

# Genomes Workflow - LBCM

01 - General guidelines and workflow  
organization

Friday 18<sup>th</sup> July, 2025

## Contents

<b>1</b>	<b>Overview</b>	<b>4</b>
1.1	Module Objectives . . . . .	4
<b>2</b>	<b>The Linux question</b>	<b>4</b>
2.1	So... What is it? . . . . .	4
2.2	Choosing a Distro . . . . .	5
2.3	Installing The Chosen One . . . . .	6
2.3.1	Before anything... . . . .	6
2.3.2	Tools . . . . .	7
2.3.3	Validating the ISO . . . . .	7
2.3.4	Process Overview . . . . .	9
2.3.5	Aftermath . . . . .	10
2.4	Command Line and Terminal operation . . . . .	10
2.4.1	Basics . . . . .	10
2.4.2	The file system tree . . . . .	11
2.4.3	Navigating with the shell . . . . .	12
2.4.4	Commands . . . . .	14
<b>3</b>	<b>Git</b>	<b>14</b>
<b>4</b>	<b>Conda</b>	<b>14</b>
<b>5</b>		<b>14</b>
<b>6</b>	<b>Tools &amp; Software</b>	<b>14</b>
6.1	Required Software . . . . .	15
6.2	Installation Guide . . . . .	15
<b>7</b>	<b>Workflow &amp; Methods</b>	<b>15</b>
7.1	Step-by-Step Protocol . . . . .	15
<b>8</b>	<b>Practical Examples</b>	<b>15</b>
8.1	Example 1: Basic Analysis . . . . .	15
8.2	Example 2: Advanced Usage . . . . .	16
<b>9</b>	<b>Results &amp; Interpretation</b>	<b>16</b>
9.1	Output Files . . . . .	16
<b>10</b>	<b>Scripts &amp; Code</b>	<b>16</b>
10.1	Helper Scripts . . . . .	16
10.2	Quality Control . . . . .	17
<b>11</b>	<b>Troubleshooting &amp; Best Practices</b>	<b>17</b>
11.1	Common Issues . . . . .	17
11.2	Best Practices . . . . .	17

12 References	17
13 Exercises & Next Steps	17
14 Research Notes	18

# 1 Overview

## 1.1 Module Objectives

This module covers:

Learning goals

- Basic Linux terminal concepts and usage.
- What is git and basic usage.
- Conda environment logic.
- Conda basic usage.
- Recommended bioinformatics project organization.

## 2 The Linux question

Taking in consideration the variety of tools available, their functionalities and general comprehension around systems and workflow organization, the choice of operational system constitutes a central point.

In multiple fronts, Linux get's the spotlight. One big advantage is the cost to install and implement the OS: Zero. With a GPL License, everyone has the right to change and redistribute it, following the condition that the code should stay available (Garrels, 2008). It's portability, security and the existence of a large and committed community all helps on the final OS choice for general Bioinformatics research projects.

In the current chapter, we intend to present basics concepts of the Linux ecosystem. Terminal operation will also be included, since it's intrinsic relation to the system core functionalities. Finally, a brief Linux Mint presentation and guidelines shall be pointed, as it's the initial distro of choice.

### Remark

For further and in depth comprehension of base Linux related topics, it is advised to check Garrels, 2008.

### 2.1 So... What is it?

With the crescent wave on tech development, the fact that some systems were directly developed for specific hardware started to weight a ton on user instruction and final cost of the products. A team of developers then started working on what would come to be the "UNIX" project. This operational system brought to the table the ability of code to be recycled, the use of a higher programming language then assembly (C) and an unique overall simplicity (Garrels, 2008).

**Definition**

UNIX refers to a family of operational systems based on the final product of the original UNIX project. They aim to be simple, elegant and allow code recycling. To achieve their goals, modular logic is applied, alongside a hierarchical file system and innate command line interface.

Then, home PC's started getting traction, in a way that running UNIX on them became a possibility. In this context, the idea of a free system directly connected to the original **UNIX** would lead to the birth of **Linux** as a freely available OS.

**Definition**

Linux is an open source OS based on UNIX. Is composed mainly by (WhatLinux2025):

**Bootloader:** Manages the boot process of the computer.

**Kernel:** The actual **Linux**. Manages the CPU, memory and peripheral devices.

**Initialization system:** Sub-system that controls daemons.

**Daemons:** Background services started up during or after boot.

**Graphical server:** Sub-system for graphical display.

**Desktop environment:** Interfaces for direct user interaction.

**Applications:** Any software that can be installed on the system and offer some service or act as a tool.

## 2.2 Choosing a Distro

A lifelong debates inside the Linux community revolves around which Linux distribution to choose. One of the biggest reasons for such is that there's no correct answer. Multiple factors interfere in the final choice, and it can be largely personal too.

Across the internet, a variety of guides and lists can be found on the subject. As it's not the central purpose of this project, we won't dive deeply in this matter. But, even though this creative freedom was taken, we will still, in a further chapter, approach one specific Linux distro, called Linux Mint. This decision was made to lay a foundation for some installation instructions, that can change based on the system of choice. Always when that's the case, it shall be pointed out.

The choice of Mint itself was based purely on the fact that it is a distro with large support, reach and with a more reasonable learning curve for Linux newcomers.

**Tip**

If you want to dive a little deeper on the Mint Distro, we recommend it's User Guide (Mint, 2024b). In it, you can find informations on the system's update manager, multiboot, upgrades from older versions and so on. The online official documentation Linux Mint webpage presents in multiple format all available guides (Mint, 2025).

## 2.3 Installing The Chosen One

An encouraged practice when making a software available is to always provide an in depth documentation on it. One can assume, then, that almost every, if not all, chosen Linux distribution has a guide to system installation and usage. Taking this in mind, the current project will describe in general lines the basic needs to install a distro, with some key concepts and necessary tools, avoiding redundancy with the material the system itself has.

### 2.3.1 Before anything...

When faced with the idea to change the current OS to a Linux distro, before major definitive modifications are made, some safety measures are encouraged. The central point is to make sure that everything runs smoothly, and avoid system breaking errors and data loss as much as possible. In this project, we joined some general measures, presented below.

**Applications list:** The majority of bioinformatics software is designed with Linux based systems in mind, or provides a compatible version. But, it's recommended to make a list of the main needs and goals of the operator with the system to facilitate the post-install and prevent incompatibility issues.

**Tip**

This list can also be used to help in choosing the distro that fits your needs the best.

**Backup:** Always backup crucial files.

**Driver support:** It's a good practice to check also if external drivers of daily and necessary use have functional drivers for Linux based systems.

**System requirements:** Check beforehand if your current hardware specifications match at least the minimum requirements of the chosen distro. Usually this information can be found in the main webpage of the OS.

**Tool and Applications support:** Although a lot of software nowadays is developed with Linux compatibility in mind, that's not always the case. Some services like the Microsoft Office Package and the Adobe Suite

don't have a secure and optimal distribution for Linux based systems, making their usage impossible, impractical or unreliable.

After considering the items listed above, you can better decide which path to take.

#### Tip

If you aren't yet secure of moving totally to Linux, you can set up a Dual Boot function with your current OS and your distro of choice. You can also use a Virtual Machine to emulate a Linux environment and use the OS from inside it!

### 2.3.2 Tools

In the majority of cases, not much is needed to proceed with a Linux OS installation. Usually, it sums up to:

**Bootable device:** A storage medium containing the core files for the installation of the chosen distro. Usually consists on the medium with the chosen system's ISO file "flashed" into.

**Image writer:** A dedicated software that is able to correctly flash a storage device with an system image, turning it into a bootable device.

**ISO file:** Is a file on the ISO format that contains the necessary structure, directions and files for the correct system installation.

#### Definition

**Flashing:** Consists on taking an ISO image and transferring all the partitions, file systems and files that were present on the device used to create the image.

**ISO format:** The ISO format can be seen as the equivalent of replicating the contents and structure of a CD, DVD or even a Blu-Ray disc.

**Storage device:** An external drive that can be used as a storage medium, e.g. an USB flash drive.

### 2.3.3 Validating the ISO

As we can infer, given the above definitions, the process basically consists on taking the target OS general structure and files, flashing it into the bootable device and using the latter to build the system on the machine. Thus, we can logically infer, the success of the operation relies heavily on the integrity of the ISO file, as it represents what will be installed.

To avoid the task of manually reviewing, we have more automated means to check integrity. For that, in the majority of cases, we can use SHA256 sums to complete this task. The usual workflow for such is:

1. Download the correct sha256sum.txt and sha256sum.txt.gpg files according to the chosen distro and version.

**Obs:** Make sure to download the files themselves, and not just copy and paste their content on a new file.

2. Open a terminal window where the downloaded sha256sum.txt file is, or navigate to the location.
3. If on a Linux system, verify the **integrity** with:

```
1 sha256sum -b chosendistro.iso
2 # Or, if the iso is on another location
3 sha256sum -b /path/to/chosendistro.iso
```

#### Remark

On macOS, the command structure is similar, although the command itself is only **shasum**. But, if you are currently using Windows, the process follows a distinct path. If that's the case, we recommend checking the guide available on the [Linux Mint Forum](#), that approaches the problem at hand.

4. If both match, the ISO image was successfully downloaded. If not, it's advisable to re-download the file.

#### Tip

Before downloading the ISO again, it's always worth the shot to check a second time if the correct hash file is being tested, that being, the hash for the same distro and same distro version chosen.

#### Remark

The shum files are usually provided with the ISO download or can be retrieved in the same page as the main download mirror.

If you wish to further comprehend what the SHA hash is and how it allows data verification, we recommend the quick post available at [How to Geek](#).

Okay, the file is intact. But, how do we know that the hash file used to test it's integrity is really authentic, and wasn't corrupted or altered? Simple, checking it's authenticity. In the case of Linux Mint, it goes as follow (Mint, 2024a):

1. Import the Linux Mint signing key.

```
1 gpg --keyserver hkp://keys.openpgp.org:80 --recv-key \
2 27DEB15644C6B3CF3BD7D291300F846BA25BAE09
```

2. Check key importation:

```
1 gpg --list-key --with-fingerprint A25BAE09
```



3. Confirm that the output from the previous command presents the key informed on the first command used and the relation to Linux Mint.
4. Check the authenticity with:

```
1 gpg --verify sha256sum.txt.gpg sha256sum.txt
```

5. If the output points that the signature is good and was made with the key imported, the test was passed.

#### Tip

**PATH:** Make sure that all files informed on the commands are on the current work directory of the opened terminal. If not, you can either change the terminal WD, the files location or specify the full path to the files on the commands.

**GPG warning:** The GPG may send a warn that the signature isn't trusted by the computer, this is expected and normal.

### 2.3.4 Process Overview

A full installation guide can usually be found on your distro of choice homepage. It usually covers all the needed steps and present FAQ's to help with following them. [Here](#) you can see the documentation for installing Linux Mint. The process has similar features for all distros:

1. Choose and download the intended version.
2. Verify the ISO integrity.
3. Create the bootable device.
4. Insert the bootable media onto your machine and boot from it (Usually accessing the BIOS and changing boot order).
5. Enter the system.
6. Start the installation setup and follow the steps pointed.
7. Reboot the machine.
8. Remove the bootable device.

The new OS is almost completely ready to use!

#### Tip

Most Linux distributions allow the user to test the system before fully installing via the bootable device.

### 2.3.5 Aftermath

## 2.4 Command Line and Terminal operation

When using Linux based OS's, one specific "feature" get's the spotlight on the eyes of newcomers: the command line. Basically all tasks done via GUI's can be also executed via command line, usually mediated with a terminal interface.

Taking this in mind, understanding the relevance and basic navigation of the command line, allows an overall better and more efficient system operation. In this sense, the present section aims to establish broad concepts and present general guidelines on the subject.

### Definition

The term command line, actually refers to the **shell**. It consists in a program whose purpose is to take keyboard commands and direct them to the OS for execution. In Linux systems, the core **shell** usually is **bash**. **Terminals** are interfaces that allow users with a GUI to interact with the **shell**(Shotts, 2024).

### 2.4.1 Basics

Nothing better to learn than getting punched by the subject. Let's experiment with basic terminal manipulation and usage. This section is an adaptation of the first chapter of **shotssLinuxCommandLine2024**, in that sense, for further comprehension on the subject, please refer to the main source.

First, open the terminal window (Ctrl+Alt+T on Linux Mint). What you are now seeing, I hope so, is something like:

```
1 [user@machine ~]$
```

Or, if using conda, probably something like:

```
1 (base)[user@machine ~]$
```

This is what we call the shell prompt, indicating that it's ready to receive orders. It informs you the current user name, the machine name and the current directory. Knowledge of the directory currently set is of extreme importance, and will be discussed on the next topic.

Now, type out on the terminal the following:

```
1 [user@machine ~]$ ls
```

The expected output, is a listing of all files and folders available **directly** on the current work directory (I said it was important). Done, you executed your first bash command! The majority of commands supports what we call **arguments**, that can be present in different forms, like **flags**, **parameters** and **positional**. Let's try it out:

```
1 [user@machine ~]$ ls -a
```

**Definition**

**Argument:** A value, whose format can vary, that informs something to the program that will be executed. In this way, he allows more interaction and control of the user with the command itself.

**Flag:** A flag consists in an argument that doesn't have a value. It usually appears in the format of a single dash followed by a letter (-l, -a, etc).

**Parameter:** A parameter is an argument with an user determined value. The base format of a parameter is two dashes followed by a space and the value (like in --help). In some cases, the space is substituted for a equality symbol (like in --output=/home).

**Positional argument:** Indicates to the command which directory or file it should seek and operate. Usually presents itself with a space from the rest of the arguments and the initial command followed by the path to needed directory or file.

Now, you can see a lot more files are being shown. That's because the **-a** flag indicates to the command that you want to see **all** possible files and directories under the current work directory. So, now your output will also include the hidden files, marked by the presence of a single dot before the file name.

**Remark**

In the majority of commands or terminal executed scripts, you can see a summary of possible flags and operational directions for command usage applying the --help parameter or the -h flag.

**Tip**

For better quality of life, you can use the up and down arrow keys to access the **terminal command history**. Try cycling through previous commands and re-running them.

This covers up our initial contact with the shell and terminal. As a way of saying goodbye, instead of closing the window in the traditional way, you can exit the shell by simply typing the below!

```
1 [user@machine ~]$ exit
```

**2.4.2 The file system tree**

Now, we are going to approach the file system tree of Linux based systems. The tree analogy says it all, the files are organized in what's called **hierarchical directory structure**(Shotts, 2024). In this sense, the system has a base for the tree, which we call the **root directory**. All others directories, that we can think as branches, and files, the leaves, will be located inside this root.

**Remark**

Unix-like systems approach storage devices in a particular way. They're treated as part of the single file system tree that begins on the root. That being, any storage device that is connected to the system, is being **mounted** in certain **point**(location) inside the main tree.

Below you can see some of the root base sub-directories (Garrels, 2008).

**Definition**

**bin:** Programs shared by the system, the system administrator and all of the users.

**boot:** Stores the startup files and the kernel.

**etc:** Stores the most important system configuration files.

**home:** Acts as a "root" for the directories of common system users.

**lib:** Library files needed by the system and it's users.

**mnt:** Standard mount point for all external file systems.

**root:** The home directory of the system administrator.

**tmp:** Used to store temporary files, being cleaned up on reboot.

**usr:** Files and directories related to all user-related programs.

**Remark**

**Caution!** The `/root` refers to the administrator home directory, while `/` refers to the root directory.

### 2.4.3 Navigating with the shell

Remember when was said that the current work directory is important? Now's the time to understand why. As was presented on the previous sub-section, Linux based systems operate on a single rooted file system. This implies that we can, from inside the terminal, access all existent sub-directories and files, but also, that a hierarchical order is present. In this sense, we can conclude that the terminal needs to know where he stands on this tree to operate in the intended way.

When opening a new terminal window, we saw that after the machine name, the **current working directory** is shown. This is where the terminal stands on the tree. The logic of this is easier comprehended by imagining as if we are at the CWD looking forward. We can see all connected branches and the current branch leaves and operate on them and theirs subsequent branches and leaves. We saw this with the `ls` command previously tested.

Let's try some basic terminal navigation. Open a new terminal window

(Ctrl+Alt+T on Linux Mint). Let's check the current work directory we are on:

```
1 [user@machine ~]$ pwd
```

#### Remark

The `~` directory is a short way to refer to the current user home directory. It's the equivalent of `/home/user/`. You can test this by opening a terminal window and typing the **pwd** command, that shows the name of the current WD.

If all worked as supposed to, the current WD was shown on the screen. Great! Now, let's create a test directory inside the current one and make it our new WD. After that, let's run again the `pwd` command.

```
1 mkdir test
2 cd test
3 pwd
```

First, we used **mkdir** with the argument 'test' to create a new sub-directory called test. Then, the **cd** command allowed us to change our WD to the one specified as an argument, in this case 'test'. The output of **pwd** shows that we did, indeed, change the current "point of view" of our shell session.

Still in this terminal, try to change again to the test folder. As you can see, the bash returns a message saying that he wasn't able to find such file or directory. This is because it really doesn't exists! At least not inside the WD. Remember, the Linux file system is hierarchical. In this sense, the command searches and operates based on the current work directory, seeing only the branches and leaves that are "inside" of it. This is extremely relevant, for when running scripts or other commands from a terminal, we need to make sure the correct WD is set, because other way, the shell call won't be able to find and operate on the intended files and directories.

#### Tip

The `cd` usage can be simplified in some cases. If we want to go back to the user home directory, we can use:

```
1 cd ~
```

Also, the parent folder can be referred as two dots (`..`). Using this we can:

```
1 cd .. # goes up one parent folder
2 cd ../.. # goes up two parent folders, and so on
```

**Remark**

Linux based systems have what is called a **PATH environment variable**. This consists on a system variable that can be changed if necessary that contains a list of system paths. Anytime a shell session begins and any command is executed, the system searches not only the WD and it's sub-directories and files, but also the paths included in this environment variable.

**2.4.4 Commands**

Aiming to provide some base commands, the user can see the cheat sheet below. In depth command usage, availability and functionality can be found on software or distribution documentation.

**Cheat Sheet**

**Command:** `ls`

**Description:** List files in current work directory.

**Command:** `pwd`

**Description:** List current work directory.

**Command:** `cd 'new-directory'`

**Description:** Changes current work directory.

**Command:** `mkdir 'new-directory'`

**Description:** Creates new directory inside current WD or on given path.

**Command:** `cp 'file' 'copy-location'`

**Description:** Creates a copy of certain files inside the given location.

**Command:** `mv 'current-location' 'new-location'`

**Description:** Moves a file inside the file system.

**Command:** `df`

**Description:** Show current amount of free space on the disk drives.

**Command:** `lsblk`

**Description:** Show current mounted disks and their mount points.

**Command:** `rm 'file-path'`

**Description:** Removes the file indicated.

**3 Git****4 Conda****5****6 Tools & Software**

## 6.1 Required Software

Installation notes

- **Primary tool:** Tool name and version
- **Dependencies:** Required libraries/packages
- **Optional:** Additional helpful tools

## 6.2 Installation Guide

```
1 # Installation commands
2 conda install -c bioconda tool_name
3 # or
4 sudo apt-get install package_name
```

Listing 1: Software installation

# 7 Workflow & Methods

## 7.1 Step-by-Step Protocol

Key parameters

1. **Data preparation:** Input requirements and formatting
2. **Quality control:** Initial data assessment
3. **Main analysis:** Core computational steps
4. **Result interpretation:** Output analysis and validation

**Example 7.1.** Practical example with real genomic data.

# 8 Practical Examples

## 8.1 Example 1: Basic Analysis

Input/output files

- **Input:** Sample data description
- **Command:** Based on approach from **example2024**
- **Output:** Expected results and file formats

```
1 # Example command with typical genomic data
2 tool_name -i input_file.fasta -o output_file.txt --parameter
  value
```

Listing 2: Basic command example

## 8.2 Example 2: Advanced Usage

Complex parameters

```
1 # Multi-step analysis pipeline
2 step1_tool input.fasta | step2_tool --param1 value1 >
   intermediate.txt
3 step3_tool intermediate.txt --param2 value2 -o final_result.txt
```

Listing 3: Advanced analysis pipeline

## 9 Results & Interpretation

### 9.1 Output Files

Common output formats and their interpretation:

File formats

- **Format 1:** Description and typical contents
- **Format 2:** When and how to use this output
- **Quality metrics:** How to assess result quality

**Remark 9.1.** Important note about result interpretation following **author2024**.

## 10 Scripts & Code

### 10.1 Helper Scripts

```
1 #!/usr/bin/env python3
2 """
3 Helper script for genomic data processing
4 Usage: python script.py input.fasta output.txt
5 """
6
7 def process_sequences(input_file, output_file):
8     """Process genomic sequences"""
9     with open(input_file, 'r') as f:
10         sequences = f.read()
11
12     # Processing logic here
13     processed = sequences.upper()
14
15     with open(output_file, 'w') as f:
16         f.write(processed)
17
18 if __name__ == "__main__":
19     import sys
20     process_sequences(sys.argv[1], sys.argv[2])
```

Listing 4: Data processing script



## 10.2 Quality Control

```
1 #!/bin/bash
2 # Quality control pipeline for genomic data
3
4 # Check file format
5 file_format_check.py $INPUT_FILE
6
7 # Basic statistics
8 sequence_stats.py $INPUT_FILE > stats.txt
9
10 # Quality assessment
11 quality_assessment_tool $INPUT_FILE --output qc_report.html
```

Listing 5: QC pipeline

## 11 Troubleshooting & Best Practices

### 11.1 Common Issues

Error solutions

- **Memory errors:** Reduce dataset size or increase available RAM
- **Format issues:** Check input file formatting and encoding
- **Parameter tuning:** Guidelines for optimization

### 11.2 Best Practices

- **Data backup:** Always keep original data copies
- **Version control:** Track analysis versions and parameters
- **Documentation:** Record all analysis steps and decisions
- **Reproducibility:** Use consistent environments and seeds

## 12 References

- Key papers: **example2024**; **author2024**; **smith2024**
- Software documentation: [Tool official docs]
- Related modules: [Other workflow modules]

## 13 Exercises & Next Steps

- **[TODO]: Practice with provided sample data**
- **[TODO]: Try different parameter settings**
- **Apply to your own genomic dataset**
- **[TODO]: Explore advanced features**

## 14 Research Notes

Additional observations and module-specific notes...

**Key insight:** Connection between this tool and genome annotation pipeline

**[IDEA]: Extension:** Integration with other bioinformatics tools in the workflow

## References

- Garrels, Machtelt (June 2008). *Introduction to Linux*. 1.27. Vol. 1. USA: Online. ISBN: 1-59682-112-4. (Visited on 07/01/2025).
- Mint, Linux (Dec. 2024a). *Installation Guide*. (Visited on 07/01/2025).
- (Dec. 2024b). *User Guide: Linux Mint*. (Visited on 07/01/2025).
- (Jan. 2025). *Linux Mint Documentation*. <https://www.linuxmint.com/documentation.php>. (Visited on 07/01/2025).
- Shotts, William (Nov. 2024). *The Linux Command Line*. 6th ed. Vol. 1. Online. (Visited on 07/01/2025).