ANALYSIS OF THE M. TUBERCULOSIS GENOME ASSEMBLY USING SPAdes and BOWTIE2

By Om Mihani, Anagha Bhangare and Siddhant Osatwal

Introduction

Motivation for research

Tuberculosis (TB) is a major public health threat and a leading cause of death worldwide caused by the bacterium Mycobacterium tuberculosis. Some key facts about tuberculosis:

- TB remains the leading cause of death by a single infectious microorganism globally and the tenth leading cause of death.
- In 2018, there were around 10 million new TB cases and 1.5 million deaths from TB worldwide.
- Brazil, along with other BRICS countries, accounts for over 40% of the global TB burden.
- HIV coinfection increases the risk of developing active TB.

Motivation for this study

The motivation for this study was to unravel the genetic diversity and transmission dynamics of M. tuberculosis strains circulating in the metropolitan region of Florianópolis, Santa Catarina state in southern Brazil, which has a higher TB incidence (46 cases per 100,000 population) compared to the overall state incidence. Previous data on circulating M.tb strains in this region was limited to lower resolution genotyping methods like spoligotyping. Whole genome sequencing (WGS) provides higher resolution data to study strain diversity, drug resistance and transmission clusters.

About the dataset

The dataset consisted of 151 M. tuberculosis clinical isolates collected during May 2014 - May 2016 from new pulmonary TB patients in the cities of Florianópolis and São José in the metropolitan region. This represented 57.4% of the total 263 new TB cases reported in this region during that period. Clinical and epidemiological data like treatment outcomes, HIV status, drug abuse history etc. was also collected from the patients.

Alignment methods

De-novo assembly of reads is performed using SPAdes(St. Petersburg genome assembler). It is designed for assembling short-read sequencing data (e.g., Illumina reads) into longer contiguous sequences (contigs). It employs a multi-step assembly algorithm that integrates various strategies such as de Bruijn graph construction, error correction, scaffolding, and gap closure to generate high-quality genome assemblies. The output of SPAdes typically includes assembled contigs, scaffolds, and assembly graphs. The results of SPAdes are typically assessed using metrics such as N50, L50, GC% etc. The results are visualised using BUSCO plots.

For alignment of short-read sequencing data to a reference genome or transcriptome, the bowtie2 tool is used. It employs the Burrows-Wheeler Transform (BWT) algorithm and FM-indexing to index the reference genome and perform rapid read alignment. The output of bowtie2 alignment file is a SAM/BAM file which can be further processed and analysed using downstream bioinformatics tools for variant detection, gene expression analysis, and other applications. The results of bowtie2 are visualised using Interactive Genomics Viewer (IGV)

Methods

Sample Selection

Four samples from the Mycobacterium tuberculosis dataset were selected (BioProject ID: PRJNA599957) for further analysis. The samples were chosen using a random number generator, resulting in the following sample IDs:

1. Sample ID: SAMN13637889 (Renamed as mtb1)
2. Sample ID: SAMN13637911 (Renamed as mtb2)
3. Sample ID: SAMN13637893 (Renamed as mtb3)
4. Sample ID: SAMN13638010 (Renamed as mtb4)

These samples were extracted using the RunSelector tool to ensure consistency and accuracy in data retrieval. The reference genome used for alignment and analysis was obtained from the NCBI Genome database. We downloaded the Mycobacterium tuberculosis H37Rv genome assembly ASM19595v2 from the NCBI website

FastQC

FastQC is a quality control tool for assessing the quality of high-throughput sequencing data. It provides comprehensive insights into various aspects of sequencing data quality, including per-base sequence quality, GC content, sequence duplication levels, overrepresented sequences, and adapter contamination.

FastQC analysis was performed on the sequencing data, with the parameter ‘length of k-mer to look for’ set to 7. This parameter determines the size of k-mers (substrings of length k) analysed by FastQC to assess sequence composition and quality characteristics.

SPAdes

The SPAdes genome assembly tool was employed with modified parameters to investigate their impact on assembly outcomes. The following parameters were adjusted across three variations:

1. Operation Mode: The "only-assembler" mode was utilised for all three variations to focus solely on the assembly process without additional error correction or scaffolding steps.
2. Reads Configuration: Despite the original data being paired-end reads as indicated in the paper, for consistency, all variations employed single-end reads.
3. K-mer Size:
   Three k-mer size configurations were tested:
   - Auto SPAdes: The default k-mer sizes (21, 33, 55, and 77) were utilised.
   - Variation 1: A k-mer size of 15 was selected.
   - Variation 2: A k-mer size of 85 was selected.
4. Phred Quality Offset:
   The Phred quality offset was set to 33 (Sanger encoding) to ensure consistency in quality score interpretation across all variations.

Bowtie2

Bowtie2 was used to map the short-end reads to the reference genome assuming the reads to be single-end.

BUSCO

A BUSCO plot visualises the results of a BUSCO analysis, displaying the percentage of complete, fragmented, and missing orthologous genes in a genome assembly.
BUSCO plots were used to analyse the scaffolds obtained from SPAdes assembly.

Mummer

MUMmer (MUMmer) facilitates the detection of similarities and differences between large genomic sequences.
After comparing the plots generated for ‘minimum match length’ values of 1000, 5000, and 10000, the optimal value is determined to be 5000.

QUAST

Quast is a tool for quality assessment of genome assemblies. The assembly mode parameter was set to "individual assembly" for Quast analysis.

VarScan

Varscan is a tool used for variant detection in next-generation sequencing data. The analysis type parameter was modified to give nucleotide variations and indels in the sequenced genome.

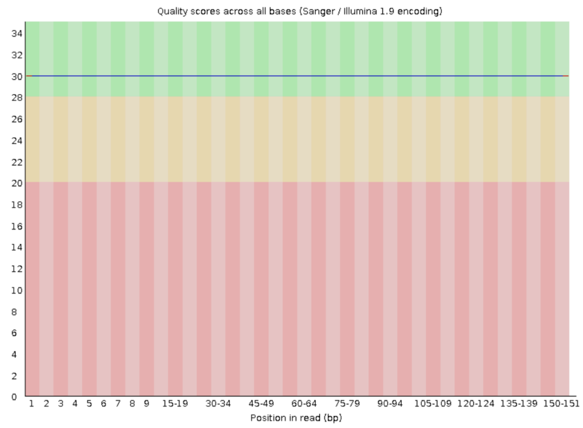
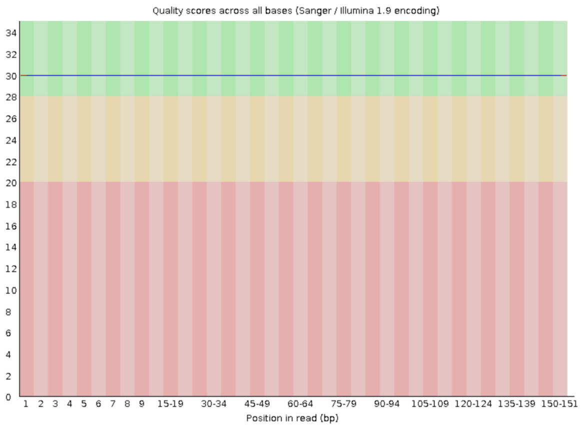
IGV

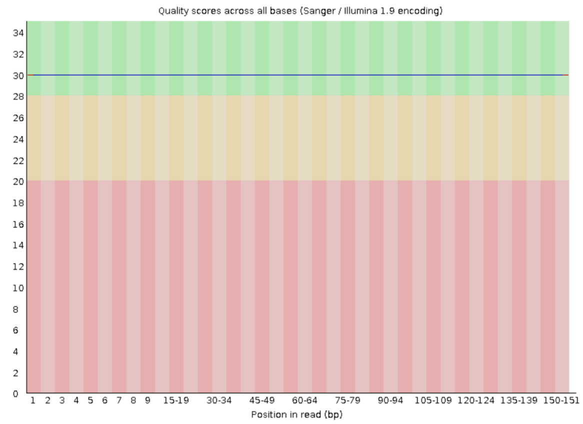
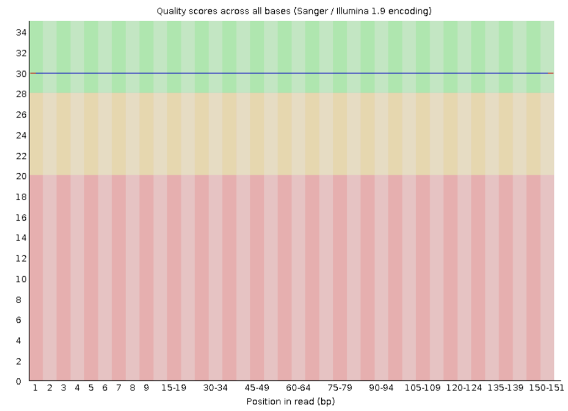
Interactive Genomics Viewer (IGV) was used to visualise the BAM file obtained from bowtie2. It highlights the single nucleotide variations and indels identified by varscan.

Results

FastQC

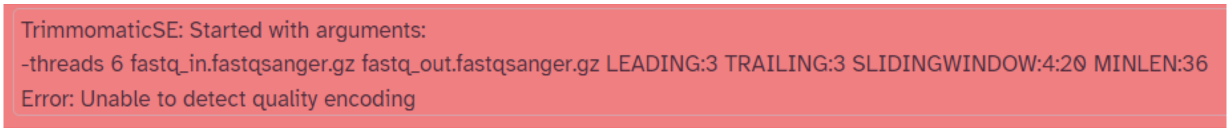
We ran FastQC on all files from mtb1 to mtb4.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>mtb1</td>
<td>mtb2</td>
</tr>
</table>

	
mtb3	mtb4

Note that, as you will observe above, the scores for all the bases are constantly = 30. Such a high phred score with no undetected base is possibly due to the high-quality sequencer used

It is possible that the sequences present here have already been trimmed, even though that wasn't the target. We still tried to trim it according to the prescribed parameters. However, this gives the following error



```
TrimmomaticSE: Started with arguments:
-threads 6 fastq_in.fastqsanger.gz fastq_out.fastqsanger.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:36
Error: Unable to detect quality encoding
```

Hence, we skip trimming and move ahead.

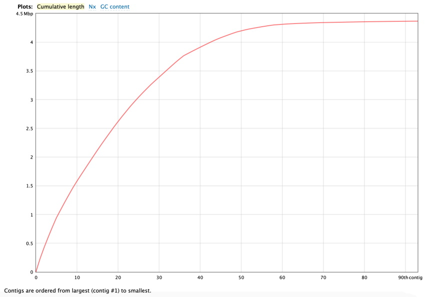
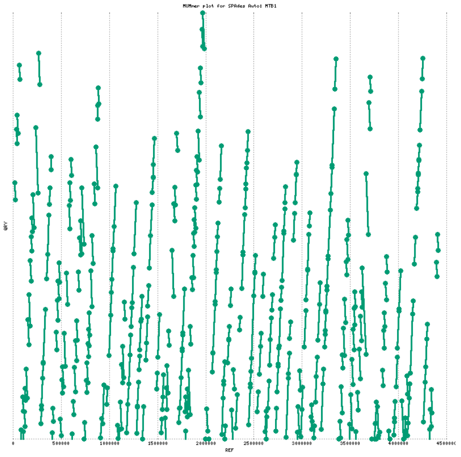
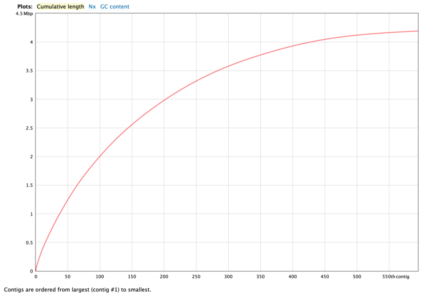
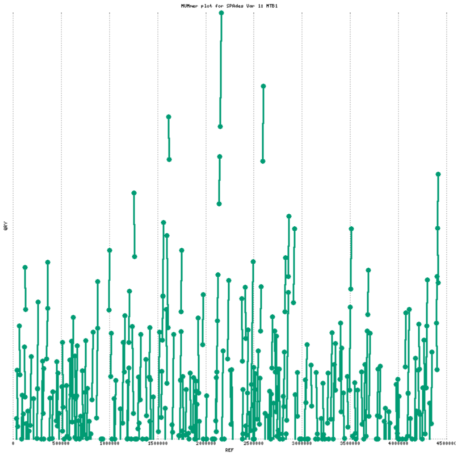
SPAdes

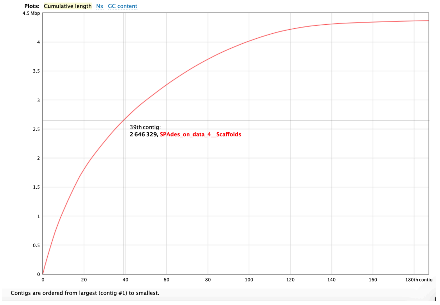
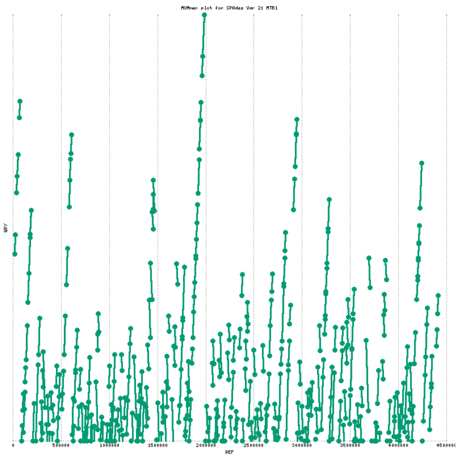
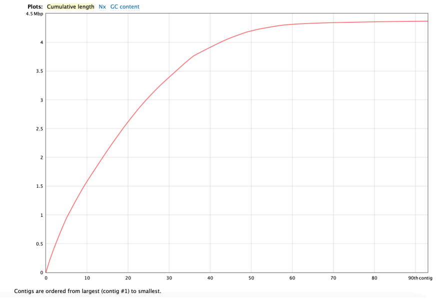
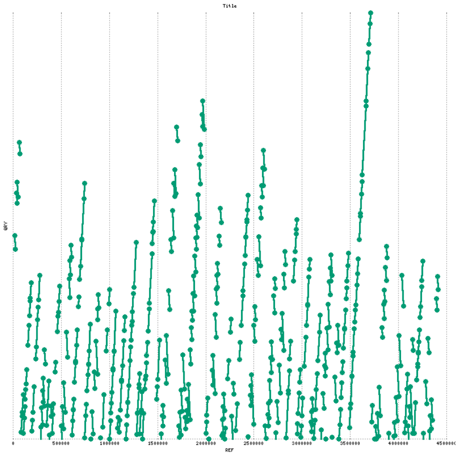
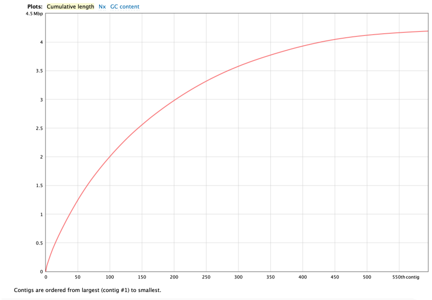
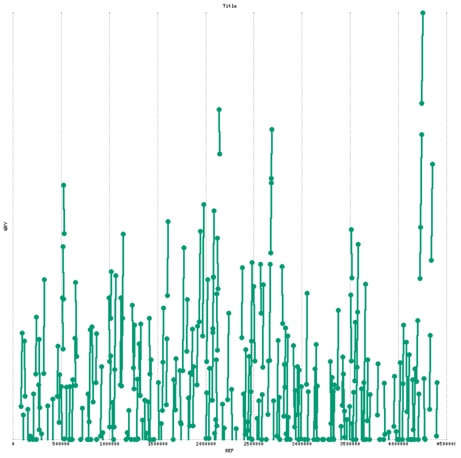
The results from FastQC analysis of the outputs obtained by SPAdes assembly are tabulated below.

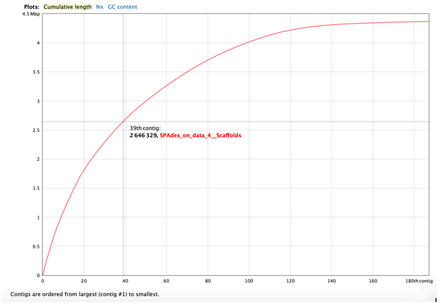
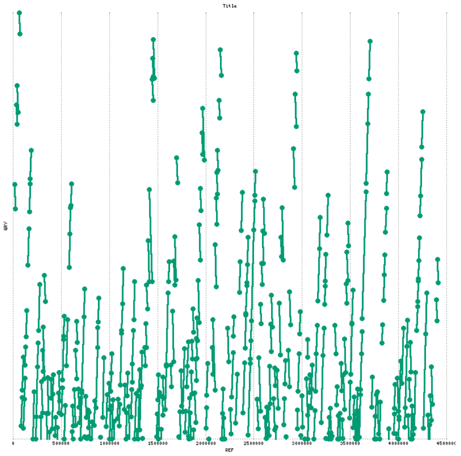
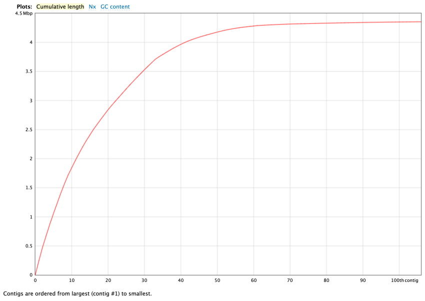
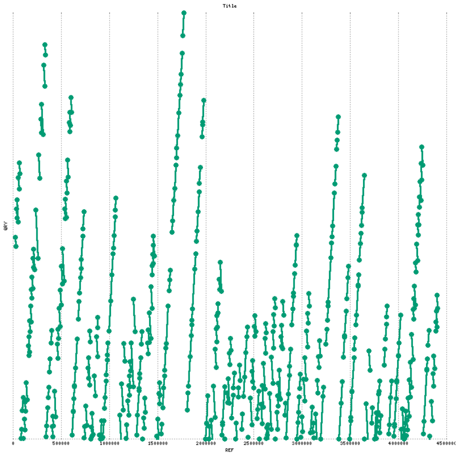
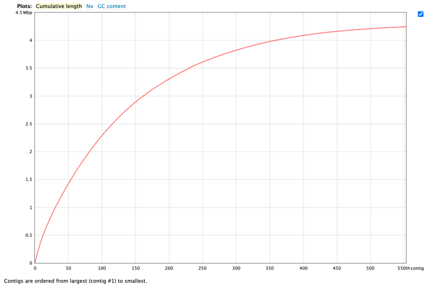
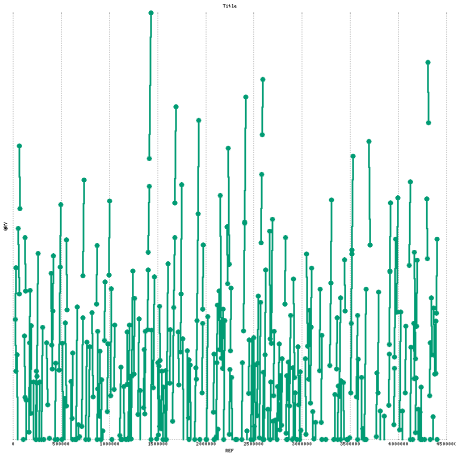
Dataset	Variation	N50	N90	L50	L90	Total length	Largest contig	Mismatches (#Ns)
mtb1	Auto	115582	33255	15	40	4375745	192810	1300
	1	13295	4051	98	316	4332456	134274	0
	2	44313	12148	30	101	4381149	153342	1100
mtb2	Auto	104032	35641	60	41	4366550	228144	800
	1	12153	3497	108	350	4191645	61712	100
	2	45140	14713	28	94	4368231	138477	800
mtb3	Auto	114916	32989	13	39	4354573	244693	1100

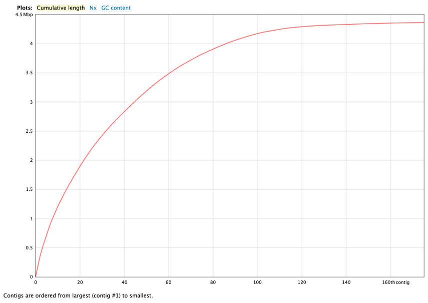
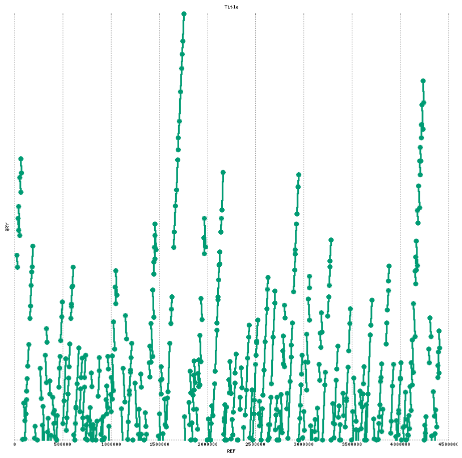
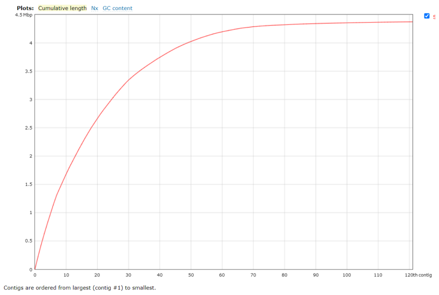
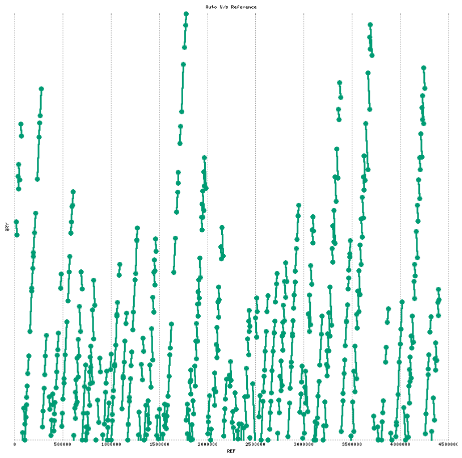
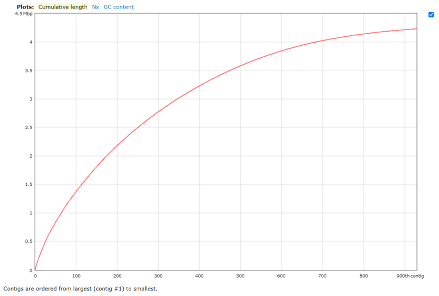
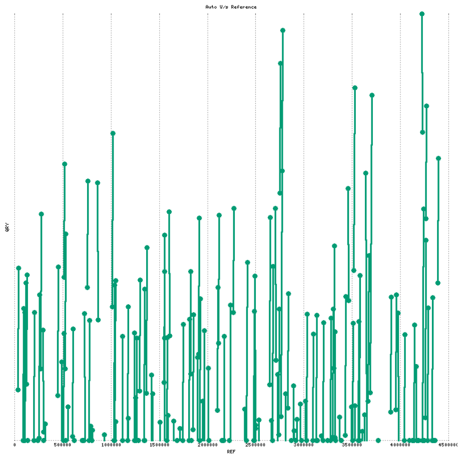
	1	15711	3702	89	299	4244619	49432	0
	2	55331	16400	25	82	4362940	184835	1100
mtb4	Auto	99848	25603	15	47	4372805	223796	900
	1	7044	2334	190	585	4230734	28969	0
	2	23809	6415	58	195	4386496	85590	700

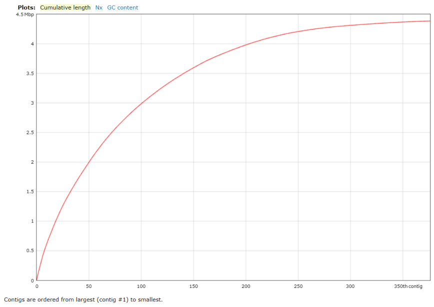
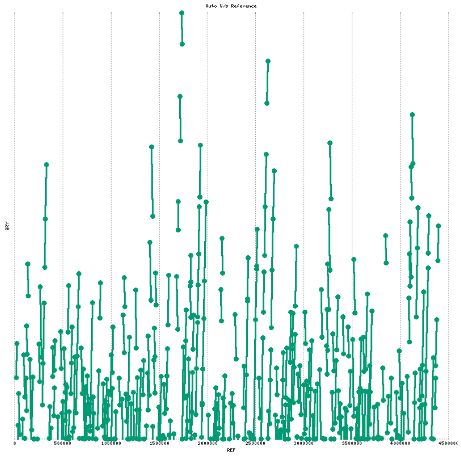
The following table summarises the cumulative length plots obtained from FastQC analysis and the MUMer plots. All MUMer plots are plotted against the reference sequence.

Dataset	Variation	Cumulative length plot	MUMmer plots
mtb1	Auto		
	1		

	2	 Plots: Cumulative length Nx GC content 4.5 Mbp 4 3.5 3 2.5 2 1.5 1 0.5 0 0 20 40 60 80 100 120 140 160 180th contig 39th contig: **2 646 329, SPAdes_on_data_4__Scaffolds** Contigs are ordered from largest (contig #1) to smallest.	
mtb2	Auto	 Plots: Cumulative length Nx GC content 4.5 Mbp 4 3.5 3 2.5 2 1.5 1 0.5 0 0 10 20 30 40 50 60 70 80 90th contig Contigs are ordered from largest (contig #1) to smallest.	
	1	 Plots: Cumulative length Nx GC content 4.5 Mbp 4 3.5 3 2.5 2 1.5 1 0.5 0 0 50 100 150 200 250 300 350 400 450 500 550th contig Contigs are ordered from largest (contig #1) to smallest.	

	2	 Contigs are ordered from largest (contig #1) to smallest.	
mtb3	Auto	 Contigs are ordered from largest (contig #1) to smallest.	
	1	 Contigs are ordered from largest (contig #1) to smallest.	

	2	 Contigs are ordered from largest (contig #1) to smallest.	
mtb4	Auto	 Contigs are ordered from largest (contig #1) to smallest.	
	1	 Contigs are ordered from largest (contig #1) to smallest.	

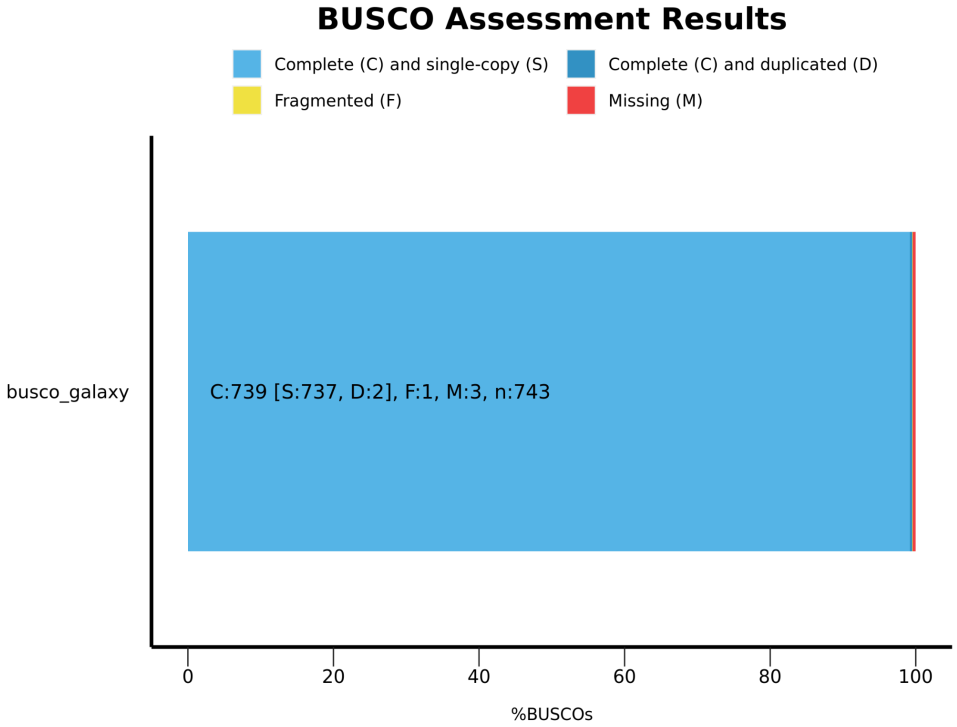
	2		

BUSCO

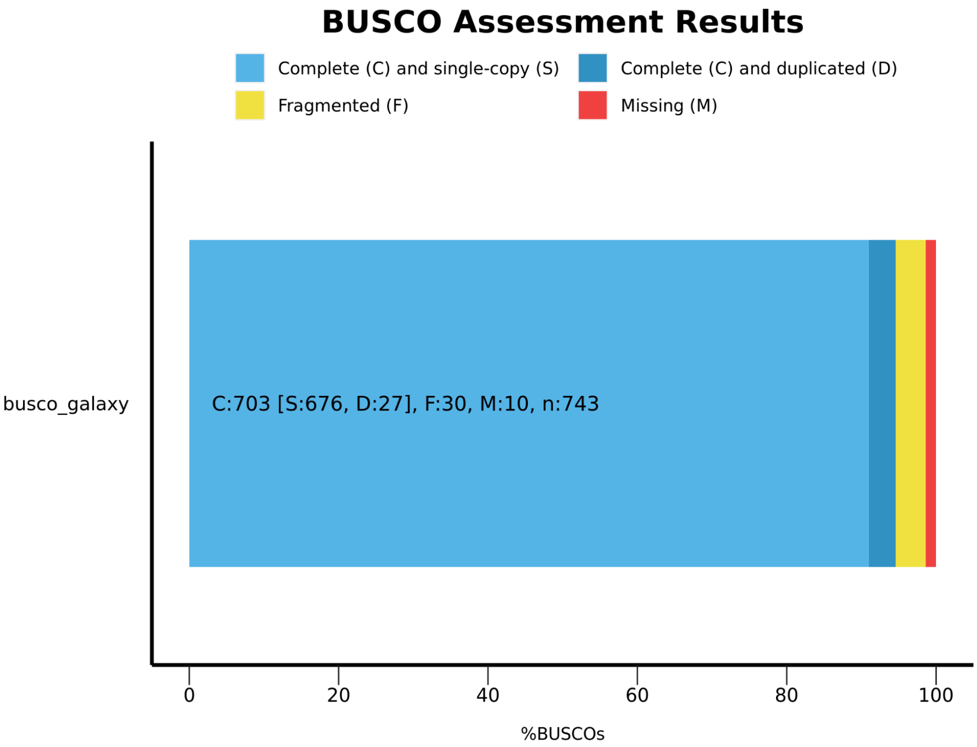
The following BUSCO plots have been obtained for the three variations of all four datasets.

1. Mtb1

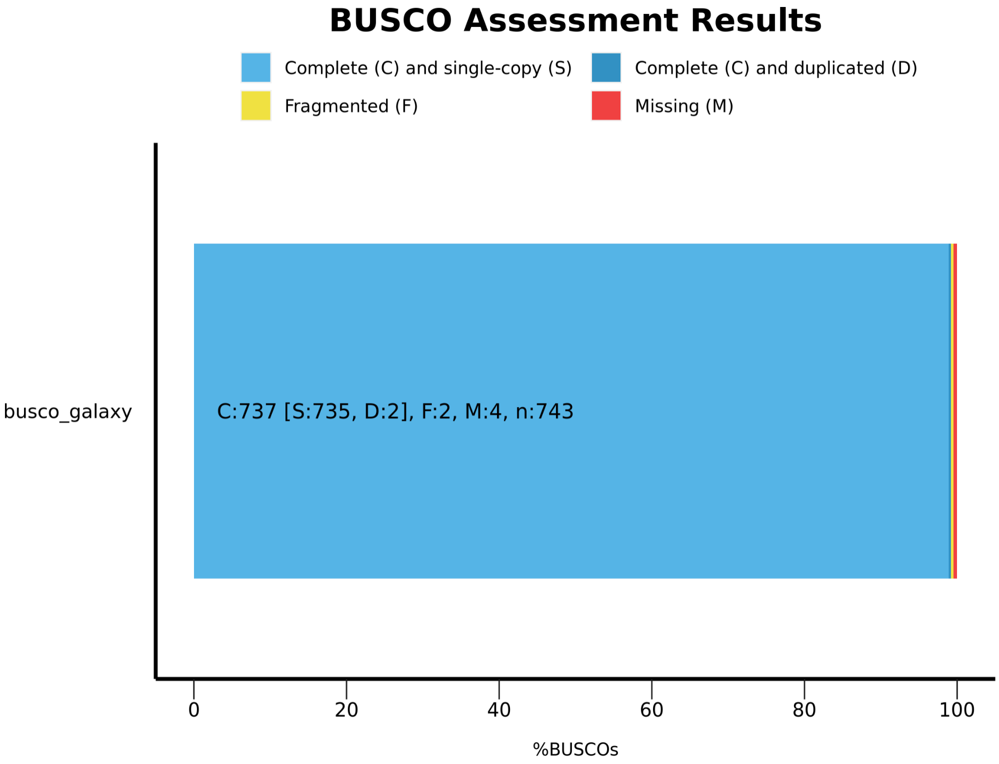
   Auto_SPAdes



   Variation 1

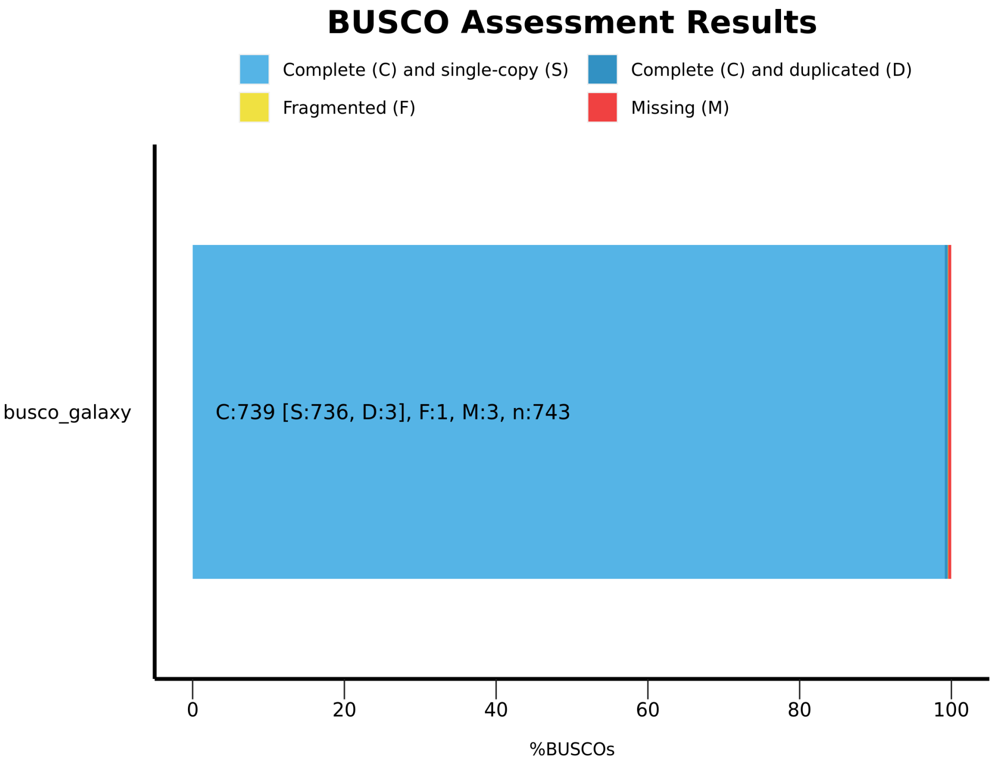


Variation 2

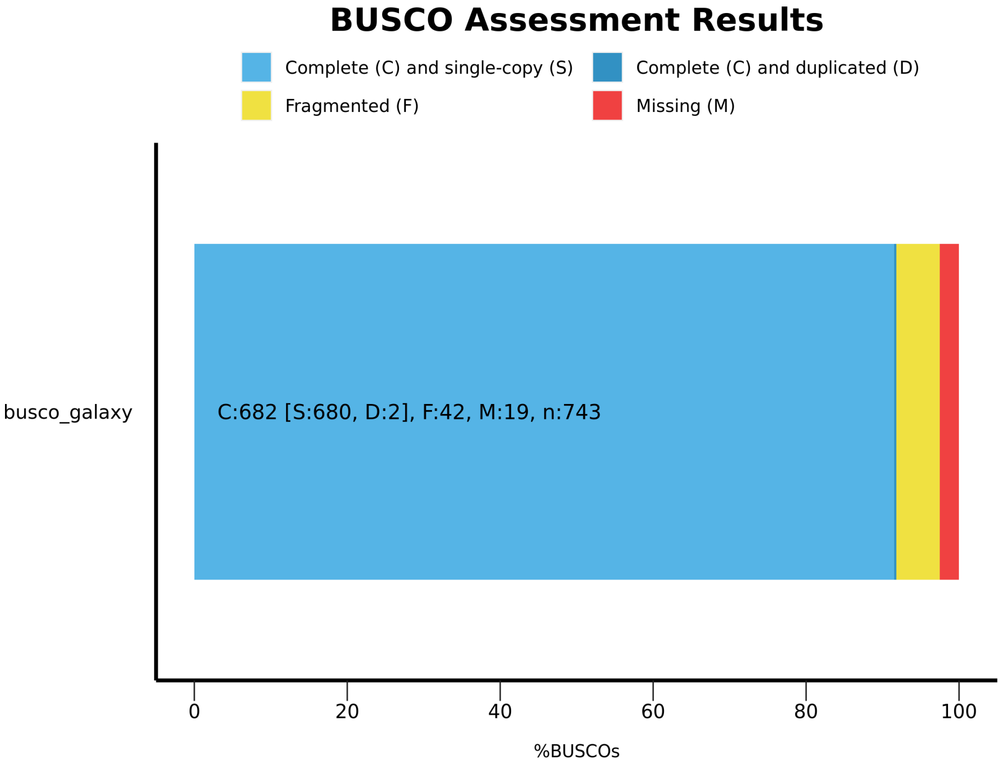


2. Mtb2

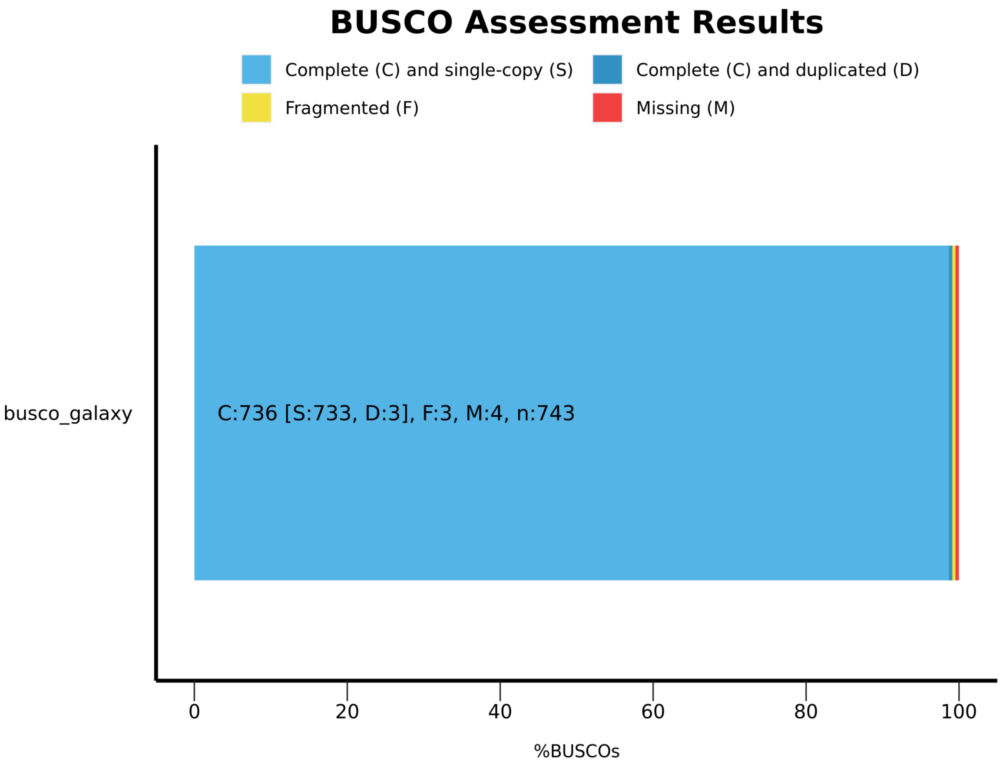
Auto_SPAdes



Variation 1

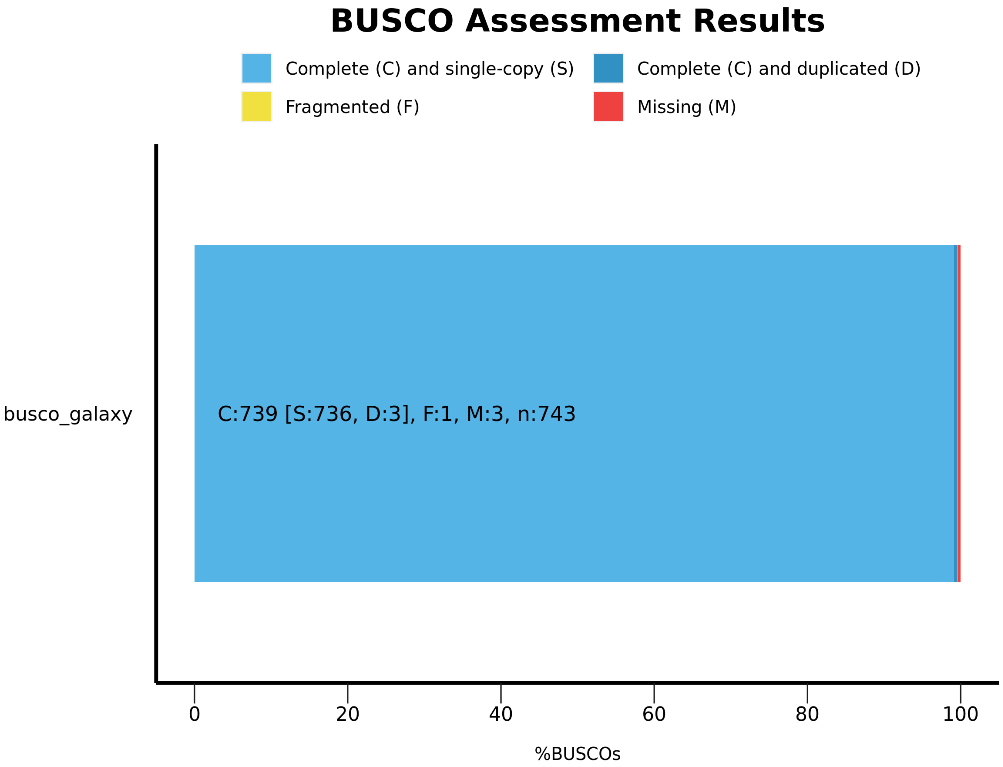


Variation 2

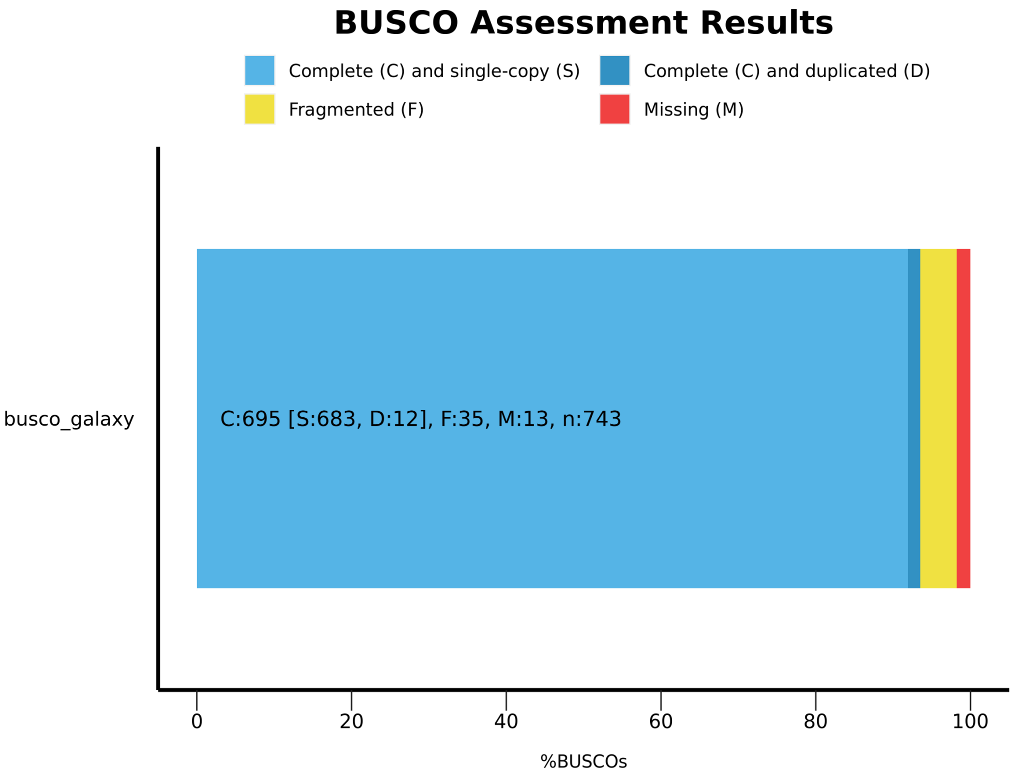


3. Mtb3

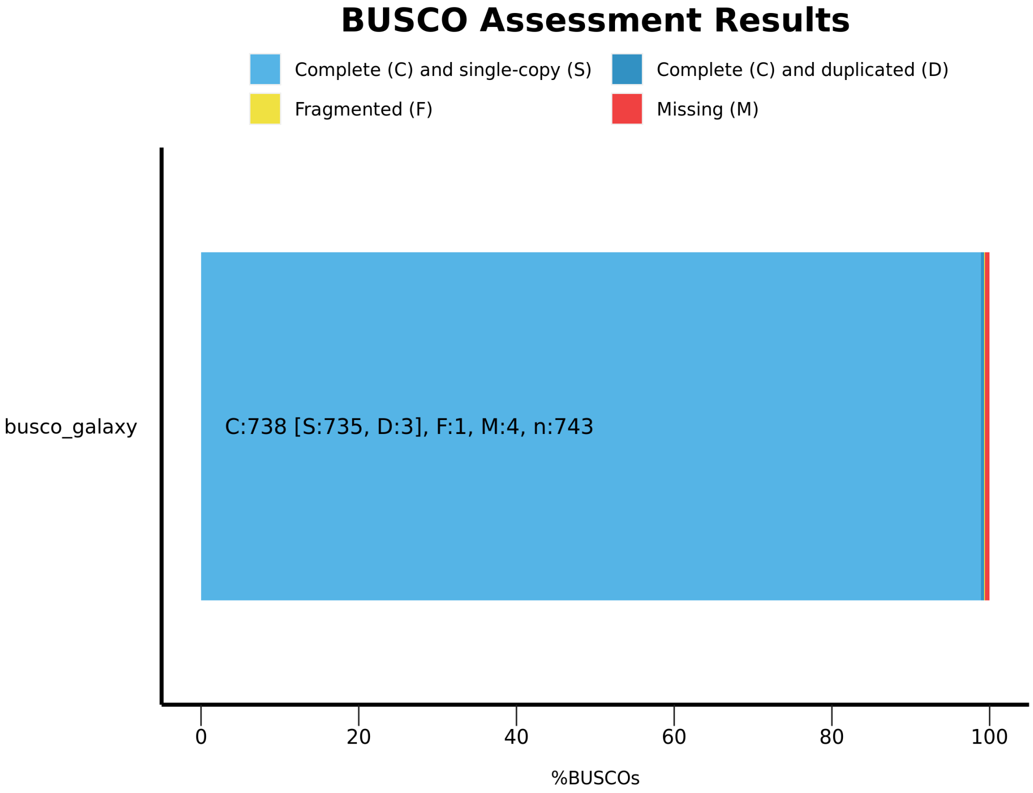
<u>Auto_SPAdes</u>



<u>Variation 1</u>

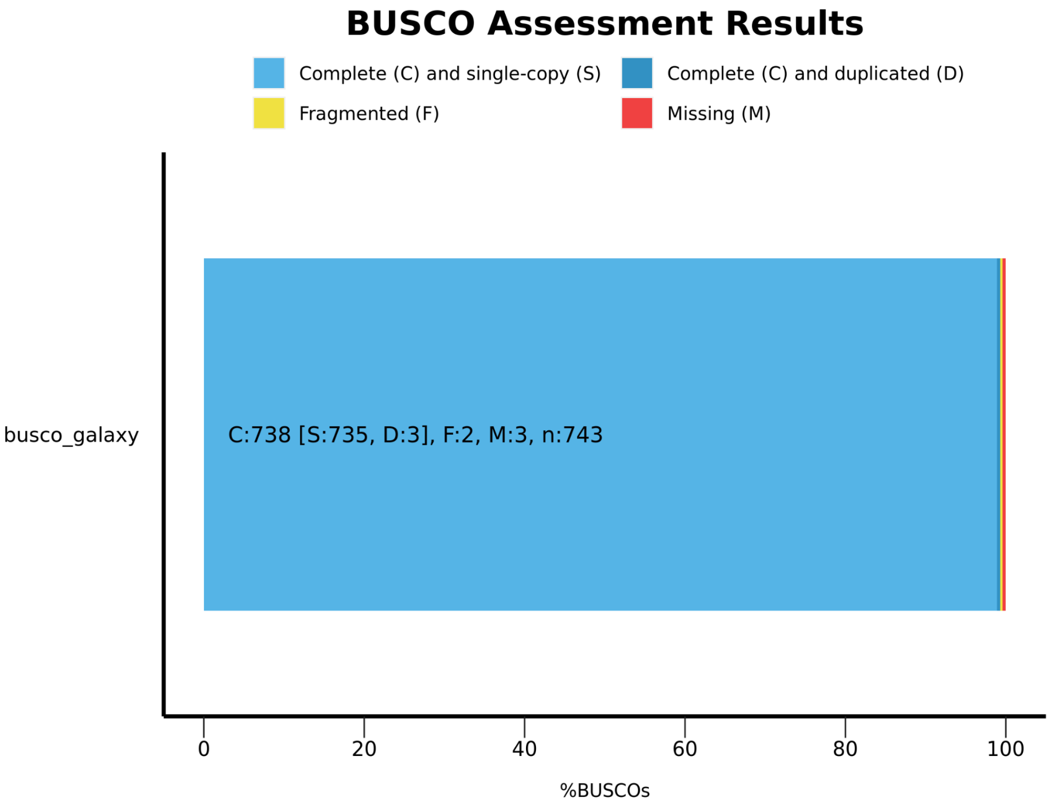


Variation 2

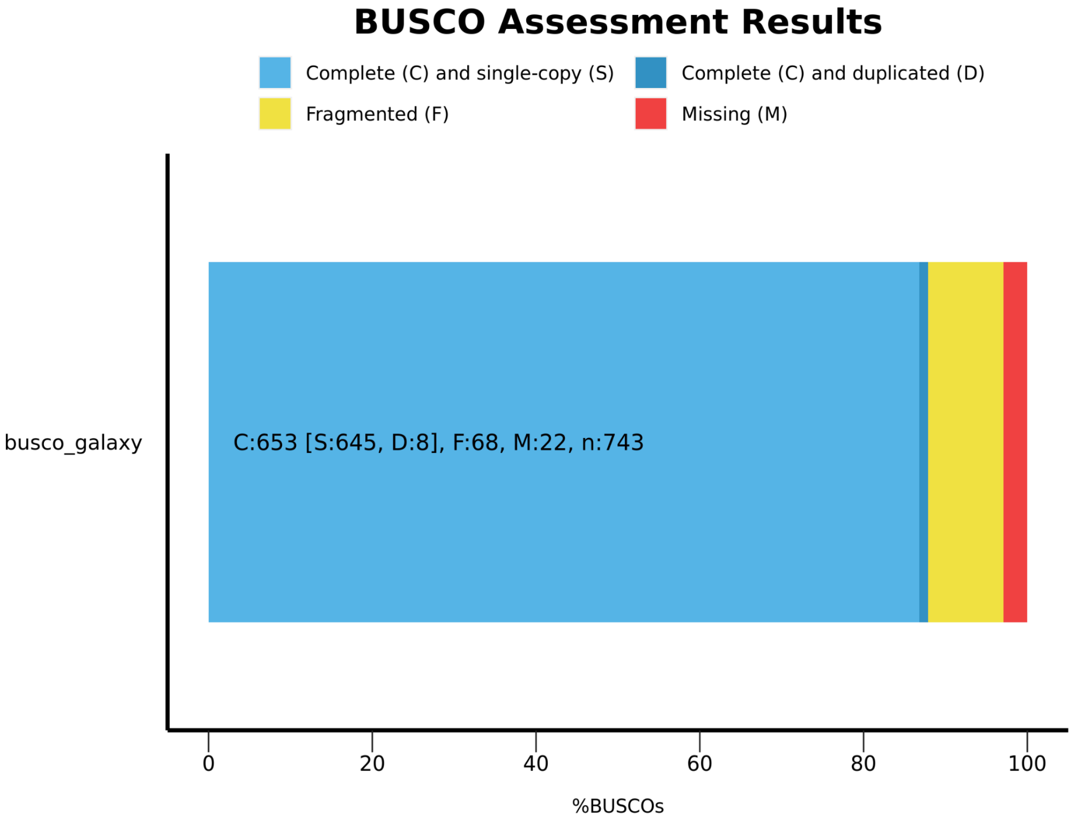


4. mtb4

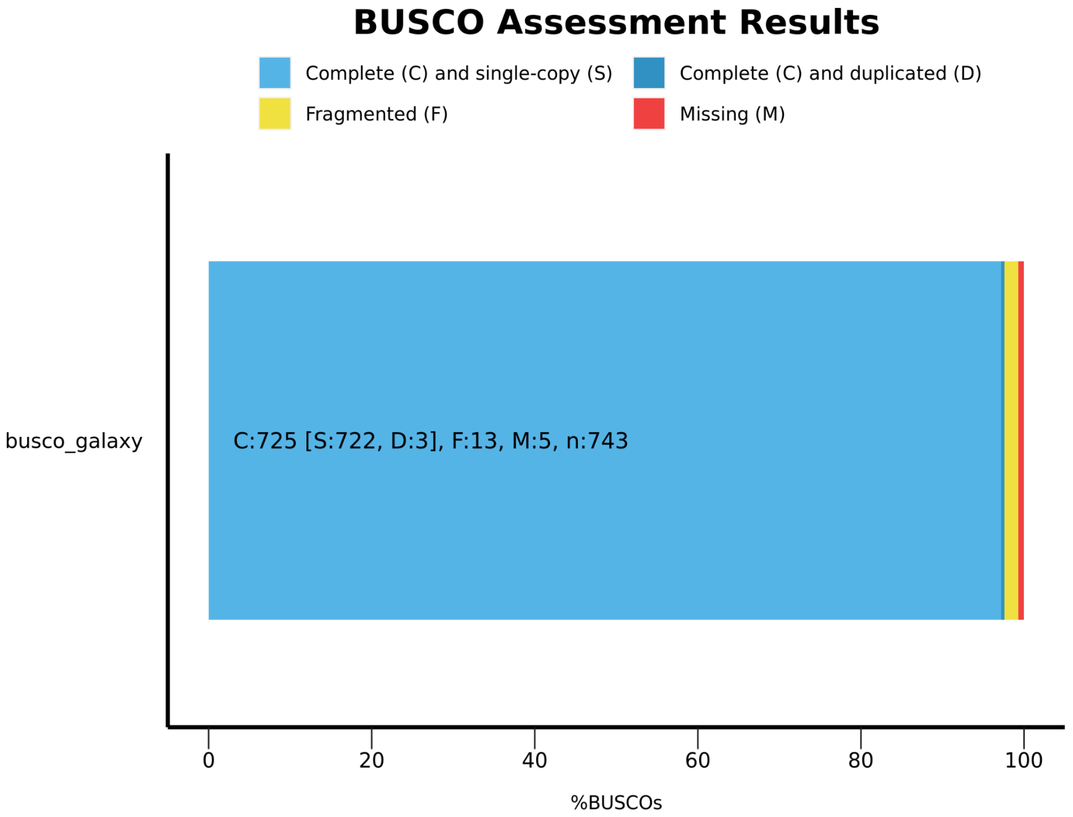
Auto_SPAdes


BUSCO Assessment Results
Complete (C) and single-copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)
busco_galaxy
C:738 [S:735, D:3], F:2, M:3, n:743
0
20
40
60
80
100
%BUSCOs

Variation 1


BUSCO Assessment Results
Complete (C) and single-copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)
busco_galaxy
C:653 [S:645, D:8], F:68, M:22, n:743
0
20
40
60
80
100
%BUSCOs

Variation 2



VarScan

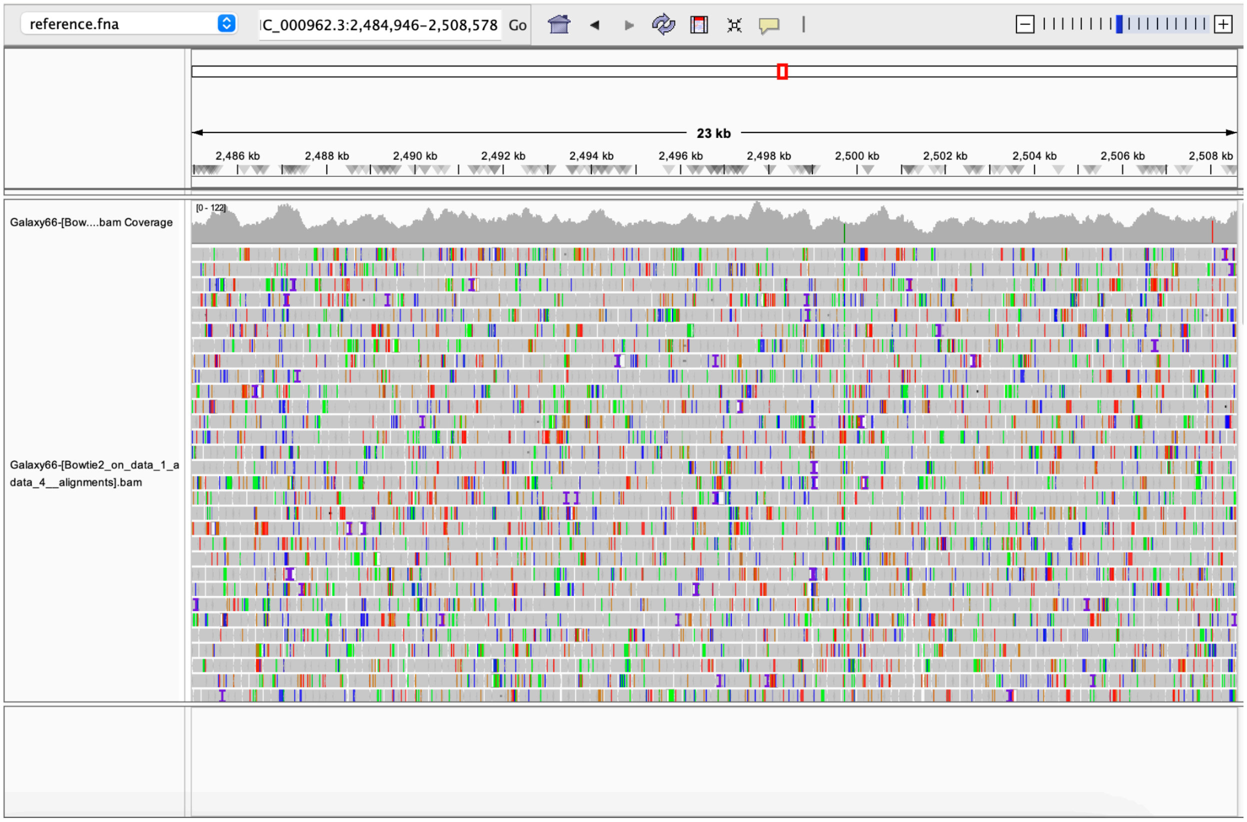
All the values obtained below are from the tool Varscan mpileup. The following parameters were set (while keeping others as default). The minimum coverage was set to be >20 and the allele frequency was set to be >0.8

1. **mtb1**
   Number of SNP's = 971
   Number of Indels = 65
2. **mtb2**
   Number of SNP's = 969
   Number of Indels = 68
3. **mtb3**
   Number of SNP's = 872
   Number of Indels = 57
4. **mtb4**
   Number of SNP's = 886
   Number of Indels = 64

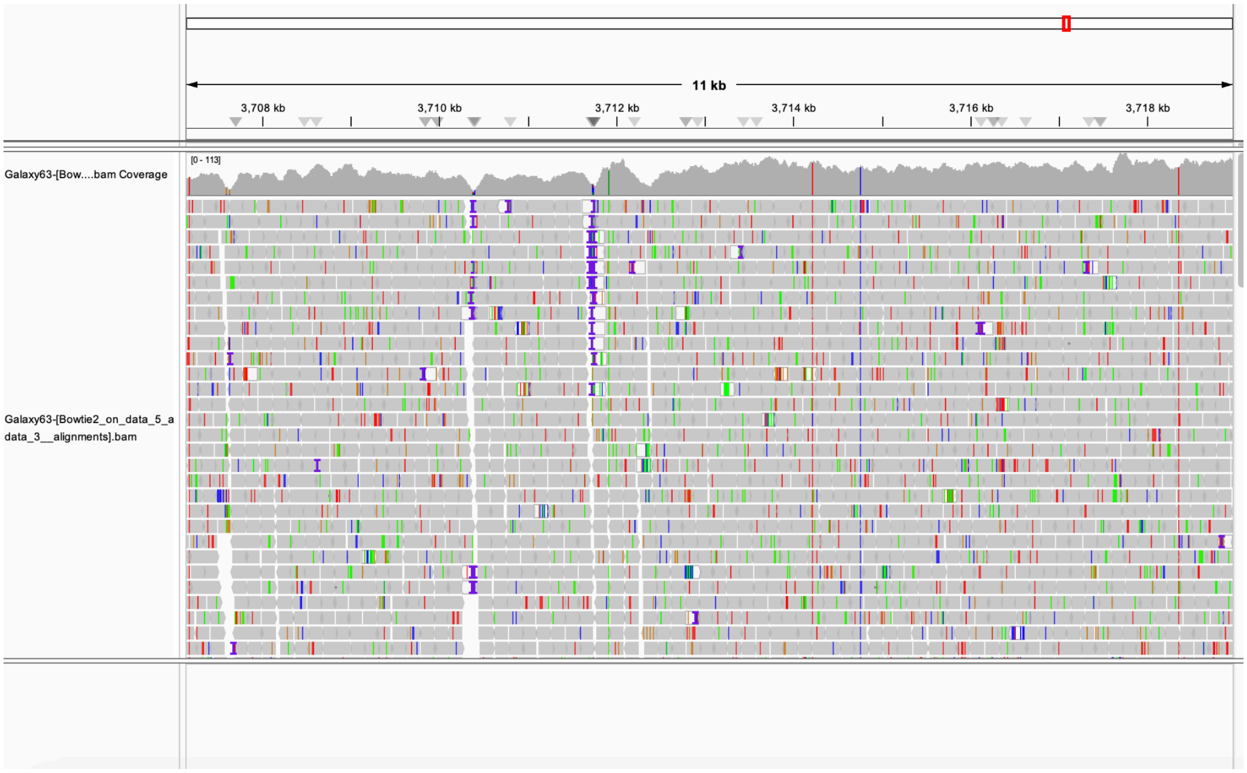
IGV

The results of bowtie2 sequencing were visualised using Integrative Genomics Viewer. The obtained results for mtb2 and mtb3 are shown below

mtb2



mtb3


11 kb
3,708 kb
3,710 kb
3,712 kb
3,714 kb
3,716 kb
3,718 kb
[0 - 113]
Galaxy63-[Bow....bam Coverage
Galaxy63-[Bowtie2_on_data_5_a data_3__alignments].bam

Discussions

We observe that all the 4 cases have a constant phred score of 30. This is hypothesised to be the property of the instruments used.

In SPAdes, we observe that Auto setting give the best results in terms of N50 and L50. But, k=15 has reduced number of mismatches. Thus, we can conclude that a lower k tries to fit the reads more stringently whereas a larger k gets a better fit at the cost of some mismatches. Note that beyond a limit of 77, the further increase in k value detoriated the N50 values.

The mummer plots show that there are lot of gaps between the reads and the reference but if we compare a low k value with a high k value, we will see that the sequence is continuosus. Moreover, we can also see that there are a huge number of INDELs as compared to the reference seen in the Bowtie2 alignment, confirming our hypothesis.

BUSCO plots are another way to analyse the goodness of the alignment. BUSCO plots replicate the results shown by the analysis of SPAdes using the QUAST tool.

Conclusions

Four Tuberculosis samples were randomly chosen and the reads were aligned using a de novo assembler (SPAdes) and a mapping based assembler(bowtie2). There were various variations chosen for the SPAdes assembler k-value and the effects were analysed. It was consluded that there is a specific range in which k-mer size is the best. Moreover, the given reads miss a lot of parts of the reference genome as shown by the mummer plots and the indels in bowtie2.

References

1. Verza, Mirela, et al. "Genomic epidemiology of Mycobacterium tuberculosis in Santa Catarina, southern Brazil." *Scientific reports* 10.1 (2020): 12891.https://doi.org/10.1038/s41598-020-69755-9
2.