245HSBD016
Yash Priya Baid

NGS Data Analysis Practical – Test 1

Part-A: Genome Assembly

Choose Staphylococcus aureus OR Drosophila melanogaster.

Tasks:

1. **Perform read QC and trimming.**
2. **Assemble the genome using the appropriate tool.**
3. **Assess assembly quality.**
4. **Annotate the assembly.**

Step-0: WGS *Staphylococcus aureus* paired-end data (SRR22796053) was downloaded before beginning the test. Following sub-directories were created in 'bacterial_genome_assembly' directory:

mkdir raw_data fastqc trimgalore spades quast

All the below operations were performed in their respective folders only, by giving the correct path of the required inputs.

Step-1: FASTQC

fastqc -o /home/ibab/NGS/bacterial_genome_assembly/fastqc
/home/ibab/NGS/bacterial_genome_assembly/raw_data/*.fastq.gz

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/fastqc$ fastqc -o /home/ibab/NGS/bacterial_genome_assembly/fastqc /home/ibab/NGS/bacterial_genome_assembly/raw_data/*.fastq.gz
Started analysis of SRR22796053_1.fastq.gz
Approx 5% complete for SRR22796053_1.fastq.gz
Approx 10% complete for SRR22796053_1.fastq.gz
Approx 15% complete for SRR22796053_1.fastq.gz
Approx 20% complete for SRR22796053_1.fastq.gz
Approx 25% complete for SRR22796053_1.fastq.gz
Approx 30% complete for SRR22796053_1.fastq.gz
Approx 35% complete for SRR22796053_1.fastq.gz
Approx 40% complete for SRR22796053_1.fastq.gz
Approx 45% complete for SRR22796053_1.fastq.gz
Approx 50% complete for SRR22796053_1.fastq.gz
Approx 55% complete for SRR22796053_1.fastq.gz
Approx 60% complete for SRR22796053_1.fastq.gz
Approx 65% complete for SRR22796053_1.fastq.gz
Approx 70% complete for SRR22796053_1.fastq.gz
Approx 75% complete for SRR22796053_1.fastq.gz
Approx 80% complete for SRR22796053_1.fastq.gz
Approx 85% complete for SRR22796053_1.fastq.gz
Approx 90% complete for SRR22796053_1.fastq.gz
Approx 95% complete for SRR22796053_1.fastq.gz
Analysis complete for SRR22796053_1.fastq.gz
Started analysis of SRR22796053_2.fastq.gz
Approx 5% complete for SRR22796053_2.fastq.gz
Approx 10% complete for SRR22796053_2.fastq.gz
Approx 15% complete for SRR22796053_2.fastq.gz
Approx 20% complete for SRR22796053_2.fastq.gz
Approx 25% complete for SRR22796053_2.fastq.gz
Approx 30% complete for SRR22796053_2.fastq.gz
Approx 35% complete for SRR22796053_2.fastq.gz
Approx 40% complete for SRR22796053_2.fastq.gz
Approx 45% complete for SRR22796053_2.fastq.gz
Approx 50% complete for SRR22796053_2.fastq.gz
Approx 55% complete for SRR22796053_2.fastq.gz
Approx 60% complete for SRR22796053_2.fastq.gz
Approx 65% complete for SRR22796053_2.fastq.gz
Approx 70% complete for SRR22796053_2.fastq.gz
Approx 75% complete for SRR22796053_2.fastq.gz
Approx 80% complete for SRR22796053_2.fastq.gz
```

Before trimming – file size

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/fastqc$ ls -lh
total 1.7M
-rw-r--r-- 1 ibab ibab 566K Aug 25 15:11 SRR22796053_1_fastqc.html
-rw-r--r-- 1 ibab ibab 259K Aug 25 15:11 SRR22796053_1_fastqc.zip
-rw-r--r-- 1 ibab ibab 566K Aug 25 15:12 SRR22796053_2_fastqc.html
-rw-r--r-- 1 ibab ibab 258K Aug 25 15:12 SRR22796053_2_fastqc.zip
```

Step-3: Trimming (default parameters) along with FASTQC

/home/ibab/NGS/Packages/TrimGalore/trim_galore –paired /home/ibab/NGS/bacterial_genome_assembly/raw_data/SRR22796053_1.fastq.gz /home/ibab/NGS/bacterial_genome_assembly/raw_data/SRR22796053_2.fastq.gz -q 25 --stringency 5 --fastqc -o /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/trimgalore$ /home/ibab/NGS/Packages/TrimGalore/trim_galore --paired /home/ibab/NGS/bacterial_genome_ass
embly/raw_data/SRR22796053_1.fastq.gz /home/ibab/NGS/bacterial_genome_assembly/raw_data/SRR22796053_2.fastq.gz -q 25 --stringency 5 --fastqc -o /home/ibab/N
GS/bacterial_genome_assembly/trimgalore/SRR22796053
Multicore support not enabled. Proceeding with single-core trimming.
Path to Cutadapt set as: 'cutadapt' (default)
Cutadapt seems to be working fine (tested command 'cutadapt --version')
Cutadapt version: 3.5
single-core operation.
Proceeding with 'gzip' for decompression
To decrease CPU usage of decompression, please install 'igzip' and run again

No quality encoding type selected. Assuming that the data provided uses Sanger encoded Phred scores (default)

Output directory /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/ doesn't exist, creating it for you...

Output will be written into the directory: /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/


AUTO-DETECTING ADAPTER TYPE
===========================
Attempting to auto-detect adapter type from the first 1 million sequences of the first file (>> /home/ibab/NGS/bacterial_genome_assembly/raw_data/SRR2279605
3_1.fastq.gz <<)

Found perfect matches for the following adapter sequences:
Adapter type    Count   Sequence        Sequences analysed      Percentage
Illumina        0       AGATCGGAAGAGC   1000000 0.00
smallRNA        0       TGGAATTCTCGG    1000000 0.00
Nextera 0       CTGTCTCTTATA    1000000 0.00
Unable to auto-detect most prominent adapter from the first specified file (count Illumina: 0, count smallRNA: 0, count Nextera: 0)
Defaulting to Illumina universal adapter ( AGATCGGAAGAGC ). Specify -a SEQUENCE to avoid this behavior).
```

After trimming – file size

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/trimgalore$ cd SRR22796053/
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/trimgalore/SRR22796053$ ls -lh
total 294M
-rw-r--r-- 1 ibab ibab 2.2K Aug 25 15:36 SRR22796053_1.fastq.gz_trimming_report.txt
-rw-r--r-- 1 ibab ibab 146M Aug 25 15:40 SRR22796053_1_val_1.fq.gz
-rw-r--r-- 1 ibab ibab 524K Aug 25 15:41 SRR22796053_1_val_1_fastqc.html
-rw-r--r-- 1 ibab ibab 242K Aug 25 15:41 SRR22796053_1_val_1_fastqc.zip
-rw-r--r-- 1 ibab ibab 2.5K Aug 25 15:40 SRR22796053_2.fastq.gz_trimming_report.txt
-rw-r--r-- 1 ibab ibab 146M Aug 25 15:40 SRR22796053_2_val_2.fq.gz
-rw-r--r-- 1 ibab ibab 526K Aug 25 15:41 SRR22796053_2_val_2_fastqc.html
-rw-r--r-- 1 ibab ibab 241K Aug 25 15:41 SRR22796053_2_val_2_fastqc.zip
```

Comparison of fastqc results – before and after trimming

SRR22796053_1

Before trimming

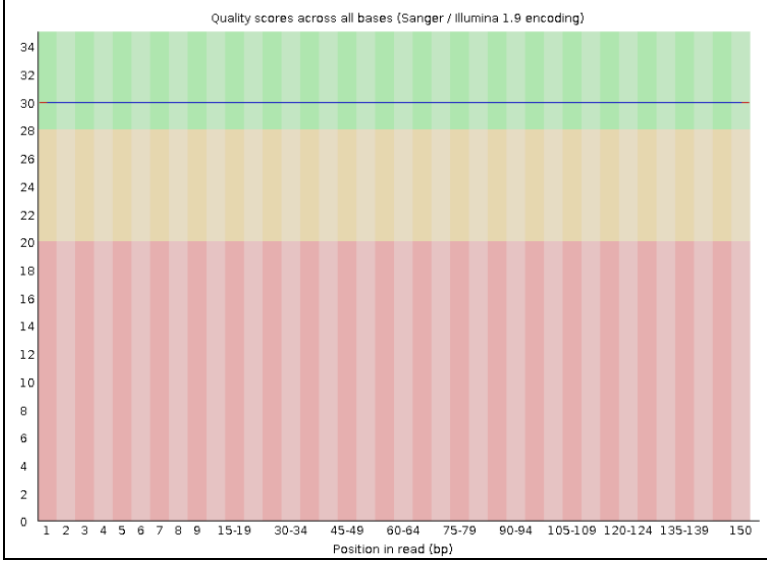
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR22796053_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3119273
Sequences flagged as poor quality	0
Sequence length	50-150
%GC	33

Per base sequence quality (PASS)



After trimming

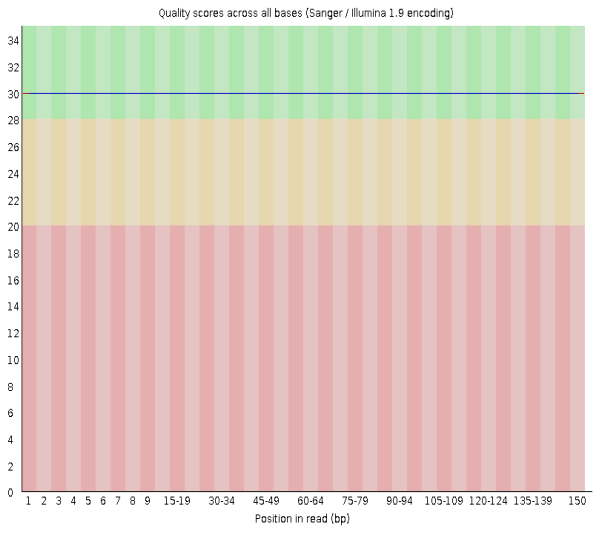
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

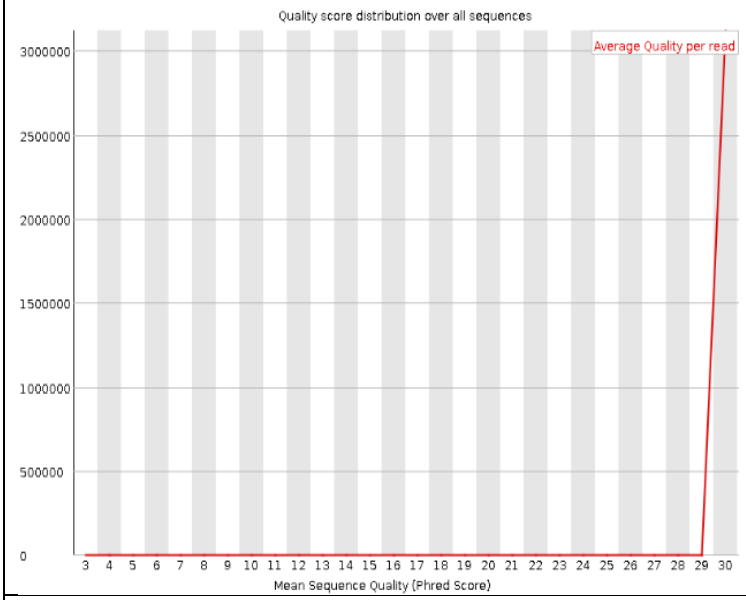
Basic Statistics

Measure	Value
Filename	SRR22796053_1_val_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3119066
Sequences flagged as poor quality	0
Sequence length	50-150
%GC	33

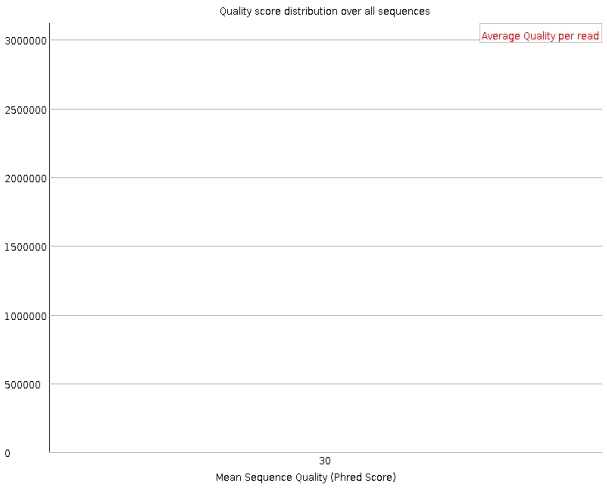
Per base sequence quality (PASS)



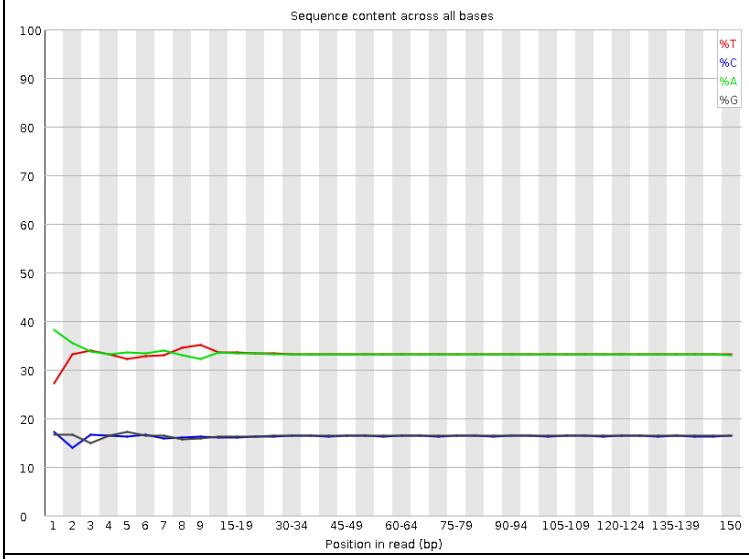
Per sequence quality scores (PASS)



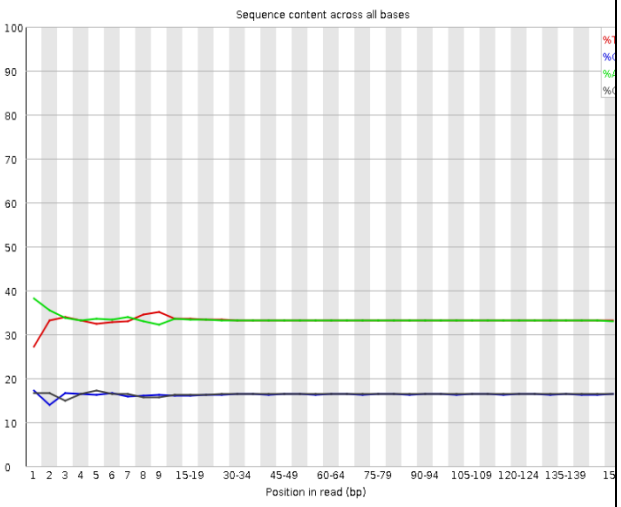
Per sequence quality scores (PASS)



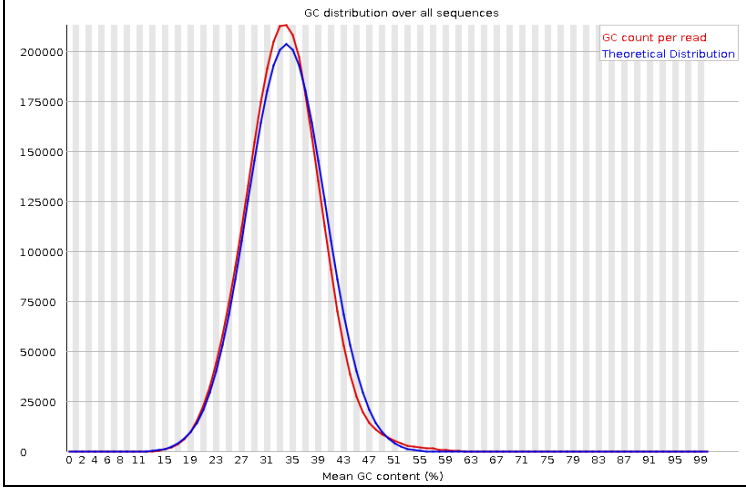
Per base sequence content (WARNING)



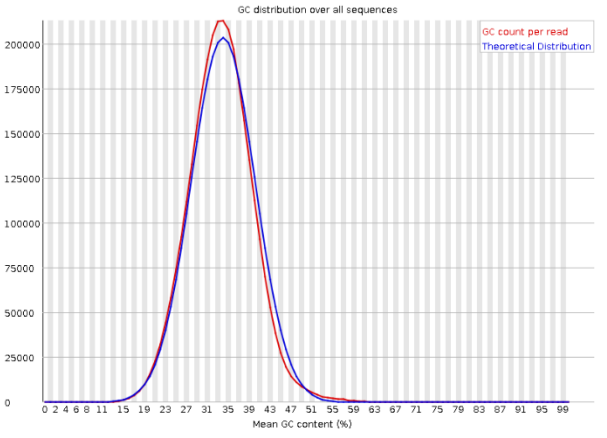
Per base sequence content (WARNING)



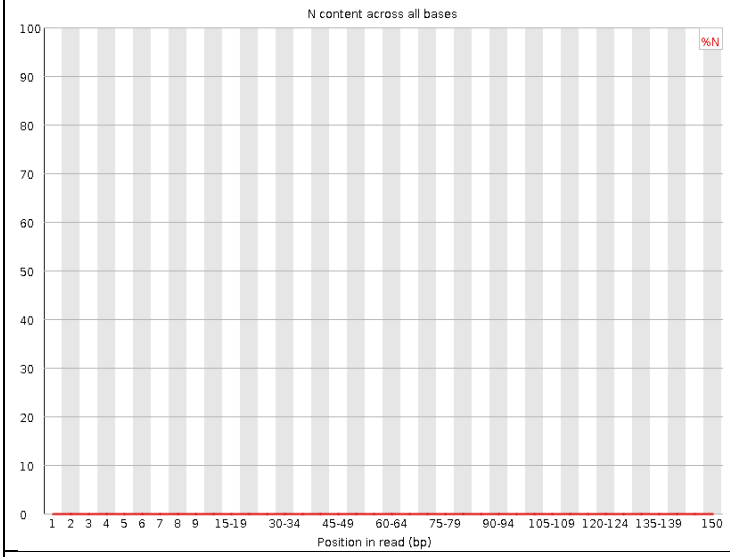
Per sequence GC content (PASS)



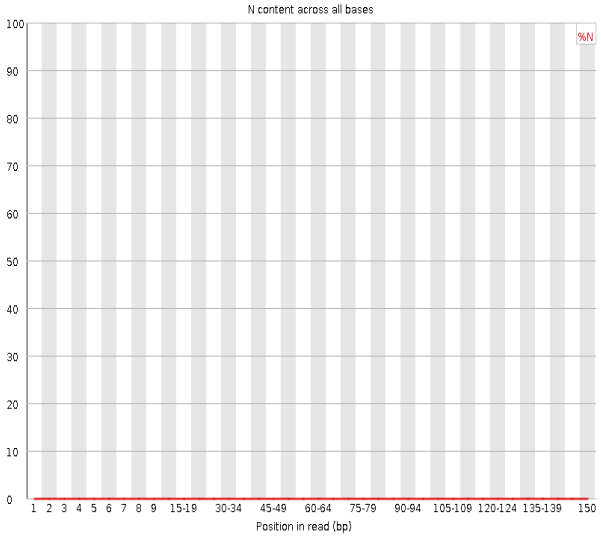
Per sequence GC content (PASS)



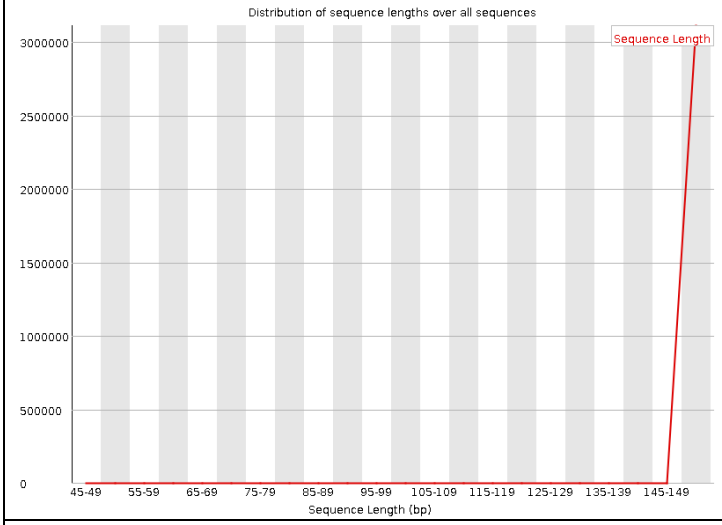
Per base N content (PASS)



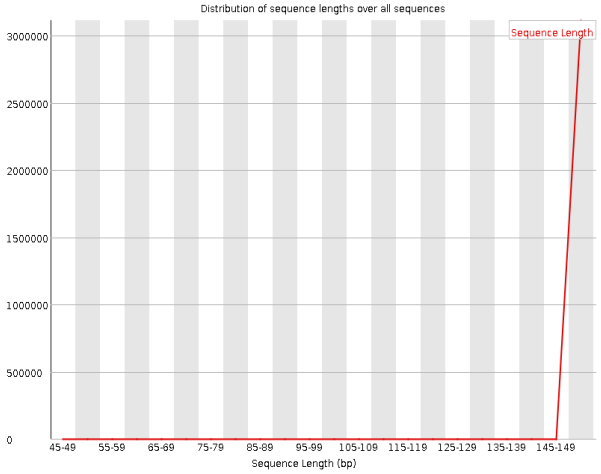
Per base N content (PASS)



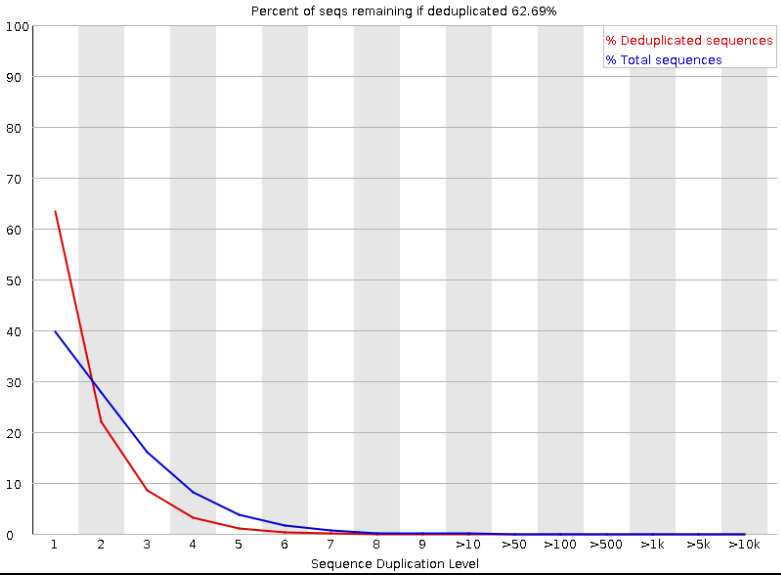
Sequence length distribution (WARNING)



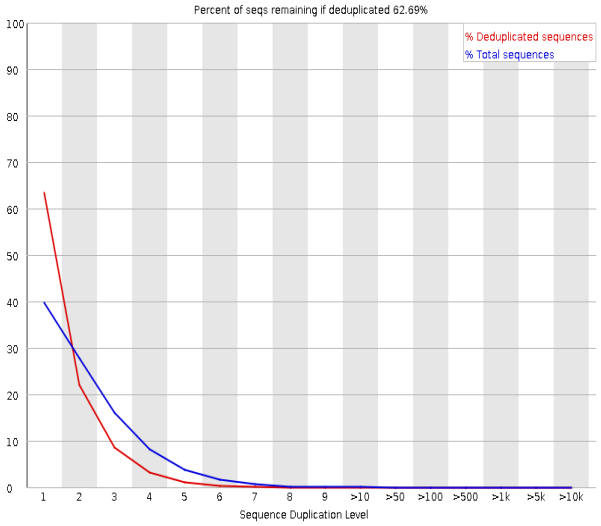
Sequence length distribution (WARNING)

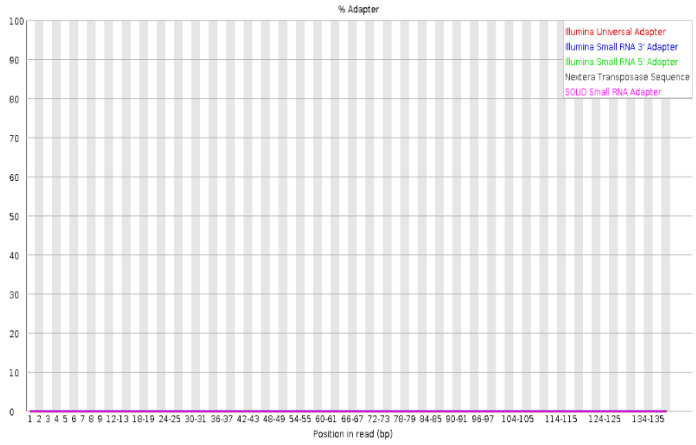
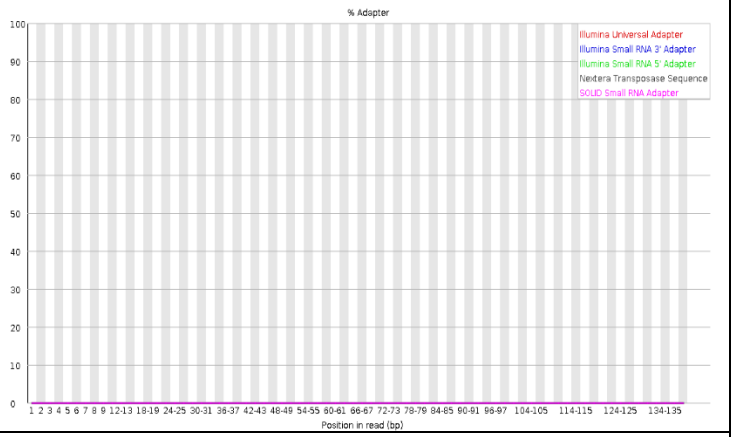


Sequence duplication levels (WARNING)



Sequence duplicate levels (WARNING)



Adapter content (PASS)	Adapter content (PASS)
	
No overrepresented sequences.	No overrepresented sequences.

Same fastqc results (before and after trimming) have been observed for SRR22796053_2 also.

Summary reports:

```
SUMMARISING RUN PARAMETERS
==========================
Input filename: /home/ibab/NGS/bacterial_genome_assembly/raw_data/SRR22796053_2.fastq.gz
Trimming mode: paired-end
Trim Galore version: 0.6.10
Cutadapt version: 3.5
Number of cores used for trimming: 1
Quality Phred score cutoff: 25
Quality encoding type selected: ASCII+33
Unable to auto-detect most prominent adapter from the first specified file (count Illumina: 0, count smallRNA: 0, count Nextera: 0)
Defaulting to Illumina universal adapter ( AGATCGGAAGAGC ). Specify -a SEQUENCE to avoid this behavior).
Adapter sequence: 'AGATCGGAAGAGC' (Illumina TruSeq, Sanger iPCR; default (inconclusive auto-detection))
Maximum trimming error rate: 0.1 (default)
Minimum required adapter overlap (stringency): 5 bp
Minimum required sequence length for both reads before a sequence pair gets removed: 20 bp
Running FastQC on the data once trimming has completed
Output file will be GZIP compressed
```

```
=== Summary ===

Total reads processed:               3,119,273
Reads with adapters:                       199 (0.0%)
Reads written (passing filters):     3,119,273 (100.0%)

Total basepairs processed:   467,752,229 bp
Quality-trimmed:                 3,210 bp (0.0%)
Total written (filtered):    467,724,282 bp (100.0%)

=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 199 times

Minimum overlap: 5
No. of allowed errors:
1-9 bp: 0; 10-13 bp: 1

Bases preceding removed adapters:
  A: 4.5%
  C: 3.0%
  G: 4.5%
  T: 6.5%
  none/other: 81.4%

Overview of removed sequences
length  count   expect  max.err error counts
9       17      11.9    0       0 17
10      8       3.0     1       0 8
11      9       0.7     1       0 9
12      1       0.2     1       0 1
38      1       0.0     1       0 1
59      1       0.0     1       0 1
146     1       0.0     1       0 1
150     161     0.0     1       0 161
```

SRR22796053_1

```
=== Summary ===

Total reads processed:               3,119,273
Reads with adapters:                        47 (0.0%)
Reads written (passing filters):     3,119,273 (100.0%)

Total basepairs processed:   467,751,957 bp
Quality-trimmed:                 3,438 bp (0.0%)
Total written (filtered):    467,747,824 bp (100.0%)

=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 47 times

Minimum overlap: 5
No. of allowed errors:
1-9 bp: 0; 10-13 bp: 1

Bases preceding removed adapters:
  A: 27.7%
  C: 17.0%
  G: 27.7%
  T: 27.7%
  none/other: 0.0%

Overview of removed sequences
length  count   expect  max.err error counts
5       2       3046.2  0       2
9       11      11.9    0       0 11
10      20      3.0     1       0 20
11      5       0.7     1       0 5
12      1       0.2     1       0 1
17      1       0.0     1       0 1
22      1       0.0     1       0 1
23      1       0.0     1       0 1
28      1       0.0     1       0 1
29      1       0.0     1       0 1
41      1       0.0     1       1
44      1       0.0     1       1
115     1       0.0     1       0 1
```

SRR22796053_2

Inference:

FastQC analysis indicated that the overall sequencing quality was high, with per-base sequence quality passing across all position both before and after trimming. Also, no adapter contamination or overrepresented sequences were reported. But we can see warnings for per-base sequence content (may be due to bias – not problematic), sequence length distribution (may be due to full-length reads being dominant) & duplication levels (may be due to high seq. depth). Hence, there is no significant change or improvement observed after trimming.

Step-4: SPADes Assembly

spades.py – 1 /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_1_val_1.fq.gz -2 /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_2_val_2.fq.gz -o . -t 8 -m 32 --careful --cov-cutoff auto --phred-offset 33

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/spades$ spades.py -1 /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_1_val_
1.fq.gz -2 /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_2_val_2.fq.gz -o . -t 8 -m 32 --careful --cov-cutoff auto --phred-off
set 33
Command line: /usr/lib/spades/bin/spades.py     -1      /home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_1_val_1.fq.gz       -2 /
home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_2_val_2.fq.gz        -o      /home/ibab/NGS/bacterial_genome_assembly/spades -t 8
-m      32      --careful       --cov-cutoff    auto    --phred-offset  33

System information:
  SPAdes version: 3.13.1
  Python version: 3.10.12
  OS: Linux-6.6.87.2-microsoft-standard-WSL2-x86_64-with-glibc2.35

Output dir: /home/ibab/NGS/bacterial_genome_assembly/spades
Mode: read error correction and assembling
Debug mode is turned OFF

Dataset parameters:
  Multi-cell mode (you should set '--sc' flag if input data was obtained with MDA (single-cell) technology or --meta flag if processing metagenomic dataset)
  Reads:
    Library number: 1, library type: paired-end
      orientation: fr
      left reads: ['/home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_1_val_1.fq.gz']
      right reads: ['/home/ibab/NGS/bacterial_genome_assembly/trimgalore/SRR22796053/SRR22796053_2_val_2.fq.gz']
      interlaced reads: not specified
```

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/spades$ ls -lh
total 17M
drwxr-xr-x 3 ibab ibab 4.0K Aug 25 18:20 K21
drwxr-xr-x 3 ibab ibab 4.0K Aug 25 18:24 K33
drwxr-xr-x 3 ibab ibab 4.0K Aug 25 18:27 K55
drwxr-xr-x 4 ibab ibab 4.0K Aug 25 18:32 K77
-rw-r--r-- 1 ibab ibab 5.5M Aug 25 18:32 assembly_graph.fastg
-rw-r--r-- 1 ibab ibab 2.7M Aug 25 18:32 assembly_graph_with_scaffolds.gfa
-rw-r--r-- 1 ibab ibab 2.8M Aug 25 18:32 before_rr.fasta
-rw-r--r-- 1 ibab ibab 2.7M Aug 25 18:37 contigs.fasta
-rw-r--r-- 1 ibab ibab  14K Aug 25 18:32 contigs.paths
drwxr-xr-x 3 ibab ibab 4.0K Aug 25 18:14 corrected
-rw-r--r-- 1 ibab ibab   79 Aug 25 18:14 dataset.info
-rw-r--r-- 1 ibab ibab  270 Aug 25 17:25 input_dataset.yaml
drwxr-xr-x 2 ibab ibab 4.0K Aug 25 18:44 misc
drwxr-xr-x 4 ibab ibab 4.0K Aug 25 18:37 mismatch_corrector
-rw-r--r-- 1 ibab ibab 1.7K Aug 25 17:25 params.txt
-rw-r--r-- 1 ibab ibab 2.7M Aug 25 18:44 scaffolds.fasta
-rw-r--r-- 1 ibab ibab  13K Aug 25 18:32 scaffolds.paths
-rw-r--r-- 1 ibab ibab 182K Aug 25 18:44 spades.log
drwxr-xr-x 2 ibab ibab 4.0K Aug 25 18:44 tmp
```

Step-5: QUAST

/home/ibab/NGS/Packages/quast/quast.py -o . /home/ibab/NGS/bacterial_genome_assembly/spades/contigs.fasta /home/ibab/NGS/bacterial_genome_assembly/spades/scaffolds.fasta

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/bacterial_genome_assembly/quast$ ls -lh
total 492K
drwxr-xr-x 2 ibab ibab 4.0K Aug 27 17:05 basic_stats
-rw-r--r-- 1 ibab ibab  53K Aug 27 17:05 icarus.html
drwxr-xr-x 2 ibab ibab 4.0K Aug 27 17:05 icarus_viewers
-rw-r--r-- 1 ibab ibab 3.6K Aug 27 17:05 quast.log
-rw-r--r-- 1 ibab ibab 364K Aug 27 17:05 report.html
-rw-r--r-- 1 ibab ibab  35K Aug 27 17:05 report.pdf
-rw-r--r-- 1 ibab ibab 1.5K Aug 27 17:05 report.tex
-rw-r--r-- 1 ibab ibab  648 Aug 27 17:05 report.tsv
-rw-r--r-- 1 ibab ibab 1.3K Aug 27 17:05 report.txt
-rw-r--r-- 1 ibab ibab 1.2K Aug 27 17:05 transposed_report.tex
-rw-r--r-- 1 ibab ibab  648 Aug 27 17:05 transposed_report.tsv
-rw-r--r-- 1 ibab ibab 1.4K Aug 27 17:05 transposed_report.txt
```

QUAST

Quality Assessment Tool for Genome Assemblies

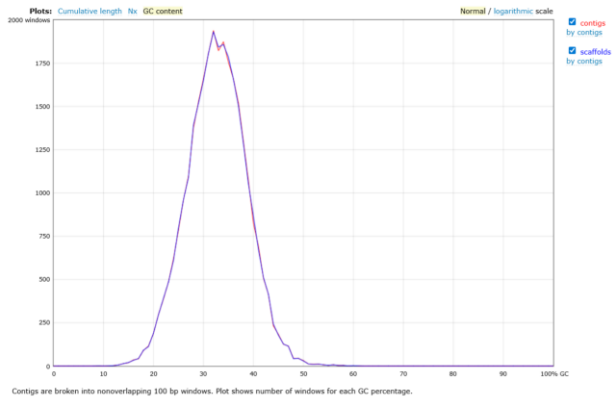
27 August 2025, Wednesday, 17:05:31

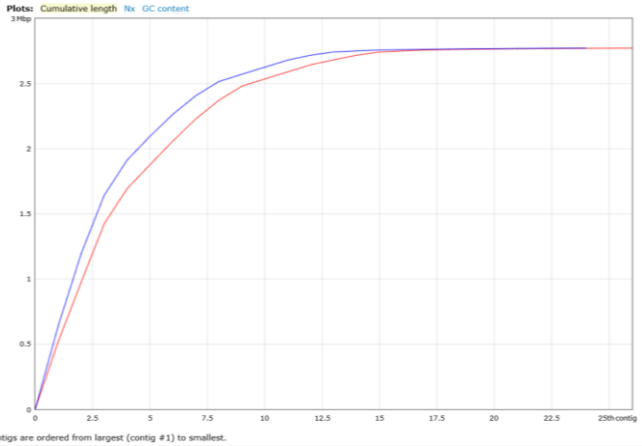
View in Icarus contig browser

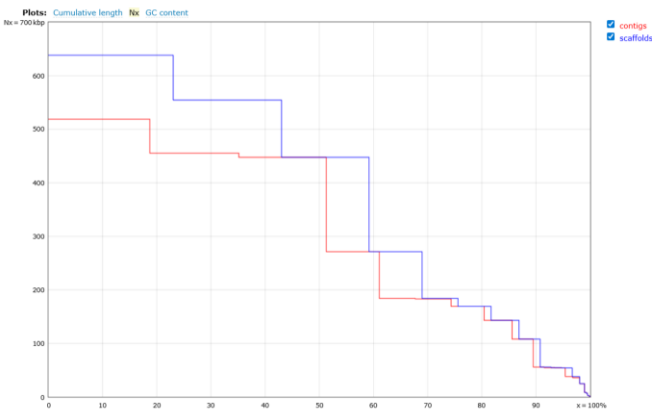
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	contigs	scaffolds
# contigs	26	24
# contigs (>= 0 bp)	76	74
# contigs (>= 1000 bp)	23	21
# contigs (>= 5000 bp)	17	15
# contigs (>= 10000 bp)	15	13
# contigs (>= 25000 bp)	14	12
# contigs (>= 50000 bp)	12	11
Largest contig	518 864	638 384
Total length	2 772 353	2 772 422
Total length (>= 0 bp)	2 780 871	2 780 940
Total length (>= 1000 bp)	2 770 275	2 770 344
Total length (>= 5000 bp)	2 758 140	2 758 209
Total length (>= 10000 bp)	2 742 958	2 743 027
Total length (>= 25000 bp)	2 718 322	2 718 391
Total length (>= 50000 bp)	2 644 709	2 680 458
N50	447 652	447 652
N90	55 620	107 982
auN	321 436	394 929
L50	3	3
L90	10	8
GC (%)	32.7	32.7
Per base quality		
# N's per 100 kbp	0	6.89
# N's	0	191







Interpretation:

The QUAST analysis report suggests that the assembly is of good quality, with a total assembly size of ~2.7 Mb. The largest contig is ~518 Kb. The N50 value is same in both; means that the length of contig or scaffold at which we've covered 50% of the genome is 447 Kb, whereas L50 value is also same in both, i.e., only 3 contigs or scaffolds needed to reach half of the genome size. This shows that the

assembly is not very fragmented. The GC content is ~32.7%, which is consistent across the assembly, as seen in the GC plot. The Nx plot shows that the scaffolding has slightly improved the continuity. Hence, the assembly is of high-quality with minimal fragmentation and good completeness, so reliable for downstream analyses.

Fill in the table below:

Metric	Result Contigs. fasta	Result Scaffols. fasta
# contigs ≥ 1 kb	23	21
N50 (bp)	447652	447652
L50	3	3
Largest contig (bp)	518864	638384
Total assembly size (Mb)	2.77	2.77
# CDS predicted	2580	2580
# rRNA genes	5	5
# tRNA genes	73	73
Example gene (locus + product)		

For gene info., look into the cds entry in .gff file or .tsv file –

contigs.fasta:

Locus Tag: PMCCGK_00002

Gene: aRO8

Product: HTH gntR-type domain-containing protein

scaffolds.fasta:

Locus Tag: OIDPMI_00001

Gene: sarT

Product: HTH-type transcriptional regulator

Step-6: BAKTA

Since, offline BAKTA tool couldn't be downloaded beforehand due to less memory / storage available. So, online BAKTA tool was used to annotate the assembly (use contigs.fasta,and scaffolds.fasta).

contigs.fasta

Bakta Web

Rapid & standardized annotation of bacterial genomes, MAGs & plasmids

Job statistics | Annotation table | Genomeviewer | Circular plot | Downloads

Input

Organism:	N.A.
Sequences:	76 contigs
Genome size:	2,780,871 bp

Runtime

Start:	27/08/2025, 17:53
Stop:	27/08/2025, 18:01
Duration:	7 minutes, 20 seconds

Statistics

N50	4,47,652 bp
N90	55,620 bp
GC-content	0.327
Coding ratio	0.85
N-ratio	0

Feature counts (Total: 2780)

tRNA:	73	**ncRNA regions:**	25	**oriC:**	4
tmRNA:	1	**CRISPR:**	0	**oriV:**	0
rRNA:	5	**CDS:**	2580	**oriT:**	1
ncRNA:	81	**sORF:**	10	**gap:**	0

scaffolds.fasta

Bakta Web

Rapid & standardized annotation of bacterial genomes, MAGs & plasmids

Job statistics | Annotation table | Genomeviewer | Circular plot | Downloads

Input

Organism:	N.A.
Sequences:	74 contigs
Genome size:	2,780,940 bp

Runtime

Start:	27/08/2025, 17:55
Stop:	27/08/2025, 18:07
Duration:	11 minutes, 51 seconds

Statistics

N50	4,47,652 bp
N90	1,07,982 bp
GC-content	0.327
Coding ratio	0.85
N-ratio	0

Feature counts (Total: 2781)

tRNA:	73	**ncRNA regions:**	25	**oriC:**	4
tmRNA:	1	**CRISPR:**	0	**oriV:**	0
rRNA:	5	**CDS:**	2580	**oriT:**	1
ncRNA:	80	**sORF:**	10	**gap:**	2