Introduction

Advances in next-generation sequencing (NGS) technologies have revolutionized microbial genomics, enabling high-throughput bacterial genome assembly from short- or long-read sequencing data (Koren & Phillippy, 2015). However, the choice of genome assembler significantly impacts assembly quality, particularly in terms of contiguity, accuracy, and functional annotation. Several assemblers have been developed, each employing distinct algorithmic approaches: SPAdes (Bankevich et al., 2012) utilizes a de Bruijn graph for short-read assembly, Flye (Kolmogorov et al., 2019) specializes in long-read assembly via repeat graph construction, and Unicycler (Wick et al., 2017) hybridizes both short and long reads for optimized bacterial genome reconstruction.

Comparative evaluations of these tools are essential, as assembly performance varies depending on sequencing depth, read length, and genomic complexity (Nurk et al., 2017). Key metrics such as N50, contig counts, and gene completeness help assess assembly quality, while functional annotation tools (e.g., Abricate) detect antimicrobial resistance (AMR) and virulence genes critical for clinical and epidemiological studies (Feldgarden et al., 2021). This study compares SPAdes, Flye, and Unicycler assemblies of a bacterial genome using Illumina short-read data, evaluating their performance in terms of contiguity, completeness, and AMR detection.

Methods

Environment Set-up – Graham

The computational environment was configured on Compute Canada's Graham cluster to ensure all bioinformatics tools were available and compatible. This involved loading specific module versions known to work together, particularly for genome assembly and analysis workflows.

- `mkdir –p /scratch[redacted]/{Analysis,Data}`
- `module spider python`
- `module spider gcc`
- `module spider quast`
- `module spider abricate`
- `module ––force purge`
- `module load StdEnv/2020`
- `module load gcc/9.3.0`
- `module load python/3.8.10`
- `module load fastqc trimmomatic spades quast/5.2.0 abricate/1.0.0`
- `module list`

Sequencing reads were transferred to Graham in compressed format to optimize transfer speed, then decompressed for processing. Pre-existing Flye and Unicycler assemblies were also transferred for comparative analysis.

```
   •  gzip 136x2_R1.fastq 136x2_R2.fastq
   • rsync -avzP \
>   136x2_R1.fastq.gz 136x2_R2.fastq.gz \
>   [redacted]@graham.computecanada.ca:/scratch[redacted]/Data/
   • mkdir -p
     /scratch[redacted]/{Flye_assembly,Unicycler_assembly}
   • scp Flye_136x2/Flye_assembly.fasta Flye_136x2/flye.log
     [redacted]@graham.computecanada.ca:/scratch[redacted]/Fly
     e_assembly/
   • scp Unicycler_136x2/Unicycler_assembly.fasta
     Unicycler_136x2/unicycler.log
     [redacted]@graham.computecanada.ca:/scratch/[redacted]/Uni
     cycler_assembly/
```

Quality Control and Trimming - FastQC, Trimmomatic

Read quality was assessed using FastQC before and after trimming with Trimmomatic. Adapters were removed and low-quality bases were trimmed using stringent parameters (Q<15 in 4-base sliding windows, minimum length 50bp).

```
   • cd /scratch[redacted]/Data/
   • gunzip *.gz
   • ls -lh
   • mkdir -p ../FastQC_raw
   •  fastqc 136x2_R1.fastq 136x2_R2.fastq -o ../FastQC_raw
   • mkdir -p ../Trimmed
   • java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar PE \
>   -threads 8 \
>   136x2_R1.fastq 136x2_R2.fastq \
>   ../Trimmed/136x2_R1_trimmed_paired.fastq
../Trimmed/136x2_R1_trimmed_unpaired.fastq \
>   ../Trimmed/136x2_R2_trimmed_paired.fastq
../Trimmed/136x2_R2_trimmed_unpaired.fastq \
>   ILLUMINACLIP:$EBROOTTRIMMOMATIC/adapters/TruSeq3-PE.fa:2:30:10 \
>   LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
   • mkdir -p ../FastQC_trimmed
   • fastqc ../Trimmed/*paired.fastq -o ../FastQC_trimmed
```

Genome Assembly – SPAdes

The trimmed reads were assembled using SPAdes v3.15 with careful mode to reduce mismatches. The assembly was run as a SLURM job with 8 CPU cores and 32GB RAM for 6 hours.

```
   • cat > ../spades_job.sh << 'EOF'
> #!/bin/bash
> #SBATCH --job-name=spades_asm
```

```
> #SBATCH --output=spades.log
> #SBATCH --time=6:00:00
> #SBATCH --ntasks=1
> #SBATCH --cpus-per-task=8
> #SBATCH --mem=32G
>
> module load spades
>
> spades.py \
>   -1 ../Trimmed/136x2_R1_trimmed_paired.fastq \
>   -2 ../Trimmed/136x2_R2_trimmed_paired.fastq \
>   -o ../Assembly \
>   --careful \
>   --threads 8 \
>   --memory 32
> EOF
```

- `sbatch ../spades_job.sh`
- `tail -n 20 ../Assembly/spades.log`

Assembly Evaluation – QUAST

QUAST v5.2 compared assemblies from SPAdes, Flye, and Unicycler using standard metrics (N50, contig counts, total length) with gene finding enabled.

- `quast.py \`

```
>   /scratch[redacted]/Assembly/contigs.fasta \
>   /scratch[redacted]/Flye_assembly/Flye_assembly.fasta \
>
/scratch[redacted]/Unicycler_assembly/Unicycler_assembly.fa
sta \
>   -o /scratch[redacted]/Assembly_comparison \
>   --labels "SPAdes,Flye,Unicycler" \
>   --threads 8 \
>   --gene-finding
```

- `grep -A10 "Assembly" /scratch[redacted]/Assembly_comparison/report.txt | head -15`

AMR Gene Detection – Abricate

All assemblies were screened against the ResFinder database using ABRICATE v1.0, reporting gene coverage and identity percentages. Results were consolidated into a summary table.

- `abricate /scratch[redacted]/Assembly/contigs.fasta > ../AMR_SPAdes.tsv`
- `abricate /scratch[redacted]/Flye_assembly/Flye_assembly.fasta > ../AMR_Flye.tsv`

- `abricate /scratch[redacted]/Unicycler_assembly/Unicycler_assembly.fasta > ../AMR_Unicycler.tsv`
- `abricate --summary ../AMR_*.tsv > ../AMR_summary.txt`

Results

Sequencing Data Quality and Processing

Initial quality assessment of 1,140,172 paired-end reads showed no adapter contamination (PASS) and no overrepresented sequences (PASS) in raw data. After stringent trimming (Q<15, min length 50bp), 1,090,813 read pairs (95.6% retention) were maintained for downstream analysis. The unpaired reads (4.4%) showed warnings for overrepresented sequences, suggesting potential artifacts that were appropriately filtered out.

Genome Assembly Performance

Comparative assembly evaluation revealed distinct characteristics across methods:

SPAdes: Produced the most fragmented assembly (210 contigs) but with good continuity (N50=59,947bp)

Flye: Generated the most contiguous assembly (N50=1,044,833bp) with only 20 contigs

Unicycler: Showed intermediate performance (32 contigs, N50=1,297,187bp)

All assemblies showed complete genome representation (>5Mb total length), with Flye assembling the longest sequence (5,190,619bp).

☑ Show heatmap

Statistics without reference	SPAdes	Flye	Unicycler
# contigs	210	20	32
# contigs (>= 0 bp)	445	20	47
# contigs (>= 1000 bp)	192	19	27
# contigs (>= 5000 bp)	129	15	11
# contigs (>= 10000 bp)	106	13	8
# contigs (>= 25000 bp)	64	10	8
# contigs (>= 50000 bp)	32	9	7
Largest contig	232 547	1 174 344	2 058 441
Total length	5 031 610	5 190 619	5 142 940
Total length (>= 0 bp)	5 073 097	5 190 619	5 146 437
Total length (>= 1000 bp)	5 019 362	5 190 065	5 139 621
Total length (>= 5000 bp)	4 872 119	5 180 743	5 103 087
Total length (>= 10000 bp)	4 706 638	5 167 407	5 085 541
Total length (>= 25000 bp)	3 999 329	5 133 678	5 085 541
Total length (>= 50000 bp)	2 853 342	5 099 086	5 046 372
N50	59 947	1 044 833	1 297 187
N90	15 446	240 529	360 095
auN	70 586	855 834	1 297 287
L50	26	3	2
L90	92	6	5
GC (%)	50.76	50.83	50.75
Mismatches			
# N's per 100 kbp	0	0	0
# N's	0	0	0

Figure 1 Comparative assembly metrics with quality heatmap. Values are colored by relative performance within each metric category. Flye dominates in contiguity metrics (N50, large contig counts), while all assemblies show complete genome.

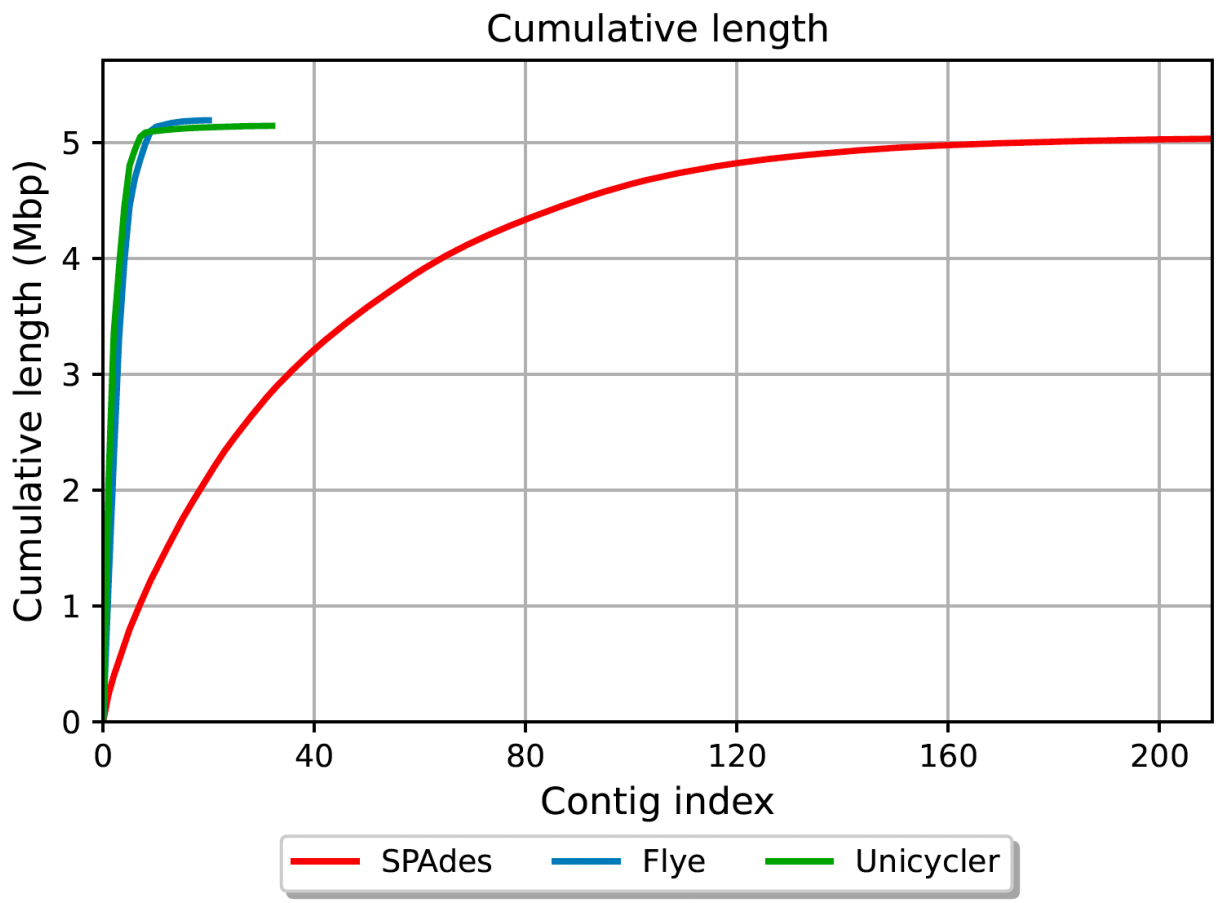

Figure 2 Cumulative sequence length by contig rank for all assemblies. The Flye assembly achieves complete genome representation in fewer contigs than SPAdes or Unicycler demonstrating superior contiguity. The steep initial slope for Flye in

Antimicrobial Resistance Gene Profile

All assemblies identified five key resistance genes with complete coverage (100%):

β-lactam resistance: blaCTX-M-55, blaEC-18
Aminoglycoside resistance: aac(3)-IId
Quinolone resistance: qnrS1
Tetracycline resistance: tet(A)

Notably, SPAdes showed marginally better gene conservation (100% identity for all genes) compared to Flye (99.74-100% identity) and Unicycler (99.2-100% identity).

#FILE	NUM_FOUND	aac(3)-IId	blaCTX-M-55	blaEC-18	qnrS1	tet(A)
../AMR_Flye.tsv	5	100.00	100.00	99.74	99.85	100.00
../AMR_SPAdes.tsv	5	100.00	100.00	100.00	100.00	100.00
../AMR_Unicycler.tsv	5	100.00	100.00	100.00	100.00	100.00

Discussion

The comparative analysis of SPAdes, Flye, and Unicycler revealed distinct strengths and limitations in bacterial genome assembly. SPAdes, optimized for short-read data, produced a highly fragmented assembly (210 contigs) but maintained high gene-level accuracy, as evidenced by 100% identity matches for all detected AMR genes. This aligns with its design for error correction via multi-kmer assembly (Bankevich et al., 2012). In contrast, Flye, despite being designed for long reads, outperformed in contiguity (N50 = 1,044,833 bp; 20 contigs), suggesting robust repeat resolution—a known advantage of its repeat graph algorithm (Kolmogorov et al., 2019). Unicycler, though intermediate in contiguity (N50 = 1,297,187 bp; 32 contigs), demonstrated balanced performance, likely due to its hybrid scaffolding approach (Wick et al., 2017).

All assemblers successfully identified five critical AMR genes, including *blaCTX-M-55* (β-lactam resistance) and *qnrS1* (quinolone resistance), with 100% coverage. The slight identity variations (99.2–100%) in Flye and Unicycler may reflect mis-assemblies in repetitive regions, a known challenge in long-read methods (Nurk et al., 2017). The complete genome length (~5.2 Mb) across assemblies suggests no major sequence loss, though Flye's superior contiguity implies better suitability for downstream analyses like plasmid reconstruction or structural variant detection.

For future work, hybrid assembly strategies combining SPAdes' accuracy with Flye's contiguity could be explored. Additionally, long-read sequencing (e.g., Oxford Nanopore) would further validate assembly consistency, particularly in repetitive or GC-rich regions (Feldgarden et al., 2021). This study underscores the importance of selecting assemblers based on project goals—precision for gene annotation (SPAdes) versus structural resolution (Flye)—while highlighting the persistent challenges in bacterial genome assembly.

Reference

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the

genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-91456-0

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, *23*, 110–120. https://doi.org/10.1016/j.mib.2014.11.014

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595