sfs/qumulo/qhome/gka6a] ml

Currently Loaded
  1) java/1.8.0


[gka6a@rivanna-v
[1] 206189
[gka6a@rivanna-v
```

Look In: fastq_raw

```
13NHet1_R1_001.fastq       17NWT3_R1_001.fastq.gz     19L-Het3
13NHet1_R2_001.fastq.gz    17NWT3_R2_001.fastq.gz     20RLHet
14RHet2_R1_001.fastq.gz    18RWT4_R1_001.fastq        20RLHet
14RHet2_R2_001.fastq.gz    18RWT4_R1_001.fastq.gz     4NWT1_F
15LWT2_R1_001.fastq.gz     18RWT4_R2_001.fastq.gz     4NWT1_F
15LWT2_R2_001.fastq.gz     19L-Het3_R1_001.fastq.gz   md5sum
```

File Name: 13NHet1_R1_001.fastq
Files of Type: Sequence Files

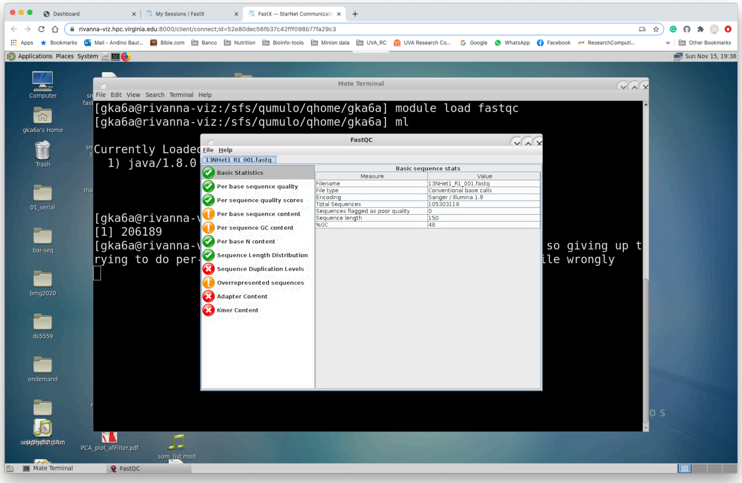

Dashboard
My Sessions | FastX
FastX -- StarNet Communicat
rivanna-viz.hpc.virginia.edu:8000/client/connect;id=52e80dec56fb37c42fff098b77fa29c3
Applications Places System
Sun Nov 15, 19:19
Mate Terminal
File Edit View Search Terminal Help
FastQC
File Help
Open
Open Cancel
Mate Terminal
FastQC

```
[gka6a@rivanna-viz:/sfs/qumulo/qhome/gka6a] module load fastqc
[gka6a@rivanna-viz:/sfs/qumulo/qhome/gka6a] ml

Currently Loaded
  1) java/1.8.0


[gka6a@rivanna-v
[1] 206189
[gka6a@rivanna-v
```

Look In: /

```
apps    lib64         nm       run      tmp
bin     local         nv       sbin     usr
boot    localscratch  oldhome  scratch  var
dev     logs          opt      sfs
etc     lost+found    proc     share
home    media         project  srv
lib     mnt           root     sys
```

File Name:
Files of Type: Sequence Files


Dashboard
My Sessions | FastX
FastX -- StarNet Communicat
rivanna-viz.hpc.virginia.edu:8000/client/connect;id=52e80dec56fb37c42fff098b77fa29c3
Applications Places System
Sun Nov 15, 19:38
Mate Terminal
File Edit View Search Terminal Help
FastQC
File Help
13NHet1_R1_001.fastq
Mate Terminal
FastQC

```
[gka6a@rivanna-viz:/sfs/qumulo/qhome/gka6a] module load fastqc
[gka6a@rivanna-viz:/sfs/qumulo/qhome/gka6a] ml

Currently Loaded
  1) java/1.8.0


[gka6a@rivanna-v
[1] 206189
[gka6a@rivanna-v                                             so giving up t
rying to do per-                                             ile wrongly
```

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Adapter Content
- ✗ Kmer Content

Basic sequence stats

Measure	Value
Filename	13NHet1_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	105303118
Sequences flagged as poor quality	0
Sequence length	150
%GC	48

RUNNING FASTQC

…another way but slower

- ssh with –Y

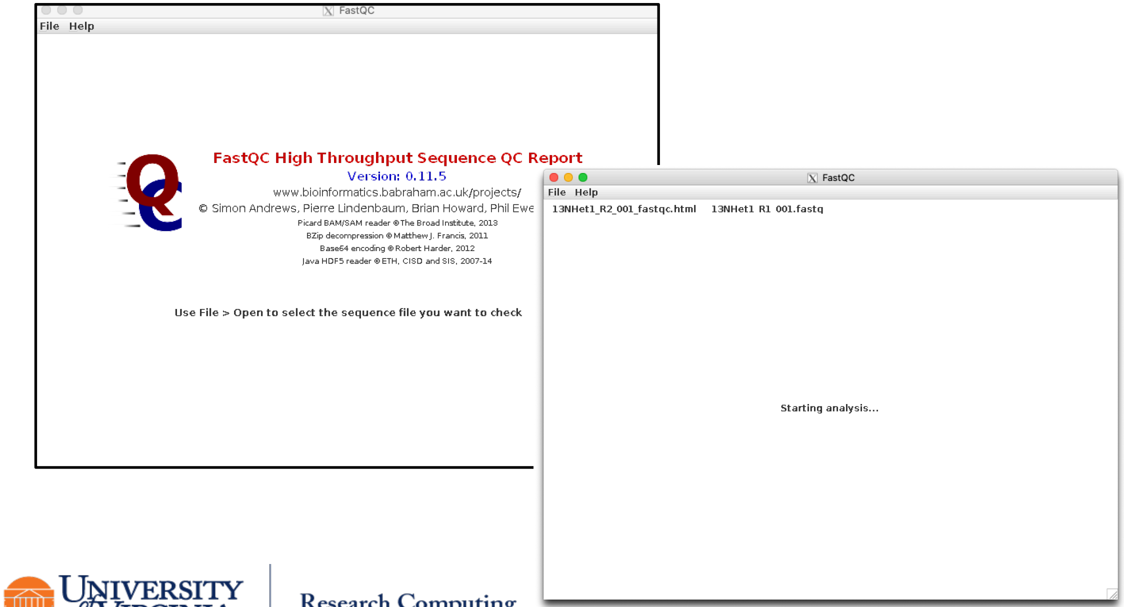
```
ssh -Y gka6a@rivanna.hpc.virginia.edu
```

- Start an interactive job

```
ijob -N1 -c 1 --ntasks=1 -J fastqc-inte -p standard  -A rivanna-training
```

- Load the module

```
module load fastqc
module list
fastqc &
```



FastQC High Throughput Sequence QC Report
Version: 0.11.5
www.bioinformatics.babraham.ac.uk/projects/
© Simon Andrews, Pierre Lindenbaum, Brian Howard, Phil Ewe
Picard BAM/SAM reader ©The Broad Institute, 2013
BZip decompression ©Matthew J. Francis, 2011
Base64 encoding ©Robert Harder, 2012
Java HDF5 reader ©ETH, CISD and SIS, 2007-14

Use File > Open to select the sequence file you want to check

Starting analysis...

RUNNING FASTQC

- Alternatively you can run FastQC in a non-interactive mode where you specify the files you want to process on the command line and FastQC will generate an HTML report for each file without launching a user interface. This would allow FastQC to be run as part of an analysis pipeline.

RUNNING FASTQC

```
$ fastqc
$ ijob -N1 -c 4 --ntasks=1 -J fastqc-inte -p standard  -A rivanna-training
salloc: Pending job allocation 5192794
salloc: job 5192794 queued and waiting for resources
salloc: job 5192794 has been allocated resources
salloc: Granted job allocation 5192794
srun: Step created for job 5192794

$ module load fastqc
time fastqc -t 4 -o fastqc-raw SRR5992812_1.fastq
Started analysis of SRR5992812_1.fastq
Approx 5% complete for SRR5992812_1.fastq
Approx 95% complete for SRR5992812_1.fastq
Analysis complete for SRR5992812_1.fastq

real   0m54.300s
user   0m51.677s
sys    0m1.051s
```

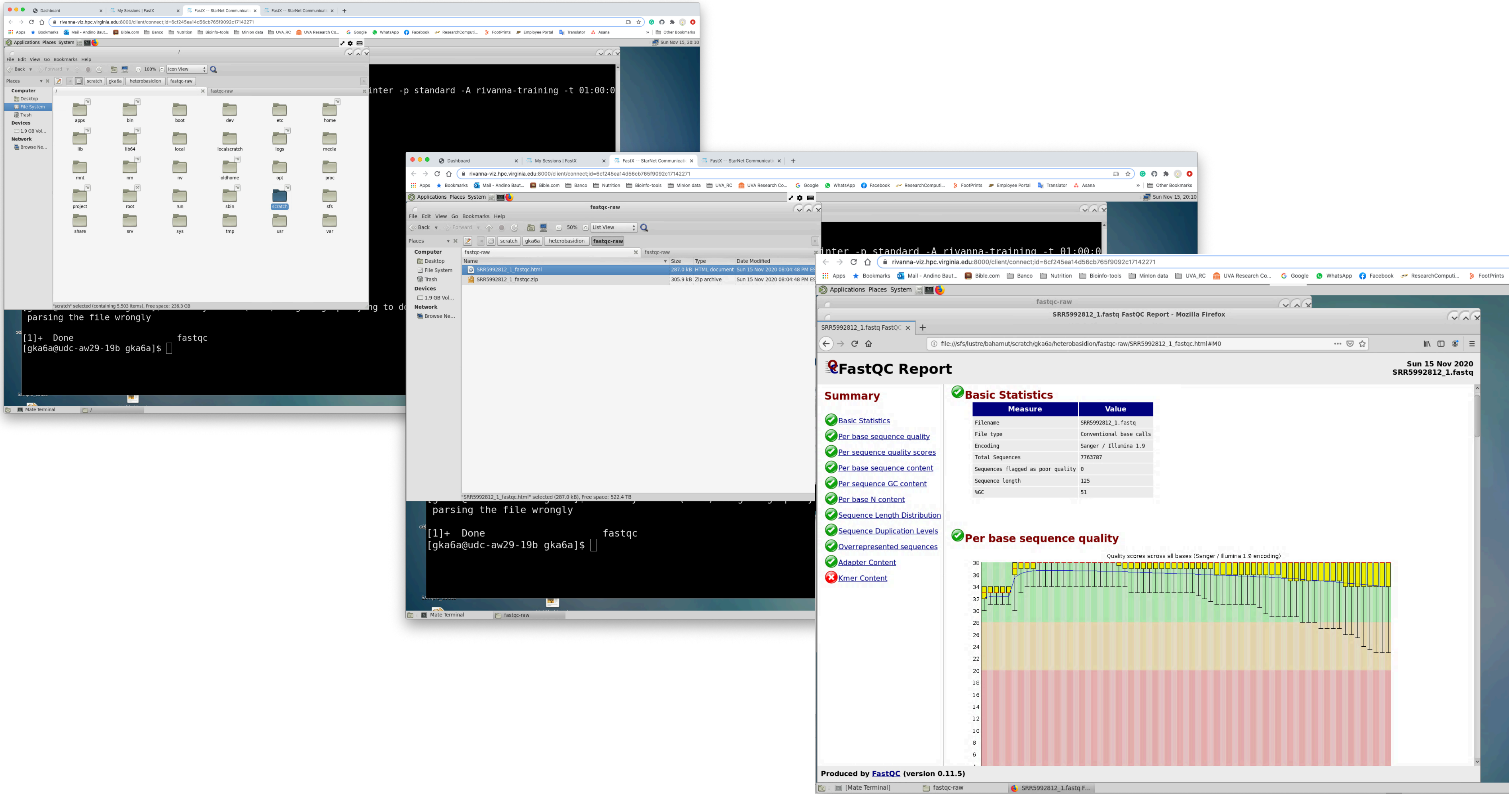
RUNNING FASTQC

```
#!/bin/bash
#SBATCH -N 1
#SBATCH --ntasks=1
#SBATCH -c 6
#SBATCH -p standard
#SBATCH -A rivanna-training
#SBATCH -t 01:00:00
#SBATCH -J fastqc
#SBATCH --output=%x_%j.out
#SBATCH --error=%x_%j.err


#  load modules
module purge
module load fastqc
module list

#  change to working directory
cd $SLURM_SUBMIT_DIR
pwd

cat $0
date +"%d %B %Y %H:%M:%S"
echo " "

# raw data, pre cleaning fastqc
# data formats .fastq,.fq,.fastq.gz

mkdir fastqc_raw
fastqc  -t $SLURM_CPUS_PER_TASK \
-o fastqc_raw  *.fastq.gz

echo " "
date +"%d %B %Y %H:%M:%S"
```

FASTQC - RESULTS

- SRR2584863_1_fastqc.html
- SRR2584863_1_fastqc.zip
- SRR2584863_2_fastqc.html
- SRR2584863_2_fastqc.zip
- SRR2584866_1_fastqc.html
- SRR2584866_1_fastqc.zip
- SRR2584866_2_fastqc.html
- SRR2584866_2_fastqc.zip
- SRR2589044_1_fastqc.html
- SRR2589044_1_fastqc.zip
- SRR2589044_2_fastqc.html
- SRR2589044_2_fastqc.zip

RUNNING - RESULTS



```
inter -p standard -A rivanna-training -t 01:00:0
```

```
 parsing the file wrongly

[1]+  Done                    fastqc
[gka6a@udc-aw29-19b gka6a]$ 
```

"scratch" selected (containing 5,503 items), Free space: 236.3 GB

fastqc-raw

Name	Size	Type	Date Modified
SRR5992812_1_fastqc.html	287.0 kB	HTML document	Sun 15 Nov 2020 08:04:48 PM ES
SRR5992812_1_fastqc.zip	305.9 kB	Zip archive	Sun 15 Nov 2020 08:04:48 PM ES

"SRR5992812_1_fastqc.html" selected (287.0 kB), Free space: 522.4 TB

SRR5992812_1.fastq FastQC Report - Mozilla Firefox

file:///sfs/lustre/bahamut/scratch/gka6a/heterobasidion/fastqc-raw/SRR5992812_1_fastqc.html#M0

FastQC Report

Sun 15 Nov 2020
SRR5992812_1.fastq

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

✓ Basic Statistics

Measure	Value
Filename	SRR5992812_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7763787
Sequences flagged as poor quality	0
Sequence length	125
%GC	51

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.11.5)

TRIMMOMATIC - PE

- Trimmomatic: A flexible read trimming tool for Illumina NGS data: http://www.usadellab.org/cms/?page=trimmomatic
- Paired End Mode:
- Single End Mode:

Usage:

```
PE [-version] [-threads <threads>] [-phred33|-phred64] [-trimlog
<trimLogFile>] [-summary <statsSummaryFile>] [-quiet] [-validatePairs] [-
basein <inputBase> | <inputFile1> <inputFile2>] [-baseout <outputBase> |
<outputFile1P> <outputFile1U> <outputFile2P> <outputFile2U>] <trimmer1>...


ILLUMINACLIP:?\
LEADING:? \
TRAILING:? \
SLIDINGWINDOW:?:? \
MINLEN:?
```

RUNNING TRIMMOMATIC - PE

```
$ module spider trimmomatic
--------------------------------------------------------------------------
trimmomatic:
--------------------------------------------------------------------------
Description
Trimmomatic performs a variety of useful trimming tasks for illumina
paired-end and single ended data.
     Versions:
        trimmomatic/0.36
        trimmomatic/0.39
--------------------------------------------------------------------------
For detailed information about a specific "trimmomatic" package (including
how to load the modules) use the module's full name.
 Note that names that have a trailing (E) are extensions provided by other
modules.
  For example:
     $ module spider trimmomatic/0.39
-----------------------------------------------------------
```

RUNNING TRIMMOMATIC - PE

```
$ module load trimmomatic/0.39
$ module show trimmomatic/0.39
--------------------------------------------------------------
   /apps/modulefiles/standard/core/trimmomatic/0.39.lua:
--------------------------------------------------------------

Description
===========
Trimmomatic performs a variety of useful trimming tasks for illumina
paired-end and single ended data.
More information
================
whatis("Homepage: http://www.usadellab.org/cms/index.php?page=trimmomatic")
setenv("EBROOTTRIMMOMATIC","/apps/software/standard/core/trimmomatic/0.39")
```

RUNNING TRIMMOMATIC - PE

```
$ head SRR2584863_1.fastq
$ tail SRR2584863_1.fastq
$ grep -c "@SRR2584863" SRR2584863_1.fastq
$ wc -l  SRR2584863_1.fastq
```

```
SRR2584863_1.fastq
SRR2584863_2.fastq
SRR2584866_1.fastq
SRR2584866_2.fastq
SRR2589044_1.fastq
SRR2589044_2.fastq
```

RUNNING TRIMMOMATIC - PE

- We are going to run Trimmomatic on the paired-end samples (PE). While using FastQC we saw that Nextera adapters were present in our samples. The adapter sequences come with the installation of trimmomatic.

```
$ ls -l  $EBROOTTRIMMOMATIC/adapters
-rw-r--r-- 1 uvacse users 239 May 16  2018 NexteraPE-PE.fa
-rw-r--r-- 1 uvacse users 538 May 16  2018 TruSeq2-PE.fa
-rw-r--r-- 1 uvacse users 142 May 16  2018 TruSeq2-SE.fa
-rw-r--r-- 1 uvacse users 259 May 16  2018 TruSeq3-PE-2.fa
-rw-r--r-- 1 uvacse users  93 May 16  2018 TruSeq3-PE.fa
-rw-r--r-- 1 uvacse users 119 May 16  2018 TruSeq3-SE.fa
```

RUNNING TRIMMOMATIC - PE

- **LLUMINACLIP**: Cut adapter and other illumina-specific sequences from the read.
- **SLIDINGWINDOW**: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- **LEADING**: Cut bases off the start of a read, if below a threshold quality
- **TRAILING**: Cut bases off the end of a read, if below a threshold quality
- **MINLEN**: Drop the read if it is below a specified length
- **TOPHRED33**: Convert quality scores to Phred-33

```
ILLUMINACLIP:adap.fa:2:40:15 \
LEADING:10 \
TRAILING:10 \
SLIDINGWINDOW:4:20 \
MINLEN:30
```

This will perform the following:

- Remove adapters (ILLUMINACLIP:illumina-adap.fa:2:40:15)
- Remove leading low quality or N bases (below quality 10) (LEADING:10)
- Remove trailing low quality or N bases (below quality 10) (TRAILING:10)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 20 (SLIDINGWINDOW:4:20)
- Drop reads below the 30 bases long (MINLEN:30)

RUNNING TRIMMOMATIC - PE

```
# brute force

java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar PE -threads 12 \
SRR2584863_1.fastq SRR2584863_2.fastq \
SRR2584863_1.paired.fastq SRR2584863_1.unpaired.fastq \
SRR2584863_2.paired.fastq SRR2584863_2.unpaired.fastq \
ILLUMINACLIP:$EBROOTTRIMMOMATIC/adapters/NexteraPE-PE.fa:2:40:15 \
LEADING:10 \
TRAILING:10 \
SLIDINGWINDOW:4:20 \
MINLEN:30

# duplicate 2 more times, changing the sample name
# error prone
```

Trimmomatic: a flexible trimmer for Illumina sequence data
Tutorial: http://www.usadellab.org/cms/?page=trimmomatic

RUNNING TRIMMOMATIC - PE

```
# this is the trimming command definition.  Each command executed
# in the order given.  Adapter trimming should go first, if used
trimmer="ILLUMINACLIP:adapter.fa:2:40:15 \
LEADING:10 \
TRAILING:10 \
SLIDINGWINDOW:4:20 \
MINLEN:30 "

samples="SRR2584863_1.fastq SRR2584866_1.fastq SRR2589044_1.fastq"

# for each sample read 1, generate the read 2 name be replacing .1. with .2.
# generate the paired and unpaired output file names by replacing .fastq with
# paired.fastq or unpaired.fastq
for r1 in $samples; do
    r2="${r1/_1./_2.}"
    r1p="${r1/.fastq/.paired.fastq}"
    r1u="${r1/.fastq/.unpaired.fastq}"
    r2p="${r2/.fastq/.paired.fastq}"
    r2u="${r2/.fastq/.unpaired.fastq}"

    command="trimmomatic PE -threads 5 \
    data/$r1 data/$r2 \
    $r1p $r1u \
    $r2p $r2u \
    $trimmer"
    echo $command
done
wait
```

TRIMMOMATIC - RESULTS

```
ls -l *.paired*
SRR2589044_1.paired.fastq
SRR2589044_2.paired.fastq
SRR2584863_1.paired.fastq
SRR2584863_2.paired.fastq
SRR2584866_1.paired.fastq
SRR2584866_2.paired.fastq
```

```
SRR2589044_1.paired.fastq
    Number of reads: 865259
    Number of bases in reads: 123340363
SRR2589044_2.paired.fastq
    Number of reads: 865259
    Number of bases in reads: 109997636
SRR2584863_1.paired.fastq
    Number of reads: 1245672
    Number of bases in reads: 177460402
SRR2584863_2.paired.fastq
    Number of reads: 1245672
    Number of bases in reads: 156393202
SRR2584866_1.paired.fastq
    Number of reads: 1997025
    Number of bases in reads: 263177758
SRR2584866_2.paired.fastq
    Number of reads: 1997025
    Number of bases in reads: 285357086
```

TRIMMOMATIC - RESULTS

- SRR2589044

```
Input Read Pairs: 1107090 Both Surviving: 865259 (78.16%) Forward
Only Surviving: 231726 (20.93%) Reverse Only Surviving: 4206 (0.38%)
Dropped: 5899 (0.53%)
```

- SRR2584863

```
TrimmomaticPE: Completed successfully

Input Read Pairs: 1553259 Both Surviving: 1245672 (80.20%) Forward
Only Surviving: 293049 (18.87%) Reverse Only Surviving: 6124 (0.39%)
Dropped: 8414 (0.54%)
```

- SRR2584866

```
TrimmomaticPE: Completed successfully

Input Read Pairs: 2768398 Both Surviving: 1997025 (72.14%) Forward
Only Surviving: 612822 (22.14%) Reverse Only Surviving: 139086
(5.02%) Dropped: 19465 (0.70%)
```

BOWTIE2

- http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#introduction
- https://www.rc.virginia.edu/userinfo/rivanna/software/bowtie2/

- Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.
- It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes.
- Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

BOWTIE2

- `bowtie2` takes a Bowtie 2 index and a set of sequencing read files and outputs a set of alignments in SAM format.
- "Alignment" is the process by which we discover how and where the read sequences are similar to the reference sequence.
- An "alignment" is a result from this process, specifically: an alignment is a way of "lining up" some or all of the characters in the read with some characters from the reference in a way that reveals how they're similar.

For example:

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

Where dash symbols represent gaps and vertical bars show where aligned characters match.

BOWTIE2

End-to-end alignment versus local alignment

- By default, Bowtie 2 performs end-to-end read alignment. That is, it searches for alignments involving all of the read characters. This is also called an "untrimmed" or "unclipped" alignment.
- When the `--local` option is specified, Bowtie 2 performs local read alignment. In this mode, Bowtie 2 might "trim" or "clip" some read characters from one or both ends of the alignment if doing so maximizes the alignment score.

BOWTIE2

End-to-end alignment example

- The following is an "end-to-end" alignment because it involves all the characters in the read. Such an alignment can be produced by Bowtie 2 in either end-to-end mode or in local mode.

```
Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG


Alignment:
Read:
              GACTGGGCGATCTCGACTTCG
              |||||  |||||||||| |||
Reference:    GACTG--CGATCTCGACATCG
```

BOWTIE2

Local alignment example

- The following is a "local" alignment because some of the characters at the ends of the read do not participate. In this case, 4 characters are omitted (or "soft trimmed" or "soft clipped") from the beginning and 3 characters are omitted from the end. This sort of alignment can be produced by Bowtie 2 only in local mode.

```
Read:          ACGGTTGCGTTAATCCGCCACG
Reference:     TAACTTGCGTTAAATCCGCCTGG
Alignment:
Read:          ACGGTTGCGTTAA-TCCGCCACG
                   ||||||||| ||||||
Reference:     TAACTTGCGTTAAATCCGCCTGG
```

BOWTIE2

Scores: higher = more similar

- An alignment score quantifies how similar the read sequence is to the reference sequence aligned to. The higher the score, the more similar they are.

RUNNING BOWTIE2

```
$ module spider bowtie
Description:
      Bowtie…

     Versions:
        bowtie2/2.1.0
        bowtie2/2.2.9

$ module spider bowtie2/2.2.9

--------------------------------------------
bowtie2: bowtie2/2.2.9
…
You will need to load all module(s) on any one of the lines below before
the "bowtie2/2.2.9" module is available to load.
      gcc/7.1.0
      gcc/9.2.0


…
      More information
      ================
Homepage: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
```

RUNNING BOWTIE2

```
$ module spider gcc/9.2.0 bowtie2/2.2.9
ml
Currently Loaded Modules:
  1) gcc/9.2.0   2) bowtie2/2.2.9


$ ls -l $EBROOTBOWTIE2
bin
doc
easybuild
example
scripts
$ ls -l $EBROOTBOWTIE2/bin
Bowtie2
bowtie2-align-l
bowtie2-align-s
bowtie2-build
bowtie2-build-l
bowtie2-build-s
bowtie2-inspect
bowtie2-inspect-l
bowtie2-inspect-s
LICENSE
MANUALMANUAL.markdown
NEWS
```

RUNNING BOWTIE2

```
$ bowtie2 -h
Bowtie 2 version 2.2.9 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

 -x <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
 -1 <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
 -2 <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
 -U <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
 -S <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.
Options (defaults in parentheses):
 Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  --qseq             query input files are in Illumina's qseq format
Performance:
  -p/--threads <int> number of alignment threads to launch (1)
```

BOWTIE2 - RESULTS

```
Building a SMALL index
10000 reads; of these:
Concordant alignment
  10000 (100.00%) were paired; of these:
    834 (8.34%) aligned concordantly 0 times
    9166 (91.66%) aligned concordantly exactly 1 time
    0 (0.00%) aligned concordantly >1 times
    ----
Discordant alignment
    834 pairs aligned concordantly 0 times; of these:
      42 (5.04%) aligned discordantly 1 time
    ----
The rest of the reads either align as singles
    792 pairs aligned 0 times concordantly or discordantly; of these:
      1584 mates make up the pairs; of these:
        1005 (63.45%) aligned 0 times
        579 (36.55%) aligned exactly 1 time
        0 (0.00%) aligned >1 times
94.97% overall alignment rate
```

BOWTIE2 - RESULTS

Result summary are divided in 3 sections:

- Concordant alignment - In your data (9166 + 0) reads align concordantly. Which is 91.66% of reads
- Discordant alignment - So now 834 reads remain which is 8.34% (100-91.66%). Of these, 792 reads align discordantly. That is to say, of the non-concordant fraction, 5.04% of reads (42 reads) align discordantly.
- The rest - Now, remember that alignment whether concord. or discord., but both are aligned in paired-end mode. The rest of the reads either align as singles (i.e. Read1 in one locus & Read2 in completely different locus or one mate aligned and the other unaligned) or may not align at all. So the reads that are in this section is Total - (Concord.+Discord.). 10000 -(9166+42) = 792
- Now to reach the overall alignment, count the mates in total (i.e. mates aligned in paired and mates aligned in single fashion). That would be: (9166 x2)+(42 x2)+579 = 18995 mates. That is 18995 mates aligned of total (10000 x2) mates, which is 94.97%.

BOWTIE2 - RESULTS

Output .sam

```
@HD   VN:1.0SO:unsorted
@SQ   SN:gi|9626243|ref|NC_001416.1| LN:48502
@PG   ID:bowtie2  PN:bowtie2  VN:2.2.9
      CL:"/apps/software/standard/compiler/gcc/9.2.0/bowtie2/2.2.9/bin/…
r5 99 gi|9626243|ref|NC_001416.1|   48010 42    138M = 48180 231
      GTCAGGAAAGTGGTAAAACTGCAACTCAATTACTGCAATGCCCTCGTAATTAAGTGAATTT…
r5 147gi|9626243|ref|NC_001416.1|   48180 42    61M = 48010 -231
      TGACCCAGGCTGACAAATTCCNGGACCCTTTTTGCTCCAGAGCGATGTTAATTTGTTCAAT…
r4 99 gi|9626243|ref|NC_001416.1|   40075 42    184M = 40211 184
      GGGCCAATGCGCTTACTGATGCGGAATTACGCCGTAAGGCCGCAGATGAGCTTGTCCATAT…
```

The first few lines (beginning with @) are SAM header lines, and the rest of the lines are SAM alignments, one line per read or mate. See the Bowtie 2 manual section on SAM output and the SAM specification for details about how to interpret the SAM file format.

RUNNING SAMTOOLS - FOR SAM/BAM FILES

```
$ module spider samtools
$ module load samtools/1.10
$ module show samtools/1.10
$ ls -l $EBROOTSAMTOOLS/bin
$ samtools --help
$ samtools view --help
   view SAM<->BAM<->CRAM conversion
$ samtools view -bS align2.sam > align2.bam
$ samtools sort align2.bam -o align2.sorted.bam
```

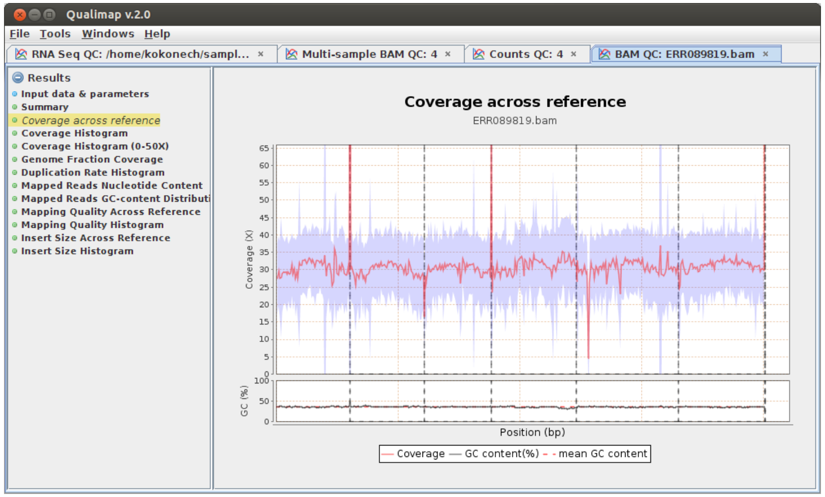
What are the options

- -b
- -S

RUNNING QUALIMAP

```
$ module spider qualimap
$ module load qualimap/2.2.1
    Files are located in $EBROOTQUALIMAP
$ ls -l $EBROOTQUALIMAP
$ qualimap -h
    Available tools:
     bamqc              Evaluate NGS mapping to a reference genome
     rnaseq             Evaluate RNA-seq alignment data
     counts             Counts data analysis (further RNA-seq data evaluation)
     multi-bamqc        Compare QC reports from multiple NGS mappings
     clustering         Cluster epigenomic signals
     comp-counts        Compute feature counts

$ qualimap bamqc -bam align2.sorted.bam
    Number of reads: 20000
    Number of valid reads: 18995
    Number of correct strand reads:0
```

Output: align2.sorted_stats



RNA-SEQ – DATA ANALYSIS

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

Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic

Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#introduction

Samtools: http://www.htslib.org/doc/samtools-merge.html

Qualimap: http://qualimap.conesalab.org/

STAR: https://github.com/alexdobin/STAR

HISAT: http://www.ccb.jhu.edu/software/hisat/index.shtml

StringTie: https://ccb.jhu.edu/software/stringtie/

Trinity: https://github.com/trinityrnaseq/trinityrnaseq/wiki

RSEM: https://deweylab.github.io/RSEM/

Salmon: https://salmon.readthedocs.io/en/latest/salmon.html

DESeq2: https://bioconductor.org/packages/release/bioc/html/DESeq2.html

edgeR: https://bioconductor.org/packages/release/bioc/html/edgeR.html