Aug 08, 2016 Version 3

Unix and Bioinformatics V.3

DOI
dx.doi.org/10.17504/protocols.io.fitbken

Benjamin Tully and Ken Youens-Clark[1]

[1]EARTHCUBE OCEANOGRAPHY AND GEOBIOLOGY ENVIRONMENTAL 'OMICS

ECOGEO

Elisha M Wood-Charlson
KBase

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.fitbken

Protocol Citation: Benjamin Tully and Ken Youens-Clark 2016. Unix and Bioinformatics. **protocols.io https://dx.doi.org/10.17504/protocols.io.fitbken**

Protocol status: Working

Created: August 08, 2016

Last Modified: March 28, 2018

Protocol Integer ID: 3379

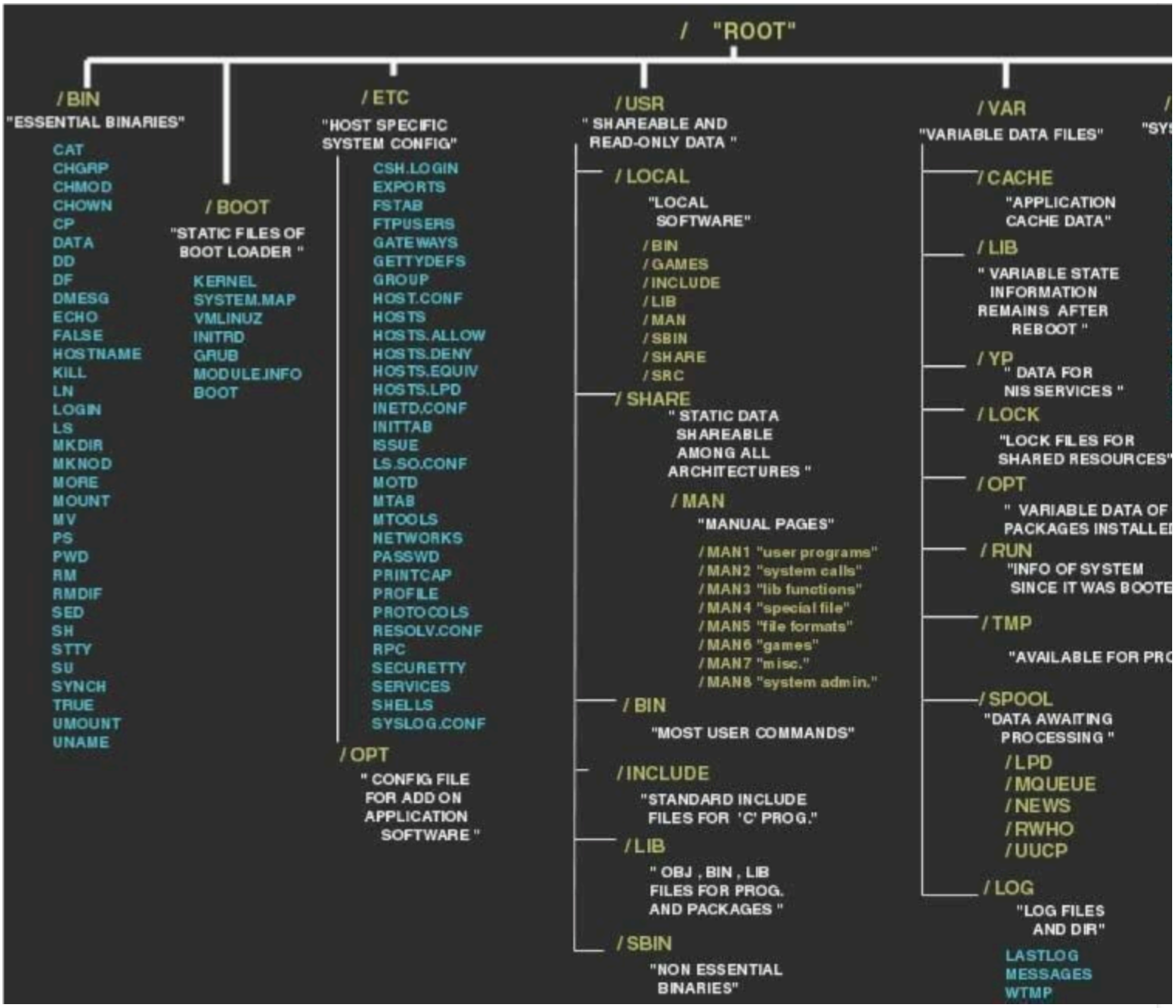
Abstract

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

Guidelines

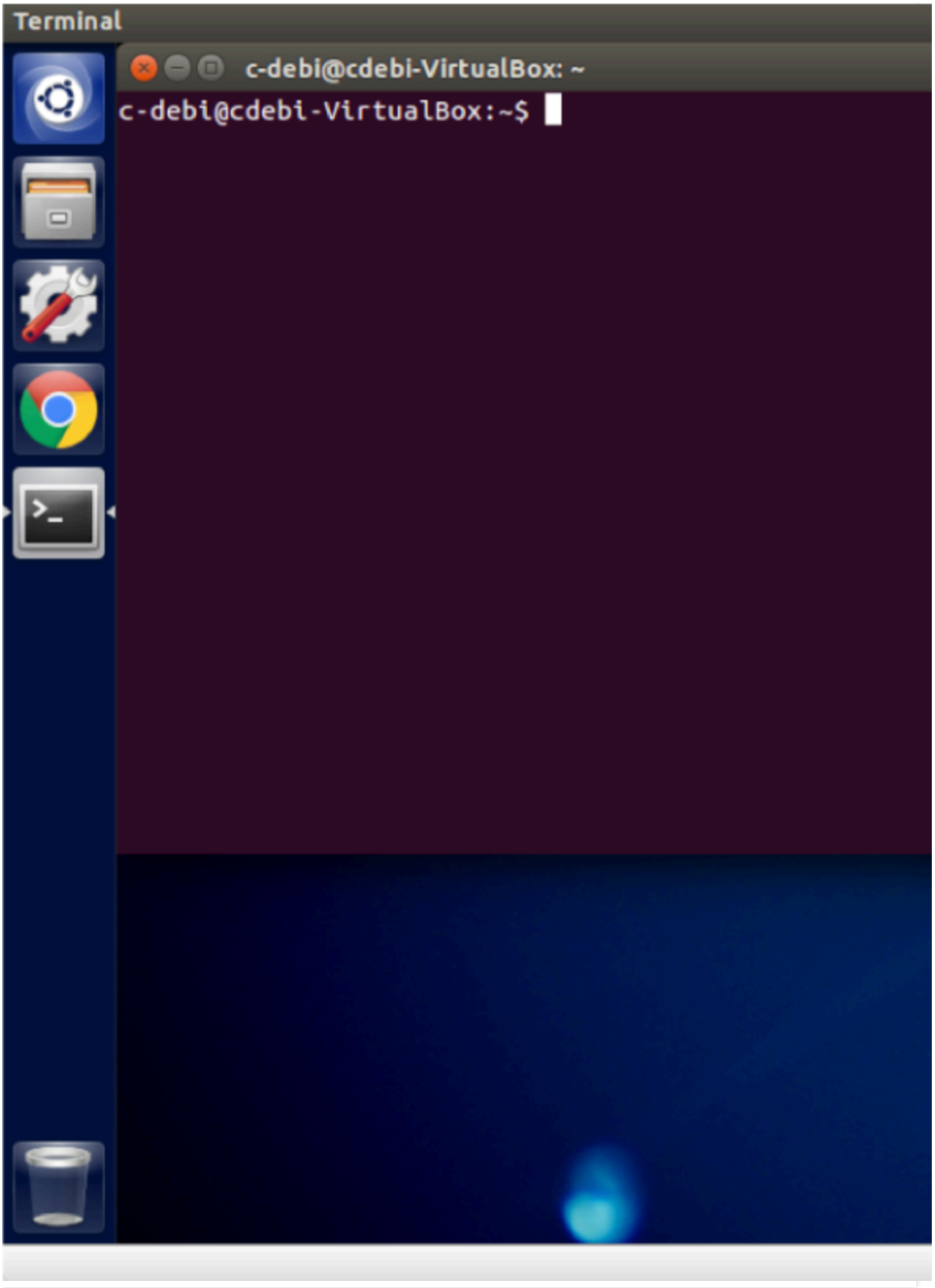
Unix Commands

	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	


/ "ROOT"
/ BIN
"ESSENTIAL BINARIES"
CAT
CHGRP
CHMOD
CHOWN
CP
DATA
DD
DF
DMESG
ECHO
FALSE
HOSTNAME
KILL
LN
LOGIN
LS
MKDIR
MKNOD
MORE
MOUNT
MV
PS
PWD
RM
RMDIF
SED
SH
STTY
SU
SYNCH
TRUE
UMOUNT
UNAME
/ BOOT
"STATIC FILES OF BOOT LOADER "
KERNEL
SYSTEM.MAP
VMLINUZ
INITRD
GRUB
MODULE.INFO
BOOT
/ ETC
"HOST SPECIFIC SYSTEM CONFIG"
CSH.LOGIN
EXPORTS
FSTAB
FTPUSERS
GATEWAYS
GETTYDEFS
GROUP
HOST.CONF
HOSTS
HOSTS.ALLOW
HOSTS.DENY
HOSTS.EQUIV
HOSTS.LPD
INETD.CONF
INITTAB
ISSUE
LS.SO.CONF
MOTD
MTAB
MTOOLS
NETWORKS
PASSWD
PRINTCAP
PROFILE
PROTOCOLS
RESOLV.CONF
RPC
SECURETTY
SERVICES
SHELLS
SYSLOG.CONF
/ OPT
" CONFIG FILE FOR ADD ON APPLICATION SOFTWARE "
/ USR
" SHAREABLE AND READ-ONLY DATA "
/ LOCAL
"LOCAL SOFTWARE"
/ BIN
/ GAMES
/ INCLUDE
/ LIB
/ MAN
/ SBIN
/ SHARE
/ SRC
/ SHARE
" STATIC DATA SHAREABLE AMONG ALL ARCHITECTURES "
/ MAN
"MANUAL PAGES"
/ MAN1 "user programs"
/ MAN2 "system calls"
/ MAN3 "lib functions"
/ MAN4 "special file"
/ MAN5 "file formats"
/ MAN6 "games"
/ MAN7 "misc."
/ MAN8 "system admin."
/ BIN
"MOST USER COMMANDS"
/ INCLUDE
"STANDARD INCLUDE FILES FOR 'C' PROG."
/ LIB
" OBJ , BIN , LIB FILES FOR PROG. AND PACKAGES "
/ SBIN
"NON ESSENTIAL BINARIES"
/ VAR
"VARIABLE DATA FILES"
/ CACHE
"APPLICATION CACHE DATA"
/ LIB
" VARIABLE STATE INFORMATION REMAINS AFTER REBOOT "
/ YP
" DATA FOR NIS SERVICES "
/ LOCK
"LOCK FILES FOR SHARED RESOURCES"
/ OPT
" VARIABLE DATA OF PACKAGES INSTALLE
/ RUN
"INFO OF SYSTEM SINCE IT WAS BOOTE
/ TMP
"AVAILABLE FOR PRO
/ SPOOL
"DATA AWAITING PROCESSING "
/ LPD
/ MQUEUE
/ NEWS
/ RWHO
/ UUCP
/ LOG
"LOG FILES AND DIR"
LASTLOG
MESSAGES
WTMP

The Start

1 Open terminal window



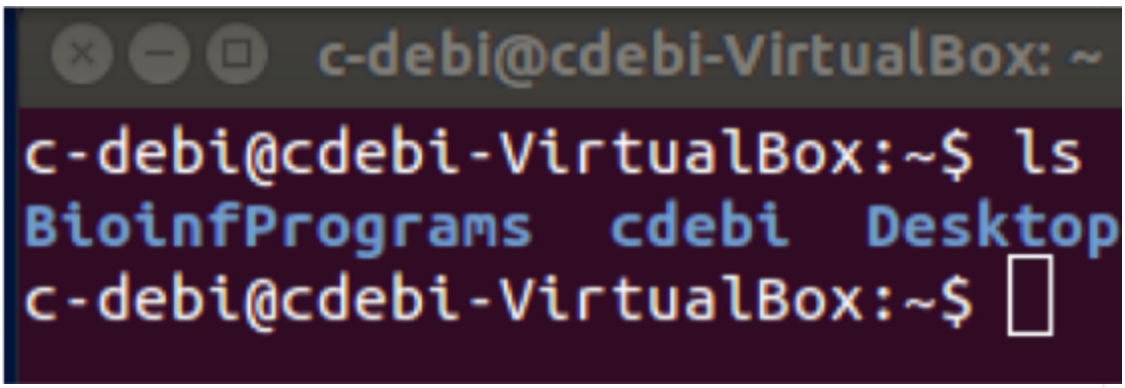
2 Use ls to list items in the current directory.

Command

lists items in the current directory

```
ls
```

Expected result



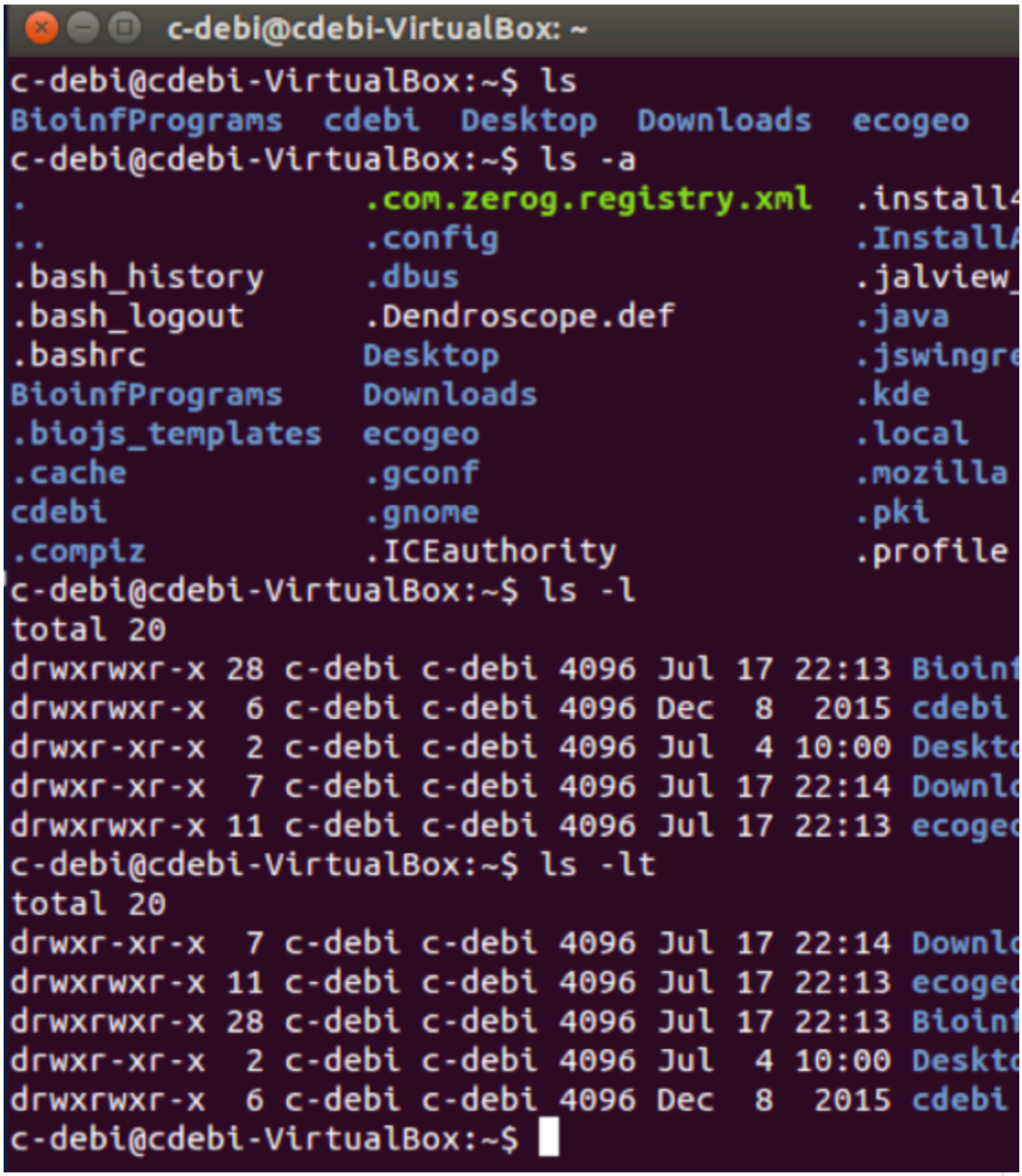
3 Many commands have additional options that can be set by a '-'

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



Directory System

4 cd - change directory

Command

```
cd ecogeo/
```

5 List the contents of the current directory.

6 Move into the directory called **unix**

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

Command

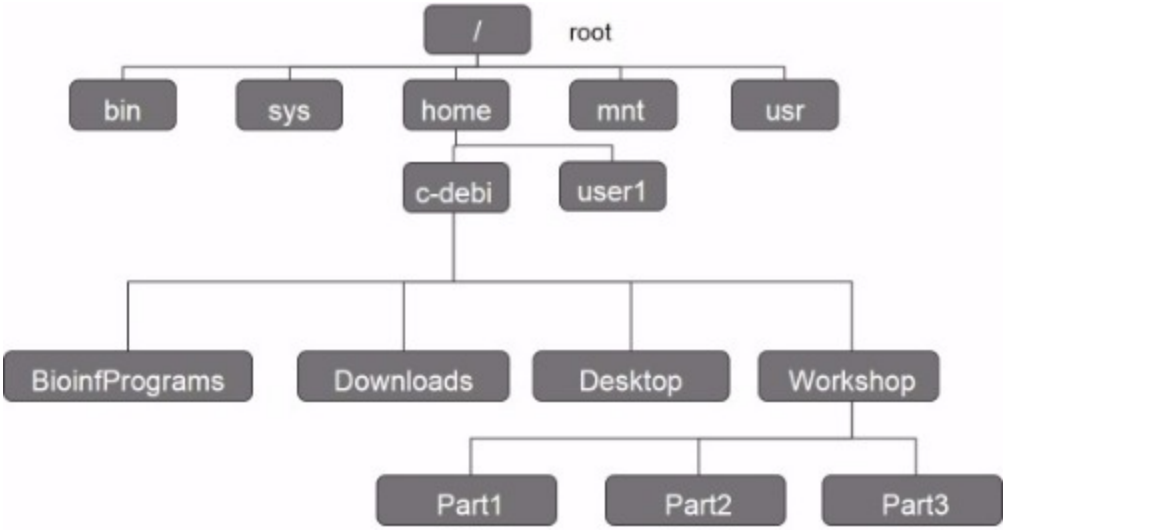
prints the path to the current directory

```
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

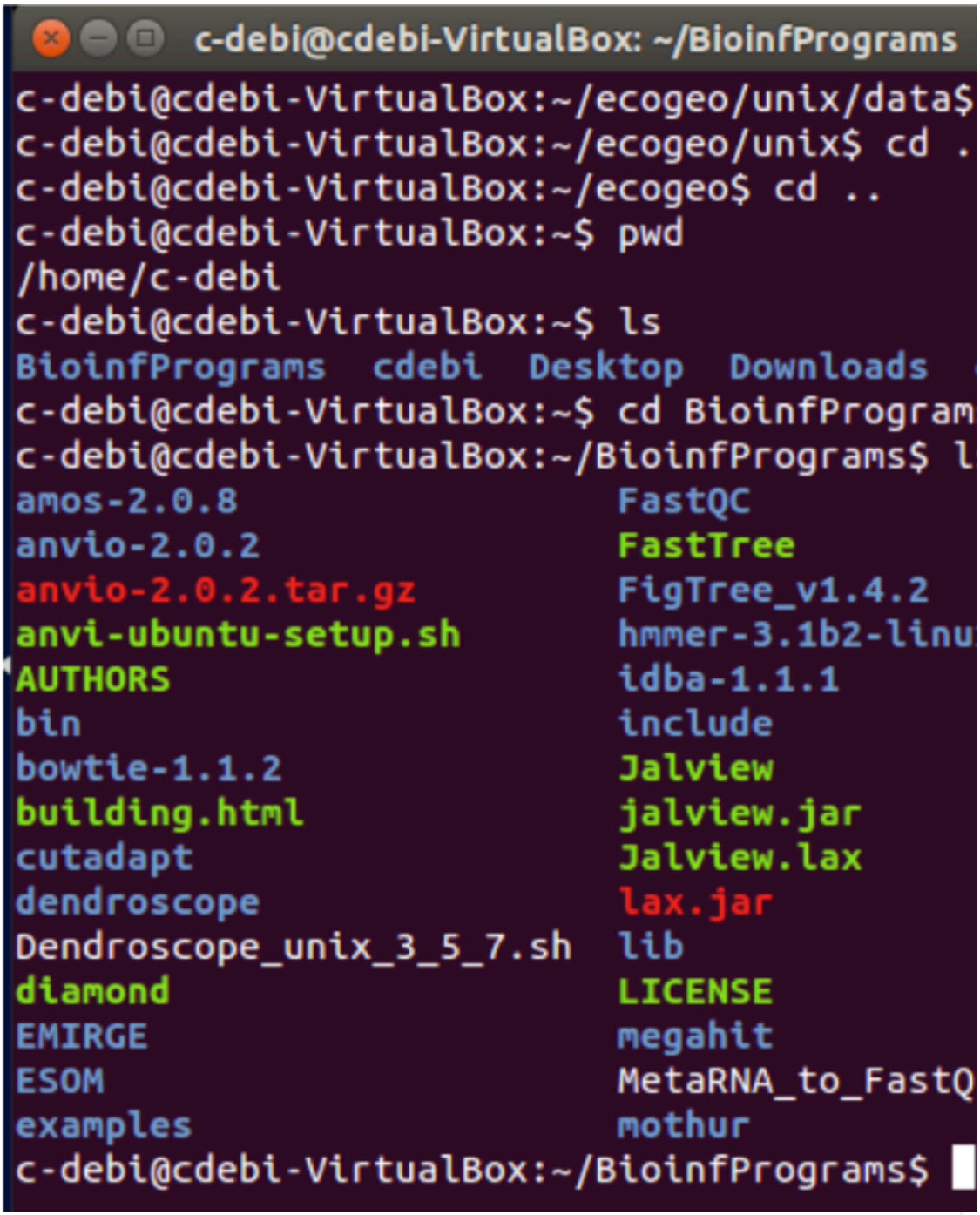
/home/c-debi/ecogeo/unix

12 Step back up to the c-debi directory.

13 Change directory to BioinfPrograms

14 List contents

Expected result



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:

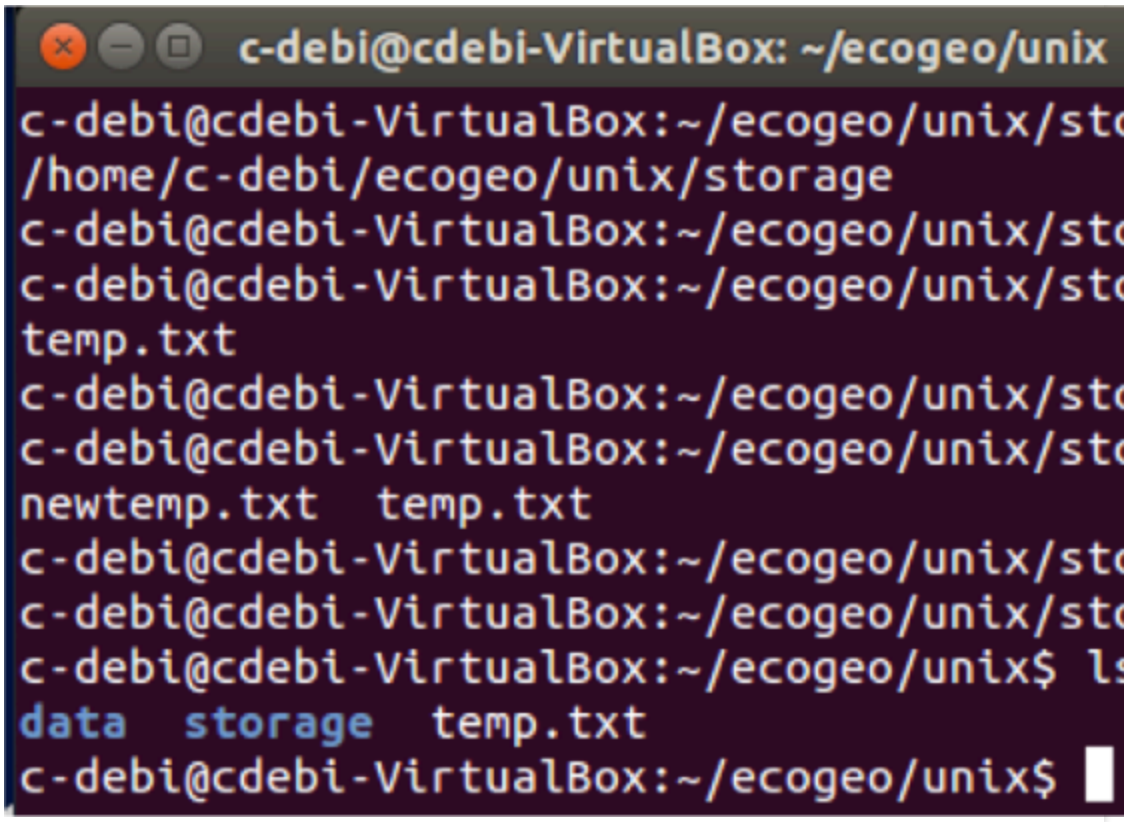
Command

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

23 Change directory up a level.

24 List contents.

Expected result



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 Change directory to data, list content.

28 List current working directory.

Command

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

29 The 'rm' remove command deleted a file PERMANENTLY

Command

```
rm oldtemp.txt
```

30 Change directory to **storage**.

31 Remove **temp.txt**

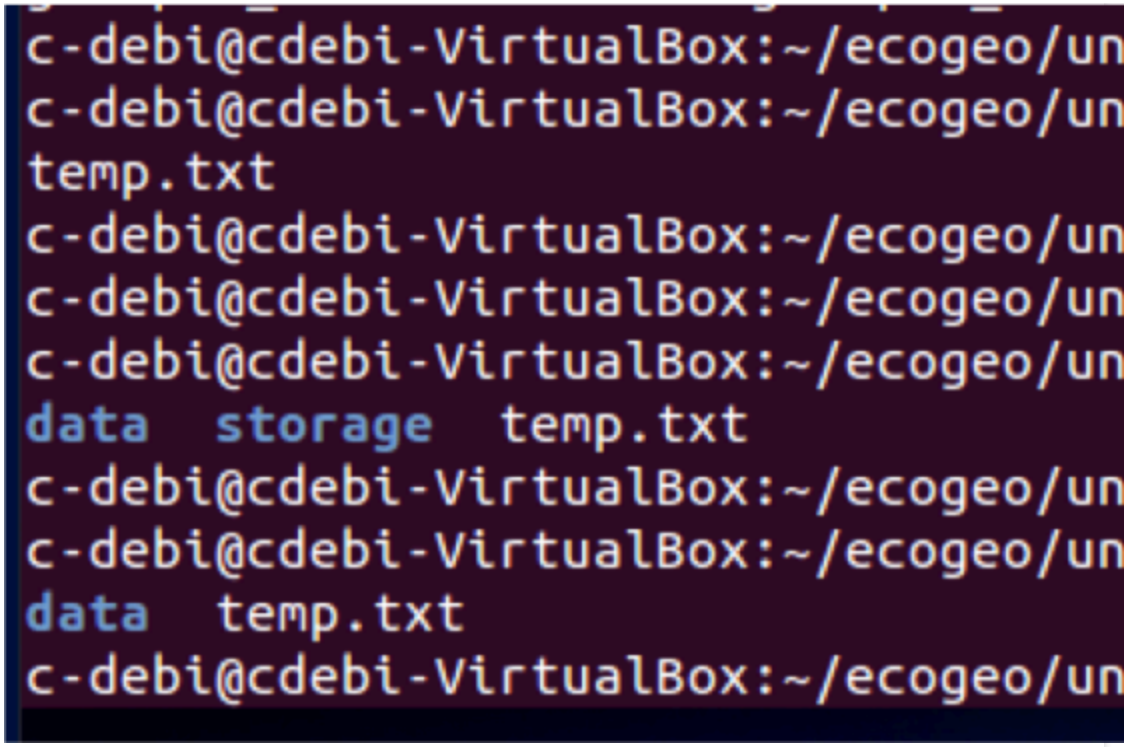
32 Change directory to **unix**

33 Remove storage directory:

Command

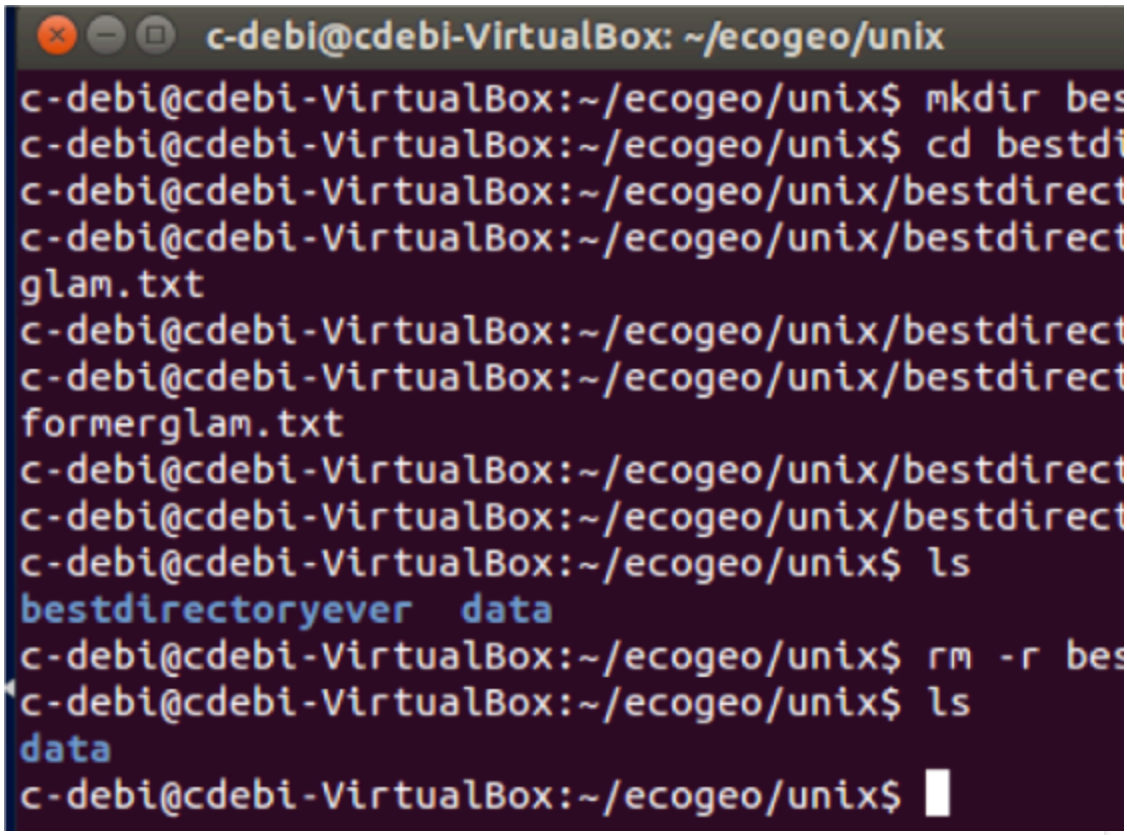
```
$ rm -r storage
```

Expected result



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 result



35 Change directory to data.

36 List contents.

37 Remove oldtemp.txt

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

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

Command

default displays the first 10 lines

```
$ head [filename]
```

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

Command

default displays last 10 lines

```
$ tail [filename]
```

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

Command

```
$ less [filename]
```

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**

43 grep - file pattern searcher

Command

```
$ grep
```

44 wc - count the number of words, lines, characters

45 Use grep on group12_contigs.fasta

46 How many? Combine grep and wc?

Use the "|" (pipe) symbol

47 Repeat but add the option -l to wc

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

49 What about the number of matches to “47” in **group12_contigs.fasta**?
Or “_47”?

Note

grep '>' group12_contigs.fasta | grep 47

50 Redirecting output to file:

51 Look at the contents of **group12_ids**

52 cat - has multiple functions:

Command

With a single input - prints file contents

```
$ cat group12_ids_with_47
```

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

Command

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

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

55 Concatenate files with cat - most important function:

56 Check contents of duplicate_ids using less or cat

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

Command

```
$ grep 51 group20_contigs.fasta
```

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

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

Command

```
$ uniq
```

60 sort - sort lines in a file alphanumerically

Command

```
$ sort
```

61 Compare **multiple_ids** before and after uniq

Command

```
$ uniq multiple_ids
```

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

63 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*
```

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

Command

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

65 Identify duplicates:

66 Identify unique entries:

67 **temp1_ids** = **group12_ids_with_47** &
temp2_ids = **group20_ids_with_51**

68 Remove all present files with temp in title

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

Command

```
$ sed
```

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

Note

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

Command

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

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

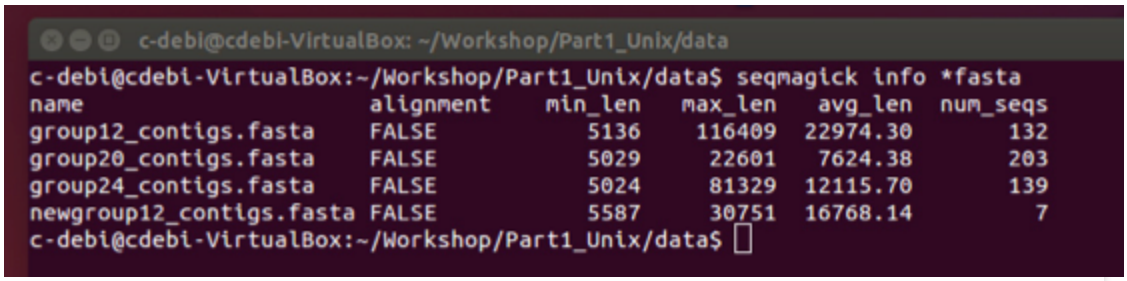
```
$ seqmagick convert --include-from-file newclean_ids
group12_contigs.fasta newgroup12_contigs.fasta
```

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

Command

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

Expected result



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

Note

chmod 755 simplebashscript

Command

```
$ chmod 775 simplebashscript
```

80 Plain text file → executable text file.

Command

```
$ ./simplebashscript
```