Genomics Projects

Project Title:

De novo Genome Assembly and Quality Analysis of *E. coli* (SRR35831031)

Project Overview:

This project performs a **complete de novo genome assembly** of an *Escherichia coli* isolate from a patient using Illumina paired-end sequencing data obtained from the NCBI SRA (SRR35831031). The workflow covers raw data retrieval, read extraction, genome assembly, quality assessment, completeness evaluation, read mapping, and genome visualization.

The final assembly is high quality, with strong QUAST metrics, 100% BUSCO completeness, and uniform read coverage confirmed through IGV.

Aim

To reconstruct and validate the whole genome of an *E. coli* isolate using a de novo assembly pipeline and assess the assembly's quality through multiple bioinformatics tools.

Background

De novo genome assembly is essential when working with organisms or strains that do not have an accurate or available reference genome, or when the strain may be genetically different from existing references. Unlike reference-based mapping—which can miss new genes, plasmids, insertions, deletions, and other structural changes—de novo assembly rebuilds the genome completely from scratch using only sequencing reads.

This avoids reference bias, captures novel genetic elements, and gives an accurate representation of the genome's true structure.

This experiment demonstrates the full microbial genome assembly workflow, including SRA data retrieval, FASTQ extraction, SPAdes assembly, QUAST quality assessment, BUSCO completeness check, BWA alignment, and IGV visualization—representing a standard professional pipeline used in microbial genomics, clinical microbiology, and research.

Dataset Information

- Accession: SRR35831031
- Organism: Escherichia coli
- Platform: Illumina NextSeq 2000
- Read Type: Paired-end
- Total Reads: ~2.6 million
- File Source: NCBI SRA

Tools & Software Used

- **SRA Toolkit** – Prefetch, fasterq-dump
- **SPAdes** – Genome assembly
- **QUAST** – Assembly quality assessment
- **BUSCO** – Genome completeness check
- **BWA-MEM** – Read alignment
- **samtools** – Sorting, indexing, BAM processing
- **IGV** – Visualizing read alignment

Methodology

1 Downloading SRA Data

>prefetch SRR35831031

2 Extracting FASTQ Files

>fasterq-dump --split-files SRR35831031.sra

3 Genome Assembly (SPAdes)

>spades.py -1 SRR35831031_1.fastq -2 SRR35831031_2.fastq -o spades_output/

4 Quality Assessment (QUAST)

>quast.py contigs.fasta -o quast_results/

5 Completeness Check (BUSCO)

>busco -i contigs.fasta -l bacteria_odb10 -o busco_r -m genome

6 Read Alignment (BWA + Samtools)

>bwa index contigs.fasta
>bwa mem contigs.fasta SRR35831031_1.fastq SRR35831031_2.fastq > aln.sam
>samtools view -Sb aln.sam | samtools sort -o aln.sorted.bam
>samtools index aln.sorted.bam

7 Visualization

Load: contigs.fasta + aln.sorted.bam into IGV.

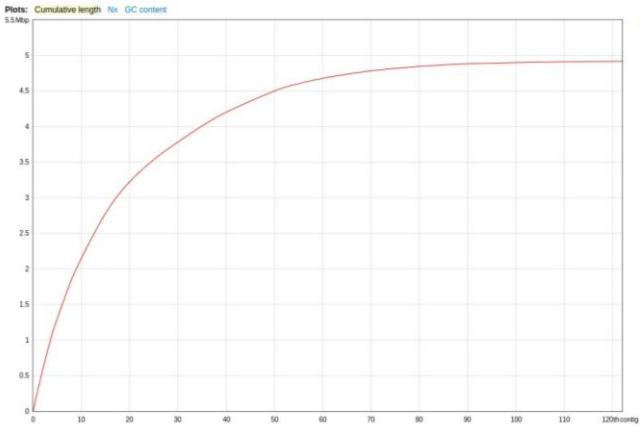

Figure 1Cumulative Genome Length Plot of Assembled Contigs (QUAST Output)

Results:

QUAST Summary

- Total contigs: 122
- Genome size: 4,918,756 bp
-
- N50: 127,576 bp
- Largest contig: 304,100 bp
- GC content: 50.63%
- N's: 0 (no gaps)

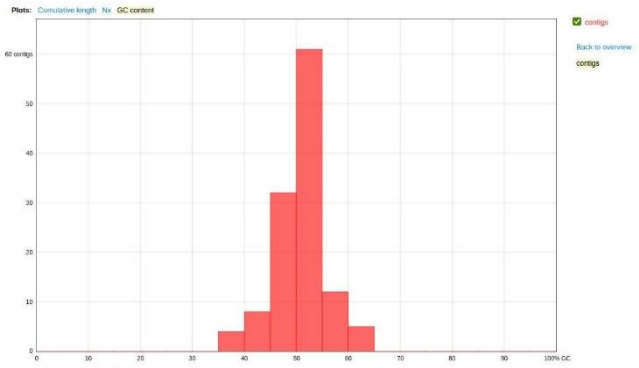

Figure 2: Distribution of GC Content Across Contigs

BUSCO Results

- Completeness: 100%
- BUSCOs detected: 124/124
- No missing or fragmented genes

```
---------------------------------------------------
|Results from dataset bacteria_odb10              |
---------------------------------------------------
|C:100.0%[S:100.0%,D:0.0%],F:0.0%,M:0.0%,n:124    |
|124    Complete BUSCOs (C)                       |
|124    Complete and single-copy BUSCOs (S)       |
|0      Complete and duplicated BUSCOs (D)        |
|0      Fragmented BUSCOs (F)                     |
|0      Missing BUSCOs (M)                        |
|124    Total BUSCO groups searched               |
---------------------------------------------------
```

Figure 3: BUSCO Assessment of Dataset bacteria_odb10 Completeness

IGV Mapping Results

- Uniform coverage across genome
- Very few mismatches
- No structural anomalies
- Confirms assembly accuracy

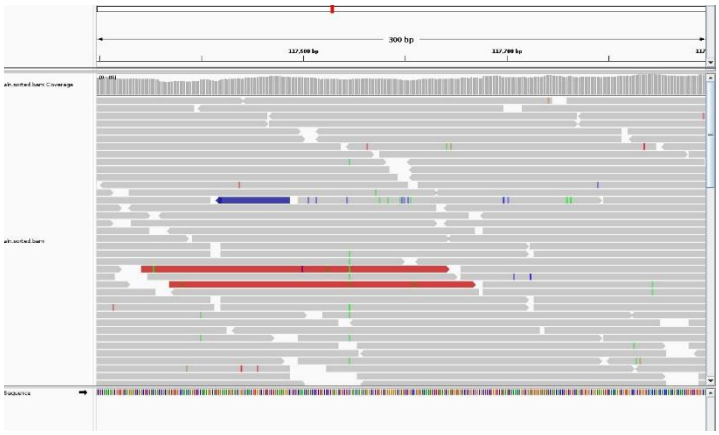

Figure 4: Visualization of a Large Structural Variant via Read Alignment

---

CONCLUSION:

The *E. coli* genome assembly (SRR35831031) produced a high-quality, near-complete assembly of ~4.9 Mb with 122 contigs and 50.6% GC content.
BUSCO analysis confirmed 100% completeness, and QUAST metrics supported excellent assembly integrity.
This demonstrates successful de novo assembly and validation of an *E. coli* isolate using Illumina paired-end reads.

This project successfully reconstructed a *de novo E. coli* genome with high accuracy using Illumina reads. Quality assessment tools (QUAST, BUSCO, IGV) confirm strong assembly integrity and completeness, demonstrating a robust microbial genome assembly workflow.