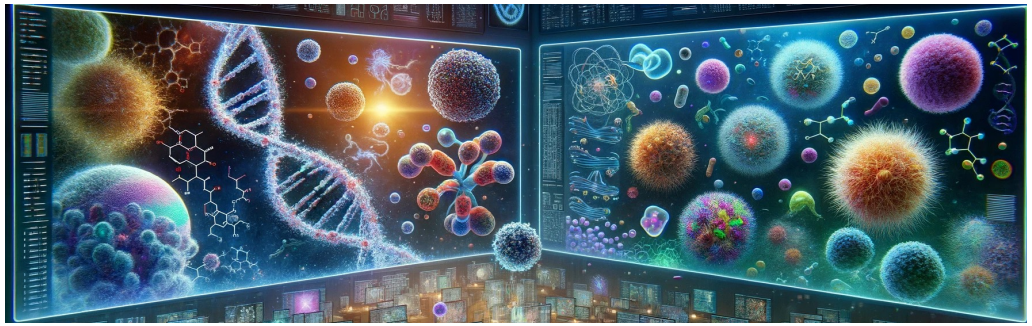


Advanced Applications of Next Generation Sequencing in Food Safety

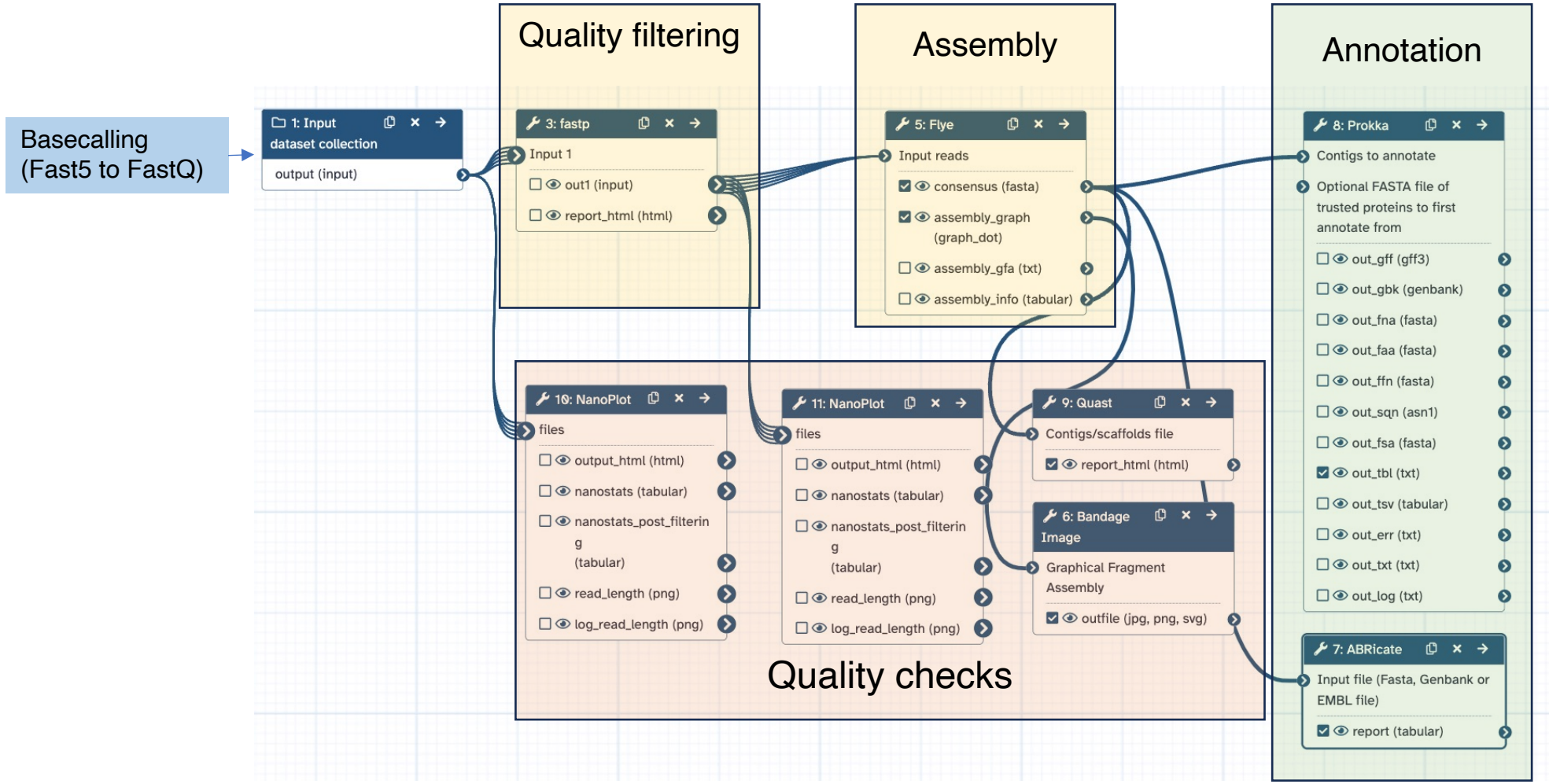
www.ucd.ie/cfs
@CfsUcd

Dr Guerrino Macori, BSc, MSc, PhD
Assistant Professor
School of Biology and Environmental Science
University College Dublin, Ireland
UCD Centre for Food Safety
guerrino.macori@ucd.ie
@guerrinomacori

Definition of pipeline

- A "pipeline" refers to a set of data processing steps or a sequence of tools/software used to analyse complex biological data.
- Each step in the pipeline performs a specific function, such as filtering data, aligning sequences, or annotating features, and passes the results on to the next step.
- Pipelines are crucial for the systematic and efficient handling of large datasets, such as those generated by high-throughput technologies like genome sequencing.
- They are designed to be automated and reproducible, allowing for the consistent analysis of multiple datasets in a similar manner.

Background



Each of these steps should be performed in sequence, ensuring that the output of one step feeds correctly into the input of the next. It's also recommended to run quality checks at various stages

Basecalling

1. **Conversion from Fast5 to FastQ**:
    1. Tool: Guppy or MinKNOW (part of the Nanopore software suite)
    2. Description: Converts the raw signal data (Fast5 files) to sequence data (FastQ files). Guppy is a basecalling software that translates the electrical signals into nucleotide sequences.
2. **Barcoding (Demultiplexing)**:
    1. Tool: Guppy barcoder or MinKNOW
    2. Description: If you have multiplexed samples, this step will sort the reads according to the barcode sequences, effectively separating different samples within the same sequencing run.
3. **Concatenation of FastQ Files**:
    1. Tool: cat (Unix Command Line)
    2. Description: Merges multiple FastQ files into one file for easier handling. Useful if you have multiple FastQ outputs from different runs or flow cells for the same sample.

Quality Checks and Read Filtering

Tool: NanoPlot

Description: Generates plots and statistics to give an overview of the dataset, including read length distribution and quality scores.

Tool: Filtlong or NanoFilt

Description: These tools are used for filtering reads based on quality and length. This step improves the quality of the dataset by removing poor-quality reads that could negatively impact the assembly.

Genome Assembly

Genome Assembly:

Tool: Flye
Description: de novo assembler for long reads that can generate high-quality genome assemblies. It is designed to work well with Nanopore data.

Quality Checks Genome Assembly and visualisation

Tool: QUAST

Description: Evaluates genome assemblies by analyzing various metrics such as N50, L50, and others. It helps to gauge the contiguity and completeness of the assembled genome.

Tool: Bandage

Description: Visualises assembly graphs generated by assembly tools. It helps in understanding the structure of the assembly and identifying issues like misassemblies or potential contaminations.

Genome Annotation

Tool: Prokka

Description: Rapidly annotates bacterial, archaeal, and viral genomes. It identifies genes, rRNA, tRNA, and other functional elements within the genome sequence.

Detection of AMR and Virulence Factors:

Tool: ABRicate

Description: Screens the assembled genome against multiple databases of antimicrobial resistance (AMR) genes and virulence factors to identify potential resistance genes and pathogenicity-related genes.

Tool: Prokka

Description: Rapidly annotates bacterial, archaeal, and viral genomes. It identifies genes, rRNA, tRNA, and other functional elements within the genome sequence.

Detection of AMR and Virulence Factors:

Tool: ABRicate

Description: Screens the assembled genome against multiple databases of antimicrobial resistance (AMR) genes and virulence factors to identify potential resistance genes and pathogenicity-related genes.