Genome Assembly Pipeline: Tools & Steps

Overview

This document outlines the tools and steps used in the **genome assembly pipeline** after data download. The steps include quality control, read trimming, assembly, evaluation, and annotation.

Table of Contents

Tools Required

The following tools must be installed before running the pipeline:

- **FastQC**: Quality control of raw and trimmed reads.
- **Trimmomatic**: Read trimming and quality filtering.
- **SPAdes**: Genome assembly and error correction.
- **QUAST**: Assembly quality assessment.
- **Prokka**: Genome annotation.

Pipeline Steps

1. Quality Control with FastQC

Command:

```
fastqc SRR9620862_1.fastq SRR9620862_2.fastq
```

- This step checks read quality, GC content, and adapter contamination.
- Generates an HTML report for visual inspection.

---

2. Read Trimming with Trimmomatic

Command:

```
trimmomatic PE SRR9620862_1.fastq SRR9620862_2.fastq \
    SRR9620862_1_paired.fastq SRR9620862_1_unpaired.fastq \
    SRR9620862_2_paired.fastq SRR9620862_2_unpaired.fastq \
    LEADING:10 TRAILING:10 SLIDINGWINDOW:5:20 MINLEN:250
```

- Trims low-quality bases and removes short reads.
- **Paired-end reads** are handled separately to ensure both ends are retained.

---

3. Error Correction with SPAdes

Command:

```
spades.py -1 SRR9620862_1.fastq -2 SRR9620862_2.fastq \
    -o spades_corrected --only-error-correction
```

- Performs read error correction before assembly.

---

4. Genome Assembly using SPAdes

Default Run

```
spades.py -1 spades_corrected/corrected/SRR9620862_100.0_0.cor.fastq.gz \
    -2 spades_corrected/corrected/SRR9620862_200.0_0.cor.fastq.gz \
    -o spades_default_assembly --only-assembler
```

Careful Run with K-mer Optimization

```
spades.py -k 21,33,55,77,99,127 --careful --only-assembler \
    -1 spades_corrected/corrected/SRR9620862_100.0_0.cor.fastq.gz \
    -2 spades_corrected/corrected/SRR9620862_200.0_0.cor.fastq.gz \
    -o spades_careful_assembly
```

- The **default run** performs standard assembly.
- The **careful run** uses optimized k-mers and error correction.

---

5. Assembly Evaluation with QUAST

Command:

```
quast -o quast_SRR9620862_out \
    -R Reference_Genome/reference_genome.fna.gz \
    -g Reference_Genome/anno_reference_genome.gff.gz \
    --labels Spades_Default,Spades_Careful \
    spades_default_assembly/contigs.fasta \
    spades_careful_assembly/contigs.fasta
```

- Compares assemblies to the reference genome.

- Reports N50, GC content, number of contigs, and misassemblies.

---

6. Genome Annotation with Prokka

Command:

```
prokka --force --outdir prokka_annotation --prefix annotation \
    spades_default_assembly/contigs.fasta
```

- Identifies genes, proteins, and functional elements in the assembled genome.
- Outputs annotation files including `.gff`, `.gbk`, and `.faa`.
