NCI/CCR Bioinformatics Training & Education Program

Website: btep.ccr.cancer.gov
Email: ncibtep@nih.gov
Amy Stonelake
Peter Fitzgerald
Carl McIntosh

Practical Bioinformatics: working at the Unix command line on Biowulf

June 5, 2019

Amy Stonelake, Ph.D.

Bioinformatics Training & Education Program (BTEP)

Today, we will…

- Log on to Unix system (NIH Biowulf)
- Work at the command line
- Transfer files back and forth from Biowulf
- Take a look at different file formats used in NGS
- Understand environment modules on Biowulf
- Run scientific software programs in interactive, batch and swarm modes
- Query scientific databases

Downloads available on the BTEP website

- RNA-Seq data
- Hand-out
- Unix/Linux Command Reference (Fosswire.com)
- Slides (pdf)

- In class you will be given a student login and password
- password is Btep5Jun2019

Practical Bioinformatics…on Biowulf

- Part 1 – Working at the Unix command line
- Part 2 – Moving files to Biowulf (and back again)
- Part 3 - Scientific analyses and databases

Part 1

Working at the command line in Unix on Biowulf

Who is “username”? (It’s you!)

- Wherever you see “username” in these slides or in the hand-out, you will type in your username
- Your username was assigned to you when you set up your helix/biowulf account
- For this class, you will use a student account “student1,student2,etc”
- But when you get back to your lab, use your “username”

What is Unix?

- An operating system just like Windows or Mac
- Has been around for a long time (1969)
- ”Linux” is a variety of Unix (ubuntu, red hat are Linux os)
- Well-suited to working with very large data files

- Bet you didn’t know – Apple computers use the Unix operating system!

DNA, a double helix…

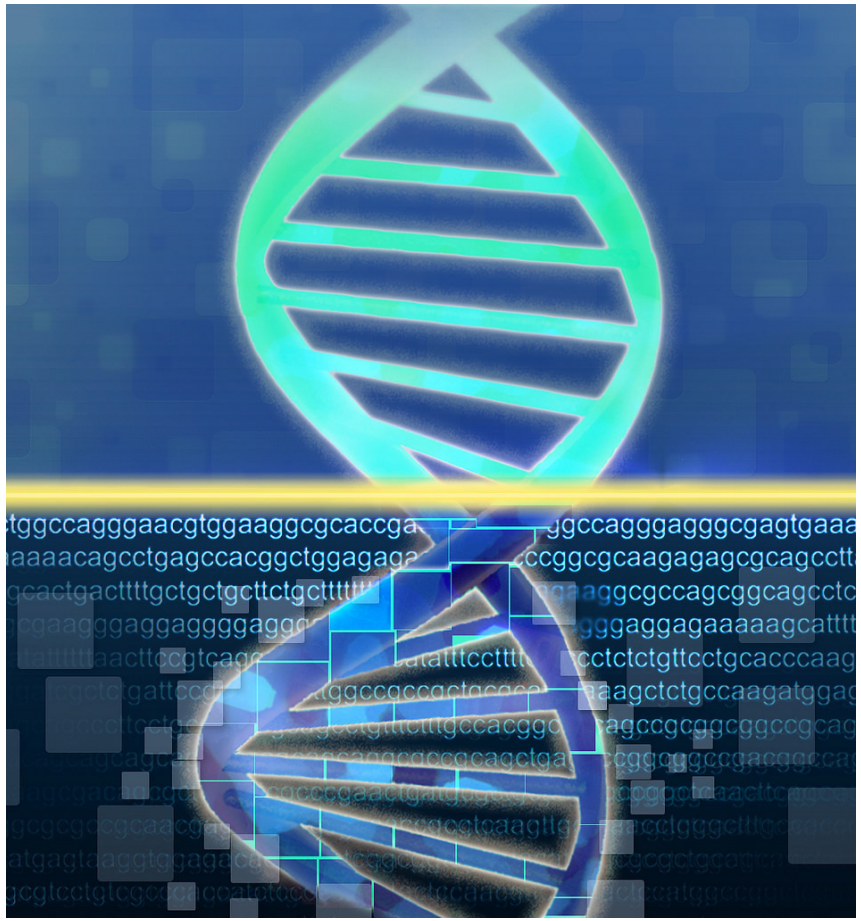
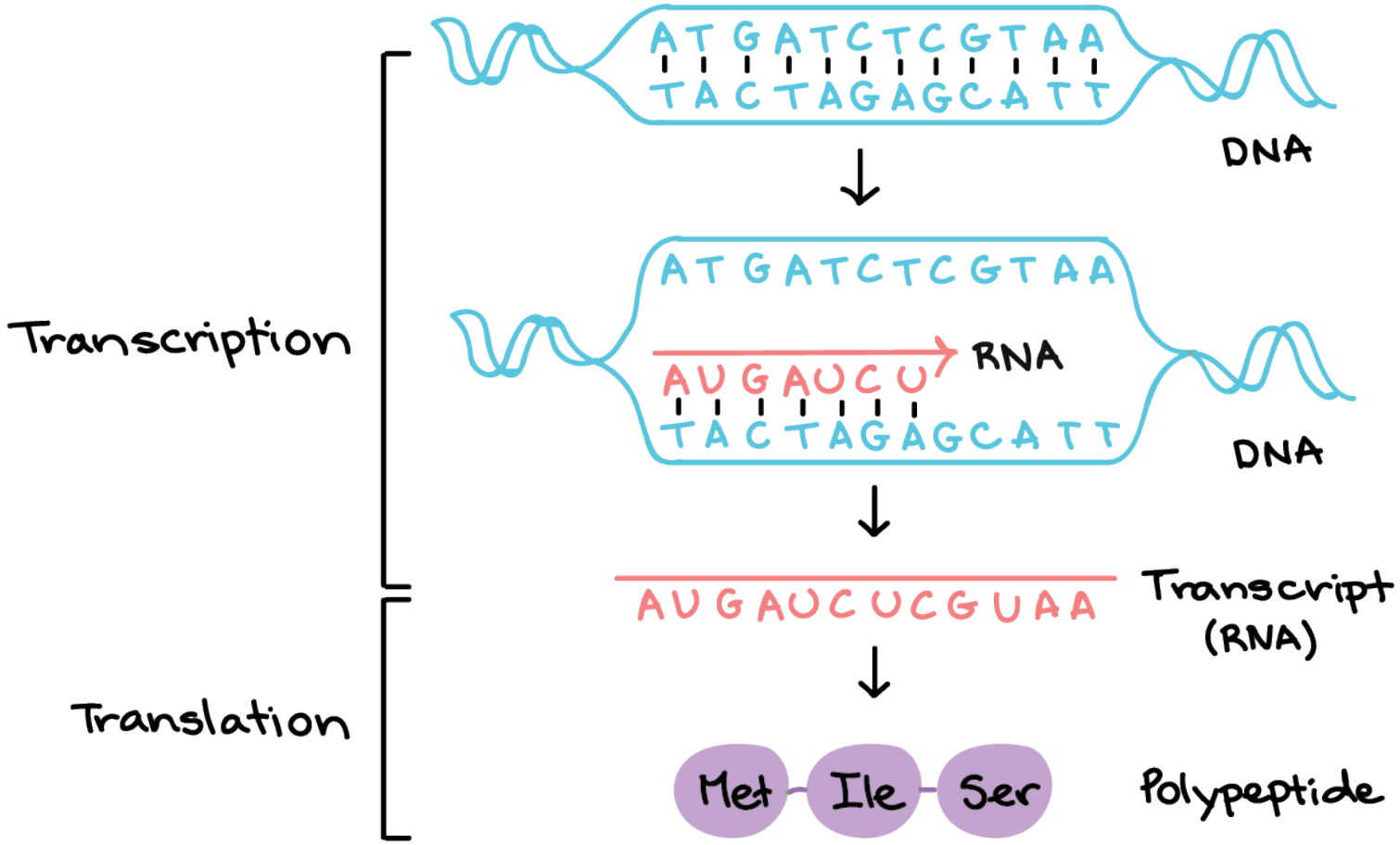


Image copyright NHGRI

Or a very large text file?

DNA to RNA to protein



Why Unix?

- Unix is well-suited to analysis of biological molecules, which can be represented as text files
- Many programs are available (and free, open-source) for biological analysis (BLAST, SAMtools, FASTQC)
- Tools (programs) can be linked together to form an analysis “pipeline”
- Programs can be centrally installed and maintained by sys admin
- Unix systems can handle “big data”
- Can run computationally intense programs (hours, days…)

Let’s get started… Introducing Biowulf

- The NIH high-performance compute cluster is known as “Biowulf”
- It is a 90,000+ processor Linux cluster
- Can perform large numbers of simultaneous jobs
- Jobs can be split among several nodes
- Scientific software (600+) and databases are already installed
- For more information see https://hpc.nih.gov
- Can only be accessed on NIH campus or via VPN
- *Do not put data with PII (personally identifiable information), patient data for example, on Biowulf*

Logging in to Biowulf

- If you're on a Mac… you can "ssh" from the "Terminal" app
- If you're on a PC… you will need to download and install "PuTTY"

- You are connecting to biowulf via a "secure shell" or "ssh" connection
- Once logged into biowulf, everything you do is running on biowulf, not your local machine
- Your local machine is just a gateway to biowulf (a Unix system)

Connecting to Biowulf with a Mac computer

- Find the “Terminal” app on your machine and open it
- You will see something like this

```
Last login: Thu Sep  6 16:10:04 on ttys000
NCI-02090676-ML:~ username $
```

- At the dollar sign “$” type the following:

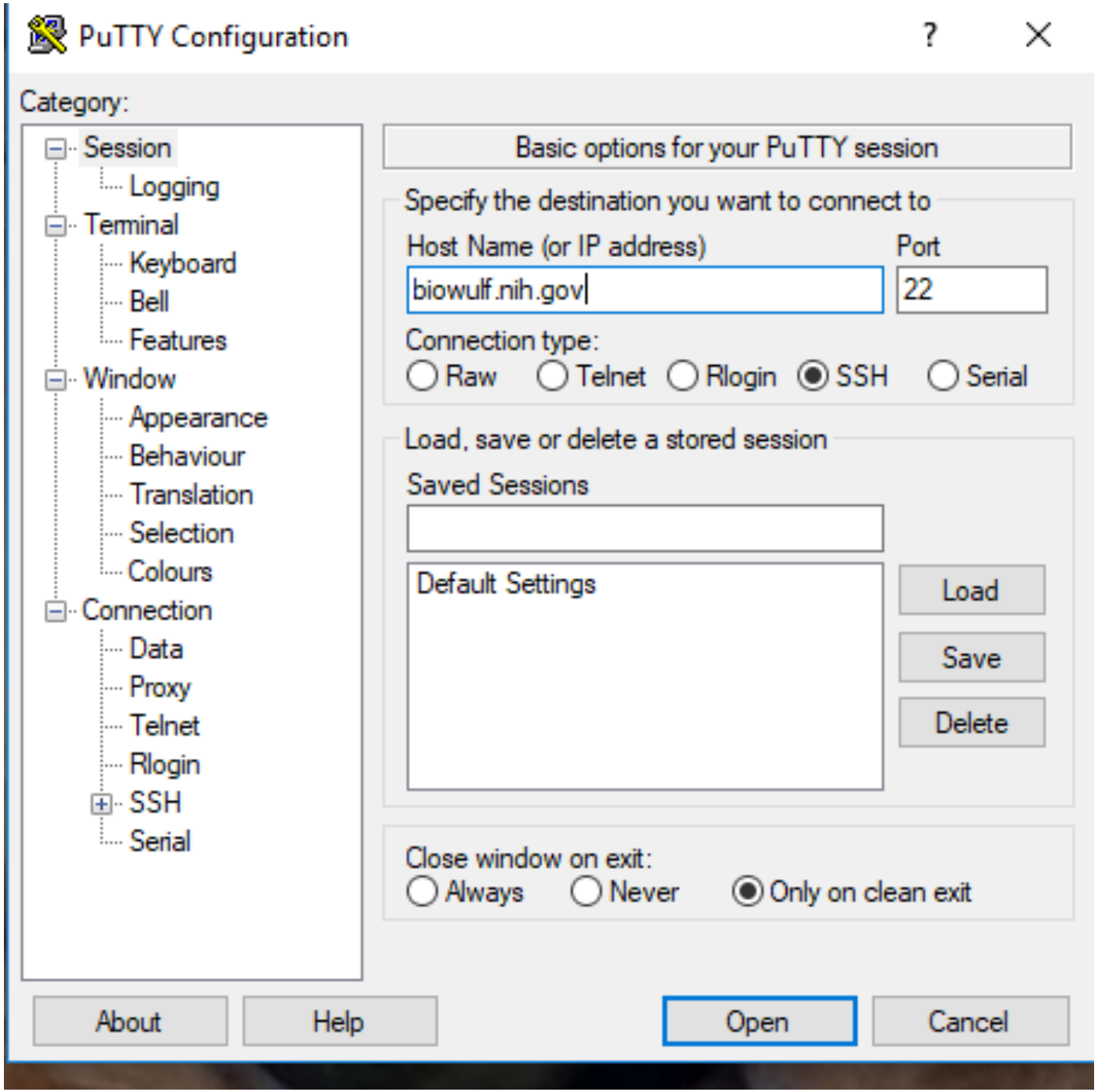
```
ssh username@biowulf.nih.gov
```

Where “username” is your username

Connecting to Biowulf with a Windows PC

- Download and install PuTTY

https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html


PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
biowulf.nih.gov
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel

Connecting to Biowulf by ssh

- Mac
- Open the Terminal window
- ssh username@biowulf.nih.gov

- PC
- Download and install PuTTY
- https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
- Host name – biowulf.nih.gov
- Connection type ”SSH”

Making the connection

- After making the “ssh” connection to biowulf, you will see a warning message about proper usage and then you will be prompted for your password
- **Be aware – the cursor does not move when typing in your password** (so if you make a mistake just hit return/enter and start over, or hit the backspace key many times)

username@biowulf.nih.gov's password: type in your password here

The command line

- Looks something like this

  [username@biowulf ~] $

  - “username” is your username
  - @biowulf means you are logged into biowulf
  - “~” indicates your home directory
  - When you see the dollar sign “$”, you know you are at the command line
  - If you don’t see the dollar sign, something is going on (running a program)

Your first Unix command…pwd

pwd means “print working directory” aka “Where am I?”

[username@biowulf ~]$ pwd

What do you see?
/home/username

You are in your “home” directory.

What do you see…ls

ls means “list the contents of the directory” aka “What’s in this folder?”

[username@biowulf ~]$ ls

What do you see?

(nothing)

Let’s create a file

[username@biowulf ~]$ touch file.txt

“touch” command creates a file

There is a space “ “ between the command, and the file name

You can name your file anything, but it has to follow the rules…

.txt file extension means this is a text file

File exists, but it is empty

Now let’s check again with “ls”

[username@biowulf ~]$ ls

What do you see? Should look something like this.

file.txt

Let’s create a folder (directory) for our file

- A ”directory” in Unix is a “folder” on other operating systems
- Directories contain files, more directories, programs, etc.
- Make a directory with the “mkdir” command
- There is a space between the “mkdir” command and the directory name

```
[username@biowulf ~]$ mkdir my_dir

[username@biowulf ~]$ ls
file.txt  my_dir
```

Moving on…the “mv” command

- How would you put the file we created inside the directory we created?
- Use the “move” command, “mv”
- [username@biowulf ~]$ mv file.txt my_dir
- There is a space “ “ between the “mv” command and ”file.txt”
- There is also a space between “file.txt” and “my_dir”

Where did our file go?

- [username@biowulf ~]$ ls

- What do you see?

  my_dir

So where is the file?

Looking inside a directory

- First, you have to “go to” the directory
- This is done with the “change directory”, or “cd” command

  [username@biowulf ~]$ cd my_dir
- There is a space “ “ between the command “cd” and ”my_dir”
- Now let’s look inside this directory, with “ls” command
- [username@biowulf ~]$ ls

  file.txt

Looking inside a file with “less”

- less to look inside a file
- quit (q) to get out of “less”

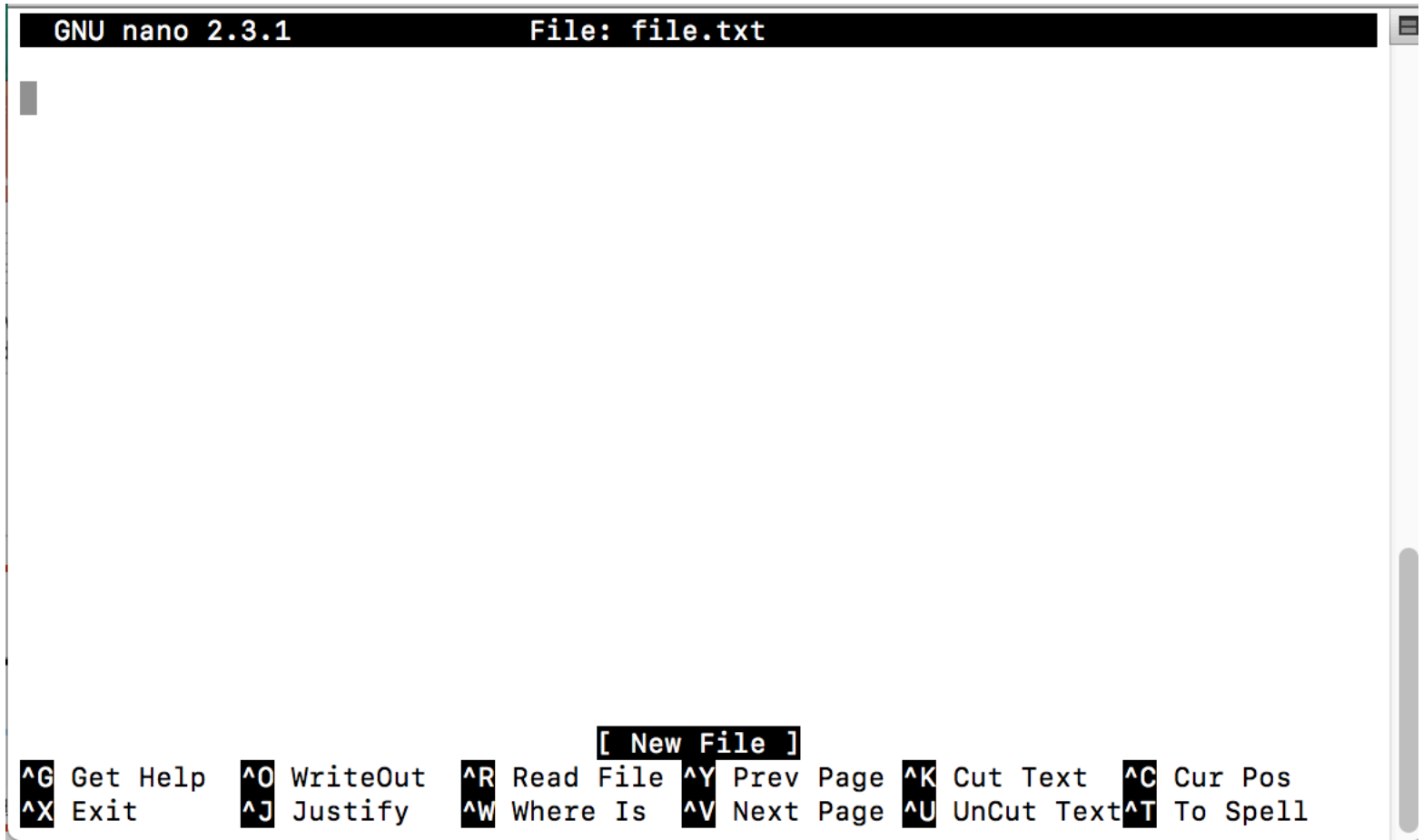
- [username@biowulf~]$ less file.txt

What do you see?

(nothing)

Let’s put something in that file using the “nano” editor

- [username@biowulf~]$ nano file.txt



The “nano” editor

[username@biowulf~]$ nano file.txt

”The quick brown fox jumped over the lazy yellow dog.”

Control X to quit

Save the buffer? Y

File name to write? file.txt (hit return/enter)

Now, what’s in the file?

```
[username@biowulf~]$ less file.txt
```

What do you see now?

“The quick brown fox jumped over the lazy yellow dog.”

Useful Unix Commands so far

- Figured out where we were with “pwd”
- Listed content of directory with “ls”
- Created text file with “touch”
- Looked inside text file with “less”
- Used the “nano” editor to put content in file
- Put file in a directory with “mv” command (can also use “mv” command to rename files)

```
[username@biowulf~]$ mv oldfilename.txt  newfilename.txt
```

- Moved from one directory to another with “cd” (change directory)

Coming up, more useful Unix skills

- Finding your path (pwd), changing your path (cd) and understanding your path
- Counting lines, words and characters with “wc”
- Using flags/options/switches
- Detailed listing of files with ”ls –alt”
- A look at permissions “rwx- - x rw-”
- Be careful removing files!
- How to name files
- Unix tricks (up arrow and tab complete)

A bit about finding your path in Unix

- pwd (print working directory)

  [username@biowulf~]$ pwd

  /home/username/my_dir (this is known as the “path”)

- cd (change directory, go home)

  [username@biowulf~]$ cd

  [username@biowulf~]$pwd

  /home/username (here is a different “path”)

Absolute vs. relative file paths

Absolute path (can be used to get anywhere)

```
cd /users/stonelakeak/Desktop/files/unix.txt
```

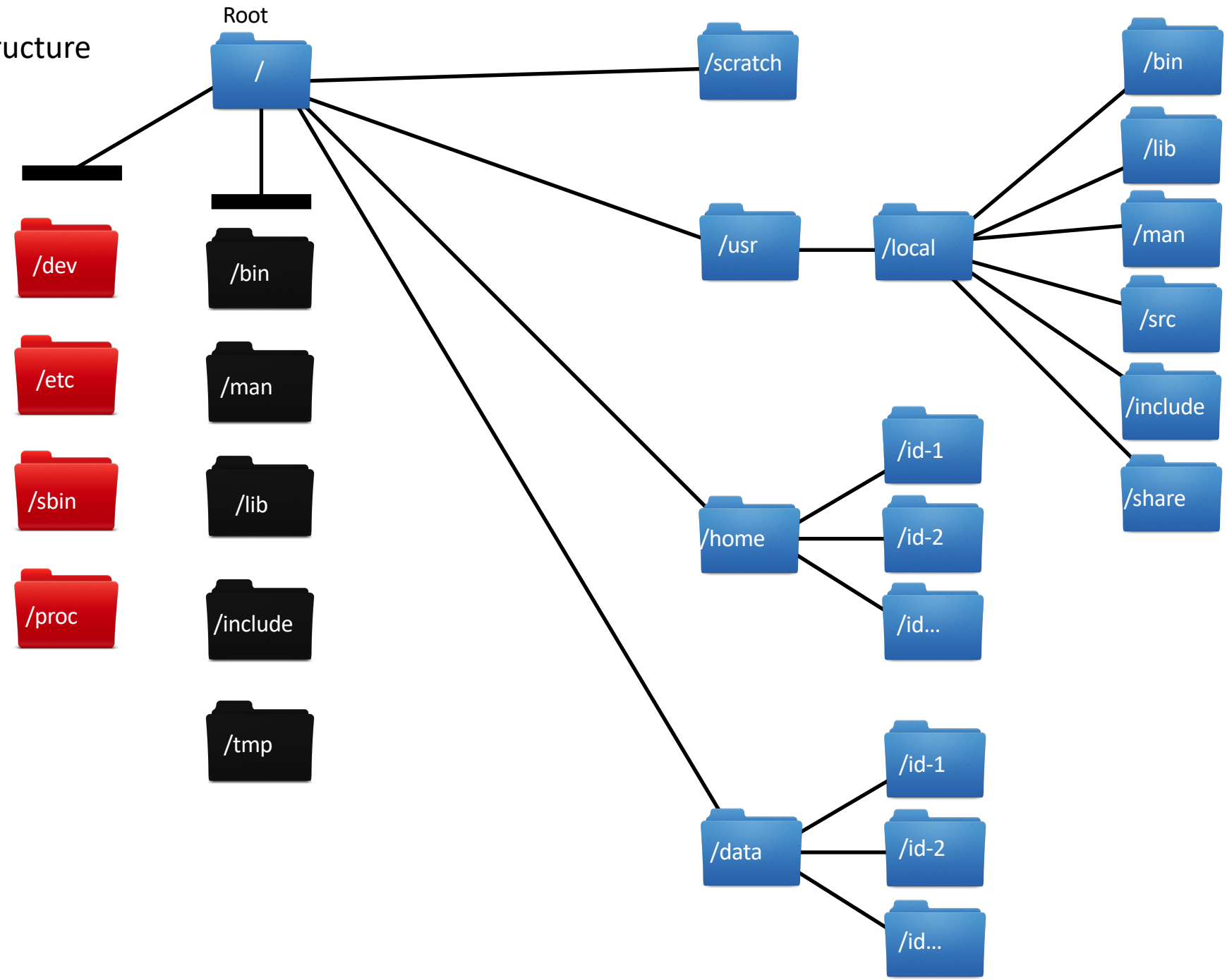
Relative path (only to get to files within the directory you are in)

If I am in /users/stonelakeak, I can just:

```
cd Desktop
```

Without typing the first forward slash (/)

Unix file structure



Counting lines, words and characters (char)

- Line count, word count, character count, “wc”

```
[username@biowulf ~] $ wc file.txt
1    10  53  file.txt

[username@biowulf ~] $ wc –l file.txt
1  file.txt
```

Line counting is very helpful when checking output, without opening the output file, which may be very large.

Using flags/options/switches

- These are used to control programs
- Go on the command line with the program

```
blast –i  input_file –db nr –o output_file
```

Flags – true or false (default), no additional info needed

Options – tells the program how to act

Switches – tells the program what to act on

More useful commands

head – outputs the first few lines of a file

tail – outputs the last few lines of a file

cat “concatenate” – used to view files, concatenate files or redirect output

$cat file1.txt file2.txt file3.txt >file4.txt

Input, output and append

input “<“

output “>”

Append ”>>”

At the command line:

head –n 40000 bigfile.fastq >smallfile.fastq

Hidden files… ls -a

```
[stonelakeak@biowulf ~]$ ls -a
.                       blast_output          .gnome2        slurm-7359593.out
..                      .cache                .java          slurm-7438680.out
3Ms_project             ccbr_pipeliner        .kshrc         slurm-7442490.out
.addressbook            .config               .lesshst       .ssh
.bash_history_biowulf   downloadfastq.swarm   mail           swarm_command_sra
.bash_history_helix     ._.DS_Store           .mozilla       teaching
.bash_logout            .DS_Store             .ncbi          .vim
.bash_profile           .emacs                .parallel      .viminfo
.bashrc                 file.txt              .pinerc        .Xauthority
bin                     .globus.cfg           project.json   .zshrc
```

Listing the details… ls -l

```
[stonelakeak@biowulf ~]$ ls -l
total 228
drwxrwxrwx 2 stonelakeak GAU         24576 Aug 15 16:39 3Ms_project
drwxr-xr-x 2 stonelakeak stonelakeak  4096 Aug  9 12:14 bin
drwx------ 4 stonelakeak stonelakeak  4096 Sep 18 15:23 blast_output
drwxr-xr-x 2 stonelakeak stonelakeak  4096 Aug 15 09:46 ccbr_pipeliner
-rw-r--r-- 1 stonelakeak stonelakeak   666 Aug 17 15:58 downloadfastq.swarm
-rw-r--r-- 1 stonelakeak stonelakeak    53 Sep 13 11:24 file.txt
drwx------ 2 stonelakeak stonelakeak  4096 Aug 15 09:48 mail
-rw-r--r-- 1 stonelakeak stonelakeak  3831 Aug 17 13:28 project.json
-rw-r--r-- 1 stonelakeak stonelakeak 51550 Aug 16 09:31 slurm-7359593.out
-rw-r--r-- 1 stonelakeak stonelakeak 51824 Aug 17 12:31 slurm-7438680.out
-rw-r--r-- 1 stonelakeak stonelakeak 51550 Aug 17 13:51 slurm-7442490.out
-rw-r--r-- 1 stonelakeak stonelakeak    65 Aug 15 10:10 swarm_command_sra
drwxr-xr-x 2 stonelakeak stonelakeak  4096 Sep 13 16:44 teaching
```

ls -alt

```
[stonelakeak@biowulf stonelakeak]$ ls -alt
total 9542816
drwxr-xr-x  377 root        root             16384 Apr 12 16:41 ..
drwxr-xr-x    2 stonelakeak stonelakeak       4096 Apr 11 10:00 array_express_data
drwxrwx---+   8 stonelakeak stonelakeak       4096 Apr  8 14:26 .
drwxr-xr-x    2 stonelakeak stonelakeak       4096 Apr  2 16:02 Mackem_pact_multiqc_report_data
-rw-r--r--    1 stonelakeak stonelakeak        699 Apr  2 16:02 slurm-23637880.out
-rw-r--r--    1 stonelakeak stonelakeak    1224167 Apr  2 16:02 Mackem_pact_multiqc_report.html
-rw-r--r--    1 stonelakeak stonelakeak        327 Apr  2 15:51 multiqc.sh
-rw-r--r--    1 stonelakeak stonelakeak        518 Apr  2 15:00 swarm_23633984_0.o
-rw-r--r--    1 stonelakeak stonelakeak       3373 Apr  2 15:00 swarm_23633984_0.e
-rw-r--r--    1 stonelakeak stonelakeak         87 Apr  2 14:56 slurm-23634014.out
-rw-r--r--    1 stonelakeak stonelakeak        279 Apr  2 14:52 fastqc.sh
-rw-r--r--    1 stonelakeak stonelakeak        240 Apr  2 14:46 fastqc.swarm
drwxr-xr-x    2 stonelakeak stonelakeak       4096 Apr  2 14:44 Mackem_pactme_multiqc_report_data
-rw-r--r--    1 stonelakeak stonelakeak        692 Apr  2 14:41 slurm-23633070.out
-rw-r--r--    1 stonelakeak stonelakeak    1127798 Apr  2 14:41 Mackem_pactme_multiqc_report.html
-rw-r--r--    1 stonelakeak stonelakeak        425 Apr  2 14:26 swarm_23631576_0.o
-rw-r--r--    1 stonelakeak stonelakeak       1143 Apr  2 14:26 swarm_23631576_0.e
```

All files, including hidden files, listed with full details, by descending time order.

A first look at permissions

```
drwxrwxrwx 2 stonelakeak GAU        24576 Aug 15 16:39 3Ms_project

drwxr-xr-x

d -> directory
r -> read
w -> write
x -> execute
User/owner, group and other

User/owner is the creator of the files, usually you
Group is a group of users having the same privileges
Other is the general public

chmod -> Unix command to change permissions (chmod 777 gives everyone full permissions).
```

A word of caution on “rm” (removing files)

rm (remove file)
rmdir (remove directory)
A directory must be empty before you can remove it.

```
cd
rmdir my_dir
cd my_dir
rm file.txt
ls
cd ..
rmdir my_dir
```

Naming files and directories on Unix

- Don’t use spaces in names

  file1.txt is ok, but not file 1.txt
- Don’t use these characters in file or dir names (/, <, >, |, :, &)
- File and directory names are case sensitive

  file.txt and FILE.txt are different
- But, files with same name can exist in different directories

  /home/file.txt and /data/file.txt are valid
- Use uppercase, lowercase, numbers, dot (.) and underscore (_)

Use underbars or CamelCase for file names

- Use underbars for multiple word file names like this
- Heres_a_multiple_word_file_name.txt

- Or CamelCase
- HeresAMultipleWordFileName.txt

- But NOT with spaces!
- “Here’s a multiple word file name.txt” (do not do this!)

Unix tricks

- Hit the “up arrow” key on your keyboard to recall previous commands
- Tab complete – type the first part of file or directory name and “tab” will complete the rest IF it is a “unique” file or directory name
- Wildcard (*) in Unix can make your life easier (or harder)
- For example…

```
mv reallylongfilename.fastq.gz  new_dir/fastq
```

OR

```
mv *.fastq.gz  new_dir/fastq
```

(this will move any file with extension .fastq.gz)

What is a ”tarball” file in Unix?

- Very large files that have been compressed verybigfile.tar
- How to “untar” a file
- At the command line, type:

  tar –xvf verybigfile.tar
- Files may also be ”zipped” using gzip/bzip, and need to be unzipped

  tar –xvzf verybigfile.tar.gz
- Or just be “zipped”, like the fastq.gz files we downloaded

  gunzip verybigfile.gz

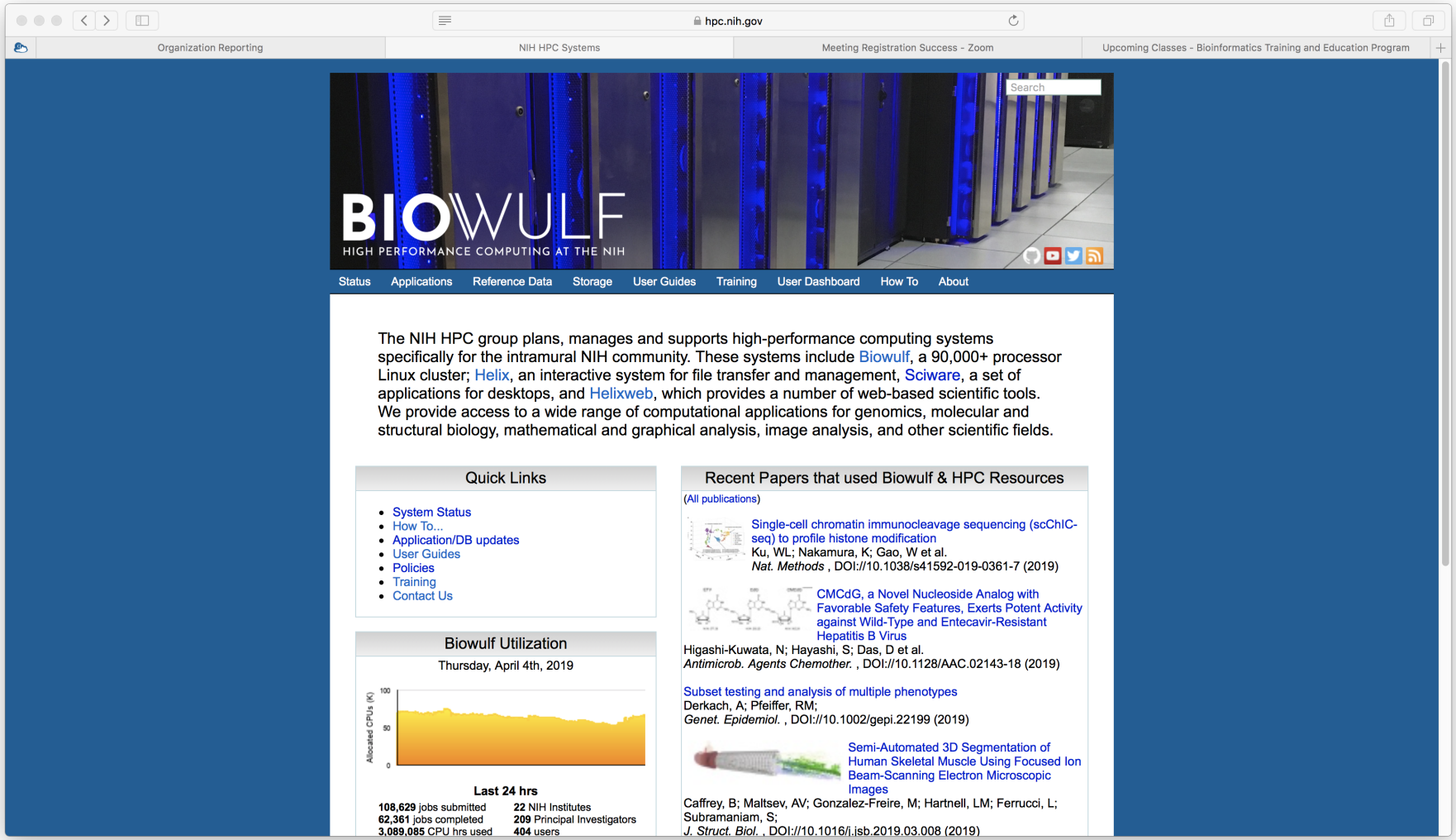
Your Biowulf account

- You have both /home and /data directories on your account
- /home is limited size
- /data is where you will do most of your work
- /lscratch is available for temp files
- To do RNA seq work, request up to 1 TB in your /data directory
- Keep an eye on your disk space!
- Do not work on the Biowulf login node!

Being a good citizen on Biowulf

- checkquota will show /home and /data
- OR
- See http://hpc.nih/gov -> User Account -> Disk Usage

User Dashboard on Biowulf



Checking Disk Usage on Biowulf

hpc.nih.gov

Organization Reporting | User Dashboard | Meeting Registration Success - Zoom | Upcoming Classes - Bioinformatics Training and Ed

BIOWULF
HIGH PERFORMANCE COMPUTING AT THE NIH

Search

Status | Applications | Reference Data | Storage | User Guides | Training | User Dashboard | How To | About

User Dashboard

last page refresh: 2019-04-04 12:45:58 PM

Accounts | Disk Usage | Job Info

Diskspace Usage

last updated: 2019-04-02 03:00:10 PM

/data/CCBR (gs4)	194.7 TB	/	200.0 TB	owner: **maggiec**
/data/GAU (gs5)	18.0 TB	/	20.0 TB	owner: **fitzgepe**
/data/elloumif (spin1)	875.9 MB	/	100.0 GB	owner: **elloumif**
/data/nelsong (spin1)	69.1 GB	/	100.0 GB	owner: **nelsong**
/data/stonelakeak (gs10)	676.8 GB	/	2.0 TB	request quota increase
/home/chenx3	5.6 GB	/	16.0 GB	
/home/lautenbj	298.2 MB	/	16.0 GB	
/home/stonelakeak	3.7 GB	/	16.0 GB	

Last modified: 4 April 2019

HPC @ NIH ~ Contact

Disclaimer ~ Privacy ~ Accessibility ~ CIT ~ NIH ~ DHHS ~ USA.gov

How many cores and how much memory should you allocate to run trimmomatic on Biowulf as a batch job?

Batch job

Most jobs should be run as batch jobs.

Create a batch input file (e.g. trimmomatic.sh). For example:

```
#!/bin/bash

ml trimmomatic || exit 1
java -Djava.io.tmpdir=. -jar $TRIMMOJAR PE -phred33 -threads $SLURM_CPUS_PER_TASK \
    SRR292678_1.fastq.gz SRR292678_2.fastq.gz \
    output_forward_paired.fq.gz output_forward_unpaired.fq.gz \
    output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz \
    ILLUMINACLIP:/usr/local/apps/trimmomatic/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 \
    SLIDINGWINDOW:4:15 MINLEN:36
```

Submit this job using the Slurm sbatch command.

```
sbatch -c 2 --mem=6g trimmomatic.sh
```

Biowulf has suggestions for running jobs with adequate resources.

Additional resources

- Datacamp.com to learn unix/R/python
- Unix Tutorial for Beginners (https//www.cs.sfu.ca/~ggbaker/reference/unix/index.html)
- Software carpentry (http://swcarpentry.github.io/shell-novice/)
- hpc.nih.gov (Biowulf)
- Unix cheat sheet (Fosswire.com)
- man pages for any command

[username@biowulf] $ man ls

Unix/Linux Command Reference

FOSSwire.com

File Commands

ls – directory listing
ls -al – formatted listing with hidden files
cd *dir* - change directory to *dir*
cd – change to home
pwd – show current directory
mkdir *dir* – create a directory *dir*
rm *file* – delete *file*
rm -r *dir* – delete directory *dir*
rm -f *file* – force remove *file*
rm -rf *dir* – force remove directory *dir* *
cp *file1 file2* – copy *file1* to *file2*
cp -r *dir1 dir2* – copy *dir1* to *dir2*; create *dir2* if it doesn't exist
mv *file1 file2* – rename or move *file1* to *file2* if *file2* is an existing directory, moves *file1* into directory *file2*
ln -s *file link* – create symbolic link *link* to *file*
touch *file* – create or update *file*
cat > *file* – places standard input into *file*
more *file* – output the contents of *file*
head *file* – output the first 10 lines of *file*
tail *file* – output the last 10 lines of *file*
tail -f *file* – output the contents of *file* as it grows, starting with the last 10 lines

Process Management

ps – display your currently active processes
top – display all running processes

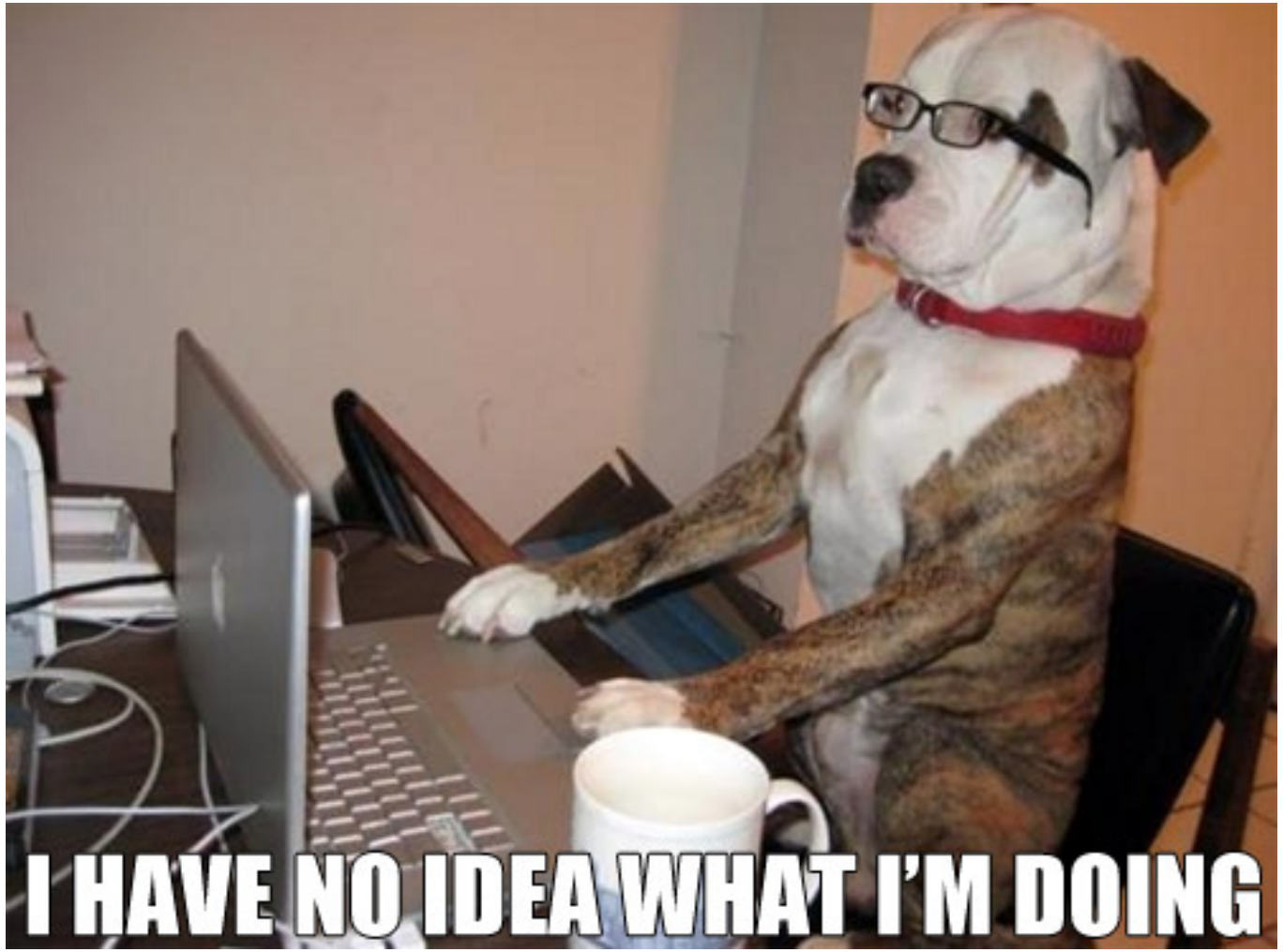
System Info

date – show the current date and time
cal – show this month's calendar
uptime – show current uptime
w – display who is online
whoami – who you are logged in as
finger *user* – display information about *user*
uname -a – show kernel information
cat /proc/cpuinfo – cpu information
cat /proc/meminfo – memory information
man *command* – show the manual for *command*
df – show disk usage
du – show directory space usage
free – show memory and swap usage
whereis *app* – show possible locations of *app*
which *app* – show which *app* will be run by default

Compression

tar cf *file.tar files* – create a tar named *file.tar* containing *files*
tar xf *file.tar* – extract the files from *file.tar*
tar czf *file.tar.gz files* – create a tar with Gzip compression
tar xzf *file.tar.gz* – extract a tar using Gzip
tar cjf *file.tar.bz2* – create a tar with Bzip2 compression
tar xjf *file.tar.bz2* – extract a tar using Bzip2
gzip *file* – compresses *file* and renames it to *file.gz*

End of Part 1 – take a short break, any ?s



Part 2: Moving files to Biowulf (and back again)

We will look at several methods to transfer files.

- The Globus service allows easy file transfers (you need to request Globus access for your Biowulf account, info at globus.org) –best when moving large files

- Mounting a drive – creates a graphical user interface (GUI) so you can drag and drop files

- Secure copy protocol (scp) or secure file transfer protocol (sftp), with WinSCP (PC) or FileZilla* (Mac,PC)

Log on to Biowulf

- To work at the command line, you need a “ssh” connection – go ahead and establish one now using Terminal (Mac) or PuTTY (PC)

- To transfer files back and forth, you need “scp or sftp” connection, there are several ways to do that

Moving files with Globus

- For transferring large files
- Need a Biowulf account to use Globus
- Setup your Globus endpoint (only need to do this one time)
- Open Globus Connect Personal (need to do this every time)

Data transfer and sharing using Globus

Globus is a service that makes it easy to move, sync, and share large amounts of data. Globus will manage file transfers, monitor performance, retry failures, recover from faults automatically when possible, and report the status of your data transfer. Globus uses GridFTP for more reliable and high-performance file transfer, and will queue file transfers to be performed asynchronously in the background.

Globus was developed and is maintained at the University of Chicago and is used extensively at supercomputer centers and major research facilities. [Globus website]

NIH scientists who wish to utilize this service to transfer data to/from their Helix/Biowulf disk space can use their NIH Login username and password to login.

No matter how you transfer data in and out of our systems, be aware that PII and PHI data cannot be stored or transferred into the NIH HPC systems.

Helix/Biowulf Endpoints

The endpoint nihhelix#helix has been shut down as of 30 Apr 2017. All users must use the nihhpc#globus endpoint. Any endpoints that were previously shared from nihhelix#helix must be re-shared from nihhpc#globus.

The Globus endpoint for transferring data to or from your Helix/Biowulf /home, /data or /scratch areas is **nihhpc#globus**. This endpoint is implemented using eight "Data Transfer Nodes" which can operate in parallel to provide 80 Gb/s of aggregate bandwidth.

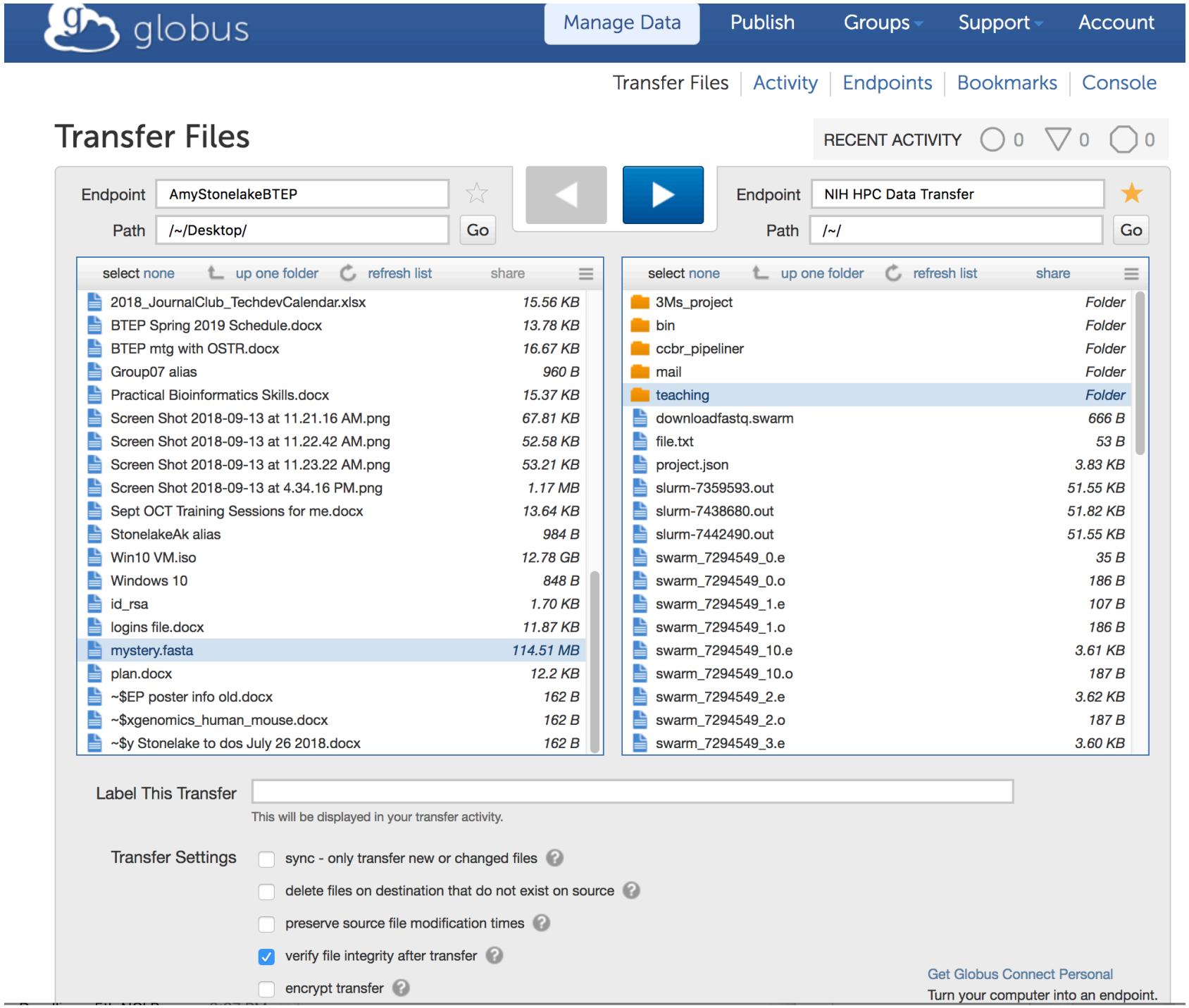
You do not need to be logged on to Helix or Biowulf to start or monitor a transfer.

Logging into Globus with your NIH login

back to top

Instructions for using Globus on hpc.nih.gov

- Under “How To”
- Choose “Transfer Files”
- Select the link “Setting up a Globus account, transferring and sharing data”
- If you have trouble setting up Globus on your laptop, contact staff@hpc.nih.gov



Go to globus.org
Choose your personal endpoint
Choose a folder on biowulf
Click the blue arrow
You get an e-mail when it's done!

Mounting a drive

Mac – “Go” -> ”Connect to server”

smb://helixdrive.nih.gov/username

PC - ”Computer”, “Tools” then “Map Network Drive” tab

See instructions on hpc.nih.gov (Biowulf) – “How To – Transfer Files”, “Transferring data to/from the NIH HPC systems”

Secure Copy Protocol (scp)

- Windows PC – download WinSCP – GUI – drag and drop files – easy!

- Mac – scp at the Mac command line

- FileZilla – be sure to get a clean copy!
  - Mac OSX:
  - http://packages.partek.com/bin/filezilla/fz-osx.app.tar.bz2
  - Windows 32-bit:
  - http://packages.partek.com/bin/filezilla/fz-win32.exe
  - Windows 64-bit:
  - http://packages.partek.com/bin/filezilla/fz-win64.exe

Summary of data transfer options on Biowulf

Platform	Application	Pros	Cons
All platforms	Globus	Best for very large files (> 256MB). Clients for all platforms, web-based. Notifications sent on completion.	The client must first be installed on the desktop.
	Filezilla v3.0	Better control over transfer during the process, fewer and simpler controls than WinSCP, fastest transfer rates by sFTP.	scp not an option.
Windows	WinSCP	Much faster transfer rates than PuTTY-pscp/psftp, but slightly faster than Filezilla for uploads using scp (rates were found to vary considerably by cipher used, in the order of Blowfish > AES >> 3DES), highly comprehensive configuration.	Cumbersome user interface for changing local and remote directories.
	pscp/psftp	Direct command line control over process.	Need to run through the command prompt, slowest transfer rates seen.
	Mapped Network Drive	Convenient.	Fairly slow transfer rates, especially very large files.
Macs	bbcp,scp,sftp	Can be used for scripting & automatic file transfers, fastest transfer rates	non-GUI interface.
	Fugu	Easy to configure and use.	Slower than command-line.
	Mapped Network Drive	Convenient drag-and-drop.	Fairly slow transfer rates, especially for large files.
Linux/Unix	scp,sftp	Same as for Macs.	Same as for Macs.
	bbcp	Fastest transfer rate.	

Last modified: 7 September 2018

Downloading files from BTEP website

- Go to the class website

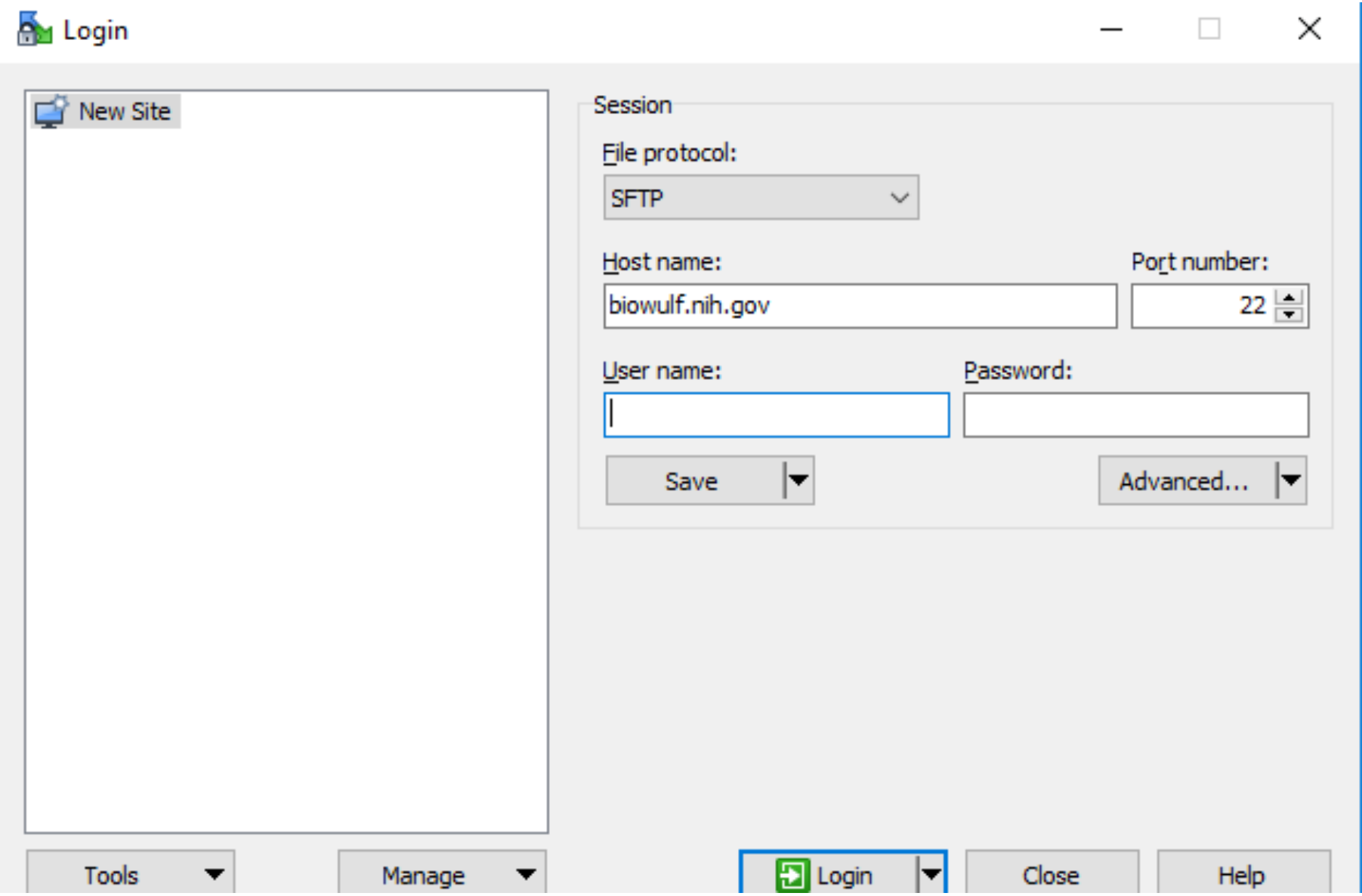
btep.ccr.cancer.gov/classes/unix-Frederick

Download all files – fastq, pdf – put them in your Downloads directory

Next…

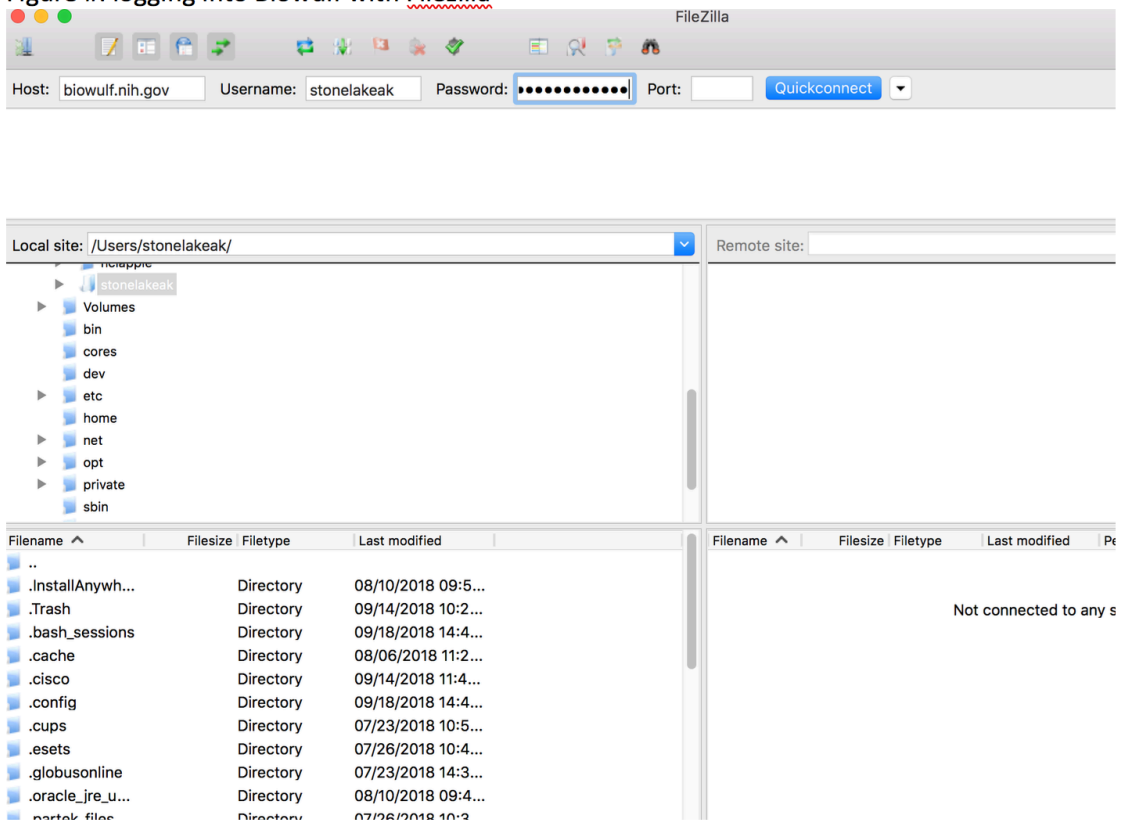
We need to transfer the fastq.gz file to Biowulf so we can work with it

Using WinSCP to transfer files



OR…Using FileZilla to transfer files

Figure #: logging into Biowulf with Filezilla



OR…Another way to transfer files… uploading files to Biowulf on a Mac computer using the command line

Open Terminal (Mac) and use cd to go to the location of the downloaded file

```
cd /Users/username/Downloads
```

Then type this (on the command line of your machine)

```
scp   filename.fastq.gz  username@biowulf.nih.gov:/data/username
```

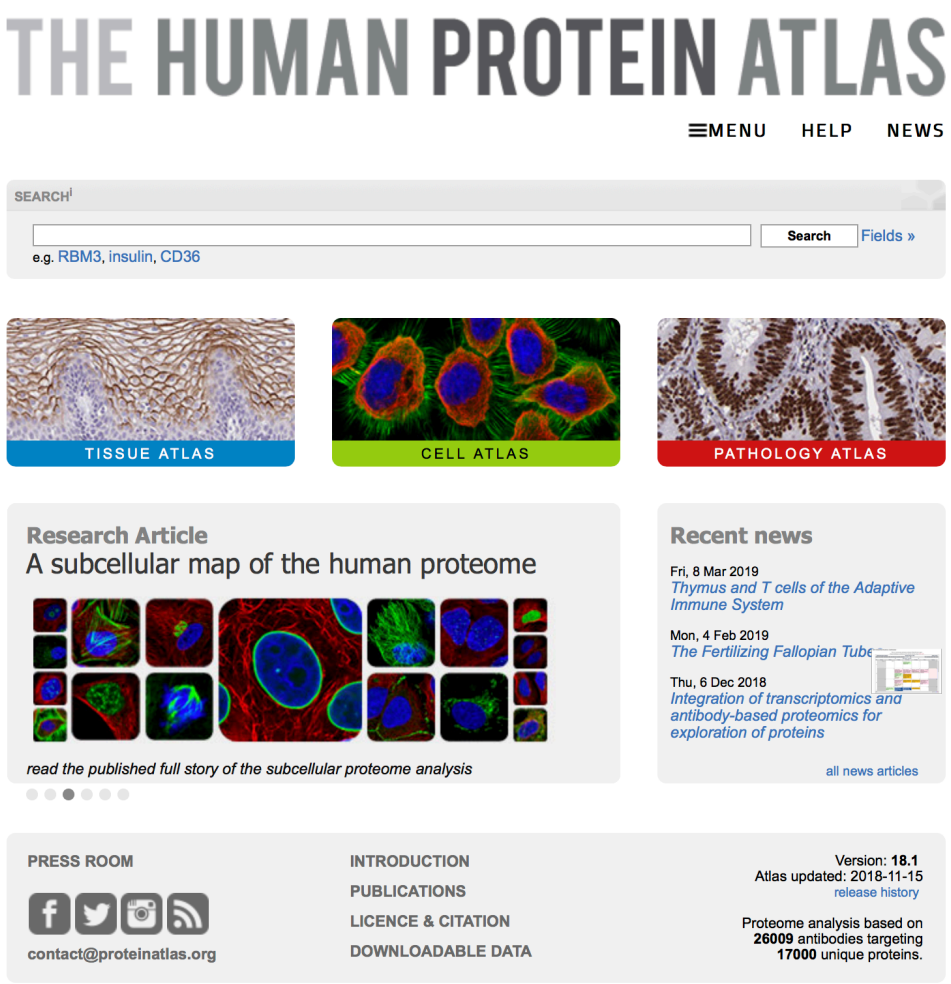
Where “filename.fastq.gz” is the name of the file
Username is your username

Mistakes you will make when uploading files from your Mac to Biowulf

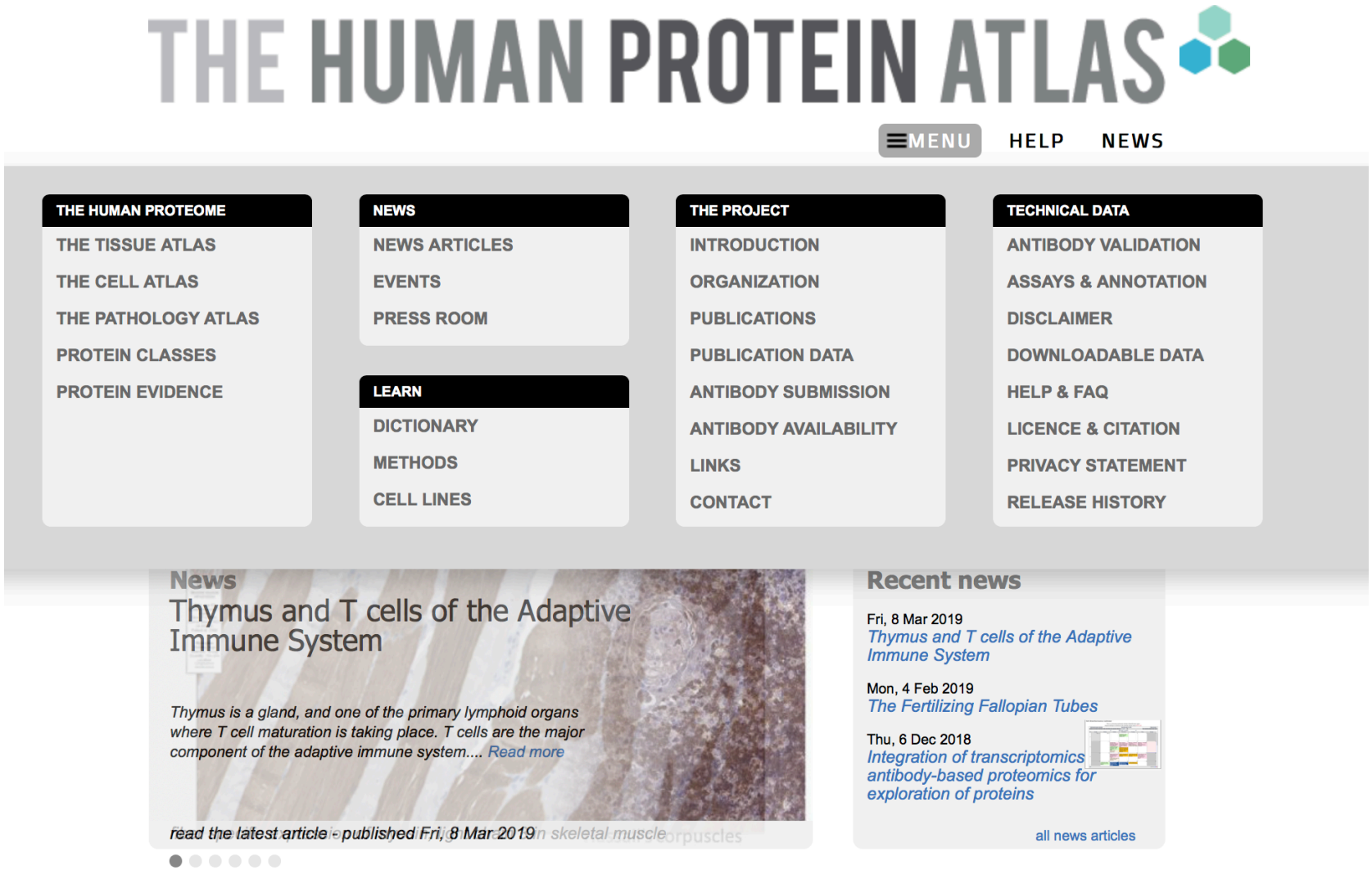
- You will forget to type the command in a terminal window **on your machine**
- You will type “username” instead of your username
- You will type “filename” instead of the name of the file
- You will not type the path correctly to the file.
- You will have a typo in the name of the file

After you do it correctly, be sure to celebrate!

Where did these files come from?
The Human Protein Atlas (proteinatlas.org)



Click on DOWNLOADABLE DATA



RNA gene data

RNA sequencing data for human tissue

ioinformatics Calendar | Bioinformatics
ng and Education Program

Search Fields »

≡MENU HELP NEWS

TECHNICAL DATA : DOWNLOADABLE DATA

4 **RNA gene data**
RNA levels in 64 cell lines and 37 tissues based on RNA-seq. The tab-separated file includes Ensembl gene identifier ("Gene"), analysed sample ("Sample") and transcripts per million ("Value" and "Unit"). The data is based on The Human Protein Atlas version 18.1 and Ensembl version 88.38.
RNA sequencing data for human tissue
RNA sequencing data for human cell lines

rna_tissue.tsv.zip
TSV-file, 3.7 MB
rna_celline.tsv.zip
TSV-file, 6.2 MB

5 **RNA isoform data**
RNA levels in 64 cell lines and 37 tissues based on RNA-seq. The tab-separated file includes Ensembl gene identifier ("Gene"), Ensembl transcript identifier ("Transcript"), analysed sample ("Sample") and transcript per million ("TPM"). The data is based on The Human Protein Atlas version 18.1 and Ensembl version 88.38.

transcript_rna_tissue.tsv.zip
TSV-file, 73.7 MB
transcript_rna_celline.tsv.zip
TSV-file, 51.9 MB

6 **Data from the Human Protein Atlas in tab-separated format**
This file contains a subset of the data in the Human Protein Atlas version 18.1 corresponding to the data seen in the search result. This data can also be downloaded for a resulting gene set when using the search function (via the TSV link on the result page).

proteinatlas.tsv.zip
TSV-file (gzip compressed), 1.5 MB

7 **Data from the Human Protein Atlas in XML format**
The XML file contains most of the data in the Human Protein Atlas version 18.1, including protein expression data (in normal and tumor tissues and in cell lines), antigen sequences, Western blot data for antibodies, protein array data for antibodies, RNA-seq data, external references such as UniProt identifiers, and more. The data is based on Ensembl version 88.38. The file structure is presented in the XSD-schema. This data can also be downloaded for a resulting gene set when using the search function (via the xml link on the result page).
The XML file presented here is compressed with gzip due to its size. It can be uncompressed with an archive program like 7-zip.

proteinatlas.xml.gz
XML-file (gzip compressed), 261.8 MB

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

advanced search

Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Login

ARRAYEXPRESS Browse ArrayExpress B-2836

E-MTAB-2836 - RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues

Status	*Submitted on 4 May 2014, released on 14 January 2015, last updated on 23 November 2018*
Organism	Homo sapiens
Samples (122)	Click for detailed sample information and links to data
Protocols (7)	Click for detailed protocol information
Description	RNA-seq was performed of tissue samples from 122 human individuals representing 32 different tissues in order to study the human tissue transcriptome. This submission contains 27 new samples and the data from E-MTAB-1733.
Experiment types	RNA-seq of coding RNA, organism part comparison design
Contact	Björn M Hallström <bjorn.hallstrom@gmail.com>
Citations	Proteomics. Tissue-based map of the human proteome. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. *Science* 347(6220) (2015), PMID:5613900

Click for detailed sample information and links to data

Can click and download data to desktop….but wait, there’s a better way!

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

advanced search

Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Login

ARRAYEXPRESS / BROWSE / E-MTAB-2836 / SAMPLES AND DATA

E-MTAB-2836 - RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues

Display full sample-data table

Export table in Tab-delimited format

Page 1 2 3 4 5 6 .. 16

Showing 1 - 25 of 400 rows

Page size 25 50 100 250 500

	Sample Attributes				Variables	Assay	Links to Data	
Source Name	organism part	organism	sex	developmental stage	organism part	Assay Name	ENA	FASTQ
adrenal_4a	adrenal gland	Homo sapiens		adult	adrenal gland	1_130213_AH07R5ADXX_P282_102B_index25		
adrenal_4a	adrenal gland	Homo sapiens		adult	adrenal gland	1_130213_AH07R5ADXX_P282_102B_index25		
adrenal_4a	adrenal gland	Homo sapiens		adult	adrenal gland	2_130213_AH07R5ADXX_P282_102B_index25		
adrenal_4a	adrenal gland	Homo sapiens		adult	adrenal gland	2_130213_AH07R5ADXX_P282_102B_index25		
adrenal_4c	adrenal gland	Homo sapiens		adult	adrenal gland	1_130213_AH07R5ADXX_P282_104B_index1		
adrenal_4c	adrenal gland	Homo sapiens		adult	adrenal gland	1_130213_AH07R5ADXX_P282_104B_index1		
adrenal_4c	adrenal gland	Homo sapiens		adult	adrenal gland	2_130213_AH07R5ADXX_P282_104B_index1		
adrenal_4c	adrenal gland	Homo sapiens		adult	adrenal gland	2_130213_AH07R5ADXX_P282_104B_index1		

Let’s explore this data…
click on Display full-sample data table

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

advanced search

Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Login

ARRAYEXPRESS / BROWSE / E-MTAB-2836 / SAMPLES AND DATA

E-MTAB-2836 - RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues

Display summary

Export table in Tab-delimited format

Page 1 2 3 4 5 6 .. 16 Showing 1 - 25 of 400 rows Page size 25 50 100 250 500

Source Name	JN]	Comment[FASTQ_URI]	Comment[MD5]	Comment[SPOT_LENGTH]	Cor
adrenal_4a		ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315335/ERR315335_1.fastq.gz	129127427adf13499dd774c6db307d78	102	209
adrenal_4a		ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315335/ERR315335_2.fastq.gz	b6ea9b491ca0a9164146f4f9d96b6483	102	209
adrenal_4a		ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315452/ERR315452_1.fastq.gz	f3375fcef92155ced216c8aea16dc0df	102	209
adrenal_4a		ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315452/ERR315452_2.fastq.gz	d8f0f1061349329b2f4de0564b773975	102	209

Scroll across to see the details for these files, they are paired RNA-seq data

See the ftp address ->ftp://ftp.sra.ei.ac.uk/

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

advanced search

Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Login

ARRAYEXPRESS / BROWSE / E-MTAB-2836 / SAMPLES AND DATA

E-MTAB-2836 - RNA-seq of coding RNA from tissue samples of 122 human individuals representing 32 different tissues

Display summary

Export table in Tab-delimited format

Page 1 2 3 4 5 6 .. 16

Showing 1 - 25 of 400 rows

Page size 25 50 100 250 500

Source Name		Comment[ENA_RUN]	Comment[FASTQ_URI]	Comment[MD5]	Comment[S
adrenal_4a	z	ERR315335	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315335/ERR315335_1.fastq.gz	129127427adf13499dd774c6db307d78	102
adrenal_4a	z	ERR315335	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315335/ERR315335_2.fastq.gz	b6ea9b491ca0a9164146f4f9d96b6483	102
adrenal_4a	z	ERR315452	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315452/ERR315452_1.fastq.gz	f3375fcef92155ced216c8aea16dc0df	102
adrenal_4a	z	ERR315452	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR315/ERR315452/ERR315452_2.fastq.gz	d8f0f1061349329b2f4de0564b773975	102

How to download files from Human Protein Atlas

- Follow the directions on the handout
- Use the ”wget” command from Biowulf

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/dir/file.fastq.gz
```

End of Part 2 – time for a break

Part 3

Scientific analyses and databases on Biowulf

Now that you've uploaded your data, you'll need an "ssh" connection to work at the command line

- Use Terminal (Mac) or PuTTY (PC) to log into Biowulf
- Look at the data you've just uploaded using unix commands learned earlier (ls, cd, less)

Analyses of fastq sequencing files

- NGS sequence results are returned to you in FASTQ format
- FastQC/MultiQC – quality check of FASTQ files
- Trimmomatic – remove adapters at ends
- Bowtie2 – align to genome
- IGV – Integrated Genome Viewer

A word about sequence formats

- FASTA – commonly used text format for downstream analysis such as sequence similarity searches
- FASTQ – output format from NGS technologies with quality scores
- SAM – tab-delimited text format with alignment data
- BAM – compressed, binary version of SAM

Sequence formats - FASTA

FASTA – has a header line that begins with “>” and a data line containing the sequence

>this_is_a_fasta_header_it_can_say_anything_here

ATCTAGGACCTGAAGACGGGACCTTTTTACGACTAC

> sequence 1 can have spaces in the header line

ATCTATGAGATAGACTATATACTAGACGATACGATGACGATAGAACATCTATGA

>can_also_be_a_protein_sequence

MPYWTGAMYAVPWTERHGPNCTAAVPMYGATRE

Sequence formats - FASTQ

FASTQ – contains both sequence data and quality score data

@whatever_the_name_of_the_sequence_is
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCA
+ whatever_the_name_of_the_sequence_is
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

FASTQ scores and ASCII codes

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			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

drive.com

SAM files are tab-delimited text files

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Wikipedia

Don’t open BAM files at the command line

- Binary files like BAM are machine-readable, not human-readable, there is no reason to open them
- But what if you wanted to know how big they were, when they were generated? What commands would you use?
- ls -alt

We are going to run programs on Biowulf

- Each program is known as a “module” and there are over 600 of them on Biowulf
- To use a module, you must load the module
- You need to load modules every time you log into Biowulf

What modules are available to us?

- module avail – to see list of available modules on Biowulf
- module spider ”filename” – to find files using part of name
- module load – adds location of program to your path
- ECHO $PATH – to see that module has loaded
- module list – to see what modules you have loaded
- module unload or module purge – not really necessary, modules will unload automatically when you logout from Biowulf
- Always need to reload modules when you log back in

module avail

- ---------- /usr/local/lmod/modulefiles ----------
- 3DSlicer/4.8.1
- ACFS/20180316
- AMOS/3.1.0
- ANNOgesic/0.7.18
- ANTs/2.2.0
- ATLAS/3.10.2
- Accurity/20180724
- AdmixTools/4.1
- Autodock/4.2.6
- AutodockVina/1.1.2
- Azimuth/2.0
- BEAST/1.8.4-2.1.2
- BEAST/1.10.0
- BEAST/2.4.7 (D)
- BOLT-LMM/v2.3
- BOLT-LMM/v2.3.2 (D)
- BRASS/6.1.2
- Beagle/4.1_08Jun17
- Bsoft/2.0.2
- CCP4/7.0.050
- CCP4/7.0.051

module spider blast

- ----------
- blast:
- ----------
- Versions:
- blast/2.2.26
- blast/2.2.30+
- blast/2.5.0+
- blast/2.7.1+
- blast/2.8.0+alpha
- Other possible modules matches:
- igblast rmblast samblaster
-
- ----------
- To find other possible module matches execute:
-
- $ module -r spider '.*blast.*'
-
- ----------
- For detailed information about a specific "blast" module (including how to load the modules) use the module's full name.
- For example:
-
- $ module spider blast/2.8.0+alpha
- ----------

Running jobs on Biowulf…correctly

1. Interactive – use the “sinteractive” command to allocate resources for an interactive job

2. Batch – need to create a batch input file and submit job using the sbatch command (Slurm – job scheduler, resource manager)

3. Swarm- create a swarmfile (myfile.swarm) and submit using the “swarm” command

Do not work on Biowulf login mode!

- If you run computationally intensive jobs on the Biowulf login node, you may lose your account
- There are instructions for running interactive jobs, creating batch files, and swarming on the Biowulf web page
- We will do one of each (sinteractive, batch, swarm) so you can get some practice

To unzip compressed files

At the command line, type:

```
$sinteractive
```

(wait)

```
$gunzip filename1.fastq.gz
```

When that has finished,

```
$gunzip filename2.fastq.gz
```

This is the Biowulf help page for fastqc

Interactive job

Interactive jobs should be used for debugging, graphics, or applications that cannot be run as batch jobs.

Allocate an interactive session and run the program. Sample session:

```
[user@biowulf]$ sinteractive
salloc.exe: Pending job allocation 46116226
salloc.exe: job 46116226 queued and waiting for resources
salloc.exe: job 46116226 has been allocated resources
salloc.exe: Granted job allocation 46116226
salloc.exe: Waiting for resource configuration
salloc.exe: Nodes cn3144 are ready for job

[user@cn3144 ~]$ module load fastqc
[user@cn3144 ~]$ fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN

[user@cn3144 ~]$ exit
salloc.exe: Relinquishing job allocation 46116226
[user@biowulf ~]$
```

At the command line, type “sinteractive”, then module load, then run your program.

Quality check with FASTQC - sinteractive

Use the “module load” command

```
[username@biowulf dir_name] $ sinteractive
[username@biowulf dir_name] $ module load fastqc
[+] Loading fastqc 0.11.6
[username@biowulf dir_name] $ fastqc filename1.fastq
[username@biowulf dir_name] $ fastqc filename2.fastq
```

Generates html report – but you can’t view an html report on a unix machine! What do we do? Transfer the file **from** Biowulf **to** your local machine (laptop). Use FileZilla, WinSCP (PC) or scp command line (Mac)

Checking on job status

$ sjobs

```
[stonelakeak@biowulf jobs]$ sjobs
User          JobId      JobName  Part  St  Reason  Runtime  Walltime  Nodes  CPUs
Memory      Dependency Nodelist
=================================================================================
==================================
stonelakeak  27646978  star.sh  norm  PD  ---        0:00   2:00:00      1    12
35GB/node
=================================================================================
==================================
cpus running = 0
cpus queued = 12
jobs running = 0
jobs queued = 1
[stonelakeak@biowulf jobs]$
```

If you want to run fastqc as a batch, you can create a batch file

Batch job

Most jobs should be run as batch jobs.

Create a batch input file (e.g. fastqc.sh). For example:

```
#!/bin/bash
set -e
module load fastqc
fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN
```

Submit this job using the Slurm sbatch command.

```
sbatch --mem=10g fastqc.sh
```

Use the nano editor to create the file on Biowulf, then use the sbatch command to run it.

Use the nano editor, create file fastq.sh

```
#!/bin/bash
set –e
module load fastq
fastqc –o /data/username/output –f fastq  filename1.fastq filename2.fastq
```

(Submit fastqc.sh using this command)

```
sbatch –mem=10g fastqc.sh
```

To run fastqc as a swarm on Biowulf

Swarm of Jobs

A swarm of jobs is an easy way to submit a set of independent commands requiring identical resources.

Create a swarmfile (e.g. fastqc.swarm). For example:

```
cd dir1;fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN
cd dir2;fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN
cd dir3;fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN
cd dir4;fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN
cd dir5;fastqc -o output_dir [-f fastq|bam|sam] -c contaminant_file seqfile1 .. seqfileN
```

Submit this job using the swarm command.

```
swarm -f fastqc.swarm -g 10 --module fastqc
```

where

`-g #`	Number of Gigabytes of memory required for each process (1 line in the swarm command file)
`--module fastqc`	Loads the fastqc module for each subjob in the swarm

Fastqc generates an html file, but we can’t view html on an “ssh” connection

- Globus (easiest once you’ve got it set up, meant for big files but you can transfer any size files)
- Drag and drop interfaces (WinSCP, FileZilla*) – PC or Mac
- Use scp at the Mac command line
- Go ahead now and use any method to move html file from Biowulf to your local machine

Using the command line to download your file from Biowulf to Mac

For example…
Type this on your local Mac, not Biowulf!

```
scp  username@biowulf.nih.gov:/data/username/dirname/filename.html .
```

where username is your username,
dirname is the path to your file
filename is the name of the file
and there is a dot “.” at the end of the command

Mistakes you will make when transferring files from Biowulf to your Mac laptop/desktop

- You will forget to type the command in a terminal window **<u>on your machine</u>**
- You will forget to type the dot "." at the end of the command (the dot means put the file "here")
- You will not type the path correctly to the file.
- You will have a typo in the name of the file

After you do it correctly, be sure to celebrate!

Let’s look at the FastQC html report

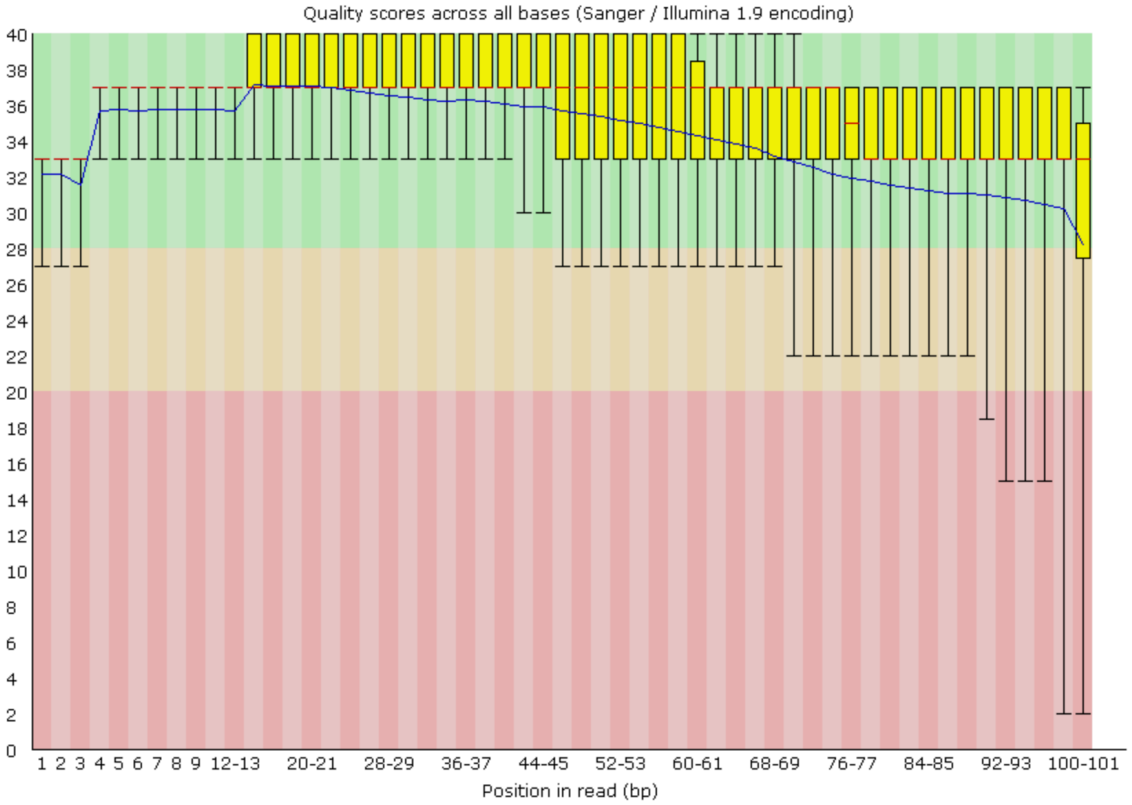
- Basic statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence length distribution
- Duplicate sequences
- Overrepresented sequences
- Adapter content
- Kmer content
- Per tile sequence quality

Basic Statistics

Measure	Value
Filename	lung_4a_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8782257
Sequences flagged as poor quality	0
Sequence length	101
%GC	50

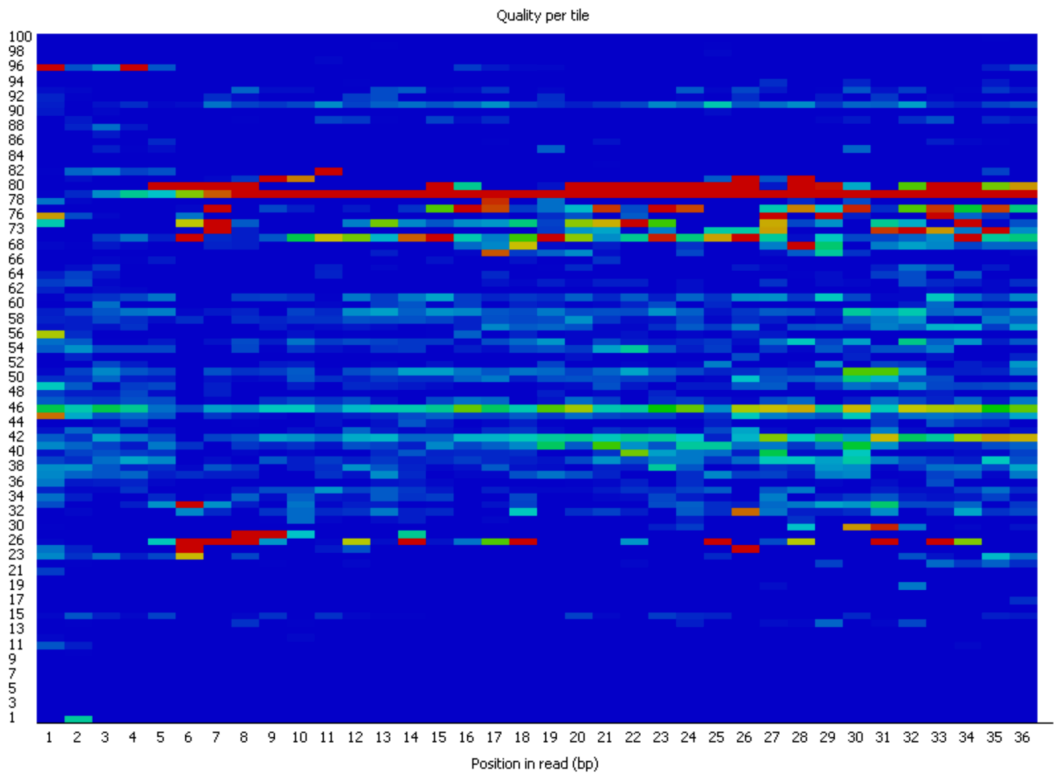
bioinformatics.babraham.ac.uk

FastQC per base sequence quality



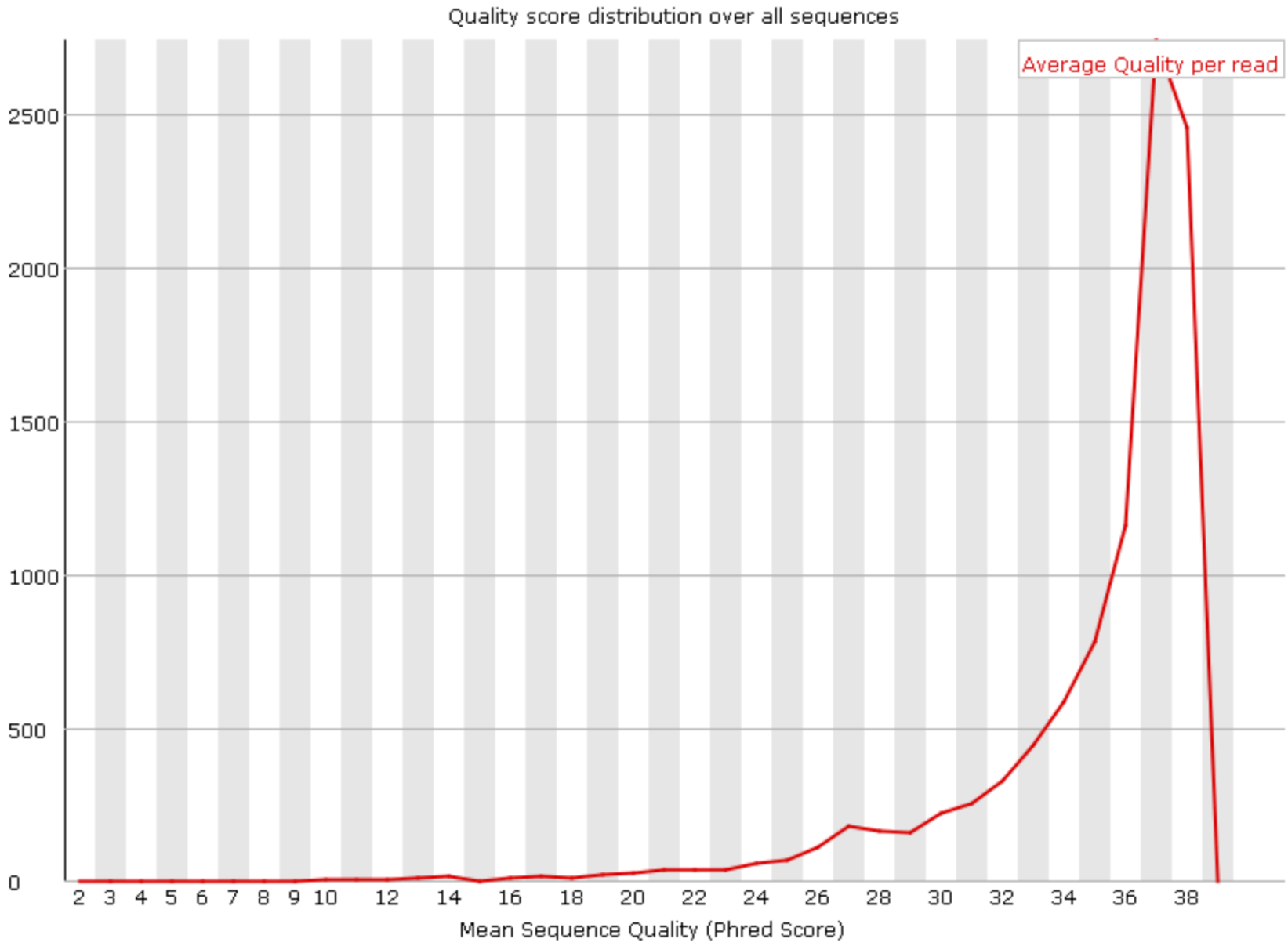
- Box whisker plot for each position
- Y-axis shows quality scores
- The higher the base the better the call
- Green – good quality, orange – reasonable,
  red – poor quality
- Quality typically degrades at the end of the read
- Red line is median
- Yellow box is inter-quartile range (25% - 75%)
- Upper and lower whiskers represent 10% and 90% points
- Blue line is the mean

Per tile sequence quality

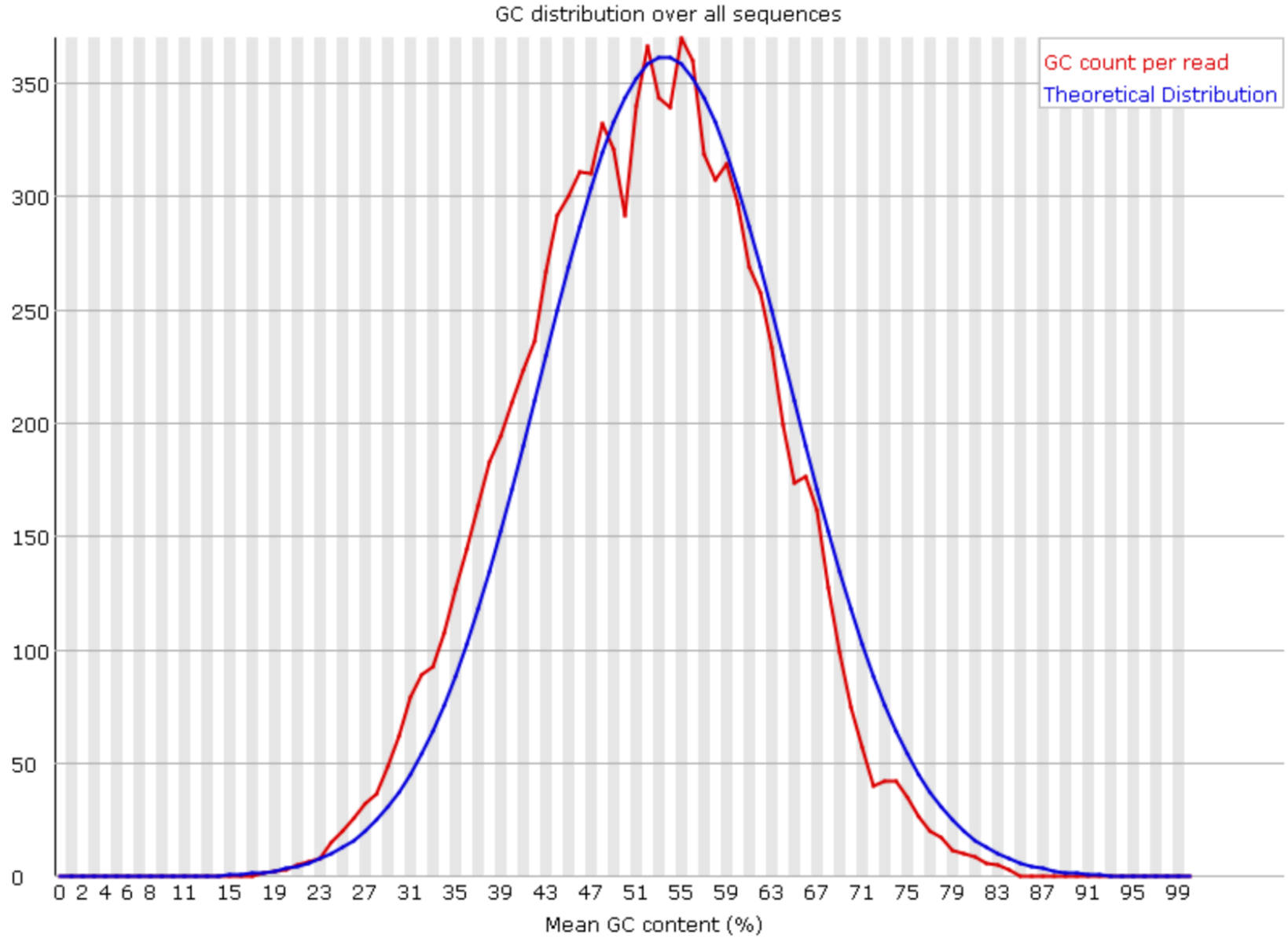


- This graph only appears if you're using an Illumina library with original sequence identifiers
- Shows the deviation from average quality for each tile
- A good plot should be blue all over
- Warnings or errors indicate issues with the flowcell

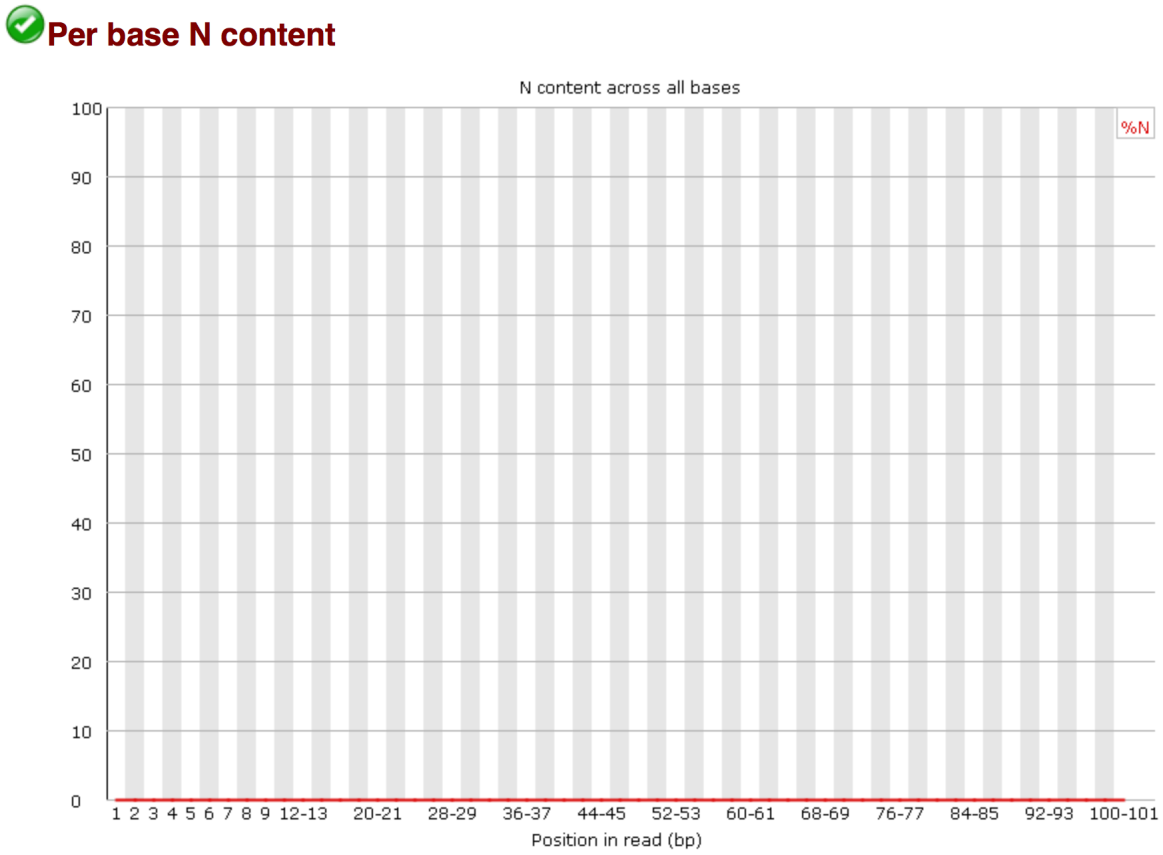
Per sequence quality scores



Per sequence GC content

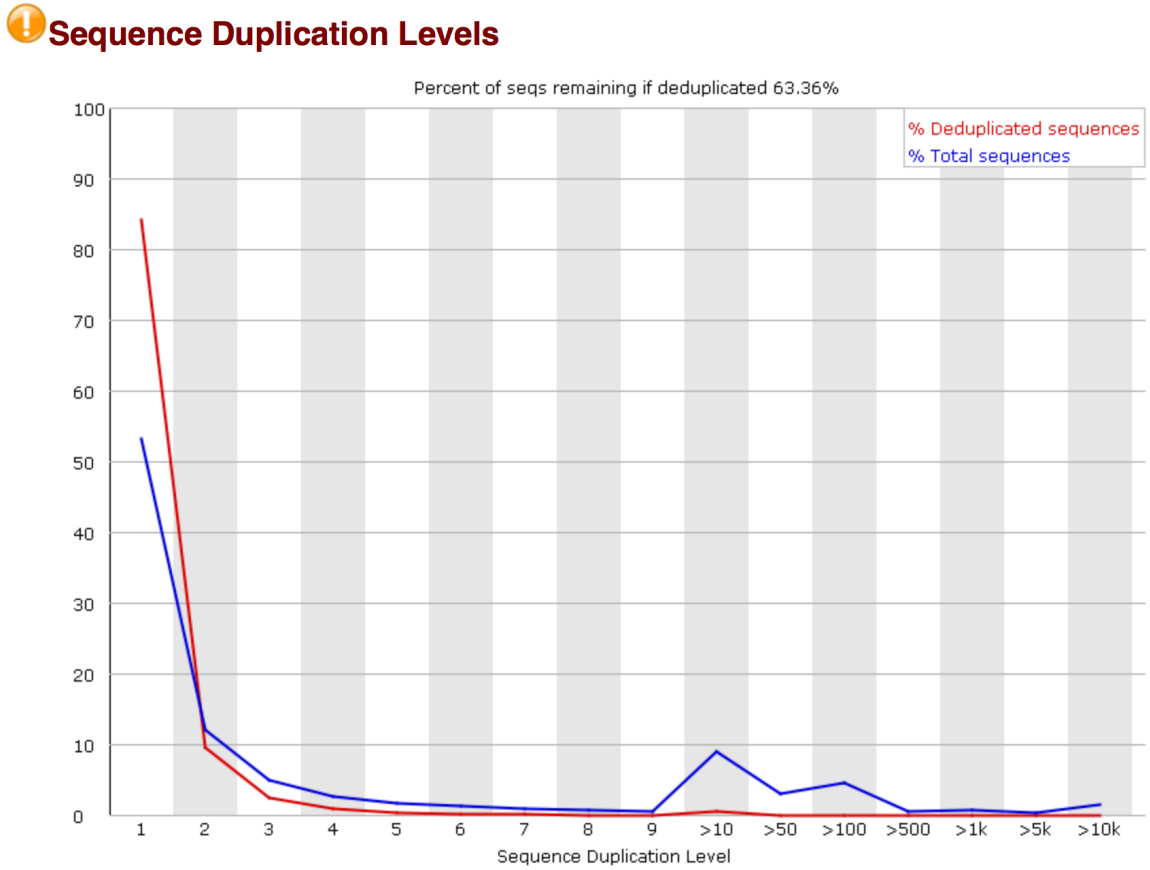


Per base N content



- If the sequencer it unable to identify the base, it will use “N”
- See the percentage of “N” base calls at each read position

Sequence duplication levels



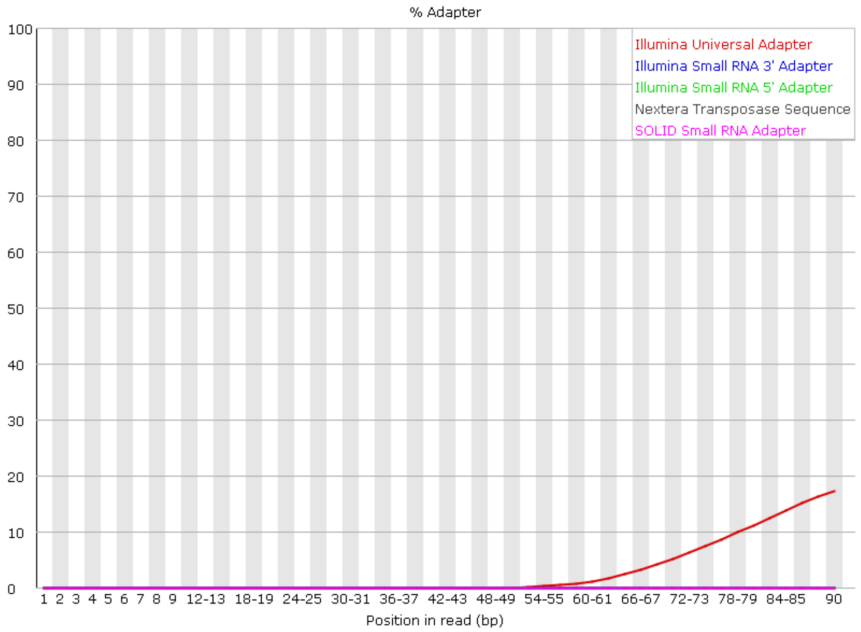
- A diverse library will have a low number of duplicate sequences
- High level of duplication can indicate an enrichment bias (PCR)
- Only sequences first 100,000 sequences in each file
- Blue line show duplication levels for full sequence set
- Red line indicates proportions of deduplicated sequences
- Most sequences should be in far left of plot in both red and blue
- Low complexity contaminants will produce spikes in the plot

Overrepresented/adapter sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGC	149221	1.6991190305635555	TruSeq Adapter, Index 6 (100% over 50bp)

Adapter Content



- Indicates overrepresented sequences
- Should these sequences be trimmed before continuing analysis?

How could you summarize FastQC results from multiple files?

- unzip .zip files
- less summary.txt
- mkdir summary
- cat */summary.txt > /data/username/summary/fastqc_summaries.txt

less fastqc_summaries.txt

```
PASS    Basic Statistics        short_lung_4a_R1.fastq
PASS    Per base sequence quality       short_lung_4a_R1.fastq
PASS    Per tile sequence quality       short_lung_4a_R1.fastq
PASS    Per sequence quality scores     short_lung_4a_R1.fastq
FAIL    Per base sequence content       short_lung_4a_R1.fastq
PASS    Per sequence GC content short_lung_4a_R1.fastq
PASS    Per base N content      short_lung_4a_R1.fastq
PASS    Sequence Length Distribution    short_lung_4a_R1.fastq
PASS    Sequence Duplication Levels     short_lung_4a_R1.fastq
FAIL    Overrepresented sequences       short_lung_4a_R1.fastq
FAIL    Adapter Content short_lung_4a_R1.fastq
PASS    Basic Statistics        short_lung_4a_R2.fastq
PASS    Per base sequence quality       short_lung_4a_R2.fastq
PASS    Per tile sequence quality       short_lung_4a_R2.fastq
PASS    Per sequence quality scores     short_lung_4a_R2.fastq
FAIL    Per base sequence content       short_lung_4a_R2.fastq
PASS    Per sequence GC content short_lung_4a_R2.fastq
PASS    Per base N content      short_lung_4a_R2.fastq
PASS    Sequence Length Distribution    short_lung_4a_R2.fastq
PASS    Sequence Duplication Levels     short_lung_4a_R2.fastq
PASS    Overrepresented sequences       short_lung_4a_R2.fastq
FAIL    Adapter Content short_lung_4a_R2.fastq
```

A program to trim adapters -> Trimmomatic

- Bolger et al.,Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (2014)
- For Illumina paired-end and single ended data

The current trimming steps are:

- ILLUMINACLIP: 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
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

Running Trimmomatic as a batch job

Batch job

Most jobs should be run as batch jobs.

Create a batch input file (e.g. trimmomatic.sh). For example:

```
#!/bin/bash

ml trimmomatic || exit 1
java -Djava.io.tmpdir=. -jar $TRIMMOJAR PE -phred33 -threads $SLURM_CPUS_PER_TASK \
    SRR292678_1.fastq.gz SRR292678_2.fastq.gz \
    output_forward_paired.fq.gz output_forward_unpaired.fq.gz \
    output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz \
    ILLUMINACLIP:/usr/local/apps/trimmomatic/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
\
    SLIDINGWINDOW:4:15 MINLEN:36
```

Submit this job using the Slurm sbatch command.

```
sbatch -c 2 --mem=6g trimmomatic.sh
```

We are not going to run Trimmomatic today

- We are going to align fastq files directly to the (human) genome
- using the aligner “bowtie2” (creates .bam files)
- “samtools” to create bam files index
- Bring the .bam and .bam.bai files to local machine
- View .bam file with IGV (Integrative Genomics Viewer)

Running Bowtie2 as a batch job

- Go to /scratch/btepclass
- Find the file “bowtie2.sh” and use the “cp” command to bring it into your student directory

```
$cp bowtie2.sh /data/username
```

Running bowtie2 as a batch job

Batch job

Most jobs should be run as batch jobs.

Create a batch input file (e.g. bowtie2.sh), which uses the input file 'bowtie2.in'. For example:

```
#!/bin/bash
module load bowtie/2 || exit 1
module load samtools || exit 1
export BOWTIE2_INDEXES=/fdb/igenomes/Mus_musculus/UCSC/mm9/Sequence/Bowtie2Index/
bowtie2 --phred64 -x genome --threads=$(( SLURM_CPUS_PER_TASK - 4 )) \
    --no-unal --end-to-end --sensitive \
    -U $BOWTIE_TEST_DATA/ENCFF001KPB.fastq.gz \
  | samtools view -q30 -u - \
  | samtools sort -O BAM -@3 -T /lscratch/$SLURM_JOB_ID/ENCFF001KPB -m 2g -o ENCFF001KPB.bam
```

Submit this job using the Slurm sbatch command.

```
sbatch --cpus-per-task=10 --mem=14g --gres=lscratch:10 bowtie2.sh
```

Here is the command line to run bowtie2

```
$ sbatch --cpus-per-task=10 --mem=14g --gres=lscratch:10 bowtie2.sh
```

To create index file (bai) from bam, use samtools

```
$module load samtools
$samtools index filename.bam
```

”$” designates the command line, do not type the “$”

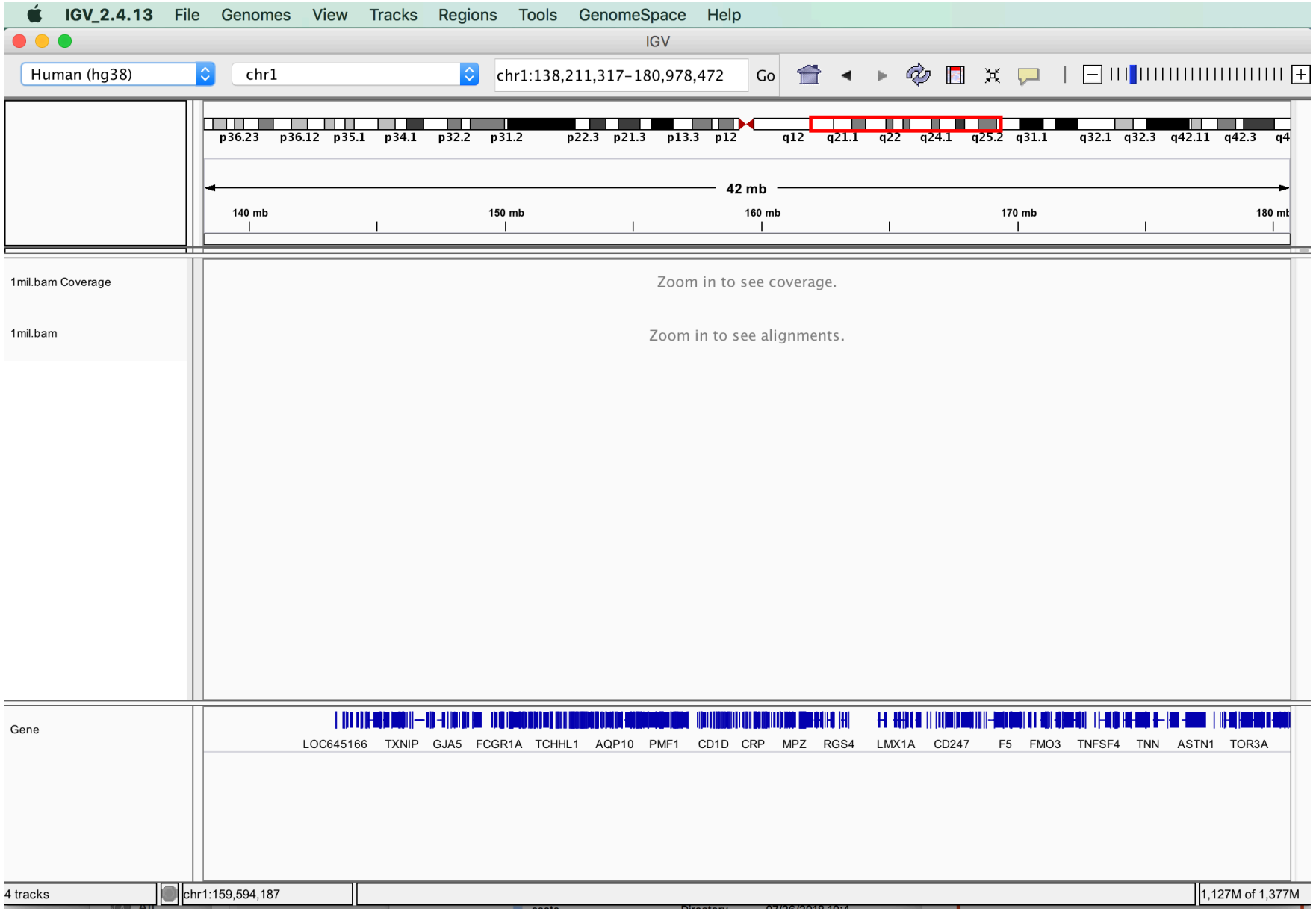
Bring both .bam and .bam.bai files to local

Use globus, WinSCP, FileZilla, or scp at the Mac command line

Open IGV (Integrative Genomics Viewer)

File -> load file.bam

IGV


IGV_2.4.13
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
IGV
Human (hg38)
chr1
chr1:138,211,317-180,978,472
Go
p36.23
p36.12
p35.1
p34.1
p32.2
p31.2
p22.3
p21.3
p13.3
p12
q12
q21.1
q22
q24.1
q25.2
q31.1
q32.1
q32.3
q42.11
q42.3
42 mb
140 mb
150 mb
160 mb
170 mb
1mil.bam Coverage
Zoom in to see coverage.
1mil.bam
Zoom in to see alignments.
Gene
LOC645166
TXNIP
GJA5
FCGR1A
TCHHL1
AQP10
PMF1
CD1D
CRP
MPZ
RGS4
LMX1A
CD247
F5
FMO3
TNFSF4
TNN
ASTN1
TOR3A
4 tracks
chr1:159,594,187
1,127M of 1,377M

You did it!

To summarize:

- Basic unix commands (ls, pwd, cd, less, nano, mkdir, rm, rmdir)
- Unix directory structure (files and folders)
- Logging in to Biowulf by SSH
- Moving files from laptop to Biowulf and back (globus, scp, mount drive)
- Running command-line programs on Biowulf (fastqc, bowtie2) in interactive, batch and swarm modes

Anyone wishing to contribute to the calendar should contact BTEP staff.

NIH Bioinformatics Calendar (Managed by BTEP) **Monday, May 6, 2019** **Public Access**

Options | Year | Month | Work Month | Week | Work Week | Day | Upcoming | Matrix(C) | 05.06.2019 | ← | Print | Upcoming | Search | Help | Log In

April 2019 - July 2019

wk	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
19		10:00AM All of Us Research Program Symposium: From Data to Discoveries: Creating a Research Program for All of Us 1:00PM NIH Library: Introduction to R Data Types Webinar	9:00AM NIH HPC: Introduction to Linux 10:00AM CBIIT: Omics Data Analysis in Partek 11:00AM NIH HPC: Neurogenetics on Biowulf: From GWAS to Machine Learning	9:00AM Woman-Led Biodata Science Hackathon 9:00AM NIH HPC: Introduction to Linux 1:00PM NIH Library: Hands-on RNA-Seq Data Analysis in Partek Flow	9:00AM Woman-Led Biodata Science Hackathon 12:00PM BUF: Object-Oriented Programming	ChIP-Seq Data Analysis: Probing DNA-Protein Interactions 9:00AM Woman-Led Biodata Science Hackathon	
20	12	13	14 9:00AM··· NIH HPC: Bash Shell Scripting for Helix and Biowulf	15 9:00AM NLM Reproducibility in Bioinformatics Workshop 9:00AM NIH HPC: Bash Shell Scripting for Helix and Biowulf	16 9:00AM NLM Reproducibility in Bioinformatics Workshop 10:00AM CBIIT: MacVector 17.0 Training Workshop 1:00PM NIH Library: DNASTAR Lasergene Demo and Training Workshop	17 9:00AM NLM Reproducibility in Bioinformatics Workshop 11:00AM CBIIT: MacVector 17.0 Training Workshop	18
21	19	20 9:00AM NICHD Bioinformatics and Scientific Programming Core: Reproducible workflows with Snakemake 9:00AM Cancer Data Science 101, Part II	21 BTEP, RNA-Seq Week: Graphical Excellence and Integrity: How to make your data sing! 9:00AM NICHD Bioinformatics and Scientific Programming Core: Reproducible workflows with Snakemake	22 10:00AM NIH Library: Writing Custom Functions in R	23 8:45AM NIA: Single-Cell Analysis in Aging and Disease	24 BTEP, RNA-Seq Week: Hands-on drop-in session on RNA-Seq 10:00AM BTEP Drop-in session: RNA-Seq analysis tools	25

RSS

powered by LuxSoft

NCI CCR Bioinformatics Training and Education Program

Website: btep.ccr.cancer.gov

OSTR: ostr.ccr.cancer.gov/bioinformatics

Email: ncibtep@nih.gov

Amy Stonelake

Peter Fitzgerald

Carl McIntosh


End of Presentation
ANY QUESTIONS?
memecreator.org