Unix and Bioinformatics

Benjamin Tully

Abstract

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

Citation: Benjamin Tully Unix and Bioinformatics. **protocols.io**
dx.doi.org/10.17504/protocols.io.eptbdnn
Published: 25 Jul 2016

Guidelines

Unix Commands

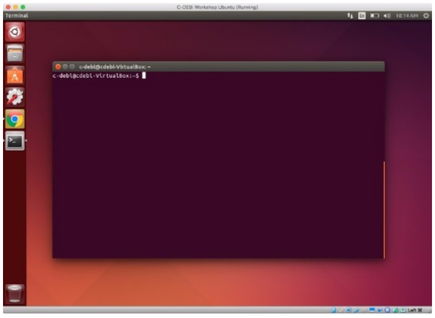
pwd
ls
cd
mkdir
touch
cp
mv
rm
'>'
cat
'<'
'|'
sort
uniq
grep
sed
nano
history
$PATH
less
head
tail
cut
top
screen
ssh
df
rsync/scp
install

Protocol

The Start

Step 1.

Open terminal window



```
c-debi@cdebi-VirtualBox:~$
```

The Start

Step 2.

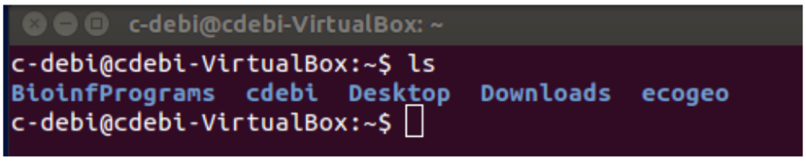
Use ls to list items in the current directory.

cmd COMMAND

```
ls
```

lists items in the current directory

EXPECTED RESULTS



```
c-debi@cdebi-VirtualBox:~$ ls
BioinfPrograms  cdebi  Desktop  Downloads  ecogeo
c-debi@cdebi-VirtualBox:~$ 
```

The Start

Step 3.

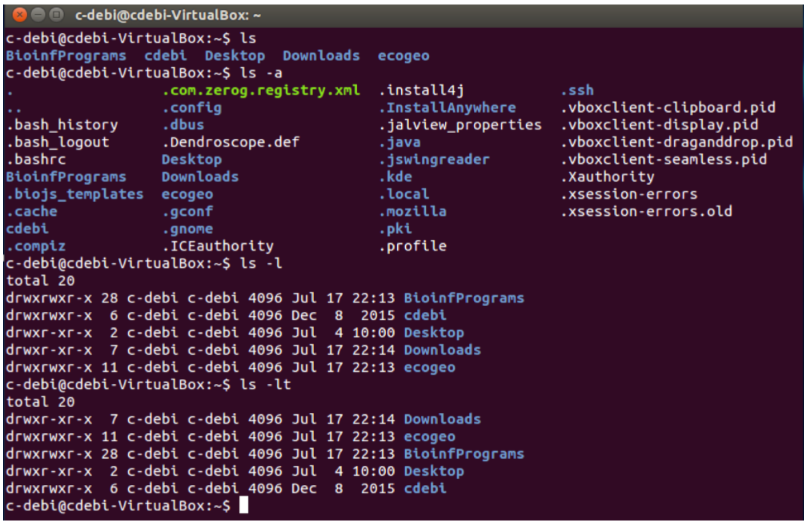
Many commands have additional options that can be set by a '-'

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:~$ ls
BioinfPrograms  cdebi  Desktop  Downloads  ecogeo
c-debi@cdebi-VirtualBox:~$ ls -a
.                  .com.zerog.registry.xml  .install4j           .ssh
..                 .config                  .InstallAnywhere     .vboxclient-clipboard.pid
.bash_history      .dbus                    .jalview_properties  .vboxclient-display.pid
.bash_logout       .Dendroscope.def         .java                .vboxclient-draganddrop.pid
.bashrc            Desktop                  .jswingreader        .vboxclient-seamless.pid
BioinfPrograms     Downloads                .kde                 .Xauthority
.biojs_templates   ecogeo                   .local               .xsession-errors
.cache             .gconf                   .mozilla             .xsession-errors.old
cdebi              .gnome                   .pki
.compiz            .ICEauthority            .profile
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

COMMAND

```
cd ecogeo/
```

Directory System

Step 5.

List the contents of the current directory.

Directory System

Step 6.

Move into the directory called Part1_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

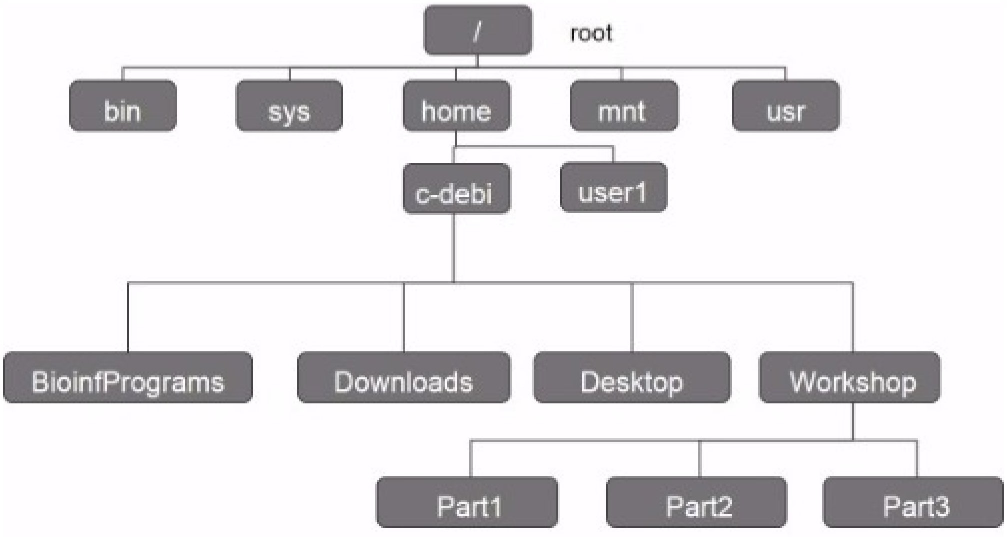
EXPECTED RESULTS

cd /home/c-debi/ecogeo/unix

Directory System

Step 8.

Move to the root directory.



cmd COMMAND

```
cd /
```

NOTES

James Thornton Jr 10 Mar 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

James Thornton Jr 10 Mar 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

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.

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.

List current working directory.

cmd COMMAND

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

Manipulating files

Step 28.

The 'rm' remove command deleted a file PERMANENTLY

cmd COMMAND

```
rm oldtemp.txt
```

Manipulating files

Step 29.

Change directory to storage.

Manipulating files

Step 30.

Remove temp.txt

Manipulating files

Step 31.

Change directory to unix

Manipulating files

Step 32.

Remove storage directory:

cmd COMMAND

```
$ rm -r storage
```

Manipulating files

Step 33.

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**

Manipulating files

Step 34.

Change directory to data.

Manipulating files

Step 35.

List contents.

Manipulating files

Step 36.

Remove oldtemp.txt

Manipulating files

Step 37.

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 38.

'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 39.

'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 40.

'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 41.

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 42.

grep - file pattern searcher

cmd COMMAND

```
$ grep
```

Looking at the contents of a file

Step 43.

wc - count the number of words, lines, characters

Looking at the contents of a file

Step 44.

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 45.

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 46.

Repeat but add the option -l to wc

Looking at the contents of a file

Step 47.

Use the same technique to determine the number of sequences in **group20_contigs.fasta**

Looking at the contents of a file

Step 48.

What about the number of matches to “47” in **group12_contigs.fasta**?

Or “_47”?

ANNOTATIONS

James Thornton Jr 25 Jul 2016

grep '>' group12_contigs.fasta | grep 47

Looking at the contents of a file

Step 49.

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 50.

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 51.

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 52.

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 53.

Double check to make sure **temp1_ids** = **temp2_ids**

Looking at the contents of a file

Step 54.

Concatenate files with cat - most important function:

cmd COMMAND

```
$ cat temp1_ids temp2_ids > duplicate_ids
```

Looking at the contents of a file

Step 55.

Check contents of duplicate_ids using less or cat

Looking at the contents of a file

Step 56.

Grab all of the contigs IDs from **group20_contigs.fasta** that contain the number “51”

ANNOTATIONS

James Thornton Jr 25 Jul 2016

grep 51 group20_contigs.fasta

Looking at the contents of a file

Step 57.

Concatenate the new IDs to the duplicate_ids file in a file called **multiple_ids**

Looking at the contents of a file

Step 58.

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 59.

sort - sort lines in a file alphanumerically

cmd COMMAND

```
$ sort
```

Looking at the contents of a file

Step 60.

Compare **multiple_ids** before and after uniq

cmd COMMAND

```
$ uniq multiple_ids
```

Looking at the contents of a file

Step 61.

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 62.

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 63.

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 64.

Identify duplicates:

cmd COMMAND

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

Uniq -d identifies only duplicates

Looking at the contents of a file

Step 65.

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 66.

temp1_ids = **group12_ids_with_47** &

temp2_ids = **group20_ids_with_51**

Looking at the contents of a file

Step 67.

Remove all present files with temp in title

Looking at the contents of a file

Step 68.

sed - modify files a file based on the issued commands

cmd COMMAND

```
$ sed
```

Looking at the contents of a file

Step 69.

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
```

Looking at the contents of a file

Step 70.

sed ‘s/C/c/’

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

Looking at the contents of a file

Step 71.

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 72.

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 73.

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

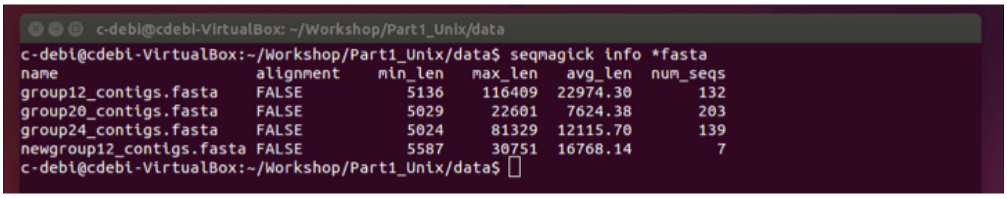
Step 74.

How many sequences are in **newgroup12_contigs.fasta**? Using grep '>':

cmd COMMAND

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

EXPECTED RESULTS



c-debi@cdebi-VirtualBox: ~/Workshop/Part1_Unix/data

```
c-debi@cdebi-VirtualBox:~/Workshop/Part1_Unix/data$ seqmagick info *fasta
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
c-debi@cdebi-VirtualBox:~/Workshop/Part1_Unix/data$ 
```

Looking at the contents of a file

Step 75.

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 76.

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 77.

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 78.

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 79.

chmod - change file modes

ANNOTATIONS

James Thornton Jr 25 Jul 2016

chmod 755 simplebashscript

Some additional tools

Step 80.

Plain text file -> executable text file.

cmd COMMAND

```
$ ./simplebashscript
```

Accessing a server

Step 81.

Logging in from the terminal:

cmd COMMAND

```
$ ssh -l USERNAME SERVERNAME.WEBADDRESS.EDU
$ ssh -l btully kuat.usc.edu
```

Accessing a server

Step 82.

Using top:

cmd COMMAND

```
$ top
```

Produces an active table of who is using the server, the number of CPUs in use and the amount of memory/RAM being utilized

Accessing a server

Step 83.

Produces a human-readable output of the storage space in use:

cmd COMMAND

```
$ df -h
$ du -h
```

Accessing a server

Step 84.

Using screen:

cmd COMMAND

```
$ screen
```

Creates an additional instance of the shell - that will not be disrupted if service is interrupted

Accessing a server

Step 85.

Detach from a screen instance - Ctrl+A, Crtl+D

cmd COMMAND

```
$ screen -ls
$ screen -r XXXX.pts-1.cdebi-VirtualBox
```

screen -a will reattach to a screen session

Accessing a server

Step 86.

Permanently end a screen session - type “exit” in the screen (The same command to log off the server)

Accessing a server

Step 87.

Kill a detached screen:

cmd COMMAND

```
$ screen -S XXXX.pts-1.cdebi-VirtualBox -X quit
```

Accessing a server

Step 88.

scp - secure copy

cmd COMMAND

```
$ scp filename.fasta btully@kuat.usc.edu://directory/destination
$ rsync
```

Accessing a server

Step 89.

rsync - transfers, making changes to existing files, maintains transfer if connection lost

cmd COMMAND

```
$ rsync filename.fasta btully@kuat.usc.edu://directory/destination
```

Installations

Step 90.

Easy installs:

1.Program manager - pip, apt-get, macports

2.Executables - mothur, Trimmomatic

3.From source

Hard installs:

1.Improperly annotated dependencies or prerequisites

Installations

Step 91.

AMOS - a software infrastructure for developing assembly tools

Installation source: http://amos.sourceforge.net/wiki/index.php/AMOS_Getting_Started

Installations

Step 92.

IDBA - iterative De Bruijn Graph De Novo Assembler for Short Reads Sequencing data with Highly Uneven Sequencing Depth

Installations

Step 93.

Change directory to **/home/c-debi/Downloads**

Installations

Step 94.

Move the compressed IDBA file to **/home/c-debi/BioinfPrograms**

Installations

Step 95.

Uncompress file:

cmd COMMAND

```
$ tar zxvf idba-1.1.1.tar.gz
```

Installations

Step 96.

Change directory to **/idba-1.1.1**

Installations

Step 97.

Examine the README contents

Installations

Step 98.

Modify a value in the file sequence.h in directory **/idba-1.1.1/src/sequence/**

Installations

Step 99.

Change directory back to **/idba-1.1.1**

Installations

Step 100.

Complete remaining install steps:

cmd COMMAND

```
$ ./configure
$ make
```

Installations

Step 101.

Want ease of access to programs by placing them OR linking them to **/usr/local/bin**

cmd COMMAND

```
$ sudo ln -s /home/c-debi/BioinfPrograms/idba-1.1.1/bin/idba /usr/local/bin/
```

Installations

Step 102.

Repeat for other files