BRAC UNIVERSITY
Department of Mathematics and Natural Sciences (MNS)

RNA-seq Read Mapping to a Bacterial Genome Assembly

Course: BTE406 – Genomics & Proteomics
Fall 2025

Submitted By:
Atkia Anika
Student ID: 22336039

Instructor:
Mahdi Moosa, PhD
Department of Mathematics and Natural Sciences

January 2, 2026

Contents

1 Species and SRA Accession Details

The bacterial species selected for this project is *Salmonella enterica* subsp. *enterica* serovar Typhimurium strain D23580, a sequence type 313 (ST313) lineage associated with invasive non-typhoidal *Salmonella* disease. Both genomic and transcriptomic datasets for this strain were obtained from the NCBI Sequence Read Archive (SRA) and analysed using the Galaxy platform, in accordance with the project instructions.

Table 1: Dataset Accession and Sequencing Details

Data Type	SRA Accession ID	Sequencing Platform	Read Type
Genomic DNA (WGS)	SRR29773379	Illumina	Paired-end
RNA-seq	SRR35569531	Illumina	Paired-end

These datasets satisfy the requirement to use a single bacterial species with separate whole-genome and RNA-seq datasets generated using paired-end Illumina sequencing.

2 Genome Assembly Statistics

Genomic paired-end reads from SRR29773379 were assembled *de novo* using the SPAdes assembler with default parameters available in Galaxy. Assembly quality metrics were generated using the QUAST tool without supplying a reference genome, consistent with the assignment's emphasis on evaluating draft genome assemblies rather than optimising genome closure.

The resulting assembly statistics were as follows:

- **Total assembly length:** 5,150,460 bp
- **Number of contigs:** 153
- **N50:** 222,743 bp
- **Largest contig:** 595,984 bp

These values indicate a fragmented but acceptable bacterial draft genome. Although the assembly is not fully closed, the overall genome size and the presence of several large contigs make it suitable for downstream RNA-seq read mapping and interpretation.

3 RNA-seq Read Mapping Results

Paired-end RNA-seq reads from SRR35569531 were mapped to the *de novo* assembled genome using Bowtie2 with default alignment parameters. The assembled contigs FASTA file was used as the reference genome, and mapping statistics were extracted using Samtools flagstat.

The alignment results were:

- **Total RNA-seq reads:** 54,655,298
- **Mapped reads:** 48,644,755
- **Unmapped reads:** 6,010,543
- **Properly paired reads:** 15,187,714 (27.79%)

4 Mapping Statistics and Calculations

Mapping fractions were calculated using the formulas provided in the assignment instructions.

$$\text{Mapped fraction} = \left(\frac{48,644,755}{54,655,298}\right) \times 100 = 89.00\% \tag{1}$$

$$\text{Unmapped fraction} = \left(\frac{6,010,543}{54,655,298}\right) \times 100 = 11.00\% \tag{2}$$

These results indicate that the majority of RNA-seq reads successfully aligned to the draft genome assembly.

5 Interpretation and Discussion

5.1 Mapping percentage and assembly quality

An overall RNA-seq mapping rate of 89.00% is high for a *de novo* bacterial genome assembly. This indicates that most expressed genes have corresponding sequences present in the assembled contigs, suggesting that the genome assembly is genomically largely complete despite being structurally fragmented into 153 contigs.

5.2 Biological factors contributing to unmapped reads

Some unmapped RNA-seq reads can be attributed to biological features of *Salmonella Typhimurium* D23580. This strain harbours plasmids and prophages that contain repetitive or mobile genetic elements, which are difficult to assemble accurately using short-read sequencing. If such regions are partially assembled or absent from the draft genome, transcripts originating from these elements may fail to map. In addition, transcripts from low-abundance or condition-specific genes may correspond to genomic regions represented by short or fragmented contigs, further contributing to the unmapped read fraction.

5.3 Technical factors contributing to unmapped or improperly paired reads

Technical limitations of the assembly and mapping process also play a major role. The fragmentation of the genome into 153 contigs significantly reduces alignment quality for paired-end RNA-seq data. As illustrated in Figure 1, the assembly reaches the total genome size of 5.15 Mb through a long tail of smaller contigs, visually confirming this structural fragmentation.

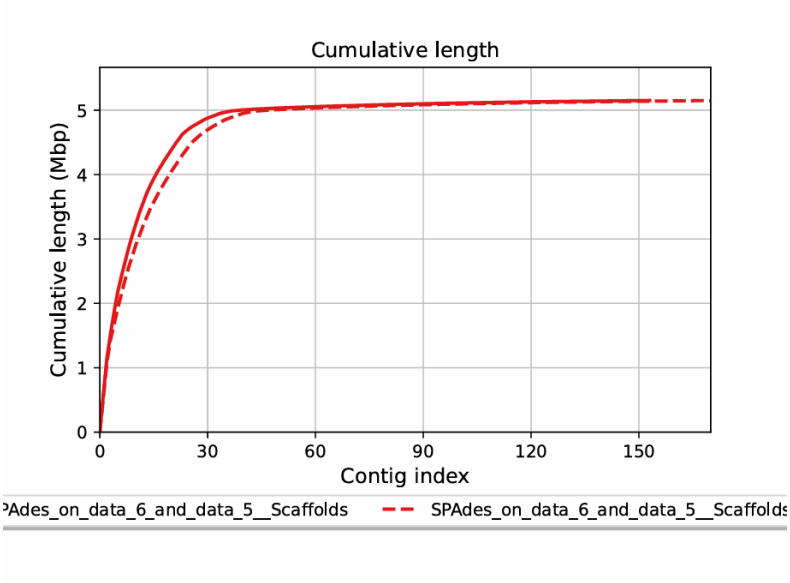

Figure 1: Cumulative length plot showing assembly fragmentation into small contigs.

This fragmentation disrupts the concordant alignment of paired-end reads, particularly those spanning contig boundaries. Notably, only **27.79%** of reads were reported as "properly paired" despite the **high 89% overall mapping rate**. This sharp contrast indicates that while most individual reads can align to the genome, the fragmented assembly frequently prevents correct concordant placement of paired-end reads. Read pairs spanning these boundaries lose their proper orientation information, leading to discordant or improperly paired alignments. Additional technical factors contributing to the unmapped fraction include sequencing errors, residual ribosomal RNA contamination, and inherent limitations of short-read alignment algorithms.

5.4 Use of a complete reference genome

If a complete reference genome were used instead of a *de novo* draft assembly, the RNA-seq mapping rate would be expected to increase. A closed reference genome resolves repetitive regions, fills assembly gaps, and typically includes plasmid and prophage sequences, providing continuous genomic coverage. Under such conditions, more RNA-seq read pairs would align concordantly across uninterrupted genomic regions, reducing both the unmapped fraction and the proportion of improperly paired reads.

Acknowledgements

I would like to express my gratitude to my instructor, Mahdi Moosa, PhD, for his guidance and support throughout this course. I also thank the Department of Mathematics and Natural Sciences at BRAC University for providing the necessary resources for this project.

Appendix: Galaxy Workflow and Analysis History

All analyses were performed using the Galaxy platform, following the structured bioinformatics pipeline illustrated in Figure 2. This workflow covers the entire process from initial data acquisition to final statistical interpretation.

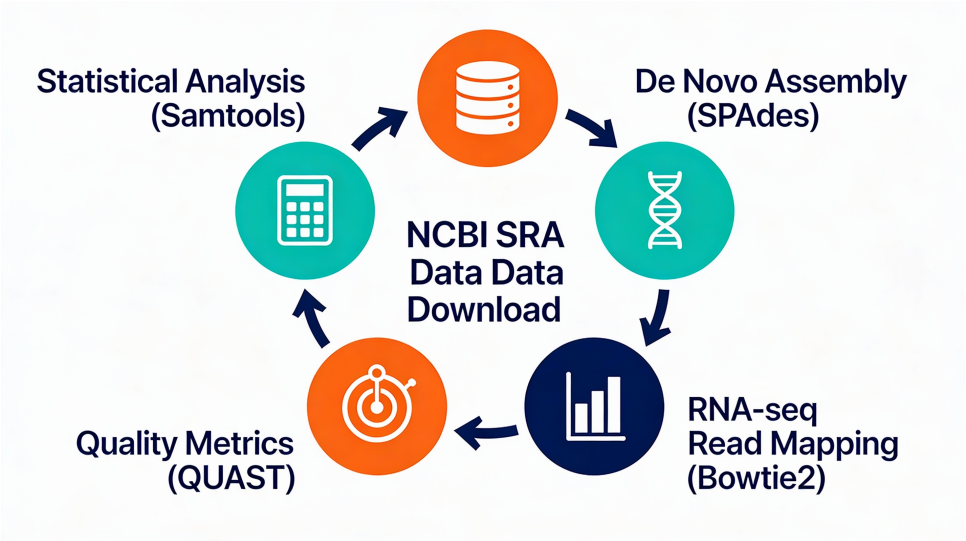

Figure 2: Galaxy workflow for *S. Typhimurium* genome assembly and RNA-seq mapping.

The complete Galaxy analysis history, containing all raw datasets, intermediate files (such as SPAdes contigs), and final mapping outputs (Samtools flagstat), is available for review at: *Publicly Available URL:* `https://usegalaxy.org.au/u/atkia_anika/h/final-406-assignment`