

# Unix and Bioinformatics version 2

Benjamin Tully and Ken Youens-Clark

## Abstract

This protocol details the use of various unix commands commonly used in bioinformatics.

**Citation:** Benjamin Tully and Ken Youens-Clark Unix and Bioinformatics. **protocols.io**

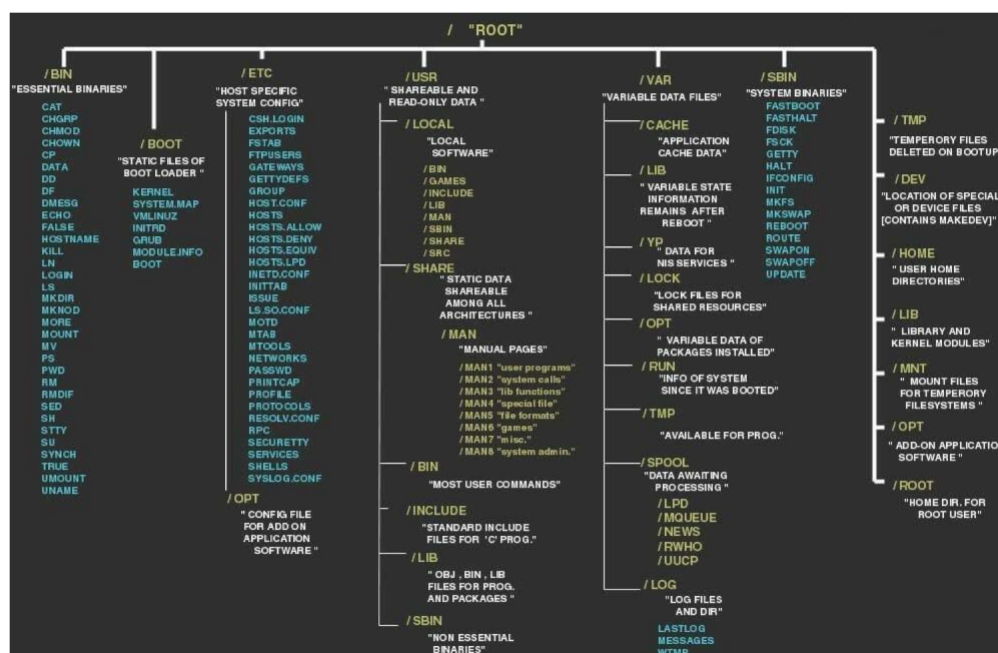
dx.doi.org/10.17504/protocols.io.ff7bjrn

**Published:** 08 Aug 2016

## Guidelines

### Unix Commands

pwd rm grep tail install  
ls '>' sed cut  
cd cat nano top  
mkdir '<' history screen  
touch '|' \$PATH ssh  
cp sort less df  
mv uniq head rsync/scp

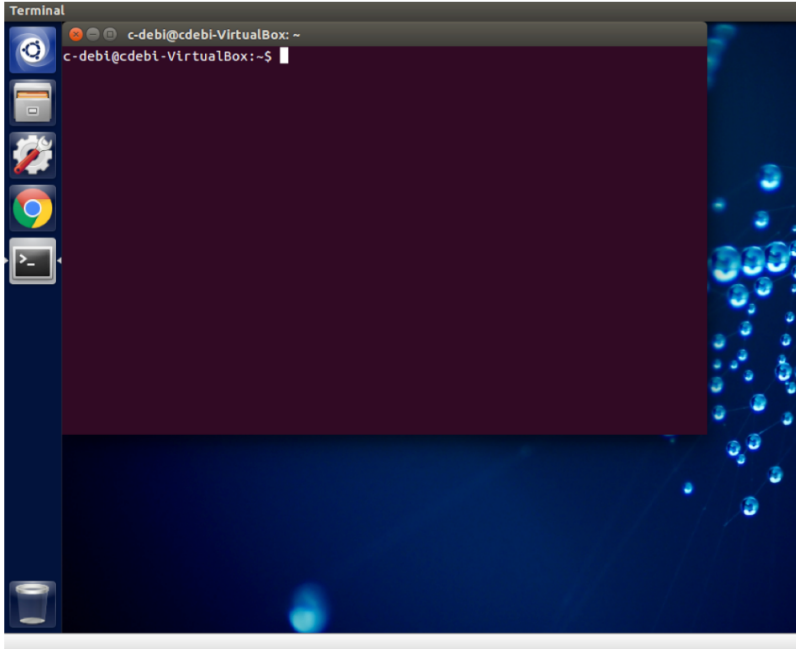


## Protocol

### The Start

#### Step 1.

Open terminal window



### The Start

#### Step 2.

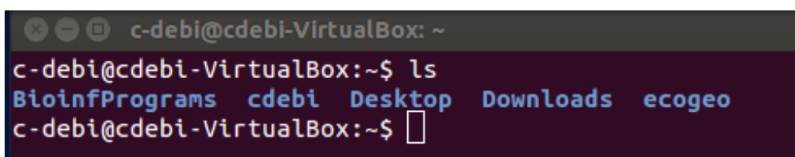
Use ls to list items in the current directory.

cmd **COMMAND**

ls

lists items in the current directory

📄 **EXPECTED RESULTS**



### The Start

#### Step 3.

Many commands have additional options that can be set by a '-'

cmd **COMMAND**

ls -a

ls -l

ls -lt

lists all files/directories, including hidden files '.' lists the long format lists the long format, but ordered by date last modified

📄 **EXPECTED RESULTS**

```

c-debi@cdebi-VirtualBox: ~
c-debi@cdebi-VirtualBox:~$ ls
BioinfPrograms cdebi Desktop Downloads ecogeo
c-debi@cdebi-VirtualBox:~$ ls -a
.
..
.bash_history
.bash_logout
.bashrc
BioinfPrograms
.biojs_templates
.cache
cdebi
.compiz
.com.zerog.registry.xml
.config
.dbus
.Dendroscope.def
.Desktop
Downloads
.ecogeo
.gconf
.gnome
.ICEauthority
.install4j
.InstallAnywhere
.jalview_properties
.java
.jswingreader
.kde
.local
.mozilla
.pki
.profile
.ssh
.vboxclient-clipboard.pid
.vboxclient-display.pid
.vboxclient-draganddrop.pid
.vboxclient-seamless.pid
.Xauthority
.xsession-errors
.xsession-errors.old
c-debi@cdebi-VirtualBox:~$ ls -l
total 20
drwxrwxr-x 28 c-debi c-debi 4096 Jul 17 22:13 BioinfPrograms
drwxrwxr-x 6 c-debi c-debi 4096 Dec 8 2015 cdebi
drwxr-xr-x 2 c-debi c-debi 4096 Jul 4 10:00 Desktop
drwxr-xr-x 7 c-debi c-debi 4096 Jul 17 22:14 Downloads
drwxrwxr-x 11 c-debi c-debi 4096 Jul 17 22:13 ecogeo
c-debi@cdebi-VirtualBox:~$ ls -lt
total 20
drwxr-xr-x 7 c-debi c-debi 4096 Jul 17 22:14 Downloads
drwxrwxr-x 11 c-debi c-debi 4096 Jul 17 22:13 ecogeo
drwxrwxr-x 28 c-debi c-debi 4096 Jul 17 22:13 BioinfPrograms
drwxr-xr-x 2 c-debi c-debi 4096 Jul 4 10:00 Desktop
drwxrwxr-x 6 c-debi c-debi 4096 Dec 8 2015 cdebi
c-debi@cdebi-VirtualBox:~$

```

## Directory System

### Step 4.

cd - change directory

cmd **COMMAND**  
cd ecogeo/

## Directory System

### Step 5.

List the contents of the current directory.

## Directory System

### Step 6.

Move into the directory called **unix**

## Directory System

### Step 7.

pwd (present working directory) can be used to show the current directory.

cmd **COMMAND**  
pwd  
prints the path to the current directory

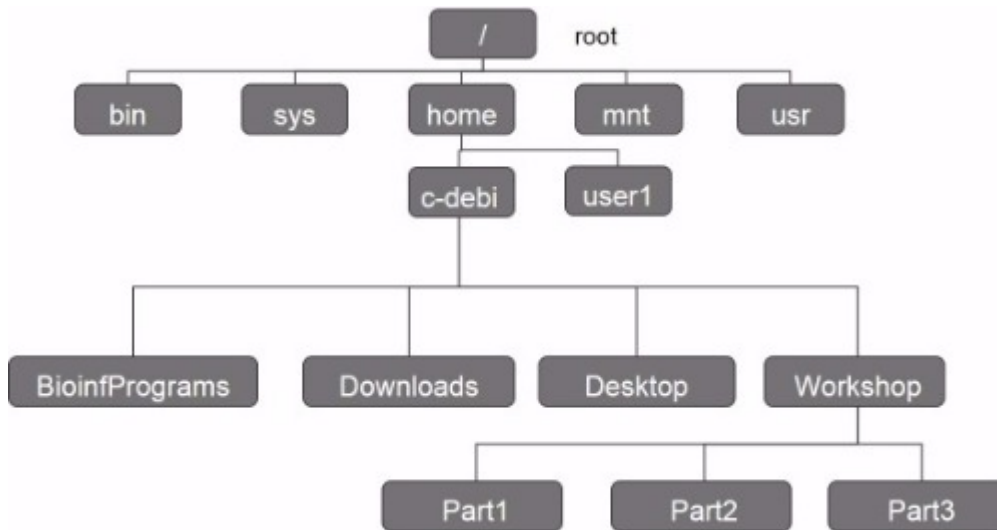
### EXPECTED RESULTS

cd /home/c-debi/ecogeo/unix

## Directory System

### Step 8.

Move to the root directory.



cmd **COMMAND**

cd /

### 📌 NOTES

**Elisha Wood-Charlson** 02 Aug 2016

This is where everything is stored in the computer. All the commands we are running live in /bin.

## Directory System

### Step 9.

Change directory to **home**

Change directory to **c-debi**

Change directory to **ecogeo**

Change directory to **unix**

List contents

Change directory to **data**

Change directory to **root**

### 📌 NOTES

**Elisha Wood-Charlson** 02 Aug 2016

Tabs can be used to auto complete names.

## Directory System

### Step 10.

Change directory to **unix/data** in one step

cmd **COMMAND**

\$ cd /home/c-debi/ecogeo/unix/data

## Directory System

### Step 11.

cd '..' allows you to step back up through the path directory. Display present working directory path.

cmd **COMMAND**

```
cd ..
```

```
pwd
```

moves back in the path directory

✓ **EXPECTED RESULTS**

```
/home/c-debi/ecogeo/unix
```

## Directory System

### Step 12.

Step back up to the c-debi directory.

## Directory System

### Step 13.

Change directory to BioinfPrograms

## Directory System

### Step 14.

List contents

✓ **EXPECTED RESULTS**

```
c-debi@cdebi-VirtualBox: ~/BioinfPrograms
c-debi@cdebi-VirtualBox:~/ecogeo/unix/data$ cd ..
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ cd ..
c-debi@cdebi-VirtualBox:~/ecogeo$ cd ..
c-debi@cdebi-VirtualBox:~$ pwd
/home/c-debi
c-debi@cdebi-VirtualBox:~$ ls
BioinfPrograms  cdebi  Desktop  Downloads  ecogeo
c-debi@cdebi-VirtualBox:~$ cd BioinfPrograms/
c-debi@cdebi-VirtualBox:~/BioinfPrograms$ ls
amos-2.0.8          FastQC          muscle
anvio-2.0.2         FastTree        ncbi-blast-2.2.31+
anvio-2.0.2.tar.gz  FigTree_v1.4.2  output.txt
anvi-ubuntu-setup.sh hmmer-3.1b2-linux-intel-x86_64 prodigal
AUTHORS            idba-1.1.1      README_IA
bin                include         rna_hmm3
bowtie-1.1.2       Jalview         samtools-1.2
building.html      jalview.jar     share
cutadapt           Jalview.lax     sickle
dendroscope        lax.jar         SPAdes-3.8.1-Linux
Dendroscope_unix_3_5_7.sh lib              THIRDPARTYLIBS
diamond            LICENSE         trimal
EMIRGE             megahit         Trimmomatic-0.35
ESOM               MetaRNA_to_FastQ.py Uninstall_Jalview
examples           mothur          usearch
c-debi@cdebi-VirtualBox:~/BioinfPrograms$
```

## Directory System

### Step 15.

Change directory to unix/

## Directory System

### Step 16.

Make a directory named "storage".

cmd **COMMAND**

```
mkdir storage
```

## Directory System

### Step 17.

List contents of directory.

## Directory System

### Step 18.

Move into the storage directory.

## Manipulating files

### Step 19.

The 'touch' command allows you to create a blank file of the input name.

```
cmd COMMAND
touch temp.txt
creates a blank file of the input name
```

## Manipulating files

### Step 20.

The 'cp' command allows you to copy a file and can be used to move a copy of a file to a directory.

```
cmd COMMAND
$ cp
```

## Manipulating files

### Step 21.

The 'mv' or move command "destroys" the original and places the content elsewhere.

```
cmd COMMAND
$ mv
```

## Manipulating files

### Step 22.

Using copy:

```
cmd COMMAND
$ cp temp.txt newtemp.txt
$ cp temp.txt ../
```

## Manipulating files

### Step 23.

Change directory up a level.

## Manipulating files

### Step 24.

List contents.

 EXPECTED RESULTS

```

c-debi@cdebi-VirtualBox: ~/ecogeo/unix
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ pwd
/home/c-debi/ecogeo/unix/storage
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ touch temp.txt
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ ls
temp.txt
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ cp temp.txt newtemp.txt
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ ls
newtemp.txt  temp.txt
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ cp temp.txt ../
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ cd ..
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ ls
data  storage  temp.txt
c-debi@cdebi-VirtualBox:~/ecogeo/unix$

```

## Manipulating files

### Step 25.

Change directory to storage.

## Manipulating files

### Step 26.

Utilize move command:

```

cmd COMMAND
$ mv newtemp.txt oldtemp.txt
$ mv oldtemp.txt /home/c-debi/ecogeo/unix/data

```

## Manipulating files

### Step 27.

Change directory to data, list content.

## Manipulating files

### Step 28.

List current working directory.

```

cmd COMMAND
/home/c-debi/ecogeo/unix/data

```

## Manipulating files

### Step 29.

The 'rm' remove command deleted a file PERMANENTLY

```

cmd COMMAND
rm oldtemp.txt

```

## Manipulating files

### Step 30.

Change directory to **storage**.

## Manipulating files

### Step 31.

Remove **temp.txt**

## Manipulating files

## Step 32.

Change directory to **unix**

### Manipulating files

## Step 33.

Remove storage directory:

cmd **COMMAND**  
\$ rm -r storage

### EXPECTED RESULTS

```
c-debi@cdebi-VirtualBox:~/ecogeo/unix/data$ cd ../storage/  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ ls  
temp.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ rm temp.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/storage$ cd ..  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ ls  
data storage temp.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ rm -r storage/  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ ls  
data temp.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$
```

### Manipulating files

## Step 34.

Create a directory called **bestdirectoryever**

Change directory to **bestdirectoryever**

Create a file called **glam.txt**

Change **glam.txt** to **formerglam.txt**

Remove **formerglam.txt**

Change directory to **unix**

Remove **bestdirectoryever**

### EXPECTED RESULTS

```
c-debi@cdebi-VirtualBox:~/ecogeo/unix  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ mkdir bestdirectoryever  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ cd bestdirectoryever/  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/bestdirectoryever$ touch glam.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/bestdirectoryever$ ls  
glam.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/bestdirectoryever$ mv glam.txt formerglam.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/bestdirectoryever$ ls  
formerglam.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/bestdirectoryever$ rm formerglam.txt  
c-debi@cdebi-VirtualBox:~/ecogeo/unix/bestdirectoryever$ cd ..  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ ls  
bestdirectoryever data  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ rm -r bestdirectoryever/  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$ ls  
data  
c-debi@cdebi-VirtualBox:~/ecogeo/unix$
```



## Manipulating files

### Step 35.

Change directory to data.

## Manipulating files

### Step 36.

List contents.

## Manipulating files

### Step 37.

Remove oldtemp.txt

## Manipulating files

### Step 38.

group12\_contigs.fasta

group20\_contigs.fasta

group24\_contigs.fasta

FASTA files - specific format

> Header line, contains ID and information about...

ATGATAGCTAGCAGCAGCTA[...] 80bp and then a newline.

## Looking at the contents of a file

### Step 39.

'head' will allow you to view the first 10 lines of a file.

```
cmd COMMAND
$ head [filename]
default displays the first 10 lines
```

## Looking at the contents of a file

### Step 40.

'tail' allows you to view the last 10 lines of a file.

```
cmd COMMAND
$ tail [filename]
default displays last 10 lines
```

## Looking at the contents of a file

### Step 41.

'less' allows you to scroll through a file using arrow keys or spacebar = advanced page | b = reverse page | q = quit

```
cmd COMMAND
$ less [filename]
```

Looking at the contents of a file

#### Step 42.

Use head to display the first 10 lines of **group12\_contigs.fasta**

Display the first 5 lines of **group12\_contigs.fasta**

Display the last 10 lines of **group12\_contigs.fasta**

Display the last 5 lines of **group12\_contigs.fasta**

Looking at the contents of a file

#### Step 43.

grep - file pattern searcher

```
cmd COMMAND  
$ grep
```

Looking at the contents of a file

#### Step 44.

wc - count the number of words, lines, characters

Looking at the contents of a file

#### Step 45.

Use grep on group12\_contigs.fasta

```
cmd COMMAND  
$ grep ">" group12_contigs.fasta  
stdout prints all matches of ">" in the file
```

Looking at the contents of a file

#### Step 46.

How many? Combine grep and wc?

Use the "|" (pipe) symbol

```
cmd COMMAND  
$ grep ">" group12_contigs.fasta | wc
```

Looking at the contents of a file

#### Step 47.

Repeat but add the option -l to wc

Looking at the contents of a file

#### Step 48.

Use the same technique to determine the number of sequences in **group20\_contigs.fasta**

Looking at the contents of a file

### Step 49.

What about the number of matches to “47” in **group12\_contigs.fasta**?

Or “\_47”?

cmd **COMMAND**

```
$ grep '>' group12_contigs.fasta | grep 47
```

📌 **NOTES**

**Elisha Wood-Charlson** 02 Aug 2016

```
grep '>' group12_contigs.fasta | grep 47
```

Looking at the contents of a file

### Step 50.

Redirecting output to file:

cmd **COMMAND**

```
$ grep ">" group12_contigs.fasta > group12_ids
```

'>' - redirects the output of STDOUT to a file

Looking at the contents of a file

### Step 51.

Look at the contents of **group12\_ids**

cmd **COMMAND**

```
$ grep "47" group12_contigs.fasta > group12_ids_with_47
```

Looking at the contents of a file

### Step 52.

cat - has multiple functions:

cmd **COMMAND**

```
$ cat group12_ids_with_47
```

With a single input - prints file contents

Looking at the contents of a file

### Step 53.

With '>' cat has the same function as cp

cmd **COMMAND**

```
$ cat group12_ids_with_47 > temp1_ids
```

```
$ cp group12_ids_with_47 temp2_ids
```

Looking at the contents of a file

### Step 54.

Double check to make sure **temp1\_ids = temp2\_ids**

Looking at the contents of a file

### Step 55.

Concatenate files with cat - most important function:

```
cmd COMMAND
$ cat temp1_ids temp2_ids > duplicate_ids
```

Looking at the contents of a file

### Step 56.

Check contents of duplicate\_ids using less or cat

Looking at the contents of a file

### Step 57.

Grab all of the contigs IDs from **group20\_contigs.fasta** that contain the number "51"

```
cmd COMMAND
$ grep 51 group20_contigs.fasta
```

Looking at the contents of a file

### Step 58.

Concatenate the new IDs to the duplicate\_ids file in a file called **multiple\_ids**

Looking at the contents of a file

### Step 59.

uniq - can be used to remove duplicates or identify lines with 1 occurrence or multiple occurrences

```
cmd COMMAND
$ uniq
```

Looking at the contents of a file

### Step 60.

sort - sort lines in a file alphanumerically

```
cmd COMMAND
$ sort
```

Looking at the contents of a file

### Step 61.

Compare **multiple\_ids** before and after uniq

```
cmd COMMAND
$ uniq multiple_ids
```

Looking at the contents of a file

### Step 62.

Why was there no change?

uniq has a weakness, can only identify duplicates in adjacent lines

cmd **COMMAND**

```
$ sort multiple_ids | uniq > clean_ids
```

\*\*note the version of sorting used by Unix

Looking at the contents of a file

### Step 63.

Clear all present files with temp in title

cmd **COMMAND**

```
$ rm temp*
```

'\*' - acts as a wildcard, so any file that starts with temp would be identified and removed, no matter the suffix

Looking at the contents of a file

### Step 64.

How do **temp1\_ids** & **temp2\_ids** compare?

cmd **COMMAND**

```
$ sort multiple_ids | uniq -d > temp1_ids
$ sort multiple_ids | uniq -u > temp2_ids
```

Looking at the contents of a file

### Step 65.

Identify duplicates:

cmd **COMMAND**

```
$ sort multiple_ids | uniq -d > temp1_ids
```

Uniq -d identifies only duplicates

Looking at the contents of a file

### Step 66.

Identify unique entries:

cmd **COMMAND**

```
$ sort multiple_ids | uniq -u > temp2_ids
```

Uniq -u identifies only unique entries

Looking at the contents of a file

### Step 67.

**temp1\_ids = group12\_ids\_with\_47 &**

**temp2\_ids = group20\_ids\_with\_51**

Looking at the contents of a file

### Step 68.

Remove all present files with temp in title

Looking at the contents of a file

### Step 69.

sed - modify files a file based on the issued commands

```
cmd COMMAND  
$ sed
```

Looking at the contents of a file

### Step 70.

Want a list of sequence IDs without the '>'?

```
cmd COMMAND  
$ sed 's/C/c/' clean_ids  
$ sed 's/_/./' clean_ids  
$ sed 's/>/' clean_ids > newclean_ids
```

### ■ ANNOTATIONS

**Elisha Wood-Charlson** 08 Aug 2016

```
sed 's/C/c/'
```

between the single quotes, **substitute** the occurrence of upper case C to lower case c

Looking at the contents of a file

### Step 71.

```
sed 's/C/c/'
```

between the single quotes, **substitute** the occurrence of upper case C to lower case c

Looking at the contents of a file

### Step 72.

seqmagick

Wrapper designed to utilize built in Biopython modules to manipulate and change FASTA files

Requires Biopython

<http://fhcrc.github.io/seqmagick/>

Looking at the contents of a file

### Step 73.

Discuss:

convert - produce a modified new file

mogrify - change the input file

info - present information of files in a directory

Additionally: backtrans-align, extract-ids, quality-filter, and primer-trim

```
cmd COMMAND
$ seqmagick
```

Looking at the contents of a file

## Step 74.

Execute seqmagick convert:

```
cmd COMMAND
$ seqmagick convert --include-from-
file newclean_ids group12_contigs.fasta newgroup12_contigs.fasta
```

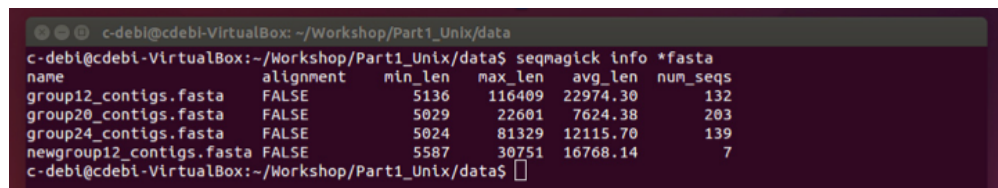
Looking at the contents of a file

## Step 75.

How many sequences are in **newgroup12\_contigs.fasta**? Using grep '>':

```
cmd COMMAND
$ seqmagick extract-ids newgroup12_contigs.fasta | wc
$ seqmagick info *fasta
```

### ✓ EXPECTED RESULTS



name	alignment	min_len	max_len	avg_len	num_seqs
group12_contigs.fasta	FALSE	5136	116409	22974.30	132
group20_contigs.fasta	FALSE	5029	22601	7624.38	203
group24_contigs.fasta	FALSE	5024	81329	12115.70	139
newgroup12_contigs.fasta	FALSE	5587	30751	16768.14	7

Looking at the contents of a file

## Step 76.

Store the information generated by 'seqmagick info' in a new file

### fasta\_info

```
cmd COMMAND
$ cut
$ cut -f 2 fasta_info
$ cut -f 2,4 fasta_info
$ cut -f 2-4 fasta_info
cut - pulling out columns from a table file -d allows for the assignment of the type of delimiter
between fields, if not TAB -f delineates which fields to preserve, starting at 1
```

## Some additional tools

### Step 77.

history - prints a sequential list of all commands in the current session

echo \$PATH - lists the directories for which the OS is checking for commands and data

## Some additional tools

### Step 78.

nano - in window text editor

cmd **COMMAND**

```
$ nano fasta_info
```

Additional text can be entered like any text editor To close out - Ctrl+X, hit 'Y', then ENTER Create a new file - nano and then enter file name after Ctrl+X

## Some additional tools

### Step 79.

Simple bash scripts: Text file with a list of commands that can be executed as a batch. Look at the contents of **simplebashscript**

## Some additional tools

### Step 80.

chmod - change file modes

cmd **COMMAND**

```
$ chmod 775 simplebashscript
```

📌 **NOTES**

**Elisha Wood-Charlson** 02 Aug 2016

```
chmod 755 simplebashscript
```

## Some additional tools

### Step 81.

Plain text file -> executable text file.

cmd **COMMAND**

```
$ ./simplebashscript
```

## Installations

### Step 82.

AMOS - a software infrastructure for developing assembly tools



Installation source: [http://amos.sourceforge.net/wiki/index.php/AMOS\\_Getting\\_Started](http://amos.sourceforge.net/wiki/index.php/AMOS_Getting_Started)