Genomes Workflow - LBCM

01 - General guidelines and workflow organization

Wednesday 16th July, 2025

Contents

1 Overview

1.1 Module Objectives

Learning goals

This module covers:

- 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 led to the birth of **Linux** as a freely available OS.

Definition

Linux is an open source OS based on UNIX. Is composed mainly by (Linux, 2025):

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 Command Line Terminal

2.3 Linux and Bioinformatics

2.4 Choosing a Distro

3 Tools & Software

3.1 Required Software

Installation notes

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

3.2 Installation Guide

```
# Installation commands
conda install -c bioconda tool_name
# or
sudo apt-get install package_name
```

Listing 1: Software installation

4 Workflow & Methods

4.1 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 4.1. Practical example with real genomic data.

5 Practical Examples

5.1 Example 1: Basic Analysis

Input/output files

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

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

Listing 2: Basic command example

5.2 Example 2: Advanced Usage

Complex parameters

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

Listing 3: Advanced analysis pipeline

6 Results & Interpretation

6.1 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 6.1. Important note about result interpretation following **author2024**.

7 Scripts & Code

7.1 Helper Scripts

```python
#!/usr/bin/env python3
"""
Helper script for genomic data processing
Usage: python script.py input.fasta output.txt
"""

def process_sequences(input_file, output_file):
    """Process genomic sequences"""
    with open(input_file, 'r') as f:
        sequences = f.read()

    # Processing logic here
    processed = sequences.upper()

    with open(output_file, 'w') as f:
        f.write(processed)

if __name__ == "__main__":
    import sys
    process_sequences(sys.argv[1], sys.argv[2])
```

Listing 4: Data processing script

7.2 Quality Control

```bash
#!/bin/bash
# Quality control pipeline for genomic data

# Check file format
file_format_check.py $INPUT_FILE

# Basic statistics
sequence_stats.py $INPUT_FILE > stats.txt

# Quality assessment
```

```
quality_assessment_tool $INPUT_FILE --output qc_report.html
```

Listing 5: QC pipeline

8 Troubleshooting & Best Practices

8.1 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

8.2 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

9 References

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

10 Exercises & Next Steps

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

11 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).

Linux (Jan. 2025). *What Is Linux*. https://www.linux.com/what-is-linux/. (Visited on 01/16/2025).