

Jul 25, 2016

Version 1

Unix and Bioinformatics V.1

DOI

<https://dx.doi.org/10.17504/protocols.io.eptbdnn>

Benjamin Tully¹

¹EARTHCUBE OCEANOGRAPHY AND GEOBIOLOGY ENVIRONMENTAL 'OMICS

VERVE Net



James E Thornton Jr

Hurwitz Lab

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.eptbdnn>

Protocol Citation: Benjamin Tully 2016. Unix and Bioinformatics. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.eptbdnn>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working



Created: March 10, 2016

Last Modified: March 28, 2018

Protocol Integer ID: 2515

Keywords: bioinformatics this protocol, various unix command, use of various unix command, bioinformatics, protocol

Abstract

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

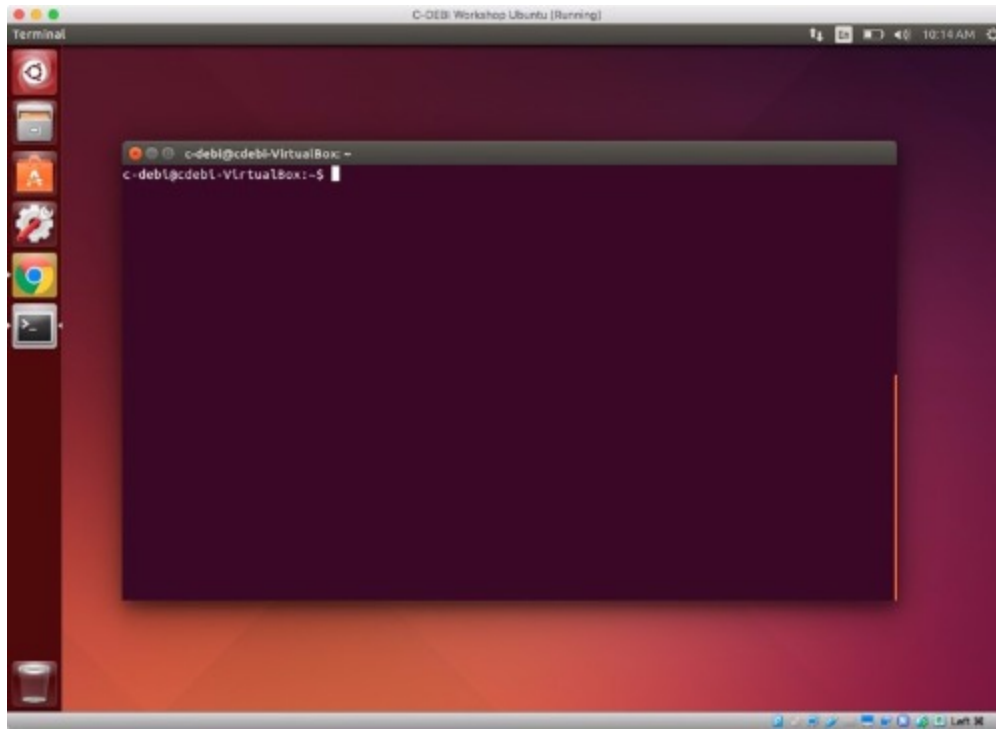
Guidelines

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	

The Start

- 1 Open terminal window



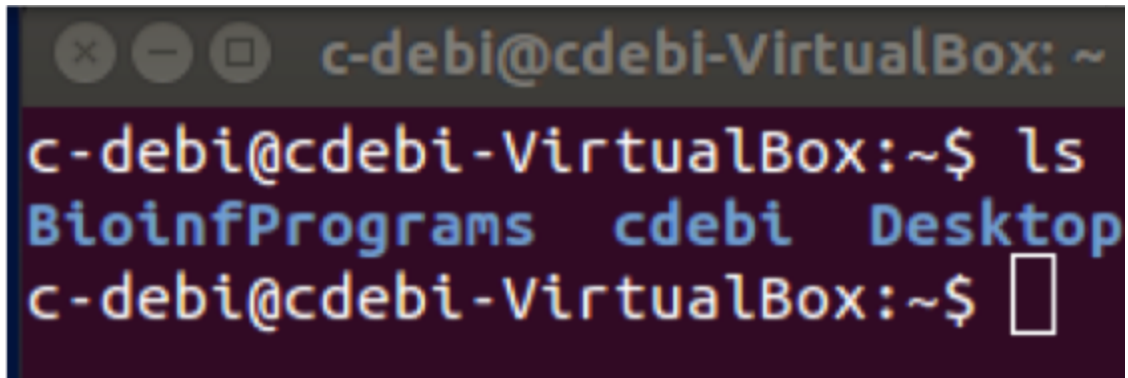
- 2 Use ls to list items in the current directory.

Command

lists items in the current directory

```
ls
```

Expected result



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

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

Command

lists all files/directories, including hidden files '.'

lists the long format

lists the long format, but ordered by date last modified

```
ls -a
```

```
ls -l
```

```
ls -lt
```



Expected result

```
c-debi@cdebi-VirtualBox: ~  
c-debi@cdebi-VirtualBox:~$ ls  
BioinfPrograms  cdebi  Desktop  Downloads  ecogeo  
c-debi@cdebi-VirtualBox:~$ ls -a  
.                .com.zerog.registry.xml  .install4j  
..               .config                  .Install4j  
.bash_history    .dbus                    .jalview_4  
.bash_logout     .Dendroscope.def         .java  
.bashrc          Desktop                  .jswingre  
BioinfPrograms  Downloads                .kde  
.biojs_templates ecogeo                   .local  
.cache           .gconf                   .mozilla  
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 Bioinf  
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 Downlo  
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 Downlo  
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 Bioinf  
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

4 cd - change directory

**Command**

```
cd ecogeo/
```

- 5 List the contents of the current directory.
- 6 Move into the directory called Part1_Unix
- 7 pwd (present working directory) can be used to show the current directory.

Command

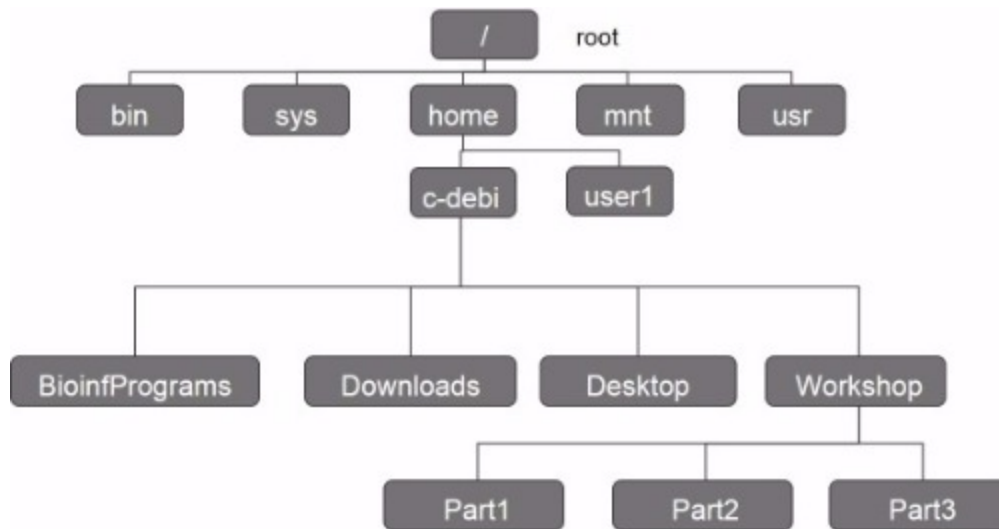
prints the path to the current directory

```
pwd
```

Expected result

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

- 8 Move to the root directory.



Note

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

Command

```
cd /
```

- 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**

Note

Tabs can be used to auto complete names.

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

**Command**

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

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

Command

moves back in the path directory

```
cd ..  
pwd
```

Expected result

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

- 12 Step back up to the c-debi directory.
- 13 Change directory to BioinfPrograms
- 14 List contents
- 15 Change directory to unix/
- 16 Make a directory named "storage".

**Command**

```
mkdir storage
```

17 List contents of directory.

18 Move into the storage directory.

Manipulating files

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

Command

creates a blank file of the input name

```
touch temp.txt
```

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

Command

```
$ cp
```

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

Command

```
$ mv
```

- 22 Using copy:

Command

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

- 23 Change directory up a level.

- 24 List contents.

- 25 Change directory to storage.

- 26 Utilize move command:

**Command**

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

27 List current working directory.

Command

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

28 The 'rm' remove command deleted a file PERMANENTLY

Command

```
rm oldtemp.txt
```

29 Change directory to storage.

30 Remove temp.txt

31 Change directory to unix



32 Remove storage directory:

Command

```
$ rm -r storage
```

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**

34 Change directory to data.

35 List contents.

36 Remove oldtemp.txt

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

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

Command

default displays the first 10 lines

```
$ head [filename]
```

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

Command

default displays last 10 lines

```
$ tail [filename]
```

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

Command

```
$ less [filename]
```

- 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**
- 42 grep - file pattern searcher



Command

```
$ grep
```

- 43 wc - count the number of words, lines, characters
- 44 Use grep on group12_contigs.fasta
- 45 How many? Combine grep and wc?
Use the "|" (pipe) symbol
- 46 Repeat but add the option -l to wc
- 47 Use the same technique to determine the number of sequences in **group20_contigs.fasta**
- 48 What about the number of matches to "47" in **group12_contigs.fasta**?
Or "_47"?
- 49 Redirecting output to file:
- 50 Look at the contents of **group12_ids**
- 51 cat - has multiple functions:

Command

With a single input - prints file contents

```
$ cat group12_ids_with_47
```

52 With '>' cat has the same function as cp

Command

```
$ cat group12_ids_with_47 > temp1_ids  
$ cp group12_ids_with_47 temp2_ids
```

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

54 Concatenate files with cat - most important function:

55 Check contents of duplicate_ids using less or cat

56 Grab all of the contigs IDs from **group20_contigs.fasta** that contain the number "51"

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

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



Command

```
$ uniq
```

59 sort - sort lines in a file alphanumerically

Command

```
$ sort
```

60 Compare **multiple_ids** before and after uniq

Command

```
$ uniq multiple_ids
```

61 Why was there no change?
uniq has a weakness, can only identify duplicates in adjacent lines

62 Clear all present files with temp in title

Command

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

```
$ rm temp*
```

63 How do **temp1_ids** & **temp2_ids** compare?

Command

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

64 Identify duplicates:

65 Identify unique entries:

66 **temp1_ids = group12_ids_with_47 &**
temp2_ids = group20_ids_with_51

67 Remove all present files with temp in title

68 sed - modify files a file based on the issued commands

Command

```
$ sed
```

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

Command

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

70 sed 's/C/c/'
between the single quotes, substitute the occurrence of upper case C to lower case c

71 seqmagick

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

Requires Biopython

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

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

Command

```
$ seqmagick
```

73 Execute seqmagick convert:

Command

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

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

Command

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



Expected result

```
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$
```

- 75 Store the information generated by 'seqmagick info' in a new file **fasta_info**

Command

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

```
$ cut
$ cut -f 2 fasta_info
$ cut -f 2,4 fasta_info
$ cut -f 2-4 fasta_info
```

Some additional tools

- 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
- 77 nano - in window text editor



Command

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

```
$ nano fasta_info
```

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

79 chmod - change file modes

80 Plain text file → executable text file.

Command

```
$ ./simplebashscript
```

Accessing a server

81 Logging in from the terminal:

**Command**

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

82 Using top:

Command

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

```
$ top
```

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

Command

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

84 Using screen:

Command

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

```
$ screen
```

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

Command

screen -a will reattach to a screen session

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

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

87 Kill a detached screen:

Command

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

88 scp - secure copy

**Command**

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

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

Command

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

Installations

- 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
- 91 AMOS - a software infrastructure for developing assembly tools
Installation source: http://amos.sourceforge.net/wiki/index.php/AMOS_Getting_Started
- 92 IDBA - iterative De Bruijn Graph De Novo Assembler for Short Reads Sequencing data with Highly Uneven Sequencing Depth
- 93 Change directory to **/home/c-debi/Downloads**
- 94 Move the compressed IDBA file to **/home/c-debi/BioinfPrograms**



95 Uncompress file:

Command

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

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

97 Examine the README contents

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

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

100 Complete remaining install steps:

Command

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

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



Command

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

102 Repeat for other files