
Sant'Anna
Scuola Universitaria Superiore Pisa
Hands-on
Svenja Mager
svenja.mager@santannapisa.it

About the Hands-on Content

1. Introduction to the Example Data

- What data we'll be using during the Pratical Part?

→ Example data from the hazelnut project

2. Introduction to Linux and the terminal

- Why do we use Linux and not Windows operating system (OS)?
- What is a Virtual Machine
- Linux OS
- Elements of a terminal
- Basic commands to navigate your computer via the terminal
- Bioinformatics tools
  - General command structure
  - Where to find information

About the Hands-on Content

3. Data Analysis

- Quality check on raw sequencing reads → fastqc
- Read alignment to a reference genome → bwa
- Processing of aligned reads (sam files) → samtools
- Variant call among a set of samples producing a vcf file → stacks
- Producing a neighborjoining tree file (Newick format) → tassel

1. Introduction to the Example Data

➢ What data will be use during the Pratical Part?

→ Example data from the hazelnut project

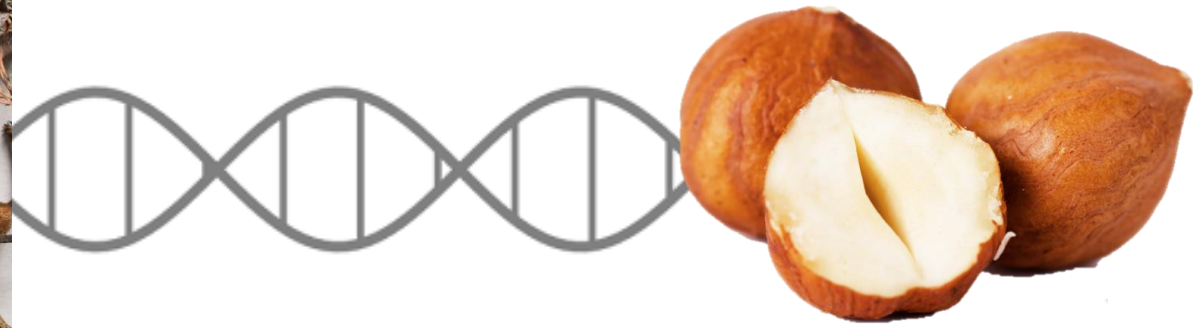
Remember our Hazelnuts Project!

We wanted to:

understand the genetic diversity to improve agronomic traits

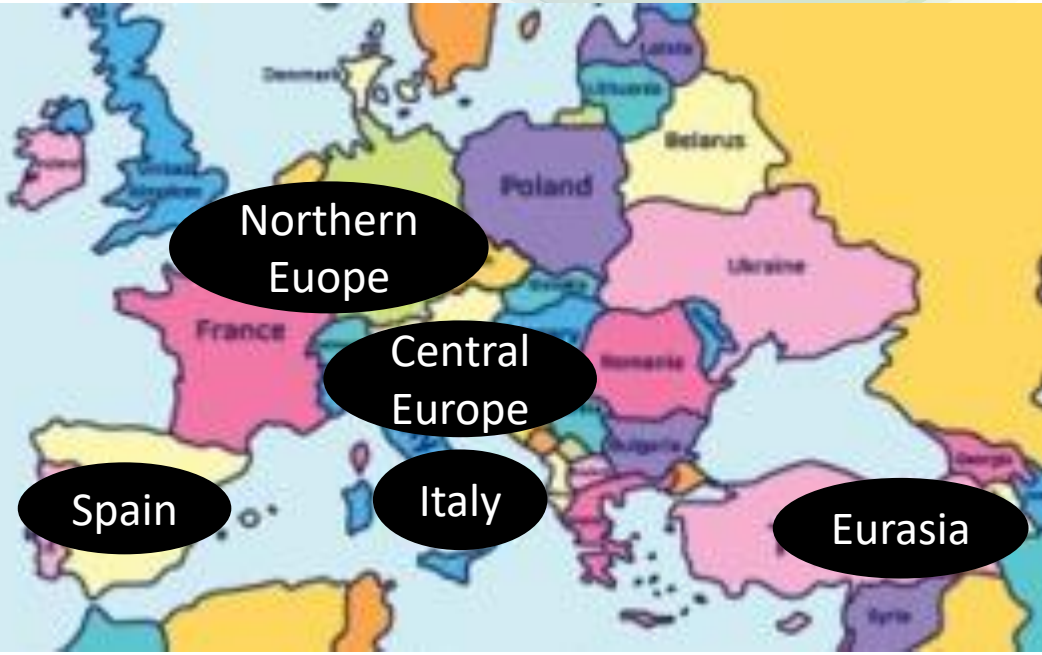
We based our analyses on:

- ❑ sequencing 141 hazelnut (*Corylus avellana*) samples from different countries in the world
- ❑ Determined differences (variants) in the DNA sequence among the samples → represented in VCF (**V**ariant **C**all **F**ormat) file



We‘ll use sequencing reads from only 6 samples and align them on chromosome 1 of our reference genome to reduce computing time

Sample1: Central Europe
Sample2: Eurasia
Sample3: Italy
Sample4: Northern Europe
Sample5: Spain
Sample6: United States

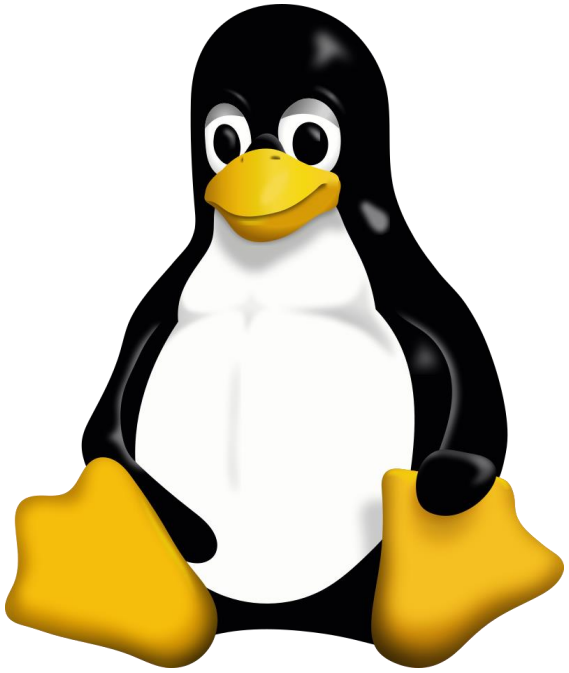


2. Introduction to Linux and the terminal

- Why do we use Linux and not Windows operating system (OS)?
- What is a Virtual Machine
- Linux OS
- Elements of a terminal
- Basic commands to navigate you computer via the terminal
- Bioinformatics tools
- Where to find information
  - General command structure

Why Linux?

- ❑ It is FREE!
- ❑ Resource efficiency (can run also on old hardware)
- ❑ More secure and private (less attacks from malware and viruses)
- ❑ Much more control over the system since open-source
  - More possibilities for programming and development than in windows (which offers few controllability)
  - Many programming languages and programs (also bioinformatics tools) were/are developed for Linux
  - No forced updates or disruptive restarts like in Windows → good for long-running tasks



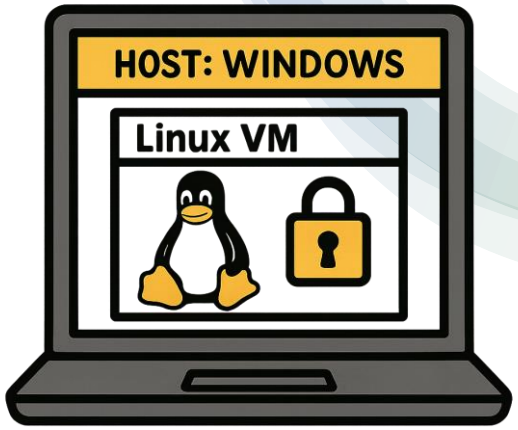
Then why is Windows used at all???

- ❑ Easier to install, often comes pre-installed on new-bought computers
- ❑ In some way easier to use: More intuitive and “beautiful“ user interface (especially in the past the Linux UI was not very user-friendly)
- ❑ Marketing (Microsoft makes much more publicity)
- ❑ Since most people use Windows, most “mainstream”/”general purpose” products target Windows
  - Many commercial Applications (Adobe Photoshop, MS Office, high-end video editors) are developed for Windows (for most programs there exist alternative freeware for Linux but sometimes less powerful or more difficult to learn and use)
  - Most Videogames run only on Windows (although more and more get available for Linux)

Windows 11

We will use Linux in Windows via a Virtual Machine Application

A virtual machine(VM) is software that acts like a separate DIGITAL computer—running its own operating system (e.g., Linux)—inside your existing PHYSICAL computer.



One of the most frequently used applications for running a VM is Oracle VirtualBox (your digital computer)

Then you need to download the OS that you want to install in the VM, e.g. Ubuntu
https://ubuntu.com/download/desktop

ubuntu

The Linux OS

- In the Ubuntu Loginscreen, you can login to your OS, it is like the start screen of windows where you enter your password

- Enter Password: **1234**

DO NOT REUSE!

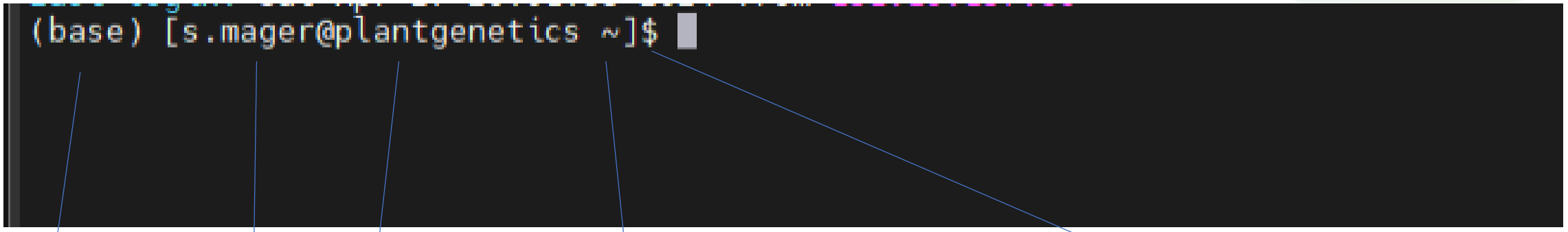
The Terminal

What is that?

A text-based way to interact with your computer
→ Instead of clicking, you write!

By writing into your terminal you can:
- navigate folders
- create folders or files
- copy/move/delete/modify files
- and much much more! For example analyse your data ☺

The Terminal



Currently activated conda environment

username

Computer-name

This is the current folder, the ~ stands for the home directory. It is the default directory in which you will be when opening a new terminal

The “$“ sign implicates that the terminal is ready to receive your command

Interacting with our terminal

Let’s write our first commands:

pwd (print working directory)

ls (list)

mkdir (make directory)

rmdir (remove directory)

Interacting with our terminal

Let’s create a file and modify it

touch test.txt (created the file test.txt)

echo “Hello World” (repeats (i.e. echoes) your input “Hello World”, printing it in the terminal itself)

echo “Hello World” > test.txt (repeats your input but instead of printing it directly into the terminal, it redirects it into the textfile test.txt)

echo “Hello Svenja” > test.txt
echo “What a lovely day” > test.txt

→ These commands overwrite the previous text in your test.txt

Interacting with our terminal

If we want to ADD lines instead of overwriting them, we can modify our command:

echo “Hello Svenja” >> test.txt

echo “What a lovely day” >> test.txt

Navigating with our terminal

To change the current folder that you are in, the command **cd** is your friend!
(cd = change directory)

Let’s enter into the Programs folder

```
cd /home/genomics/Programs/
```

And then go back to our Home folder

```
cd /home/genomics/
```

Navigating with our terminal

There are two ways of specifying the path to another folder:

Full paths and Relative paths

Full path (like we just used)

cd /home/genomics/Programs

You can copy the full path from the file browser (pressing Ctrl + l)

Relative path (Navigating to a subfolder or parentfolder relative to the current directory)

cd ./Programs

The "." tells the terminal that the specified path is relative to the the current working directory; after the slash you therefore have to specify a folder that is INSIDE the current folder

Navigating to parent folder relative to current directory:

cd ../

The first "." tells the terminal that the specified path is relative to the the current working directory!
The second "." tells the terminal that you want to go back one step, to the parent folder (only one direct parentfolder, therefore no need to specify its name)

Using Programs with the Terminal

We can not only navigate, create or modify folders and files, but also run programs and analyse data

→ In windows, when wanting to run a program, you usually download it, install it and then **start it by clicking on its symbol**

→ When using the terminal to run a program, you download and install it and then **start it by writing the program name in the terminal**

There are different ways to install programs in Linux:

- Opening internet browser, search the program, download and install (similar to windows)
- Or you can directly install it via the terminal → this usually adds the program to the so-called **PATH!**

PATH is a place on your computer (a list of folders) that the terminal searches to find program names.

Using Programs with the Terminal

Some programs have a user interface; some are text-based (some offer both)

Often, the text-based version has more options!

Let's look at fastqc!

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

`fastqc` (opens the user interface of the program)

`fastqc --help` (shows the text-based options of the program in the terminal)

Using Programs with the Terminal

Let‘s run fastqc with OPTIONS

fastqc -o /home/genomics/Documents/fastqc/ sample3.1.fq.gz sample3.2.fq.gz

Using Programs with the Terminal

Some programs have only one purpose

fastqc has only one tool
→ it checks quality parameteres of fastq read file

Other programs have several tools that you can choose from

Samtools has 40 tools to choose from
→ e.g. converting files to other formats, filtering aligned reads, indexing files etc.

Reference genomes

Plant Reference Genomes can be found here:
Ensembl Plants https://plants.ensembl.org/index.html

You can do analyses also online (free):
https://galaxy-main.usegalaxy.org/

3. Data Analysis

- Quality check on raw sequencing reads → fastqc
- Read alignment to a reference genome → bwa
- Processing of aligned reads (sam files) → samtools
- Variant call among a set of samples producing a vcf file → stacks
- Producing a neighborjoining tree file (Newick format) → tassel
- Simple tree visualization with an online tool → itol