Genomes Workflow - LBCM

02 - Data pre-processing

Thursday 24th July, 2025

Contents

1 Overview

Once the base introduction and general guidelines were covered, we can start working directly on the pipeline itself. For starters, it's always advised to pre-process our data, that being: understanding it's origin, how we expect it to behave and what special treatment does it need. Following such general stone path can help we obtain better results later on, securing data integrity and usability.

1.1 Module Objectives

- Present the logic behind data pre-processing.
- Approach tools that operate on such function.
- Analyze output examples.
- Comprehend the importance of such step on the general workflow.

1.2 Module Required Tools

- Conda (recommended)
- Git (recommended)
- FASTP
- FASTQC

1.3 Module Structure

The current module is divided in two big sections:

Theoretical Framework: Approaches the theory behind the process that's being mainly covered on the current module. The idea is to allow the user to consult for general concepts comprehension and, if needed, guide him to sources for a more in depth coverage.

Practical approach: In this section, we shall explain in practice how this step can be applied. For such, NCYC357.fastq, a paired-end Illumina set of reads, will be used, although the intrinsic logic and observations can be applied to other samples.

2 Theoretical Framework - Data pre-processing

2.1 What is it?

2.2 Why it's important?

2.3 Base assumptions

3 Practical approach

3.1 Data analysis

3.1.1 Gathering the data

3.1.2 Understanding our data

3.1.3 Taking decisions

3.2 Data pre-processing

3.2.1 Choosing the right tools

3.2.2 How results can become parameters

3.2.3 Pre-processing

3.2.4 Understanding what's found