NEXT GENERATION SEQUENCING (NGS)

Introduction to NGS data analysis

Part I - Linux

Carina Silva

Learning outcomes

To learn the basics to understand and be able to conduct a standard NGS data analysis from scratch in a Linux environment.

To know and understand the different data analysis pipelines and formats (FASTQ, SAM/BAM, VCF).

To preprocess and perform QC of raw and processed data.

To learn and use tools to perform NGS data analysis and visualization.

Introduction to Linux/Unix Shell

Introduction to Linux shell

What is Unix/Linux

● GNU/Linux is a free and open-source operating system developed by thousands of contributors and led by Linus Torvalds since the beginning in 1991.

● Linux is widely used in research and super computers, more than 96% of super computers use Linux.

●It's an essential tool for bioinformatics and big data analysis and research.
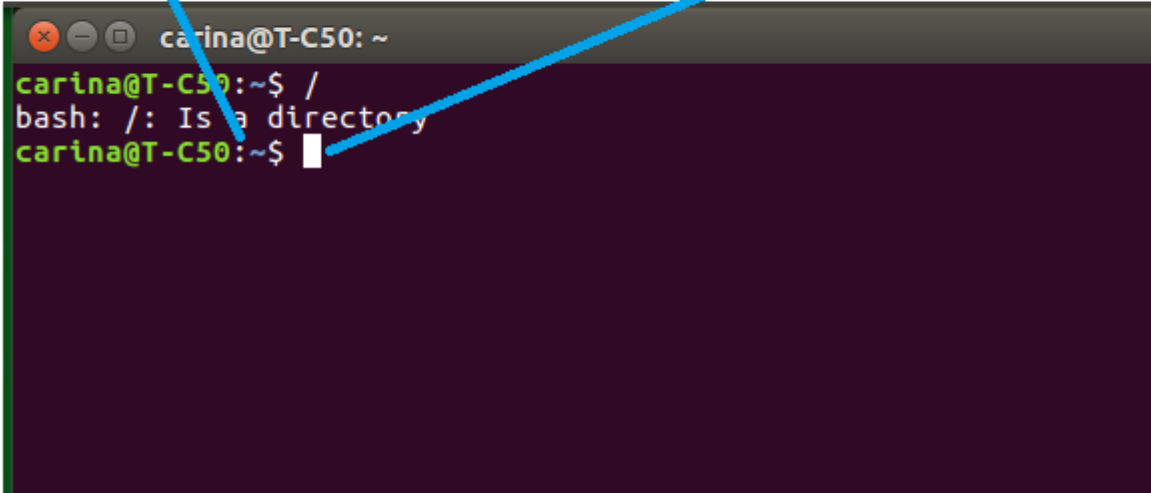
Introduction to Linux shell

Linux offers:

– Ease to program and work with big data

–Stability

–Security

–Low price

–High-performance computing

–...

●Usually visual graphical interfaces to software does not exist

What is the computer shell?

The shell is an interpreter (a program) that lets you interact with the Linux (operating system)

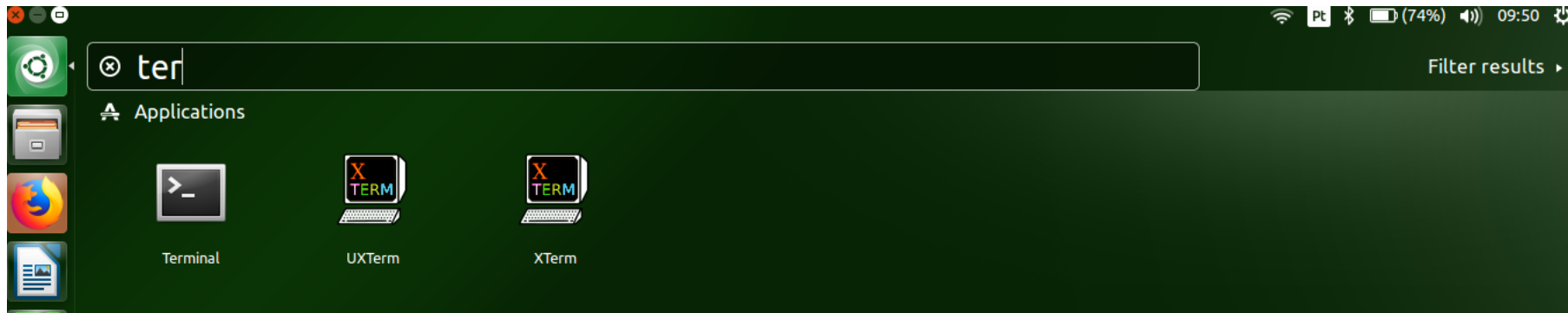The commands are themselves programs: when they terminate, the shell gives the user another prompt (~$ on our systems)

prompt

where you write your commands

```
carina@T-C50: ~
carina@T-C50:~$ /
bash: /: Is a directory
carina@T-C50:~$ █
```

6

Open a Linux terminal

Or just type in the keyboard   *crtl* + *alt* + T

Introduction to Linux shell

History

The shell keeps a list of the commands you have typed in. If you need to repeat a command, use the cursor keys to scroll up and down the list or type history for a list of previous commands.

Filename Completion

By typing part of the name of a command, filename or directory and pressing the [**Tab**] key, the shell will complete the rest of the name automatically. If the shell finds more than one name beginning with those letters you have typed, it will beep, prompting you to type a few more letters before pressing the tab key again.

Files and Processes

Everything in UNIX/LINUX is either a file or a process.

A process is an executing program identified by a unique PID (process identifier).

A file is a collection of data. They are created by users using text editors, running compilers etc.

9

The Directory Structure

All the files are grouped together in the directory structure. The file-system is arranged in a hierarchical structure, like an inverted tree. The top of the hierarchy is traditionally called **root** (written as a slash /).
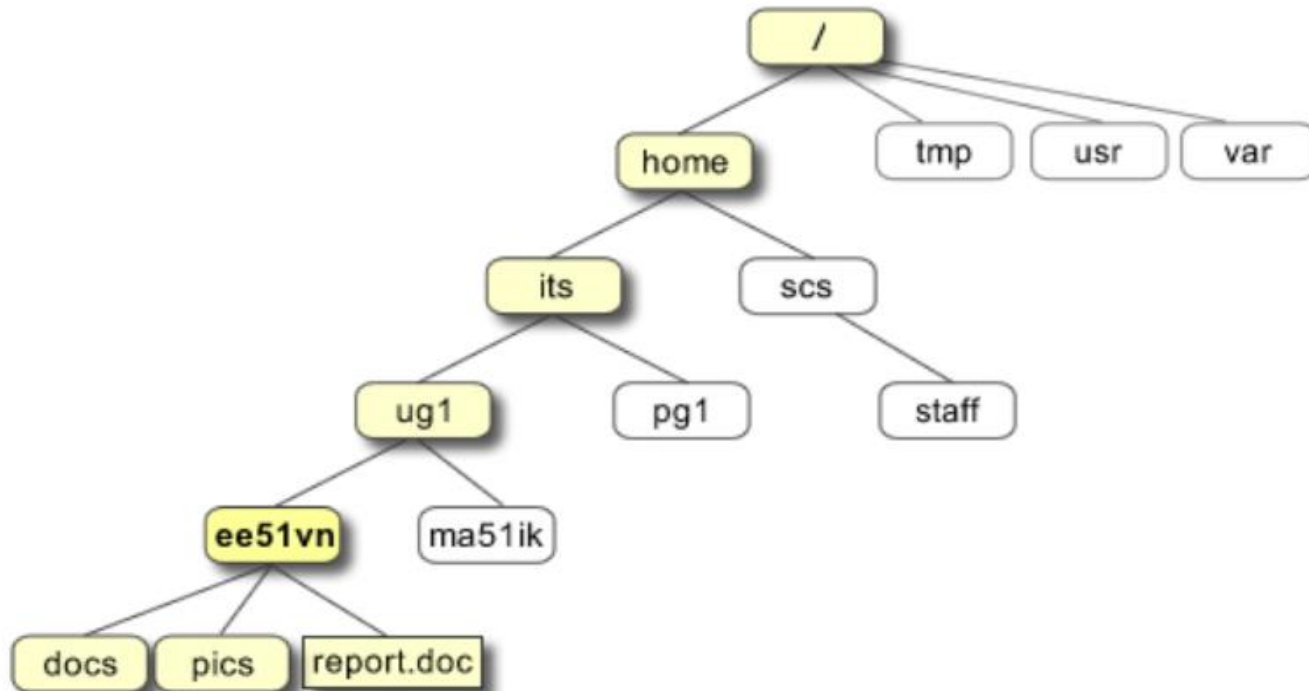
The full path to the file **report.doc** is
"/home/its/ug1/ee51vn/report.doc"

10

Task 1

How can you navigate the filesystem?

What do you need to navigate the filesystem?
Make a list of the actions needed to navigate
(example: change directory)

Solution

Find out where you are in the filesystem

Change directory

Find your way home

Identify the location of a file/directory

Move one directory up

Curso Introdutório Next Generation Sequenting

Task 2

What do you need to be able to do/manage stuff in the filesystem? (example: Make a new directory) Write a list of actions.

(*Tip Think of what you need in, e.g., Windows or Mac OSX)*

Introduction to Linux shell

Solution

Make a new directory

Remove a directory

Copy a file to another file

Rename a file/directory

Create a file

Open/close a file

Remove a file

Run programs

Listing files and directories – ls (list)

When you first login, your current working directory is your home directory.

To find out what is in your home directory, type

`~$ ls     [return]`

Making Directories - mkdir (make directory)

We will now make a subdirectory in your home directory to hold the files you will be creating and using in the course. To make a subdirectory called *yourname* in your current working directory type.

```
~$ mkdir yourname      [return]  #you will store all course
                                  #stuff in this directory
~$ ls [return]   # you will see your new directory
```

16

Changing to a different directory - cd (change directory)

The command cd *directory* means change the current working directory to *'directory'*. The current working directory may be thought of as the directory you are in, i.e. your current position in the file-system tree.

```
~$ cd yourname [return]
~/yourname$ ls [return]
```

Make another directory inside the directory *yourname* called *linuxstuff*

```
~/yourname$ mkdir linuxstuff [return]
```

Pathnames – pwd (print working directory)

To find out the absolute pathname of your home-directory, type `cd` to get back to your home-directory and then type

`~$ pwd [return]`

Task 3

Go to the url: `https://github.com/CarinaSilva/Curso-Int-NGS`

Download the zip file `Exercises.zip` who is inside the Linux folder to the directory `linuxstuff` that you created.

Unzip the Exercises.zip file

```
~/yourname/linuxstuff$ unzip Exercises.zip [return]
~/yourname/linuxstuff$ ls [return]
```
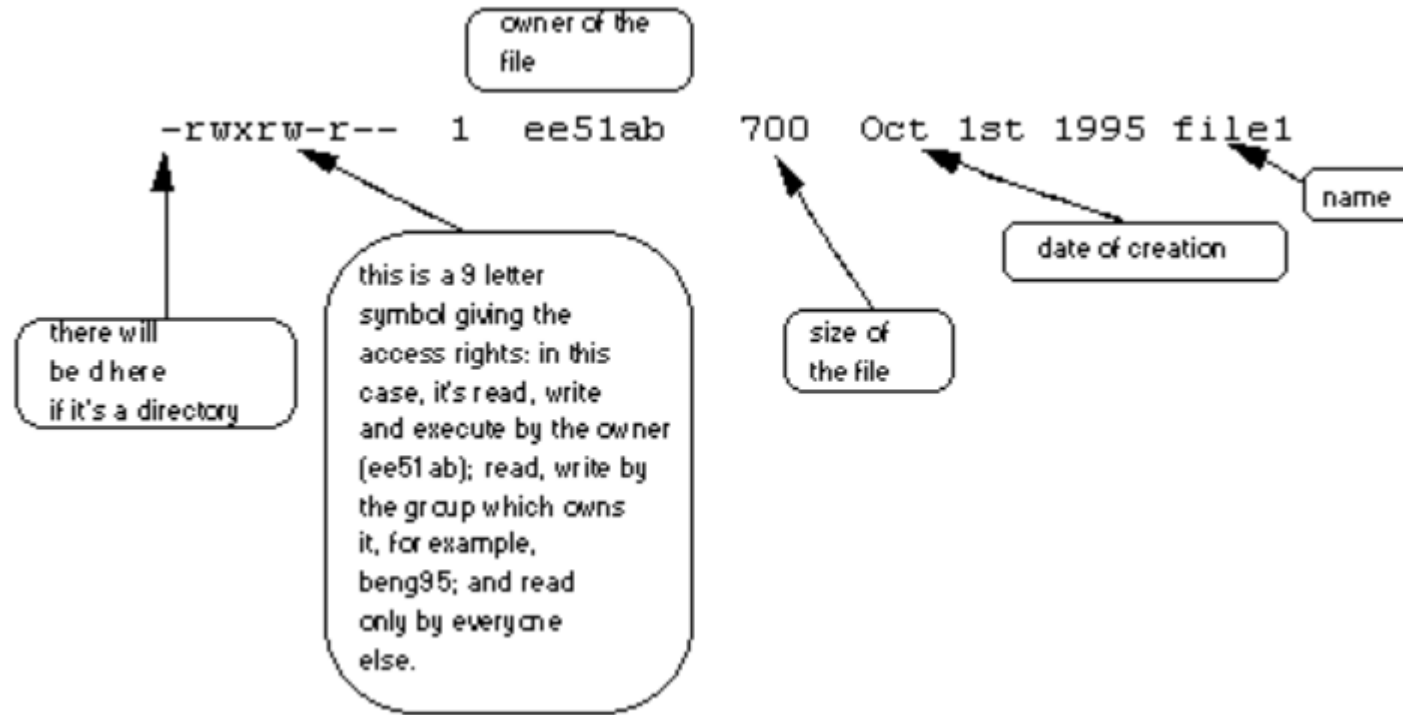
19

How can we run programs in Linux?

Prerequisites to run a program

1. The program must be somewhere on your computer.

2. The program must be executable.

Is my script executable? Each file (and directory) has associated access rights, which may be found by typing `ls -l`.

Introduction to Linux shell

owner of the file

```
-rwxrw-r--   1   ee51ab   700   Oct 1st 1995 file1
```

name

date of creation

there will be d here if it's a directory

this is a 9 letter symbol giving the access rights: in this case, it's read, write and execute by the owner (ee51ab); read, write by the group which owns it, for example, beng95; and read only by everyone else.

size of the file

Access rights on files.

r (or -), indicates read permission (or otherwise), that is, the presence or absence of permission to read and copy the file.

w (or -), indicates write permission (or otherwise), that is, the permission (or otherwise) to change a file.

x (or -), indicates execution permission (or otherwise), that is, the permission to execute a file, where appropriate.

Access rights on directories.

r allows users to list files in the directory;
w means that users may delete files from the directory or move files into it;
x means the right to access files in the directory. This implies that you may read files in the directory provided you have read permission on the individual files.

21

Some examples

-rwxrwxrwx	a file that everyone can read, write and execute (and delete).
-rw-------	a file that only the owner can read and write - no-one else can read or write and no-one has execution rights (e.g. your mailbox file).

Changing access rights – chmod (changing a file mode)

Only the owner of a file can use chmod to change the permissions of a file.

Symbol	Meaning
u	user
g	group
o	other
a	all
r	read
w	write (and delete)
x	execute (and access directory)
+	add permission
-	take away permission

For example:

~$ chmod go-rwx <filename> [return]

~$ chmod a+rw <filename> [return]

23

Task 5

Go to Git Hub:

https://github.com/CarinaSilva/Curso-Int-NGS/tree/master/Linux

Do the exercises in file `first-steps-on-the-linux-command-file.pdf`

Introduction to Linux shell

Some useful links

Unix Frequently Asked Questions (FAQs)

http://www.faqs.org/faqs/unix-faq/faq/part1/preamble.html

Ubuntu

https://www.ubuntu.com/

Ubuntu is based on the GNOME desktop environment, which is intended to provide a free, simple and intuitive interface while offering a full range of modern desktop applications.