

# UNIX/LINUX INTRODUCTION

The purpose of this document is to provide a brief introduction to the LINUX/UNIX operating system in order to run a bioinformatics analysis on viral NGS data. UNIX/LINUX is the operating system of choice for a growing number of scientific applications, especially in genomics and bioinformatics. Note that most of the commands listed here work for both UNIX and Linux.

## Can I learn UNIX/LINUX?

---

- YES! Absolutely. It is just another way of operating your computer, no more difficult than learning Word, Excel or Powerpoint.
- Biggest difference:
  - In Linux: you type the command to execute
  - In Windows and Mac: you use your mouse to execute a command
- Remember: in the terminal, DON'T USE THE MOUSE!

## The Linux/UNIX operating system

---

Linux and UNIX are operating systems (OS), like Windows (for PCs), Mac OS (for Apples). An operating system allows users to issue commands to a computer without having to deal with the lowest-level machine language that the computer hardware actually uses. UNIX/Linux are machine independent, a major advantage compared to other operating systems.

UNIX was originally developed by AT&T Bell Laboratories in the late 1960s, and other versions have since been developed by computer scientists at UC Berkeley. Linux is an open-source OS based on Unix developed by Linus Torvalds in 1991.

## The Linux Environment

---

### The terminal

There are several "layers" in the Linux environment, from the "lowest" (farthest removed from the users), which is called the kernel, to the "highest", the graphical user interface (GUI). We

use terminal (command line interface) to interact with the OS. The terminal runs one of the "shells". Shell is a program that sits between the user and the kernel and translates user commands into machine code. The advantages of using command line are greater control and flexibility over the system or software and multiple commands can be saved in a file (a script) and executed as a program.

The most common shells are: - bourne shell - bourne again shell (a variant is called Z shell) - C shell - K shell

Among these bourne again shell (BASH) is the most popular one. This is the default shell on the system and we will be using this throughout the course.

## Connecting to Linux server

In this course we will be using Ubuntu Linux using the VM (Virtual Machine software). You just need to open "Terminal" and start to type!

If you are going to work on a remote server (i.e. university computer cluster) you can connect to it:

```
ssh user@ServerName  
ssh user@ipaddress
```

## Linux command structure

---

When you open a terminal, you will see command prompt ready to take commands. The default location on the terminal is your "home directory". It is represented with a ~ (tilde) symbol.

All Linux commands are single words, with optional parameters followed by arguments. Command options are also called "flag" and are single letters. For example, if I type:

```
ls -l software
```

I can see:

```
ls -l software
total 5416
drwxrwxr-x  4 ubuntu ubuntu    4096 Dec 10 10:18 SPAdes-3.13.0-Linux
drwxrwxr-x  7 ubuntu ubuntu    4096 Dec  9 16:24 Tablet
drwxr-xr-x  2 ubuntu ubuntu    4096 Feb 17  2015 Tanoti-1.2-Linux
drwxrwxr-x  4 ubuntu ubuntu    4096 Nov 19 10:59 TrimGalore-0.6.5
drwxrwxr-x  6 ubuntu ubuntu    4096 Dec 10 17:17 assembly-stats
drwxrwxr-x  9 ubuntu ubuntu   12288 Nov 14 00:04 bbmap
drwxrwxr-x 10 ubuntu ubuntu    4096 Nov 22 15:41 bcftools
drwxr-xr-x  5 ubuntu ubuntu    4096 Apr 17  2019 bowtie2-2.3.5.1-linux-
drwxrwxr-x  3 ubuntu ubuntu    4096 Dec  9 15:35 fastqc_v0.11.8
drwxrwxr-x  9 ubuntu ubuntu    4096 Dec  6 15:58 htsjdk
drwxrwxr-x  9 ubuntu ubuntu    4096 Nov 22 15:39 htlib
-rwxr-xr-x  1 ubuntu ubuntu 3799726 Jul  5  2014 kmc
-rwxr-xr-x  1 ubuntu ubuntu 1653250 Jul  5  2014 kmc_dump
drwxrwxr-x  4 ubuntu ubuntu    4096 Mar 12  2019 ncbi-blast-2.9.0+
drwxr-xr-x  2 ubuntu ubuntu    4096 Jan 10 11:32 picard-tools-1.119
drwxr-xr-x  7 ubuntu ubuntu    4096 Dec 10 16:39 quast-5.0.2
drwxrwxr-x 12 ubuntu ubuntu    4096 Nov 22 15:39 samtools
drwxrwxr-x  3 ubuntu ubuntu    4096 Dec 10 16:41 seqtk
drwxrwxr-x  6 ubuntu ubuntu    4096 Dec 10 14:21 shiver
drwxr-xr-x  2 ubuntu ubuntu    4096 Dec 10 16:29 smalt-0.7.6-bin
drwxrwxr-x  3 ubuntu ubuntu    4096 Dec  9 16:52 varscan
drwxrwxr-x  4 ubuntu ubuntu    4096 Dec 10 16:53 weeSAM
```

Here the "ls" is the command to list the contents of the directory, "-l" is the option for long listing and "software" is the input, which is optional in this case. Without the input, "ls" shows the contents of the current directory.

## First commands

### ls

List the information about the files/directories. Default is current directory.

Commonly used options are:

- -l long list
- -a show all files (including hidden files)
- -t sort based on last modified time

```
ls -lt
```

```
total 20
drwxrwxr-x  2 ubuntu ubuntu 4096 Jan 13 15:33 Course_Data
drwxrwxr-x  3 ubuntu ubuntu 4096 Dec 12 15:39 users
drwxrwxr-x 22 ubuntu ubuntu 4096 Dec 11 17:30 software
drwxrwxr-x  3 ubuntu ubuntu 4096 Dec 10 14:53 MEGA X
drwxrwxr-x  2 ubuntu ubuntu 4096 Nov 22 15:28 data
```

Left to right columns we have:

- file permissions
- number of links
- owner's name
- group's name
- number of bytes
- last modified time
- file/directory name

## man

the man command provides reference informations, for example about commands All Linux commands have manual pages. The manual page gives a detailed explanation of the command, all available options and sometimes, also provides examples.

```
man ls
```

## pwd

```
pwd
```

will return current directory name This command will tell you where you are

```
/home/ubuntu
```

## cd

change directory

```
cd software/
```

We are now in "software" directory. You can check that typing "pwd"

Command "cd .." takes you back. Entering "cd" command will bring you to home directory.

## mkdir

make directory This command creates a directory if no file/directory exists with that name.

```
mkdir Practice
```

```
ls -lt
```

```
total 24
drwxrwxr-x  2 ubuntu ubuntu 4096 Jan 16 11:01 Practice
drwxrwxr-x 22 ubuntu ubuntu 4096 Jan 16 11:01 software
drwxrwxr-x  2 ubuntu ubuntu 4096 Jan 13 15:33 Course_Data
drwxrwxr-x  3 ubuntu ubuntu 4096 Dec 12 15:39 users
drwxrwxr-x  3 ubuntu ubuntu 4096 Dec 10 14:53 MEGA X
drwxrwxr-x  2 ubuntu ubuntu 4096 Nov 22 15:28 data
```

## pwd

path of working directory

```
pwd
```

## rmdir

Remove directory This command removes an empty directory

```
rmdir Practice/
```

```
ls -lt
```

```
total 20
drwxrwxr-x 22 ubuntu ubuntu 4096 Jan 16 11:01 software
drwxrwxr-x  2 ubuntu ubuntu 4096 Jan 13 15:33 Course_Data
drwxrwxr-x  3 ubuntu ubuntu 4096 Dec 12 15:39 users
drwxrwxr-x  3 ubuntu ubuntu 4096 Dec 10 14:53 MEGA X
drwxrwxr-x  2 ubuntu ubuntu 4096 Nov 22 15:28 data
```

Let's create a directory for the Linux practical and work inside that directory

```
mkdir linux_play
```

```
cd linux_play
```

```
ls -lt
```

## touch

It is file's timestamp changing command. However, it can be used for creating an empty file. This command is generally used for checking write permission for the user.

```
touch temp-file
```

```
ls -lt
```

```
-rw-rw-r-- 1 ubuntu ubuntu 0 Jan 16 11:19 temp-file
```

## rm

remove rm is used for removing files and directories. Be careful!! There is no "Recycle bin" in Linux, there is no way to recover a file/directory once deleted.

```
rm temp-file  
ls -lt
```

## cp

Copy files or directories To copy directories, use "-r" options

```
touch temp1  
cp temp1 temp2  
ls -lt  
  
-rw-rw-r-- 1 ubuntu ubuntu 0 Jan 16 11:20 temp2  
-rw-rw-r-- 1 ubuntu ubuntu 0 Jan 16 11:20 temp1
```

## mv

to move file or directory to rename

```
mkdir new  
mv temp1 new/.  
mv temp2 Newfilename  
  
ls -lt new  
-rw-rw-r-- 1 ubuntu ubuntu 0 Jan 16 11:20 temp1  
  
ls -lt  
drwxrwxr-x 2 ubuntu ubuntu 4096 Jan 16 11:25 new  
-rw-rw-r-- 1 ubuntu ubuntu 0 Jan 16 11:20 Newfilename
```

In the first example "temp1" is moved to "new" directory. The "." at the end will retain the file name. In the second example temp2 is renamed to Newfilename

## File viewers

First we are going to copy a few files to our Linux training directory

```
cp ../Course_Data/All_HIV_Ref.fas All_HIV_Ref.fas
```

## cat

concatenate combines files and prints on the screen

```
cat All_HIV_Ref.fas
```

## pg/more/less

these commands are used for viewing the files

```
less All_HIV_Ref.fas
```

```
>Ref.01_AE.AF.07.569M.GQ477441
GATGGGTGCGAGAGCGTCAGTATTAAGTGGGGGAAAATTAGATGCATGGGAAAAAATTCG
GTTACGGCCAGGAGGAAAGAAAAAATATAAGATGAAACATTTAGTATGGGCAAGCAGAGA
GTTAGAAAAGATTTCGACTTAATCCTGGCCTATTAGAAACAGCAGAAGGATGTCAACAAAT
AATAGAACAGTTACAGTCAACTCTCAAGACAGGATCAGAAGAACTTAAATCCTTATATAA
TACAGTAGTAACCCTCTGGTGCGTACACCAAGGATAGAGGTAAAAGACACCAAGGAAGC
TTTAGATAAAATAGAGGAAGAACAAAAGAAGAGCCGGCAAAAGACACAGCAGGCAGCAGC
TGGCACAGGAAACAGCAGCCAAGCCAGCCAAAATTACCCTATAGTGCAAATGCACAAGG
GCAAATGGTACATCAGCCTTTATCACCTAGAACTTTGAATGCATGGGTGAAGGTAGTAGA
AGAAAAGGGTTTTAACCCAGAAGTAATATCCATGTTCTCAGCATTATCAGAGGGAGCCAC
CCCACAAGATTTAAATATGATGCTAAATATAGTAGGGGGACACCAGGCAGCAATGCAAAT
GTTAAAAGACACCATCAACGAGGAAGCTGCAGAATGGGATAGGACACATCCAGTACATGC
AGGGCCTATTCCACCAGGCCAAATGAGGGAACCAAGGGGAAGTGACATAGCAGGAACACTAC
TAGTACCCTTCAAGAACAAATAGGATGGATGACAAACACTCCTCCTATCCCAGTGGGAGA
AATCTATAAAAGGTGGATAATCCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTGT
```

Press "q" to come out from the program.

## head/tail

These commands show first and last 10 lines respectively from a file

```
tail All_HIV_Ref.fas
```

```
CACTCCAGGGCCAGGAGAAAGATTCCCCTGACCTTTGGGTGGCTGTTTAACTAGTACC
AGTCTCTGAGGTAGAAGTTGAGGAAATGGGAGATGAGCAGGAGAAAGCTAAGCTGCTACA
TCCAGCCTGCACCTTATGGGGATTCAGATCATCATAGGGAGATCTTAATGTGGAAGTTTGA
CAGATCACTGGGAGAAAGACATGTTGCCTTACAAAAGCACCCGGAAGTGTACTAAAGA
CTAAAGTGCTGACGCTACGCAGCTACTAAGTTGCTGACACTGCAGGGACTTTCCGGGGAC
GGAAAGTCCCGAGGGCGGAACAAGGGGAGGAGCAGGGGAGTGGTTTACCCTCAGAGCTG
CATATAAGCAGCTGCTTCACGCTTGACTGGGTCTCTGTGACAGACCAGATTAGAGCCTG
GGAGCTCTCTGGCCTTAAGCAGAACCCACTGCTTGGAAGCTCAATAAAGCTTGCTTGAG
TGAGTAAAGCAGTGTGTGCTCATCTGTTCAGAAGTCTGGTGTCTAGAGATCCCTCAGAGC
ACTTCTAGCCGAGTGGAATCTCTAGCA
```

## File editors

There are many non-graphical text editors like vi, emacs and nano available on Linux. Here we will be learning about nano editor.

Nano can be used to edit any kind of text file from the command line. All commands are executed through the use of the keyboard, using the key modifier. You need to hold down the

CTRL key, and then press the corresponding letter for each command you wish to use.

Useful commands:

Ctrl+X: Exit the editor. If you've edited text without saving, you'll be prompted as to whether you really want to exit.

Ctrl+O: Write (output) the current contents of the text buffer to a file. A filename prompt will appear; press Ctrl+T to open the file navigator shown above.

Ctrl+R: Read a text file into the current editing session. At the filename prompt, hit Ctrl+T: for the file navigator.

Ctrl+K: Cut a line into the clipboard. You can press this repeatedly to copy multiple lines, which are then stored as one chunk.

Ctrl+J: Justify (fill out) a paragraph of text. By default, this reflows text to match the width of the editing window.

Ctrl+U: Uncut text, or rather, paste it from the clipboard. Note that after a Justify operation, this turns into unjustify.

Ctrl+T: Check spelling.

Ctrl+W: Find a word or phrase. At the prompt, use the cursor keys to go through previous search terms, or hit Ctrl+R to move into replace mode. Alternatively you can hit Ctrl+T to go to a specific line.

Ctrl+C: Show current line number and file information.

Ctrl+G: Get help; this provides information on navigating through files and common keyboard commands

Try to create a new file and explore the commands above:

A terminal window with a blue title bar and a light gray border. Inside, the text 'nano file.tab' is displayed in a monospaced font.

## Commands for text processing

### grep

search input for a given pattern Some options are: • -A: after context • -B: before context • -C: before and after context • -c: count • -l: file with match • -i: ignore case • -o: only match • -v: invert match • -w: word match

to get all references used in *AllHIVRef.fas*

---



```
grep Ref All_HIV_Ref.fas
```

We can put everything in a new file with ">"

```
grep Ref All_HIV_Ref.fas > allreflist.tab
```

## cut

The cut command will cut a section from a file. To cut a section of a file use "-c" option.

```
cut -c 1-10 All_HIV_Ref.fas
```

What's the difference compared to:

```
cat All_HIV_Ref.fas
```

Useful flags are:

- -c: cut based on character position
- -d: cut based on delimiter
- -f: field number

To get only the genbank id from the list of references

```
cut -d "." -f6 allreflist.tab
```

Here "." is a delimiter and 6th field is our choice of interest. Next try to print the reference genotype

## sort

It is used for sorting input content Some options are: • -t: field separator • -n: numeric sort • -k: sort with a key (field) • -r: reverse sort • -u: print unique entries

## I/O control in Linux

When you run a command the output will be sent to standard out (stdout) ie. terminal.

However, we can send the stdout to a file using ">" redirection. This will redirect the standard output to a file.

```
ls > list  
cat list
```

```
All_HIV_Ref.fas  
Newfilename  
allreflist.tab  
file.tab  
list  
new
```

This will create a new file called list with all the file names in the directory. Remember, if there is a file exists with a name “list”, its content will be over- written by the stdout. Instead, we can append to a file using “>>” redirection.

## Pipes

Piping in Linux is very powerful and efficient way to combine commands. Pipes (|) in Linux act as connecting links between commands. Pipe makes a previous commands output as next commands input. We can nest as many commands as we want using pipes.

To print the sequence names of the first 10 sequence in AllHIVRef.fas

```
grep Ref All_HIV_Ref.fas | head
```

## If you have time..

---

You can test your Linux skills online at <https://cmdchallenge.com/>

## Resources

---

<https://seankross.com/the-unix-workbench/introduction.html>

<https://bioinformaticsworkbook.org/Appendix/Unix/unix-basics-1.html>