Welcome to Bioinformatics Crash Course! This document will serve as a guide for the workshop. The mission of The Core is to provide users with the bioinformatic services, support, and education necessary to advance their research program. This workshop will focus on teaching you the necessary tools and techniques so that you can turn raw data into meaningful results.

## Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-10</td>
<td>Introduction and Setup</td>
<td>NGS Introduction</td>
<td>Assembly</td>
<td>Genome Annotation</td>
<td>AWS</td>
</tr>
<tr>
<td>10-11</td>
<td>Linux Commands</td>
<td>RNAseq – Steve Mount</td>
<td>Assembly</td>
<td>Genome Annotation</td>
<td>Variant Detection</td>
</tr>
<tr>
<td>11-12</td>
<td>Linux Commands/HPCC</td>
<td>Sailfish – Steve Mount</td>
<td>Assembly</td>
<td>PacBio – Matt Conte</td>
<td>Variant Detection</td>
</tr>
<tr>
<td>12-1</td>
<td>Lunch</td>
<td>Lunch</td>
<td>Lunch</td>
<td>Lunch</td>
<td>Lunch</td>
</tr>
<tr>
<td>1-2</td>
<td>Linux/Python</td>
<td>Tuxedo Analysis</td>
<td>Metagenomics – Mihai Pop</td>
<td>PacBio – Matt Conte</td>
<td>AWS</td>
</tr>
<tr>
<td>2-3</td>
<td>Python</td>
<td>Tuxedo Analysis</td>
<td>Metagenomics – Mihai Pop</td>
<td>PacBio – Matt Conte</td>
<td>AWS</td>
</tr>
<tr>
<td>3-4</td>
<td>Python</td>
<td>Tuxedo Analysis</td>
<td>Metagenomics – Mihai Pop</td>
<td>Genome A &amp; A Review</td>
<td>AWS</td>
</tr>
</tbody>
</table>

Some of the examples and tasks in this Manual have been heavily influenced by:

Unix and Perl for Biologists: http://korflab.ucdavis.edu/Unix_and_Perl/
Basic UNIX for Biologists: https://www.bioinformatics.purdue.edu/discoverypark/cyber/bioinformatics/services/workshops.php
The Terminal

This is the common name for the application that gives you text-based access to the operating system of the computer. Basically, it allows you to type input into the computer so that you can receive output from the programs you call. On Unix based machines there is always a ‘terminal’ program. On the Mac you’re using this program is called ‘Terminal’.

1) Using the spotlight feature type the word terminal.

![Spotlight search for Terminal](image1)

2) Click the application.

3) You will see something like this on your monitor.

![Terminal window](image2)

4) Your window will appear different, as the text you see will be different for each of you since you’re using different computers.
5) Some brief tips before we move further:
   a. You will need to frequently resize this window as we do the hands-on activities. This is done as with any other window.
   b. You can change the size of the text by holding ‘command’ and hitting ‘+’ or ‘-’ keys.
   c. You can have multiple terminal windows open at the same time.
   d. You can also have multiple tabs open in each terminal window.
   e. Everything you type in the terminal is case sensitive. ‘grep –F’ is not the same as ‘grep -f’. This will be a very important thing to remember when using the terminal and creating files or folders.

6) You will see that your terminal line will look different then this example.

   Ians-MacBook-Pro:~ UMD-Bioinformatics$

7) The ‘Ians-MacBook-Pro:~ UMD-Bioinformatics$’ text is the Unix command prompt. It contains the username (UMD-Bioinformatics), the name of the machine that the user is working on (Ians-MacBook-Pro) and the name of the current directory (~). Note the command prompt may look different on different machines. In this example the $ represents the end of the prompt.

8) What is your username?

9) What is the name of your machine?

10) What is the current directory?

11) This application is going to be your friend for the entire week. Get used to making new windows (command + n), new tabs (command + t), and resizing this window.
Logging into DeepThought2

SSH

Logging into DeepThought2, or any remote computer, requires a secure connection. This is typically done using a Secure Shell (SSH) protocol. Linux and Mac computers have SSH clients built into their systems. There are options on Windows platforms to install SSH clients (PuTTY).

NOTE that anytime in this document you see <username> replace it with your login username.

1) To login to DeepThought2 you will need your temporary account username and password that I provided to you.
2) On the terminal type `ssh -X <username>@login.deepthought2.umd.edu` and hit enter. You will see the following screen prompting you for a password.

3) Enter your password and hit enter. NOTE the cursor will not move as you type.
4) You should now see the following on your terminal.
Welcome to DeepThought2! You are now logged into the University’s newest High Performance Computer Cluster (HPCC). Specifically, you are going to be using the Bioinformatics Condo. This is hardware built into DT2 that is designed to deal with biological data. UMD affiliates can get access to this resource through the Bioinformatics Core for doing your own projects and data analysis. As members of this workshop you have been given temporary access to this hardware. During this workshop we will be utilizing this resource for all of our bioinformatic needs.

Before we get you to that we need to get everyone up to speed with using Linux/Unix, HPCC, and some basic programming tools. However you’re going to need some data and files to work with. Type the following command (we will explain this later).

```
nohup rsync -av /lustre/imasner/Workshop_Data /lustre/<username>/Workshop_Data &
```
Unix Tutorial

Where am I?

As you will learn everything in Unix is relative to where you are in the file system. Therefore, knowing where you are before launching a command is valuable information. Luckily, there are built in commands for this type of information. Understanding the location of files will be a key part of success.

1) To find out where you are in the file system type `pwd` in the terminal window.
2) This will return your current working directory (print working directory)

```
login-1:~: pwd
/homes/imisner
login-1:~:
```

3) As you can see above the working directory is: `/homes/imisner`
4) Write down your working directory:

5) The present directory is represented as . (dot) and the parent directory is represented as .. (dot dot)

File Structure

Unix files are arranged in a hierarchical structure, or a directory tree. From the root directory (/) there are many subdirectories. Each subdirectory can contain files or other subdirectories, etc, etc. Whenever you’re using the terminal you will always be ‘in’ a directory. The default behavior of opening a terminal window, or logging into a remote computer, will place you in your ‘home’ directory. This is true when we logged into DT2. The home directory contains files and directories that only you can modify, we will get to permissions later.
To see what files, or directories, you have in your home directory we will use the `ls` command.

1) Type `ls` and hit enter.

```
login-2:~$ ls
Mail glue_home priv welcome
```

2) You should see the list of files and directories in your current folder.
3) After the `ls` command finishes it produces a new command prompt that is ready for your next command.
4) The `ls` command can be used to list the contents of any directory not necessarily just the one you are currently in.
5) Type `ls /local/`

```
login-2:~$ ls /local/
X X1105 clamav fortune gain gnome lib mdgs request snmp windows xdg
X11 x11e6 etc freedesktop.org ganglia kde mail news scripts ssl www xmes
login-2:~$
```

### Changing Directories

To move between directories (folders) we use the `cd` (change directory) command. We are currently in our home directory on DT2. Lets move to `/lustre/<username>/Workshop_Data/Unix`
The `cd` command uses the format:

```bash
$ cd DIRECTORY
```

1) Type `cd /lustre/<username>/Workshop_Data/Unix`
2) Type `ls`
3) Type `pwd`

```
$ cd /lustre/imisner/Workshop_Data/Unix/
$ ls
Code  Data
$ pwd
/lustre/imisner/Workshop_Data/Unix
```

4) You can see that using the `cd` command moved us to a different directory.
5) By typing `ls` we can see that there is different stuff in this directory. (Nothing in your case)
6) Finally, using `pwd` shows us what directory we are not located in.

7) We could have done the previous example in separate steps
8) Type `cd /lustre`
9) Type `cd <username>`
10) Type `cd Workshop_Data`
11) Type `cd Unix`
12) Note that we needed to type `/lustre` but not `/<username>` When using a `DIRECTORY` you are specifying a directory that is below the root directory. Without the leading `/` the system will look below the current directory.
13) Type `cd lustre`
14) Type `cd /lustre`
15) What happened with the first command?

16) You will frequently need to move up a level to a parent directory. Remember that two dots `..` are used to represent the parent directory. Every directory has a parent except the root level.
17) Type `cd ..`
18) Type `pwd`
19) You can move multiple levels at the same time
20) Type `cd /lustre/<username>/Workshop_Data/Unix`
21) Type `cd ../../../`
22) Type `pwd`

23) When using `cd` everything is relative to your current location. However you can always use the *absolute* location to change directories. Let's move into the *Code* directory and look at two ways to switch to the *Data* directory.
24) Type `cd /lustre/<username>/Workshop_Data/Unix/Code`
25) Option 1: Type `cd ../Data`
26) Type `pwd`

27) Type `cd /lustre/<username>/Workshop_Data/Unix/Code`
28) Option 2: Type `cd /lustre/<username>/Workshop_Data/Unix/Data`
29) Type `pwd`
As you can see both options get us to the same place but Option 1 will only work from within a directory below Data. Option 2 will work from any location on DT2.

**Task:**

Remember the command prompt shows you the current directory you’re in, and when you’re in your home directory it shows a tilde (~). This is Unix’s short-hand method of specifying a home directory.

Type the following commands and record the results. Follow each command with `pwd` and note how your location changes.

```
$ cd / 
$ cd ~ 
$ cd / 
$ cd 
```

What were the locations after each command?

**UNIX TIP:** Tab complete is your friend. This is the best tip to learn early on as it will save you keystrokes and time. Type enough letters to uniquely identify the name of a file, directory, or program and press tab, Unix will autocomplete the word. E.g. type `tou` and press tab, Unix will autocomplete the word `touch` (we will learn about this soon). Tab completion occurred because there is no other command that starts with `tou`. If pressing tab doesn’t work you have not typed enough unique characters, pressing tab twice will show you a list of everything that starts with the characters you’ve typed thus far.

Change to your home directory, then use `cd` to move to `/lustre/<username>/Workshop_Data/Unix/Code/` directory. Use tab complete to do this.

How many keystrokes did it take?

How many does it take to do the whole thing?

In my case it takes 16 keystrokes versus the 43 it would take to type the whole string.

```
$ cd /lustre/imisner/Workshop_Data/Unix/Code/ 
```
Making Directories

Creating directories in Unix is done with the `mkdir` (**m**ake **d**irectory) command.

\$ mkdir DIRECTORY NAME

Unlike making directories on your desktop using spaces is not advised in the Unix file system. This is why you see the use of `_` in place of spaces. You can *escape* a space in Unix but it creates unnecessary typing and can create issues executing certain programs. Generally, using spaces in file and directory names is something to avoid.

1) Type `cd /lustre/<username>/Workshop_Data/Unix/`
2) Type `mkdir Work`
3) Type `ls`
4) Type `mkdir Temp1`
5) Type `cd Temp1`
6) Type `mkdir Temp2`
7) Type `cd Temp2`
8) Type `pwd`

9) In the previous example we created two temp directories but it took two steps. We could have done this in one step with by adding an *option/flag* to the `mkdir` command.

10) Type `cd ../..`
11) Type `mkdir -p Temp1.1/Temp2.1`
Task

Practice creating at least 4 directories within or below the `Work/` directory.

Move between them using `cd` by using the *absolute* and *relative* paths. Remember tab completion is your friend!

**UNIX TIP:** You may remember earlier we used a flag with the `mkdir` command that created two directories at once. How did we find out about the `-p` flag? Google? Well, you could do that, but Unix has built in manuals for each command that give you all of the details. Simply type `man COMMAND`, E.g. `$ man cd` `$ man ls` or `$ man mkdir`. Even `$ man man`

To navigate in a `man` page use the arrow keys to move down one line at a time, space bar will page down. To exit a `man` type `q`. 
In this section you will learn the basics of making files and putting things into those files. There are a variety of ways we can accomplish this as Unix has built in multiple editors for these tasks. We will review a few here.

$ touch FILENAME
This will create a new, empty file.

$ nano FILENAME
This is a built in text editor that will allow us to put information into a file.

1) Create two files in Unix/
   a. $ touch earth.txt
   b. $ touch heaven.txt
2) Type ls

```
login-2:Unix$ touch earth.txt
login-2:Unix$ touch heaven.txt
login-2:Unix$ ls
Code  Data  Temp1  Temp1.1  Work  earth.txt  heaven.txt
login-2:Unix$
```

3) We have now created two empty files called ‘earth.txt’ and ‘heaven.txt’
4) Type cd Work
5) Type touch basic info.txt
6) Type nano basic info.txt
7) We are now using an internal text editor that we can use to alter the contents of this file.
8) Add your name, email address, and favorite food to this file on separate lines.
9) Press control + x to exit and then y to save the file
10) We can also create and edit files with `nano`
a. `$ nano onestep.txt`
11) Add a line of text and save the file.

**Task**

In the **Unix/Work/** directory create a new directory called **Stuff**. Add a file to **Stuff** called **things**. Using **nano** type some text into **things** and save the file. Now edit **things** again in **nano** but save the file as **things.txt**. Finally, type **ls** what files do you see?

Record all your commands.
Moving Directories and Files

To move a file or directory the **mv** (move) command is used. This is the first command we have used that requires two arguments. You need to specify the source and the destination for the moving.

\[ \texttt{mv} \ \textit{SOURCE} \ \textit{DESTINATION} \]

1) Lets move **heaven.txt** and **earth.txt**
2) $ cd /lustre/<username>/Workshop_Data/Unix
3) $ mv heaven.txt Work/
4) $ mv earth.txt Work/
5) $ ls
6) $ ls Work/

```
login-2:Unix$ cd /lustre/imisner/Workshop_Data/Unix/
login-2:Unix$ mv heaven.txt Work/
login-2:Unix$ mv earth.txt Work/
login-2:Unix$ ls
Code   Data   Temp1 Temp1.1  Work
login-2:Unix$ ls Work/
  Stuff  basic_info.txt earth.txt heaven.txt onestep.txt
login-2:Unix$
```

7) We could have moved these files all at once using wildcards. An asterisk (*) means match anything.
8) $ mv * .txt Work/ #This will move any file that ends with .txt
9) $ mv * t Work/ # This moves any file or directory that ends with a t
10) $ mv *ea* Work/ #This works because only heaven and earth contain ‘ea’
11) $ mv can also be used to rename files
12) $ touch rags
13) $ ls
14) $ mv rags Work/ riches
15) $ ls Work/

```
login-2:Unix$ touch rags
login-2:Unix$ mv rags Work/ riches
login-2:Unix$ ls Work/
  Stuff  basic_info.txt earth.txt heaven.txt onestep.txt riches
```

16) Here we move and renamed the file **rags** to **Work/ riches**
17) We can rename it without moving the file
18) $ mv Work/riches Work/rags
19) $ ls Work/

```
login-2:Unix$ mv Work/riches Work/rags
login-2:Unix$ ls Work/
Stuff basic_info.txt earth.txt heaven.txt onestep.txt rags
```

20) The `mv` command is used to rename files or directories, as there is no ‘rename’ command in Unix.

**Task**

Move to your home directory and create a new directory called **Here**.

*Without changing directories* move **Here** to your **Work** directory on `/lustre`. After moving **Here** to **Work** move **Here** back to home but rename it **AndBackAgain**. Record your commands.

**UNIX TIP:** You can retrieve previous commands by accessing your history. On the command prompt pressing the up arrow will cycle through your previous commands. Typing `history` will list the last few hundred commands you entered. We will learn how to search through this type of information later. But using the up arrow to modify previous commands or commands with typo’s is another way to save time and keystrokes.

**Copying Directories**

To copy a file or directory `cp (copy)` command is used. Just like `mv` you will need a source and a destination to copy something.

```
$ cp SOURCE DESTINATION
1) Copying files is similar to moving them
2) $ cd /lustre/<username>/Workshop_Data/Unix/Work
3) $ mkdir Copy
4) $ cd Copy
5) $ touch file1
6) $ cp file1 file2
```
7) $ ls

```
login-2:Work$ cd /lustre/imisner/Workshop_Data/Unix/Work/
login-2:Work$ mkdir Copy
login-2:Work$ cd Copy/
login-2:Copy$ ls
file1 file2
```

8) Remember we do not have to be in a directory to make, move, or copy files.
9) $ touch ~/file3
10) $ ls
11) $ cp ~/file3 . # here we represent the current directory with a . (dot)
12) $ ls

```
login-2:Copy$ touch ~/file3
login-2:Copy$ ls
file1 file2
login-2:Copy$ cp ~/file3 .
login-2:Copy$ ls
file1 file2 file3
```

13) The `cp` command can also move directories using a flag
14) $ mkdir Example
15) $ mv file* Example/
16) $ ls

```
login-1:Copy$ mkdir Example
login-1:Copy$ mv file* Example/
login-1:Copy$ ls
Example
```

17) $ cp -R Example/ Example2
18) $ ls Example Example2/

```
```
19) What happens without the \texttt{–R} flag?

20) $ \texttt{cp Example2/ Example3}$

21) The error occurs because the \texttt{–R} flag means copy recursively. Since Example2 is not empty \texttt{cp} (without \texttt{–R}) does not descend into Example2 and copy those files it simply tries to move a directory without moving the things in the directory.

\textbf{Task}

Lets dig into the \texttt{man} for \texttt{cp}.

Are there other ways to tell \texttt{cp} to copy recursively?

How could we use \texttt{cp} to preserve the newest version of a file?

Besides \texttt{$ man cp} how else could we get help on this, or other, Unix commands?

How can we be sure \texttt{cp} will not overwrite files?

\underline{Viewing the contents of a directory}

To view the contents of directories we use the \texttt{ls (list segments)} command.
$ ls DIRECTORY

If no directory is provided `ls` will list the contents of the current directory.

1) We have been using `ls` frequently to check directory contents. However, there are many options for using `ls`. As the previous example noted we can use `ls` on multiple directories at the same time.

2) $ ls -l /lustre/<username>/Workshop_Data/Unix

3) $ ls -l --color /lustre/<username>/Workshop_Data/Unix

4) $ ls -p /lustre/<username>/Workshop_Data/Unix

```
login-1:~$ ls -l /lustre/imisner/Workshop_Data/Unix/
total 20K
drwxr-xr-x 2 imisner clfshpc 4.0K Jul 13 2009 Code
drwxr-xr-x 6 imisner clfshpc 4.0K Jun 24 13:20 Data
drwxr-xr-x 3 imisner clfshpc 4.0K Jun 25 11:51 Temp1
drwxr-xr-x 3 imisner clfshpc 4.0K Jun 25 11:55 Temp1.1
drwxr-xr-x 4 imisner clfshpc 4.0K Jun 25 16:24 Work
```

```
login-1:~$ ls -l --color /lustre/imisner/Workshop_Data/Unix/
total 20K
```

```
login-1:~$ ls -p /lustre/imisner/Workshop_Data/Unix/
```

```
Code/ Data/ Temp1/ Temp1.1/ Work/
```

5) As you can see these flags/options change the way `ls` displays the contents of the directory, giving us more or less information.

6) Notice the changes that the `-p` and `--color` flags make to the output.

**Task**

Try the following commands on any directory of your choosing.

```
$ ls -l
```
```
$ ls -R
```
Look through the man and determine what each of these flags does.

How can you display hidden files with `ls`?

Can you sort files by extension? How?

When sorting by extension what is another good flag to use?

---

**The most dangerous Unix command! Proceed with extreme caution.**

If you run `ls` on your Unix/Work/ directory I’m sure it is full of lots of empty files and directories by this point. Wouldn’t it be nice if there were a way to clean that up? Of course there is a way, however it can be dangerous.

**Please read this section carefully. Misuse of the `rm` command can delete your entire computer. Seriously.**

To delete directors and files from the system we have two options `rmdir (remove directory)` and `rm (remove)` command.

```
$ rmdir DIRECTORY
```

This command will remove empty directories only. Any directory with contents of any kind will not be deleted by `rmdir`.

```
$ rm FILE
```

This command will remove files, PERMANENTLY. There is no trashcan, recycle bin, or archive. Files or directories removed by `rm` are gone, forever. With `rm` it is possible to delete everything in your home directory, everything. This is why it is such a potentially dangerous command.

One more time just to be clear. It is possible to delete EVERY file you have ever created with the `rm` command. Thankfully there is a way to make `rm` a bit safer, and on DT2, this is the default setting. Using the `-i` flag `rm` will ask for confirmation before deleting anything.

1) $ cd /lustre/<username>/Workshop_Data/Unix/Temp1/
2) $ ls
3) If you remember the Temp2 directory is empty therefore we can use `rmdir` to delete it.
4) `$ rmdir Temp2/
5) $ ls

6) `$ rmdir` only removes empty directories, so don’t worry about deleting anything you may need with this command
7) We can now move up a level and remove Temp1
8) `$ cd ..
9) `$ rmdir Temp1/
10) $ ls

11) Now try `$ rmdir Temp1.1

12) See the error message letting us know the directory is not empty. However, we know that the only thing in Temp1.1 is Temp2.1 and that is empty so we can use `rm` and some flags to remove this directory and subdirectory in one step
13) `$ rm -r Temp1.1

```bash
login-1:Temp1$ cd /lustre/imisner/Workshop_Data/Unix/Temp1
login-1:Temp1$ ls
Temp2

login-1:Temp1$ rmdir Temp2/
login-1:Temp1$ ls

login-1:Temp1$ cd ..
login-1:Unix$ rmdir Temp1/
login-1:Unix$ ls
Code  Data  Temp1.1  Work
login-1:Unix$ 

login-1:Unix$ rmdir Temp1.1/
rmdir: failed to remove `Temp1.1/': Directory not empty
login-1:Unix$ 

login-1:Unix$ rm -r Temp1.1
```
14) Note the default behavior of `rm` on DT2 is to ask if you’re sure about removing files and directories. You have to press `y` or `n` before `rm` will proceed.

15) This feature can be useful for safety but it can be extremely cumbersome if you’re deleting directories that contain hundreds or thousands of files and subdirectories.

16) We can forcefully delete with `rm` and not be prompted for each file or subdirectory.

17) $ mkdir -p Temp1/Temp2/Temp3/Temp4

18) $ cp -r Temp1 Temp1.1

19) $ ls

20) $ rm -r Temp1

21) $ rm -rf Temp1.1

22) $ ls

<table>
<thead>
<tr>
<th>Code</th>
<th>Data</th>
<th>Temp1</th>
<th>Temp1.1</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```bash
login-1:Unix$ mkdir -p Temp1/Temp2/Temp3/Temp4
login-1:Unix$ cp -r Temp1/ Temp1.1
login-1:Unix$ ls
Code Data Temp1 Temp1.1 Work
login-1:Unix$ rm -r Temp1
/bin/rm: descend into directory `Temp1'? y
/bin/rm: descend into directory `Temp1/Temp2'? y
/bin/rm: descend into directory `Temp1/Temp2/Temp3'? y
/bin/rm: remove directory `Temp1/Temp2/Temp3'? y
/bin/rm: remove directory `Temp1/Temp2'? y
/bin/rm: remove directory `Temp1'? y
login-1:Unix$ rm -rf Temp1.1/
login-1:Unix$ ls
Code Data Work
login-1:Unix$ ls
```
23) Using \(-f\) flag saves keystrokes but can be dangerous since you’re not prompted before deleting files. ONLY USE THE \(-f\) FLAG IF YOU ARE POSITIVE YOU WANT TO REMOVE THOSE FILES OR DIRECTORIES.

**Task**

Use **`rm`** or **`rmdir`** or any combination to remove the **`Work/`** directory and all of its contents. Take some time to review **`man rm`**.

---

**Display the contents of a file**

There are various commands available to display/print the contents of a file. The default of all these commands is to display the contents of the file on the terminal. These commands are **`less`, `more`, `cat`, `head`, and `tail`**.

**$ less FILENAME**

Displays file contents on the screen with line scrolling (to scroll you can use ‘arrow’ keys, ‘PgUp/PgDn’ keys, ‘space bar’ or ‘Enter’ key). Press ‘q’ to exit.

**$ more FILENAME**

Like less command, also, displays file contents on the screen with line scrolling but uses only ‘space bar’ or ‘Enter’ key to scroll. Press ‘q’ to exit.

**$ cat FILENAME**

Simplest form of displaying contents. It **`cat`**alogs the entire contents of the file on the screen. In case of large files, entire file will scroll on the screen without pausing.

**$ head FILENAME**

Displays only the 10 starting lines of a file by default. Any number of lines can be displayed with the \(--n\) flag followed by the number of lines.

**$ tail FILENAME**
As the name implies the opposite of head this displays the last 10 lines. Again –n option can be used to change this.

1) Lets work through these commands.
2) $ cd /lustre/<username>/Workshop_Data/Unix/
3) $ less Data/Arabidopsis/At_proteins.fasta

```
$ cat /lustre/<username>/Workshop_Data/Unix/
```

1) $ cd /lustre/<username>/Workshop_Data/Unix/

2) $ less Data/Arabidopsis/At_proteins.fasta

3) This is quite a big file. You can see at the bottom `less` displays we are looking at lines 1-32 of 269,463 and we are 0% through the file.

4) We can use `h` to get help commands for `less`.
SUMMARY OF LESS COMMANDS

Commands marked with * may be preceded by a number, N.
Notes in parentheses indicate the behavior if N is given.

h H Display this help.
q :q Q :Q ZZ Exit.

MOVING

e A E j A N CR * Forward one line (or N lines).
y A Y k A K A P * Backward one line (or N lines).
f A F N SPACE * Forward one window (or N lines).
b A B ESC-v * Backward one window (or N lines).
z * Forward one window (and set window to N).
x * Backward one window (and set window to N).
ESC-SPACE * Forward one window, but don't stop at end-of-file.
d A D * Forward one half-window (and set half-window to N).
u A U * Backward one half-window (and set half-window to N).
ESC-) RightArrow * Left one half screen width (or N positions).
ESC-< LeftArrow * Right one half screen width (or N positions).
f A R N L * Repoint screen.
R * Repoint screen, discarding buffered input.

Default "window" is the screen height.
Default "half-window" is half of the screen height.

SEARCHING

HELP -- Press RETURN for more, or q when done.

6) Page forward using ‘space’, move a line at a time with ‘j’ (forward) or ‘k’ (backward) or N lines.
7) Hit ‘q’ to exit the help
8) Navigate around using the various commands
9) Try hitting ‘j’ ‘enter’ ‘100’ ‘enter’
10) Press ‘q’ when ready to exit less.
11) Similar to $ less $ more displays files to the screen but only uses ‘space’ (page) and ‘enter’ (line) to navigate.
12) From the Unix/ directory $ more Data/Arabidopsis/At_proteins.fasta
13) Navigate the file using the more command, press ‘q’ to exit.
14) $ cat is the simplest form of viewing and file. cat prints all of the file to the screen from start to finish.

15) $ cat Data/Arabidopsis/At_genes.gff.short

16) Did you get all of that?

17) $ cat is most useful when combined with other commands using | (pipes). We will cover this later.

18) The last two commands head and tail are fantastic when you need to look at a file and make sure things are in order.

19) $ head Data/GenBank/E.coli.genbank

20) $ head Data/GenBank/Y.pestis.genbank
21) We can change how many lines we see using the –n flag

22) $ head –n 1 Data/GenBank/E.coli.genbank Data/GenBank/Y.pestis.genbank

23) $ tail Data/GenBank/E.coli.genbank Data/GenBank/Y.pestis.genbank
24) This shows us the end of a file. This can be important when transferring files or data and needing to make sure everything transferred completely.
$ zip -r OUTDIR.zip DIRECTORY
Compress all files in a DIRECTORY into one archive file (OUTDIR.zip)

$ zip -r OUTFILE.zip . -i *.txt
Compress all txt files in a DIRECTORY into one archive file (OUTFILE.zip)

$ unzip SOMEFILE.zip
Decompress a file

tar utility saves many files together into a single archive file, and restores individual files from the archive. It also includes automatic archive compression/decompression options and special features for incremental and full backups.

tar -cvf OUTFILE.tar INFILE
archive INFILE

tar -czvf OUTFILE.tar.gz INFILE
archive and compress file INFILE

tar -tvf SOMEFILE.tar
list contents of archive SOMEFILE.tar

tar -xvf SOMEFILE.tar
extract contents of SOMEFILE.tar

tar -xzvf SOMEFILE.tar.gz
extract contents of gzipped archive SOMEFILE.tar.gz

tar -czvf OUTFILE.tar.gz DIRECTORY
archive and compress all files in a directory into one archive file

tar -czvf OUTFILE.tar.gz *.txt
archive and compress all *.txt files in current directory into one archive file

GZIP compression utility designed as a replacement for compress, with much better compression and no patented algorithms. The standard compression system for all GNU software.

gzip SOMEFILE
compress SOMEFILE (also removes uncompressed file)

gunzip SOMEFILE.gz
uncompress SOMEFILE.gz (also removes compressed file)
Task

A. Using **zip**, compress the **At_proteins.fasta** file in `/Unix/Data/Arabidopsis/`. Note the file size before and after compression.

File size before compression

File size after compression

B. Using **tar** archive and compress the **Unix** directory. You will have to be above the **Unix** directory in the directory tree to do this compression.

Directory size before compression

Directory size after compression

Look over the man page and identify what the flags you used are doing.

Extract the tar.gz file you made and remove it. Keep the original directory unchanged.

C. Using **gzip** archive and compress the **At_proteins.fasta** file.

File size before compression

File size after compression

Do you still have the original file?

Extract the **At_proteins.fasta** file and move it back to its original location `Data/Arabidopsis/`

D. Of the three methods for file compression which do you prefer? Why?
File permissions

All files in any operating system have a set of permissions associated with the file that define what can be done with the file and by whom. What = read, write (modify), and/or execute a file. Whom = user, group, or public.

These permissions are denoted with the following syntax:

<table>
<thead>
<tr>
<th>Permissions</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read r</td>
<td>Owner o</td>
</tr>
<tr>
<td>Write w</td>
<td>Group g</td>
</tr>
<tr>
<td>Execute x</td>
<td>Others o</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Permissions</th>
<th>Owner</th>
<th>Group</th>
<th>Size</th>
<th>Last modified</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>rwx</td>
<td>1 ase</td>
<td>cri</td>
<td>3976</td>
<td>Feb 20 2013</td>
<td>jobq_beta.sh</td>
</tr>
<tr>
<td>drwx</td>
<td>4 ase</td>
<td>cri</td>
<td>533</td>
<td>Jun 6 10:57</td>
<td>lib</td>
</tr>
<tr>
<td>drwx</td>
<td>3 ase</td>
<td>cri</td>
<td>22</td>
<td>Jun 6 10:57</td>
<td>man</td>
</tr>
<tr>
<td>-rwx-</td>
<td>1 ase</td>
<td>cri</td>
<td>28786</td>
<td>Jul 15 15:00</td>
<td>models_table.txt</td>
</tr>
<tr>
<td>-rwx-</td>
<td>1 ase</td>
<td>cri</td>
<td>38878</td>
<td>Jul 16 17:18</td>
<td>my bashrc</td>
</tr>
<tr>
<td>-rwx-</td>
<td>1 ase</td>
<td>cri</td>
<td>86</td>
<td>Jul 30 14:09</td>
<td>my bash_profile</td>
</tr>
<tr>
<td>drwxr-x</td>
<td>2 ase</td>
<td>cri</td>
<td>623</td>
<td>Jul 15 14:29</td>
<td>pilot2</td>
</tr>
<tr>
<td>drwx-r-x</td>
<td>3 ase</td>
<td>cri</td>
<td>199</td>
<td>Feb 18 2013</td>
<td>results</td>
</tr>
<tr>
<td>drwx-r-x</td>
<td>4 ase</td>
<td>cri</td>
<td>42</td>
<td>May 23 16:22</td>
<td>share</td>
</tr>
<tr>
<td>drwxr-x</td>
<td>6 ase</td>
<td>cri</td>
<td>105</td>
<td>Jul 2 10:12</td>
<td>software</td>
</tr>
<tr>
<td>drwx-r-x</td>
<td>1 ase</td>
<td>cri</td>
<td>692</td>
<td>Aug 13 11:49</td>
<td>template.sh</td>
</tr>
<tr>
<td>drwx-r-x</td>
<td>1 ase</td>
<td>cri</td>
<td>689</td>
<td>Aug 13 11:45</td>
<td>test.sh</td>
</tr>
</tbody>
</table>

Changing permissions is done via chmod (CHange MODe) command

$ chmod [Options] RELATIONS [+ or -] PERMISSIONS FILE

1) Lets make a new directory and add some files.
2) From the Unix/ directory
3) $ mkdir Allow
4) $ cd Allow/
5) $ touch read.txt write.txt execute.go all.txt
6) $ ls
7) $ ls -l
8) We have created some files but we need to change the permission for these files in order to share these or execute them as programs.
9) Since you created these files you’re the owner and have the ability to change their permissions with `chmod`.
10) From this you can see the default is for the user to have `rw` access and the group and others to have `r` access.
11) Let's add execute permissions for everyone on `execute.go`
12) `$ chmod a+x execute.go`
13) `$ ls -l`
18) Others still cannot modify this file but now members of the group will be able to modify the contents.
19) If we want a file to be completely public we need all of the flags active.
20) \$ chmod a+rwx all.txt
21) \$ ls -l

22) Now the file all.txt can be read, written, or executed by anyone on this system.
23) We can also remove permissions using this same command.
24) \$ chmod a-rwx all.txt
25) \$ ls -l
26) Now we have removed all access to the all.txt file even the owner’s access.
27) Finally we can change the permissions of all the files in a directory with the –R flag.
28) $ cd ..
29) $ chmod -R a+rwx Allow/
30) $ ls -l Allow/

```
login-1:Allow$ cd ..
login-1:Unix$ chmod -R a+rwx Allow/
login-1:Unix$ ls -l Allow/
total 0
-rwxrwxrwx 1 imisner clfshpc 0 Jul 2 10:22 all.txt
-rwxrwxrwx 1 imisner clfshpc 0 Jul 2 10:22 execute.go
-rwxrwxrwx 1 imisner clfshpc 0 Jul 2 10:22 read.txt
-rwxrwxrwx 1 imisner clfshpc 0 Jul 2 10:22 write.txt
```

31) This made all of these files public in one step.

**Task**

What group does your user account belong to?

Make a directory and some files. Change the permissions using `chmod`. Now using –R flag make the whole folder *rwx*. Make a new directory below your test directory with a new file. What are the permissions on that new file?

**Profiles**

You may have noticed that my screen output looks different than what you’ve been looking at. This is due to customizations I have made to my bash profile. Basically, using a `.profile` I’ve changed the way bash displays `ls` output as well as other options. On DT2 you have built in profiles that automatically load when you log in. However, you haven’t seen them yet because they are ‘hidden’ files. In Unix files that start with a dot are hidden to see hidden files use `ls -a`

1) from Unix/ directory `ls -a`
2) Now we can look at this file with any number of options (head, tail, cat, less, more, nano)
3) Let’s use nano to make a change to the file and talk about what these different lines mean.
4) $ nano .bashrc.mine.unix
5) We need to change the username to your login id

```
# Modify default shell behavior. This is the place to set shell
# variables, and tap commands.
# Note: If you use tap commands, you may want to use "tap -q" to
# keep the tap output from breaking the scp/sftp commands. I.e.
# tap -q matlab

# uncomment the following line to increase the command history kept
# by the shell
history="500"

# Set up paths
# you have to change <username> to your login name
PATH=$PATH"":/lustre/<username>/Workshop_Data/Unix/Code"

# this pre loads the bioinformatics modules
module load dept/bioinfo
module load bioinfo

# sets the colors for ls
LS_COLORS="di=1;35:ex=0,31:fi=0,97"
export LS_COLORS
```

6) If we want the system to “use” this file we have to use source
7) $ source .bashrc.mine.unix
8) Now to see if the changes worked
9) $ ls -l --color
10) In truth you could have used the --color flag at any point with ls and gotten a colored output. The LS COLORS options simply set custom colors over the default values.
11) We can make ls provide colors every time by setting an alias for ls in ~/.bash_aliases
12) $ nano ~/.bash_aliases
13) Add the following: alias ls="ls -h --color"
14) Now source that file: source ~/.bash_aliases
15) Now typing ls will actually be interpreted as ls -h --color
16) If you want to make this profile permanent for your DT2 login you will have to move this file to your home directory and replace the `.bashrc.mine` file in the home directory. DT2 will automatically source this file every time you login.

17) If you don’t/don’t want to move this to your `~/.` file then you will have to source this file each time you open a new terminal or login to DT2.

---

**Your first script**

Just like Perl, Python, R, etc. Unix can be used as a programming language. Depending upon the task a shell script might be all you really need to get your task completed.

To make a script we simply write shell commands into a file and then treat that file like any other program or command.

1) From the **Unix/Code/ directory**
2) `$ nano hello.sh`
3) Type the following two lines

```bash
#This is my first shell script
echo "Hello World"
```

4) Save the file and exit `nano`
5) `$ chmod u+x hello.sh`
6) `$ hello.sh`
7) Voilà! It’s that simple.
8) Now move to another directory and see if you can still run `hello.sh`
9) This works because we’ve added the **Code** directory to the **PATH**. Basically Unix knows to look in this directory for commands we type.

---

**Task**

From your home directory (execute all commands from your home directory)

1) Move `hello.sh` to `Unix/`
2) run `hello.sh`
3) Now move `hello.sh` back to `Unix/Code`, it should work again
Useful shell scripts

Look in the **Data/Unix_test_files** directory. You should see several files (all are empty) and four directories.

1) Put the following information into a shell script and save it as `cleanup.sh`.

```bash
#!/bin/bash
mv *.txt Text
mv *.jpg Pictures
mv *.mp3 Music
mv *.fa Sequences
```

2) Now return to **Data/Unix_test_files**
3) `$ ls -l`
4) `$ cleanup.sh`
5) `$ ls -l`
6) It should place the relevant files in the correct directories. This is a relatively simple use of shell scripting. As you can see the script just contains regular Unix commands that you might type at the command prompt.
7) Did you notice the `#!/bin/bash` line in this script? There are several different types of shell script in Unix, and this line makes it clearer that a) that this is actually a file that can be treated as a program and b) that it will be a bash script (bash is a type of Unix).
Task

Copy this information into a file called `change_file_extension.sh` and again place that file in the `Code` directory.

```bash
#!/bin/bash

for filename in *.$1
do
    mv $filename ${filename%$1}$2
done
```

Now go to the `Data/Unix_test_files/Text` directory. Run the following command:

```
$ change_file_extension.sh txt text
```

Now run the `ls` command to see what has happened to the files in the directory?

Try using this script to change the file extensions of other files.

It’s not essential that you understand exactly how this script works at the moment (things will become clearer as you learn Python), but you should at least see how a relatively simple Unix shell script could be potentially very useful.
Unix Power Commands

The commands that you have learned so far are essential for doing any work in Unix, but they don’t really let you do anything that is very useful. The following section will introduce new commands that will start to show you the power of Unix.

Pipes and redirects

Everything we have done so far has sent the result of the command to the screen. This is feasible when the data being displayed is small enough to fit the screen or if it is the endpoint of your analysis. But for large data outputs, or if you need a new file, printing to the screen isn’t very useful. Unix has built in methods to hand output from commands using > (greater than) or < (less than) or >> signs.

< redirects the data to the command for processing

> redirects the data from the command’s output to a file. The file will be created if it is non-existing and if present it will overwrite the contents with the new output data (you will lose the original file).

>> unlike > this redirection lets user append the data to an already existing file or a new file

Another special operator | (called pipe) is used to pass the output from a command to another command (as input) before sending it to an output file or display.

Some examples:

$ cat FILE1 > FILE2

Creates a new file (file2) with same contents as old file (file1)

$ cat FILE1 >> FILE2

Appends the contents for file1 to file2, equivalent to opening file1, copying all the contents, pasting the copied contents to the end of the file2 and saving it!

$ cat FILE1 | less
Here, `cat` command displays the contents of the file, but instead of sending it to standard output (screen) it sends it through the pipe to the next command `less` so that contents of the file are now displayed on the screen with line scrolling.

1) From the `Unix/Data/` directory
2) $ `cat seq.fasta`
3) $ `head seq.fasta > new.txt`
4) $ `cat new.txt`
5) $ `tail seq.fasta > new.txt`
6) Did you get an error?
7) This is because DT2 has set some environmental variables to prevent redirection from overwriting existing files. However, with a `>` we can override the override.
8) $ `tail seq.fasta >| new.txt`
9) $ `cat new.txt`
10) Now let's try that with the append option.
11) $ `head –n 1 seq.fasta >| new.txt`
12) $ `tail –n 1 seq.fasta >> new.txt`
13) Notice we did not get an error and we did not need the `|`. This is because we have simply appended an existing file not overwritten the file.

**Task**

The `Data/` directory contains a few fasta files with the extension `.fa`. Combine all of these files into a single file called `sequences.fasta` using redirects.

**Grep**

The `grep` (globally search a regular expression and print) is one of the most useful commands in Unix and it is commonly used to filter a file/input, line by line, against a pattern.

```
grep [OPTIONS] PATTERN FILENAME
```

Like any other command there are various options available `man grep` for this command. Most useful options include:
-v inverts the match or finds lines NOT containing the pattern.
--color colors the matched text for easy visualization
-i ignore case for the pattern matching.
-l lists the file names containing the pattern
-n prints the line number containing the pattern
-c counts the number of matches for a pattern

Some typical scenarios to use **grep**:

- Counting number of sequences in a multi-fasta sequence file
- Get the header lines of fasta sequence file
- Find a matching motif in a sequence file
- Find restriction sites in sequence(s)
- Get all the Gene IDs from a multi-fasta sequence files and many more.

You might already know that fasta files header must start with a '>' character, followed by a DNA or protein sequence on subsequent lines. To find only those header lines in a fasta file, we can use **grep**.

1) From **Unix/Data/Arabidopsis**
2) `$ grep ">" intronIME_data.fasta`
3) Did you get that?
4) Remember the default for a program is to output to the screen.
5) We can fix this with a redirect or a pipe.
6) `$ grep ">" intronIME_data.fasta | less`
7) This takes the output from **grep** and sends it as input to **less**
8) What if we want to know how many sequences are in a file?
9) `$ grep -c ">" intronIME_data.fasta`
10) We can also get lines that don’t match our string.
11) `$ grep -v ">" intronIME_data.fasta | less`
12) Given a the fasta file structure we can use **grep** to separate this information
13) `$ grep ">" intronIME_data.fasta > intron headers.txt`
14) `$ grep -v ">" intronIME_data.fasta > intron_sequences.txt`
15) We can even get some biological information from **grep**
16) `$ grep --color "GAATTC" chr1.fasta`
17) GAATTC is the EcoRI cut site. The **--color** option highlights the matches in this sequence.
Task

Let’s use `grep` to get some information that we might be interested in knowing from these files. Using `grep` and these files lets determine:

1) How many EcoR1 cut sites are there in chr1.fasta?
2) Can you find a transcription factor in At_proteins.fasta? How many?
3) What does the following command do?
   $ grep –c –w “ATP” At_proteins.fasta
   $ grep –l “AAA” ../*.fa
4) Let’s find all of the transcription factor annotations?
   $ grep –i “transcription factor” At_proteins.fasta | less
5) What if we don’t want to see genes on chromosome 1?
6) What if we only want to genes on chromosome 5?

Regular expressions

grep + regular expressions = power! Before we get into this let’s start with a task.

Task

The `. ` and `*` characters are also special characters that form part of the regular expression. Try to understand how the following patterns all differ. Try using each of these patterns with `grep –c` against any one of the sequence files. Can you predict which of the five patterns will generate the most matches?

ACGT
AC .GT
AC*GT
AC .*GT

*The asterisk in a regular expression is similar to, but NOT the same, as the other asterisks that we have seen so far. An asterisk in a regular expression means: 'match zero or more of the preceding character or pattern’*
Try searching for the following patterns to ensure you understand what `.` and `*` are doing:

- A.T
- AG*T
- A*C*G*T*

When working with the sequences (protein or DNA) we are often interested to see if a particular feature is present or not. This could be various things like a start codon, restriction site, or even a motif. In Unix all strings of text that follow some pattern can be searched using some formula called regular expressions. e.g. As you learned above regular expressions consist of normal and metacharacters. Commonly used characters include:

<table>
<thead>
<tr>
<th>Expression</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>matches any single character</td>
</tr>
<tr>
<td>$</td>
<td>matches the end of a line</td>
</tr>
<tr>
<td>^</td>
<td>matches the beginning of a line</td>
</tr>
<tr>
<td>*</td>
<td>matches one or more character</td>
</tr>
<tr>
<td>\</td>
<td>quoting character, treat the next character followed by this as an ordinary character.</td>
</tr>
<tr>
<td>[ ]</td>
<td>matches one or more characters between the brackets</td>
</tr>
<tr>
<td>[range]</td>
<td>match any character in the range</td>
</tr>
<tr>
<td>[^range]</td>
<td>match any character except those in the range</td>
</tr>
<tr>
<td>{N}</td>
<td>match N occurrences of the character preceding (sometimes simply +N) where N is a number.</td>
</tr>
<tr>
<td>{N1,N2}</td>
<td>match at least N1 occurrences of the character preceding but not more than N1</td>
</tr>
<tr>
<td>?</td>
<td>match 1 occurrence of the character preceding</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some common patterns for Nucleotide/Protein searches.

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>^ATG</td>
<td>Find a pattern starting with ATG</td>
</tr>
<tr>
<td>TAG$</td>
<td>Find a pattern ending with TAG</td>
</tr>
<tr>
<td>^A[TGC]G</td>
<td>Find patterns matching either ATG, AGG or ACG</td>
</tr>
<tr>
<td>TA[GA]$</td>
<td>Find patterns matching either TAG or TAA</td>
</tr>
<tr>
<td>[YXN] [MPR]_{[0-9]{4,9}}</td>
<td>Find patterns matching NCBI RefSeq (eg XM_012345)</td>
</tr>
<tr>
<td>(NP|XP)_{[0-9]{4,9}}</td>
<td>Find patterns matching NCBI RefSeq proteins</td>
</tr>
</tbody>
</table>

Let’s use `grep` to find a zinc finger motif. For simplicity let’s assume zinc finger motif to be `CXXCXXXXXXXXXXXXHXXXH`. Either you can use dots to represent any amino acids or use complex regular expressions to come up with a more representative pattern.

```
$ grep --color "C..C.........H...H" At_proteins.fasta
```
$ grep --color "C.{2}\{12\}H.{3}" At_proteins.fasta


These all do the exact same thing. As you can see regular expressions can be very useful for finding patterns of all kinds.

**UNIX TIP:** You can use regular expression in grep, sed, $ awk, less, perl, python, certain text editors almost any programming language or tool can utilize the power of regex.

---

**tr**

The `tr` (transliterate) command is used to translate the input file and produce a modified output. It uses two sets of parameters and replaces the occurrence of the characters in the first set with the elements from the other set.

```bash
$ tr [options] "String1" "String2" < INFILE > OUTFILE
```

This makes changing things in files very easy.

1) From the `Data/Arabidopsis/` directory
2) `$ head -n 2 chr1.fasta`
3) `$ head -n 2 chr1.fasta | tr 'A-Z' 'a-z'`
4) Using `tr` we changed the fasta from uppercase to lowercase *as well as the headers*
5) Here are some useful `tr` commands:
6) `$ tr "[:lower:]" "[:upper:]"`
7) `$ tr "ATCG" "AUCG"` # Turns cDNA into RNA
8) `$ tr -s /n` #Single spaces a document
9) In the previous command \n is the syntax for a newline Unix. Sadly, Mac, PC’s, and Unix **DO NOT** use the same newline character.
10) From the `Unix/Data/Misc` directory
11) `$ cat -v excel_data.csv`
12) Do you see all of those `^M` characters? This is how a Windows (Dos) represents the newline, Unix does not like this.
13) `$ tr -d `"^M" < excel_data.csv`
14) `$ tr -d `"^M" < excel_data.csv > excel.fixed`
15) Newer versions of Unix (including this one) have built in commands to deal with this exact issue mac2unix, dos2unix, and unix2dos. These programs will edit the file and save it without any redirects.
16) $ dos2unix excel_data.csv
17) Anytime you bring a table or file from your PC or Mac to the command line I would recommend you fix the newline character.

---

**sed**

`tr` lets you change a range of characters but what if you want to change a specific string? Enter `sed`. `sed` is a stream editor that reads one or more text files and makes changes or edits then writes the results to standard output. The simple syntax for `sed` is:

```
$ sed "OPERATION/REGEXP/REPLACEMENT/FLAGS" FILENAME
```

Above, `/` is the delimiter but you can use `_ |` or `:` as well.

**OPERATION** = the action to be performed, the most common being `s` which is for substitution.

**REGEXP** and **REPLACEMENT** = the search term and the substitution for the operation be executed.

**FLAGS** = additional parameters that control the operation, common **FLAGS** include:

- **g** replace all the instances of `REGEXP` with `REPLACEMENT` (globally)
- **n** (n=any number) replace nth instance of the `REGEXP` with `REPLACEMENT`
- **p** If substitution was made, then prints the new pattern space
- **i** ignores case for matching `REGEXP`
- **w** If substitution was made, write out the result to the given file
- **d** when specified without `REPLACEMENT`, deletes the found `REGEXP`

1) From Unix/Data/Arabidopsis
2) $ head -n 1 chr1.fasta
3) $ sed 's/Chr1/Chromosome_1/g' chr1.fasta | head -n 1
4) $ sed 's:Chr1:Chromosome_1:g' chr1.fasta | head -n 1
5) As you can see these two commands do the same thing with different delimiters. We Changed “Chr1” to “Chromosome_1” in the file. However this was not done permanently. To do that we would have to write to a new file or use a flag within `sed`.
6) $ touch greetings.txt
7) $ echo “Hello there” >> greetings.txt
8) $ head greetings.txt
9) Now we have our file to manipulate with `sed`. We have three options for altering and saving the file

10) Option 1: Make a new file

11) $ sed 's/Hello/Hi/g' greetings.txt > greetings_short.txt

12) $ head greetings*

```
login-1:Arabidopsis$ sed 's/Hello/Hi/' greetings.txt > greetings_short.txt
login-1:Arabidopsis$ head greetings*
  Hello there
  => greetings_short.txt
  => greetings_short.txt
  => greetings_short.txt
  Hi there
  => greetings.txt
login-1:Arabidopsis$
```

13) Option 2: Edit in place but make a backup of the original with the given extension

14) $ sed -i.bak 's/Hello/Hi/g' greetings.txt

15) $ head greetings*

```
login-2:Arabidopsis$ sed -i.bak "s/Hello/Hi/g" greetings.txt
login-2:Arabidopsis$ head greetings*
  => greetings.txt
  => greetings.txt
  => greetings.txt
  => greetings_short.txt
  => greetings_short.txt
  => greetings_short.txt
  Hi there
```

16) Option 3: Edit in place but without a backup. NOTE if you run out of system memory or have an error this will rewrite the original file. You will not get that file back.

17) $ sed -i 's/Hello/Hi/g' greetings.txt.bak

18) $ head greetings*
Unlike other Unix commands, `awk` is a structured language by itself. `awk` stands for the names of its authors Aho, Weinberger, and Kernighan. Many bioinformatics programs generate rows and columns of information. `awk` is an excellent tool for processing these rows and columns, and it is easier than most conventional programming languages.

The syntax for `awk` is:

```
$ awk 'PATTERN {ACTION}' FILENAME
```

`awk` then works by reading the input file one line at a time, matching the given PATTERN and performing the corresponding ACTION for the matches. If there is no PATTERN, then the ACTION will be performed on each line. But if there is no ACTION then the default ACTION (printing all lines) on the matching PATTERN will be performed (empty braces `{}` without any ACTION turns off default printing).

Some inbuilt variables of `awk` include:

- **FS**  Field Separator (default SPACE)
- **OFS** Output Field Separator (default SPACE)
- **NF**  Number of Fields in the input
- **NR**  Number of Records (lines) in the input
- **RS**  Record Separator (default NEWLINE)
- **ORS** Output Record Separator (default NEWLINE)
- **FNR** File line number
- **N**   Nth field of the line where N can be any number (eg. $0 = entire line, $1 = First field, $2 = second field and so on)

`awk` accepts all standard patterns (regular expression and expression) plus some special patterns.
BEGIN Special PATTERN that is executed before the INPUT is read
END Special PATTERN that is executed after the INPUT is read
empty nonexistent PATTERN that matches every input record

1) From Unix/Data/Arabidopsis
2) The simplest $ awk command:
3) $ $ awk '{print;}' chr1.fasta
4) Here, since there is no PATTERN, the print ACTION will be performed on each line (equivalent to cat INFILE).

Task

Some simple commands using $ awk (you can try these commands with ‘At_genes.gff’)

$ awk NF FILE Deletes all blank lines
$ awk 'NF > 0' FILE Deletes all blank lines
$ awk 'NF > 4' FILE Prints the 4th field of every line
$ awk '$NF > 4' FILE Prints all lines with value of the 4th field > 4
$ awk 'END { print $NF }' FILE Prints value of the last field of the last line
$ awk 'NR==25,NR==100' FILE Prints lines between 25 and 100
$ awk 'NR==50' FILE Prints 50th line of input
$ awk 'NR < 26' FILE Prints first 25 lines
$ awk 'NR > 25' FILE Prints file after 25th line
$ awk 'END { print NR }' FILE Prints the last line of the file
$ awk '{ print NF ":" $0 }' FILE Prints number of fields in front of every line
$ awk '{ print FNR ":" $0 }' FILE Prints line number in front of every line
$ awk '$5 == "abc123"' FILE Prints lines which have ‘abc123’ in 5th field
$ awk '{ print $1, $2 }' FILE Prints only 1st and 2nd field
$ awk '{ print $2, $1 }' FILE Prints only 2nd and 1st field (swapping columns)
$ awk '{ $2 = ""; print }' FILE Prints the file without 2nd column
$ awk '/gene/' FILE Prints all the lines having “gene”
$ awk '!/gene/' FILE Prints all the files not having “gene”
$ awk '/UTR|CDS|RNA/' FILE Prints all the files having either UTR, CDS, or, RNA

Try to understand the following commands (and record your results, where applicable):

$ awk 'END { print $NF }' At_genes.gff
$ awk 'NR==30,NR==35' At_genes.gff
$ awk 'NR==25' At_genes.gff $ awk 'NR<25' At_genes.gff
$ awk 'END { print NR }' At_genes.gff
$ awk '{ print NF "$" $0 }' At_genes.gff > with_fields.txt
$ awk '{ print NR "$" $0 }' At_genes.gff > with_Line_num.txt
$ awk '{ print $1, $3 }' At_genes.gff
$ awk '{print $1"\t"$3"\t"$2}' At_genes.gff
$ awk '{print $1,$2,$(NF-4),$\{NF-3\}}' At_genes.gff.
$ awk '/Chr1/' At_genes.gff.

---

word count

wc (word count) is a useful command in bioinformatics because it can quickly identify how many lines or words are in a file.

$ wc FILENAME
1) From Data/Arabidopsis
2) $ wc At_genes.gff
3) Here we have the total number of lines, words, and bytes in this file

```
login-2:Arabidopsis$ wc At_genes.gff
531497 4783473 39322356 At_genes.gff
login-2:Arabidopsis$
```

4) $ wc -l At_genes.gff

```
login-2:Arabidopsis$ wc -l At_genes.gff
531497 At_genes.gff
login-2:Arabidopsis$
```

5) This prints out just the line count for the input file.
6) $ wc is best used in with pipes but it can be useful to count things as well
7) $ ls /lustre/<username>/Workshop_Data/Unix/Data/Sequences | grep ".fa" | wc -l
8) This command tells you how many .fa files there are in the Sequences directory.

---

Sort

$ sort command can be used to arrange things in a file. Simplest way to use this command is:
$ sort FILE1 > SORTED_FILE1

-n  numerical sort
-r  reverse sort
-k N,N  sort the Nth field (column), where N is a number. Sorting can also be done on the exact character on a particular field eg. -k 4.3, 4.4 sorts based on 3rd and 4th character of the 4th field. Additionally you can supply additional -k for resolving ties.
-t  specify the delimiters to be used to identify fields (default is TAB) -t ':' to use ':' as delimiter

Task

The Unix/Data/Sequences directory consists of numerically labeled files. Unix can sort either alphabetically or numerically (not both) and hence they are arranged in Seq1.fa, Seq10.fa, Seq11.fa etc. In order to sort them in an easy to read way, try using

$ ls |sort -t 'q' -k 2n

This command lists all the files in sequences directory and then passes it to sort command. Sort command then sorts it numerically but only using 3rd and 4th letters of the first field (file name)

Try using sort on Data/Arabidopsis/At_genes.gff

$ sort -r -k 1,1 At_genes.gff
$ sort -r -k 4,4 At_genes.gff

uniq (unique) command removes duplicate lines from a sorted file, retaining only one instance of the running matching lines. Optionally, it can show only lines that appear exactly once, or lines that appear more than once. uniq requires sorted input since it compares only consecutive lines.

$ uniq [OPTIONS] INFILE OUTFILE

Useful options include:
-c  count; prints lines by the number of occurrences  
-d  only print duplicate lines  
-u  only print unique lines  
-i  ignore differences in case when comparing  
-s N  skip comparing the first N characters (N=number)

**Task**

From Data/

1) `cat uniq.txt`
2) Using the above options do the following:
   a. Count the occurrence of each unique line
   b. Print only duplicated lines.
   c. Print only unique lines.

---

**Comparing two files**

`diff` (difference) reports differences between two files.

$ diff [OPTIONS] FILE1 FILE2

Useful options include

- `b`  ignore blanks
- `w`  ignore white space (spaces and tabs)
- `i`  ignore case
- `r`  recursively compare all files (when comparing folders)
- `s`  list all similar files (when comparing folders)
- `y`  side by side comparison of files

Differences are reported line by line using a (addition), c (changed), and d (deleted) along with < or > indicating the direction of the change.

1) From Unix/Data
2) $ diff uniq.txt diff.txt > diff_test.txt  
3) $ cat diff_test.txt
4) He we see the output from `diff` that covers each change.

5) `2d1 = line 2 in file 1 is deleted from file 2 < aa`

6) `5,12d3 = lines 5-12 in file 1 are deleted at line 3 in file 2`

7) `15a7,8 = at line 15 in file 1 there is an addition of lines 7,8 in file 2.`

`comm` (comm on) command compares two sorted files line by line.

$ comm [OPTIONS] FILE1 FILE2

-1 suppress lines unique to FILE1
-2 suppress lines unique to FILE2
-3 suppress lines that appear in both files

Task

Use `comm` with all three options on uniq.txt and diff.txt
Dividing files

**cut** divides the file into several parts and displays selected columns or fields from each line of a file. Normally **cut** command requires how the fields are separated and what fields need to be displayed.

\$ **cut** –f 1 FILE  
\$ **cut** –d ‘,’ –f 2-4 FILE  
**cut** –d ‘|’ –f 1,9 FILE  

**split** generates output files of a fixed size (bytes or lines). Useful when a huge file needs to be processed.

\$ **split** -d –l 100 FILENAME SUFFIX  

here –d specifies numeric suffix only (suffix00, suffix01, suffix02 etc.) while –l specifies number of lines in each file (100 in this case). If you want to split based on bytes, you can use –b option (–b 1k or –b 1m for 1 KB and 1 MB respectively)

You can join these files back using

\$ **cat** suffix0[0-2] >> joinedfile

**Task**

1) Display only first column of the **At_genes.gff** file using **cut**
2) Now can you display that in a way so you can actually see what the output looks like?
3) What if you want column 1, 4, and 5?
4) Let’s split **At_genes.gff** into 10,000 line segments. Use “gff_split” as the suffix for the new files.
5) How many files did that make?
6) Can you put them back together?
Let's say that we want to extract five sequences from `intron_IME_data.fasta` that are: a) from first introns, b) in the 5' UTR, and c) closest to the TSS. Therefore we will need to look for FASTA headers that contain the text 'i1' (first intron) and also the text '5UTR'. Every intron sequence in this file has a header line that contains the following pieces of information:

- gene name
- intron position in gene
- distance of intron from transcription start site (TSS)
- type of sequence that intron is located in (either CDS or UTR)

We can use `grep` to find header lines that match these terms, but this will not let us extract the associated sequences. The distance to the TSS is the number in the FASTA header which comes after the intron position. So we want to find the five introns which have the lowest values.

Before I show you one way of doing this in Unix, think for a moment how you would go about this if you didn't know any Unix... would it even be something you could do without manually going through a text file and selecting each sequence by eye? Note that this Unix command is so long that --- depending on how you are viewing this document --- it may appear to wrap across two lines. When you type this, it should all be on a single line:

```
$ tr '
' '@' < intron_IME_data.fasta | sed 's/#/\#/g' | tr '#' '
' | grep "i1_.*5UTR" | sort -nk 3 -t '_' | head -n 5 | tr '@' '
'
```

**Task**

Break down the above command and figure out what it is doing.

See if you can write your own one-liner. In the `At_genes.gff` file there are multiple types in column 3 (gene, CDS, mRNA, etc) how many of each type are there? This can be done with a one-liner in 3 commands.
<table>
<thead>
<tr>
<th>Command</th>
<th>Function</th>
<th>Syntax/example usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Navigation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ls</td>
<td>list contents</td>
<td><code>ls [OPTIONS] DIRECTORY</code></td>
</tr>
<tr>
<td>pwd</td>
<td>print working directory</td>
<td><code>pwd</code></td>
</tr>
<tr>
<td>cd</td>
<td>change directory</td>
<td><code>cd ~</code> or <code>cd #home directory</code>&lt;br&gt;<code>cd ..</code> #previous (parent directory)</td>
</tr>
<tr>
<td><strong>File/Directory operations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mkdir</td>
<td>make directory</td>
<td><code>mkdir DIRECTORY</code></td>
</tr>
<tr>
<td>cp</td>
<td>copy files/directories</td>
<td><code>cp SOURCE DESTINATION</code></td>
</tr>
<tr>
<td>man</td>
<td>manual page (help)</td>
<td><code>man COMMAND</code></td>
</tr>
<tr>
<td>mv</td>
<td>move files/directories</td>
<td><code>mv SOURCE DESTINATION</code></td>
</tr>
<tr>
<td>touch</td>
<td>create file</td>
<td><code>touch FILE</code></td>
</tr>
<tr>
<td>nano</td>
<td>edit file</td>
<td><code>nano FILE</code></td>
</tr>
<tr>
<td>less</td>
<td>view file (with more options)</td>
<td><code>less FILE</code></td>
</tr>
<tr>
<td>more</td>
<td>view file (with less options)</td>
<td><code>more FILE</code></td>
</tr>
<tr>
<td>cat</td>
<td>catalog file contents</td>
<td><code>cat FILE</code></td>
</tr>
<tr>
<td>head</td>
<td>show first few lines of a file</td>
<td><code>head FILE</code></td>
</tr>
<tr>
<td>tail</td>
<td>show last few lines of a file</td>
<td><code>tail FILE</code></td>
</tr>
<tr>
<td>rmdir</td>
<td>remove empty directory</td>
<td><code>rmdir DIRECTORY</code></td>
</tr>
<tr>
<td>rm</td>
<td>remove file(s)</td>
<td><code>rm FILE</code></td>
</tr>
<tr>
<td><strong>Compression/archiving</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zip</td>
<td>zip compress</td>
<td><code>zip OUTFILE.zip INFILE.txt</code>&lt;br&gt;<code>zip -r OUTDIR.zip DIRECTORY</code></td>
</tr>
<tr>
<td>unzip</td>
<td>decompress zipped file</td>
<td><code>unzip ANYTHING.zip</code>&lt;br&gt;<code>tar -czvf OUTFILE.tar.gz DIRECTORY</code>&lt;br&gt;<code>tar -xzvf OUTFILE.tar.gz</code></td>
</tr>
<tr>
<td>tar</td>
<td>archive and compress files/directories</td>
<td><code>gzip SOMEFILE</code>&lt;br&gt;<code>gzip SOMEFILE.gz</code></td>
</tr>
<tr>
<td>gzip</td>
<td>gzip files</td>
<td><code>gunzip SOMEFILE.gz</code>&lt;br&gt;<code>gunzip SOMEFILE</code></td>
</tr>
<tr>
<td>gunzip</td>
<td>decompress gzipped files</td>
<td><code>gzip SOMEFILE.gz</code></td>
</tr>
<tr>
<td><strong>File permissions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chmod</td>
<td>change permissions for files/directories</td>
<td><code>chmod [OPTIONS] RELATIONS[+ or -]PERMISSIONS FILE</code></td>
</tr>
<tr>
<td><strong>File manipulations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grep</td>
<td>search a pattern</td>
<td><code>grep [OPTIONS] &quot;PATTERN&quot; FILENAME</code></td>
</tr>
<tr>
<td>sed</td>
<td>stream edit a file</td>
<td><code>sed 's/search/replace/g' FILENAME</code></td>
</tr>
<tr>
<td>awk</td>
<td>multi-purpose command</td>
<td><code>awk 'PATTERN {ACTION}' FILENAME</code></td>
</tr>
<tr>
<td>tr</td>
<td>translate or transliterate a file</td>
<td><code>tr [OPTIONS] &quot;STRING1&quot; &quot;STRING2&quot; &lt;INFILE</code></td>
</tr>
<tr>
<td>wc</td>
<td>word count</td>
<td><code>wc FILENAME</code>&lt;br&gt;<code>wc FILENAME</code></td>
</tr>
<tr>
<td>sort</td>
<td>sort files</td>
<td><code>sort FILE1 &gt; SORTED_FILE1</code></td>
</tr>
<tr>
<td>uniq</td>
<td>display unique lines</td>
<td><code>uniq [OPTIONS] INFILE &gt; OUTFILE</code></td>
</tr>
<tr>
<td>diff</td>
<td>display difference</td>
<td><code>diff [OPTIONS] FILE1 FILE2</code></td>
</tr>
<tr>
<td>comm</td>
<td>display common lines among files</td>
<td><code>comm [OPTIONS] FILE1 FILE2</code></td>
</tr>
<tr>
<td>cut</td>
<td>break files vertically based on fields</td>
<td><code>cut -d &quot;DELIMITER&quot; -f NUMBER FILE</code></td>
</tr>
<tr>
<td>split</td>
<td>break files horizontally</td>
<td><code>split [OPTIONS] FILENAME</code></td>
</tr>
</tbody>
</table>
## SLURM Guide for DT or DT2

<table>
<thead>
<tr>
<th>Task</th>
<th>User Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job submission</td>
<td>sbatch [script file]</td>
</tr>
<tr>
<td>Job deletion</td>
<td>scancel [job id]</td>
</tr>
<tr>
<td>Job status (by job)</td>
<td>squeue [job id]</td>
</tr>
<tr>
<td>Job status (by user)</td>
<td>squeue –u [username]</td>
</tr>
<tr>
<td>Queue list</td>
<td>squeue</td>
</tr>
<tr>
<td>Node list</td>
<td>sinfo –N</td>
</tr>
<tr>
<td>Cluster status</td>
<td>sinfo</td>
</tr>
</tbody>
</table>

### Job Specifications

<table>
<thead>
<tr>
<th>Script directive</th>
<th>#SBATCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queue</td>
<td>-q [queue]</td>
</tr>
<tr>
<td>Node Count</td>
<td>-N [min[-max]]</td>
</tr>
<tr>
<td>CPU Count</td>
<td>-n [count]</td>
</tr>
<tr>
<td>Wall Clock Limit</td>
<td>-t [min] OR –t [days-hh:mm:ss]</td>
</tr>
<tr>
<td>Standard Out</td>
<td>-o [file name]</td>
</tr>
<tr>
<td>Standard Error</td>
<td>-e [file name]</td>
</tr>
<tr>
<td>Combine stdout/err</td>
<td>use –o without –e</td>
</tr>
<tr>
<td>Copy Environment</td>
<td>--export=[ALL</td>
</tr>
<tr>
<td>Event notification</td>
<td>--mail-type=[events]</td>
</tr>
<tr>
<td>Email address</td>
<td>--mail-user=[address]</td>
</tr>
<tr>
<td>Job name</td>
<td>--job-name=[name]</td>
</tr>
<tr>
<td>Memory size</td>
<td>--mem=[mem][M</td>
</tr>
<tr>
<td>Tasks per node</td>
<td>--tasks-per-node=[count]</td>
</tr>
<tr>
<td>CPUs per task</td>
<td>--cpus-per-task=[count]</td>
</tr>
</tbody>
</table>
Python

For the python portion of this course we will be using “Python for Biologists” by Dr. Martin Jones.