GARDEN CITY UNIVERSITY
DEPARTMENT OF LIFE SCIENCE

MASTER OF SCIENCE IN BIOINFORMATICS

BACTERIAL GENOME ASSEMBLY USING SPADES

ABHISHEK S R
24MSBI155

INTRODUCTION TO SPADES

SPAdes (St. Petersburg genome assembler) is a versatile toolkit designed for assembling small genomes, particularly from bacteria, using Illumina, IonTorrent, PacBio, and Oxford Nanopore sequencing data.

First released in 2012, SPAdes has become one of the most widely used bacterial genome assemblers due to its accuracy, flexibility, and continual development.

FEATURES

Multi-platform support: Accommodates various sequencing technologies and hybrid assemblies

Specialized modes: metaSPAdes, plasmidSPAdes, rnaSPAdes, biosyntheticSPAdes

Built-in error correction: BayesHammer algorithm for Illumina reads

De Bruijn graph approach: Multi-sized k-mer assembly for improved contiguity

Single-cell capabilities: Designed to handle MDA-amplified single-cell bacterial data

WHY NEXTFLOW FOR SPADES?

1.Ensures automation of multiple samples.

2. Provides scalability (local, HPC, cloud).

3. Ensures reproducibility with Docker/Singularity.

4. Tracks dependencies and file outputs.

Nextflow Pipeline Workflow

Tools integrated in the workflow:
FastQC – Read quality check
Trimmomatic / fastp – Read preprocessing
SPAdes – Genome assembly
QUAST – Assembly quality assessment
MultiQC – Summary reports

PREREQUISITES / INSTALLATION

1.Install Nextflow:

curl -s https://get.nextflow.io | bash

2. Install SPAdes (via conda):

- **conda install -c bioconda spades**

3. Other tools (FastQC, QUAST, MultiQC):

- **conda install -c bioconda fastqc quast multiqc**
- **(Optional) Use Docker/Singularity for reproducibility.**

REQUIRED INPUT FILES

Input format: FASTQ/FASTA files from Illumina, IonTorrent sequencing platforms (paired-end, mate-pair, single-end)

Metadata (optional): Sample sheet with IDs and file paths.

Source of data:

Public repositories (NCBI SRA, ENA)

In-house sequencing labs

OUTPUT FILES

FastQC: HTML + ZIP reports

Trimmomatic/fastp: Cleaned FASTQ files

SPAdes:

contigs.fasta (main assembly)

scaffolds.fasta

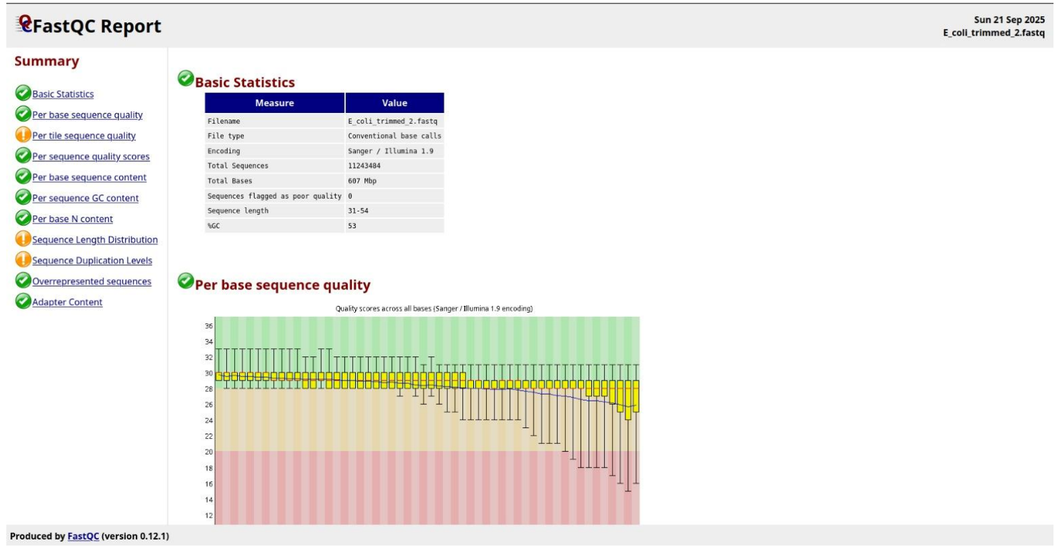
assembly_graph.gfa

QUAST: HTML, TSV, PDF assembly statistics

MultiQC: Final aggregated HTML report

INPUT SAMPLE & FASTQC

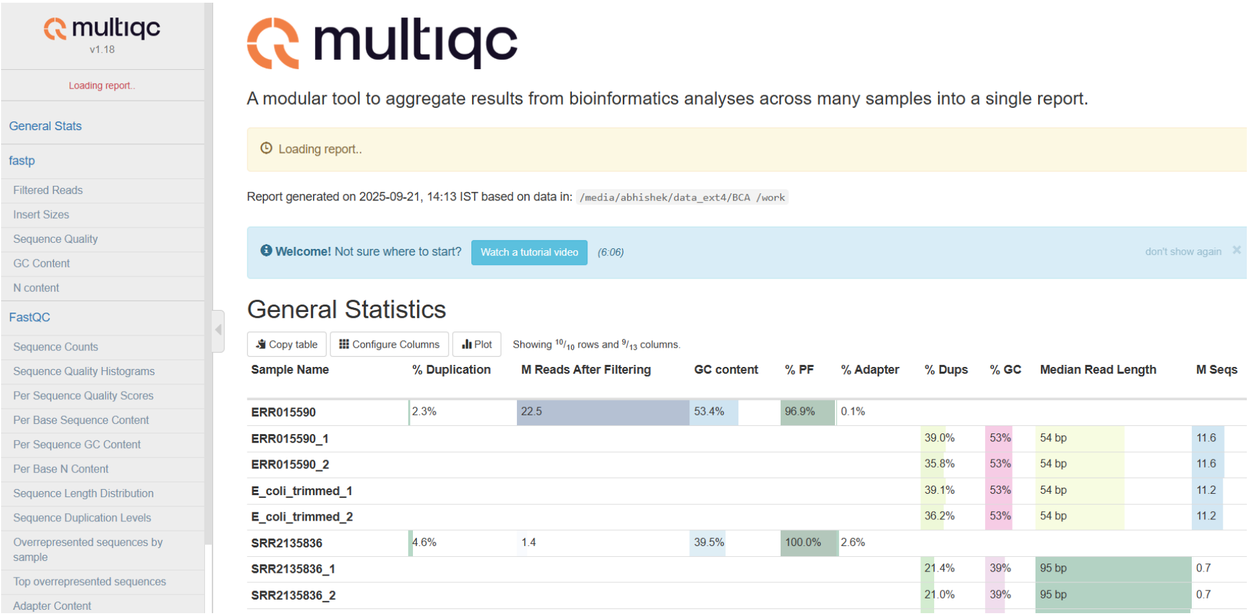
	A	B	C	D	E	F
1	sample	fastq_1	fastq_2			
2	E_coli	data/ERR0	data/ERR015590_2.fastq.gz			
3	S_aureus	data/SRR2	data/SRR2135836_2.fastq.gz			
4						
5						
6						



OUTPUT – NEXTFLOW PIPELINE

```
bhishek@sai-ram:/media/abhishek/data_ext4/BCA $ nextflow run main.nf

N E X T F L O W  ~  version 25.04.7

aunching `main.nf` [jolly_leibniz] DSL2 - revision: 89954c0063

xecutor >  local (20)
65/b4a60e] process > fastqc_raw (3)      [100%] 4 of 4 ✔
60/556e5a] process > trim_reads (2)      [100%] 2 of 2 ✔
44/d4d836] process > fastqc_trimmed (3)  [100%] 4 of 4 ✔
e4/3b6c34] process > spades_assembly (2) [100%] 2 of 2 ✔
2a/dc8f6d] process > multiqc (8)         [100%] 8 of 8 ✔
ompleted at: 21-Sep-2025 14:03:43
uration    : 14m 19s
PU hours   : 1.0
ucceeded   : 20

bhishek@sai-ram:/media/abhishek/data_ext4/BCA $
```

```
Sep 21  17:29
abhishek@sai-ram: /media/abhishek/data_ext
(base) abhishek@sai-ram:/media/abhishek/data_ext4/BCA$ nextflow clean -f
Removed /media/abhishek/data_ext4/BCA/work/12/f861120a2031bdf56e2de92ffa213a
Removed /media/abhishek/data_ext4/BCA/work/a6/04e416216ef3f21de197c78cd186d0
(base) abhishek@sai-ram:/media/abhishek/data_ext4/BCA$ nextflow run main.nf -resume

N E X T F L O W  ~  version 25.04.7

Launching `main.nf` [sick_mccarthy] DSL2 - revision: 43db40d116

executor >  local (15)
executor >  local (15)
[df/7632b6] process > fastqc_raw (3)      [100%] 4 of 4 ✔
[ae/c43c35] process > trim_reads (1)      [100%] 2 of 2 ✔
[e0/9761a4] process > fastqc_trimmed (3)  [100%] 4 of 4 ✔
[62/58ba2d] process > spades_assembly (2) [100%] 2 of 2 ✔
[74/80f33c] process > quast (2)           [  0%] 0 of 2 ✘
[64/32d56b] process > multiqc             [  0%] 0 of 1
```



APPLICATIONS OF THE PIPELINE

Genome assembly objectives:

- Draft genome reconstruction of bacterial isolates
- Comparative genomics (strain-level differences)
- Antimicrobial resistance gene identification
- Plasmid and mobile element analysis
- Metagenomic bacterial genome recovery (MAGs)

Broader studies:

- Evolutionary studies
- Functional annotation & pathway analysis
- Vaccine/therapeutic target discovery

Thank You