PROGRESS IN ENGINEERING APPLICATION AND TECHNOLOGY
e-ISSN: 2773-5303

PEAT

Vol. 6 No. 1 (2025) p. 574-588
https://publisher.uthm.edu.my/periodicals/index.php/peat

Food-Borne Pathogen Detection from Direct Nanopore Sequencing Data Using Galaxy Platform

Loh Will Han[1], Rama Yusvana[1*]

[1] *Department Of Chemical Engineering Technology, Faculty of Engineering Technology, Universiti Tun Hussein Onn Malaysia, 84600, Pagoh, Johor, MALAYSIA*

*Corresponding Author: rama@uthm.edu.my
DOI: https://doi.org/10.30880/peat.2025.06.01.061

Article Info

Received: 18 January 2025
Accepted: 04 February 2025
Available online: 30 April 2025

Keywords

Galaxy Platform, Food-Borne Pathogen Detection, Nanopore sequencing, Simulation

Abstract

Food safety is a critical global concern, with food-borne pathogens posing significant threats to public health and economic stability. Rapid and accurate detection methods are essential to mitigate these risks. This study explored the use of direct nanopore sequencing data combined with the Galaxy platform for the identification of food-borne pathogens. The National Center for Biotechnology Information (NCBI) database was used to find samples used for the data analysis in fulfilment of objectives 2 and 3 as well as reference genome. The workflow integrated real-time sequencing capabilities and bioinformatics tools available in Galaxy to analyze raw sequencing data for pathogen identification. The tools used included but not limited to FastQC, Nanoplot, Fastp, Porechop, MultiQC, Kraken2, Krona Pie Chart, Pavian Visualisation, ABRicate, Flye, etc. Results demonstrated the method's effectiveness in detecting and classifying various food-borne pathogens such as *Staphylococcus aureus* based on numerous methods that rely on the metagenome's taxonomy profile, pathogenetic gene composition and allele variation with high sensitivity and specificity. In conclusion, this approach was able to accurately determine the food-borne pathogens existing within the samples using the tools in the Galaxy platform.

1. Introduction

Food-borne pathogens, including bacteria such as *Salmonella sp.* and *Listeria monocytogenes*, present serious threats to public health, causing illnesses that range from mild symptoms to severe, life-threatening conditions. The World Health Organization [1] reports 600 million cases of food-borne diseases annually, leading to 420,000 deaths. Despite advancements in food processing designed to eliminate pathogens, resilient microorganisms like *L. monocytogenes* can survive harsh processing environments, necessitating effective detection methods to ensure food safety.

Conventional pathogen detection techniques, including immunology-based methods, culture and colony counting, and polymerase chain reaction (PCR), are reliable but have significant drawbacks [2]. These methods are often expensive, time-intensive, and unsuitable for real-time detection, limiting their effectiveness in fast-paced food safety applications. The evolution of sequencing technologies has introduced alternative approaches for pathogen detection. First-generation sequencing (Sanger sequencing) is highly accurate, with an error rate of less than 0.001% [3], but suffers from low throughput, exorbitant costs (~$10 million per gigabase), and lengthy processing times of 4–6 weeks. Second-generation sequencing (Next-Generation Sequencing, NGS) improved throughput [4] and reduced costs (around $5000 per genome) but introduced trade-offs, including shorter read lengths (50–500 bp), reduced accuracy, and continued reliance on PCR, which delays real-time detection. Third-generation sequencing (TGS) which nanopore sequencing is a part of addresses these limitations by enabling

Progress in Engineering Application and Technology Vol. 6 No. 1 (2025) p. 574-588

single-molecule sequencing without PCR amplification [5]. This approach supports real-time, on-site experiments without the need for extensive sample preparation, eliminating PCR-related biases [6]. Additionally, TGS can process long DNA reads (up to 30 kb), significantly surpassing the read lengths of earlier technologies. These features make TGS a transformative tool for rapid and efficient detection of food-borne pathogens, offering a robust solution for food safety monitoring.

Nanopore sequencing offers a promising solution with its capability for rapid, sensitive, and real-time pathogen detection. However, analysing direct nanopore sequencing data presents challenges, including high error rates, long read lengths, and hence the need for specialized bioinformatics workflows. Aside from that, there has been minimal emphasis of the necessity of workflow in food-borne pathogen detection, often lacking in detailed insights and absence of standardisation. With the Galaxy platform with its user-friendly interface and advanced bioinformatics capabilities, it offers potential for streamlining the analysis of nanopore sequencing data. This report explores the feasibility and effectiveness of utilizing Galaxy for foodborne pathogen detection to improve food safety and public health outcomes. Using the multitude of the tools the platform provides, it is possible to develop a robust bioinformatics pipelines optimised for the unique features of the nanopore sequencing data, thus enabling a comprehensive detection and characterisation of food-borne pathogens directly from complex food metagenome.

2. Methodology

This topic reviewed the details of the approach to developing a bioinformatics pipeline for pathogen detection consisting of several key steps including the data acquisition, preprocessing, taxonomy profiling, gene-based, allele-based pathogenic identification and pathogen tracking as well as visualization. By following a systematic methodology, it can be ensured that the pipelines are precise, efficient and replicable in identifying pathogens from nanopore sequencing data through k-mer matching, identification of virulence factor and antimicrobial resistance genes and variant calling.

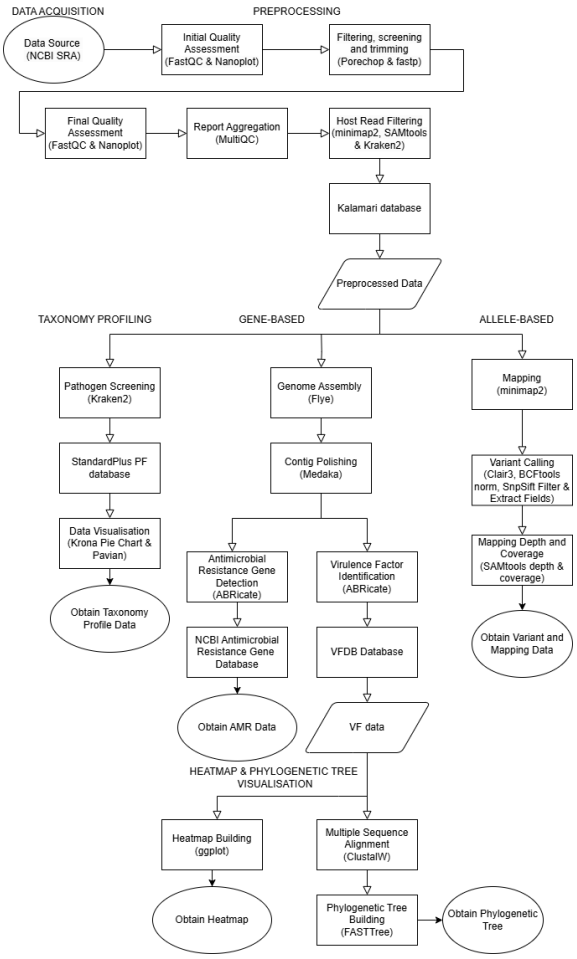


***Fig. 1*:** Foodborne Pathogen Detection Pipeline in Nanopore Sequencing Data

2.1 Data Acquisition

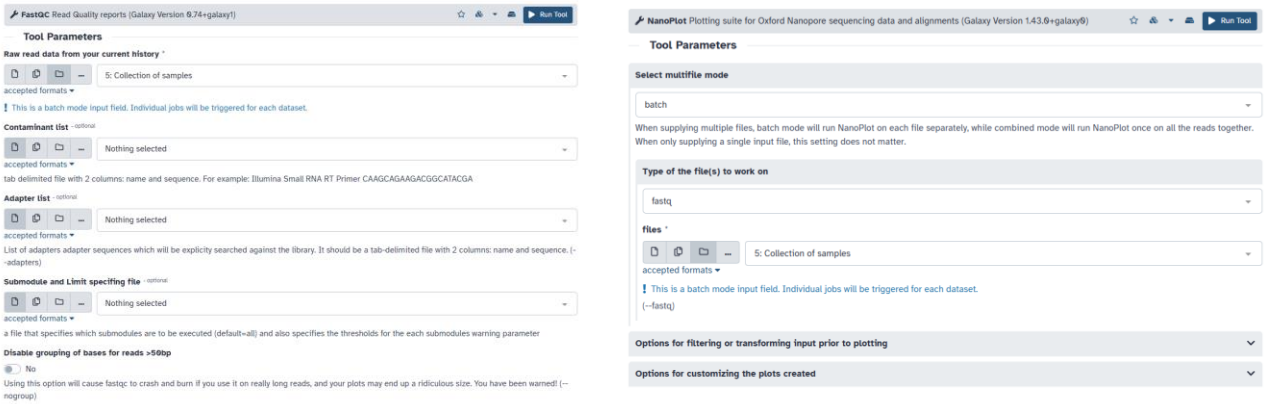
Data sources included simulated datasets from the Galaxy tutorial and the NCBI database. Samples represented environments where foodborne contamination was possible, such as farmlands, swamps, and food processing industries. Specific datasets used included simulated data [7], potato samples spiked with *S. aureus* [8], and random cow milk samples [9]. All data were uploaded and processed on the public server of Galaxy bioinformatics platform at *usegalaxy.eu* as per the study's scope.

2.2 Preprocessing Steps

Preprocessing ensured data quality through steps like quality assessment (FastQC, NanoPlot), adapter trimming (Porechop), and quality control (Fastp) [10]. These tools assessed and enhanced read quality while MultiQC aggregated the results into an interactive HTML report. Host reads were filtered out using Minimap2 and SAMtools to remove DNA from the host organism, ensuring that downstream analyses focus solely on microbial data. Kraken2 with the Kalamari database detected the remaining contamination, removed them and verified that host reads have been removed.

2.2.1 Quality Control, Filtering, and Assessment

Initial quality assessment was performed using FastQC and NanoPlot, which provided summary graphs and statistics. Subsequent steps included trimming and filtering reads using Porechop, and Fastp. Final quality assessments were performed with FastQC and NanoPlot, and then aggregated into a comprehensive report with MultiQC. The figures below showcased the interface of FastQC (Fig. 2A) and Nanoplot (Fig. 2B) used on *usegalaxy.eu.*



(a) **(b)**

***Fig. 2*:** Interface of Preprocessing Tools (a) FastQC; (b) Nanoplot

2.2.2 Host Read Filtering

Host DNA was removed by mapping reads to a reference genome using Minimap2, separating mapped (host) and unmapped (microbial) reads with SAMtools. Unmapped reads were processed further with Kraken2 to further remove any leftover host gene with Kalamari database.

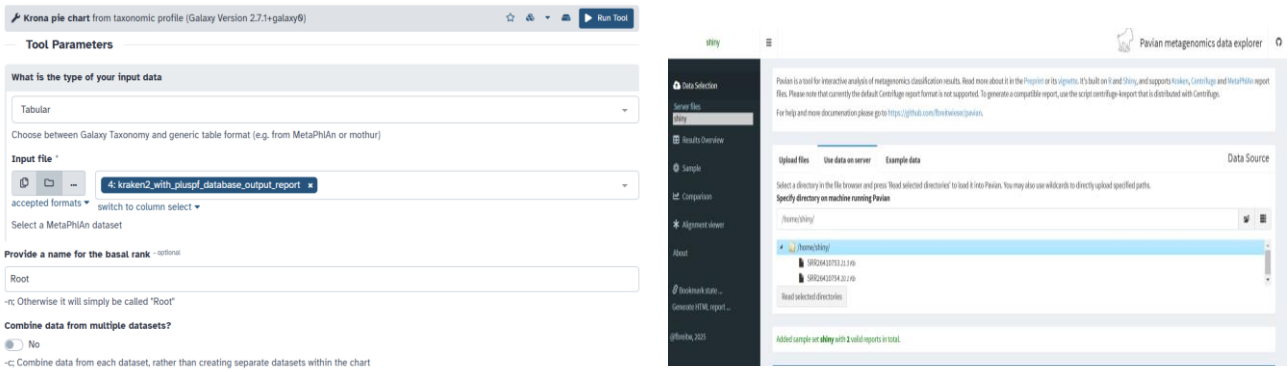
2.3 Taxonomy Profiling

Taxonomy profiling identified and classified microbial species in the sample. Kraken2 assigned taxonomic labels using the PlusPF database. The results were visualized with Krona Pie Charts and Pavian, providing interactive insights into microbial composition, including percentages and classifications down to the species level.

Progress in Engineering Application and Technology Vol. 6 No. 1 (2025) p. 574-588

2.3.1 Pathogen Screening and Visualization

Kraken2 processed the preprocessed data by assigning labels to the taxa based on exact alignments of k-mers, using the PlusPF database. The results were then visualized with Krona Pie Charts for hierarchical taxonomic views and Pavian for Sankey flow diagrams, offering detailed insights into microbial diversity. The figures below showcased the visualization tools interface for Krona Pie Chart and Pavian.



(a) **(b)**

Fig. 3: Interface of Taxonomic Visualisation Tools (a) Krona Pie Chart; (b) Pavian

2.4 Gene-Based Identification

Pathogens were identified by analyzing genes associated with pathogenicity, such as antimicrobial resistance genes and virulence factors. De novo genome assembly was conducted using Flye, which assembled the reads into contigs for mass screening and identification of pathogenic gene with ABRicate [11]. Assembly graphs were visualized with Bandage, while Medaka polished contigs for accuracy. ABRicate was then used to screen the polished contigs for identification of pathogenic genes by using relevant DNA sequence databases.

2.4.1 Genome Assembly and Visualization

Flye assembled the sequence reads into contigs for analysis. The assembly graph was visualized using Bandage, allowing the observation of the formation of contigs before polishing. Medaka further refined the contigs through consensus sequence generation and variant calls, allowing a more refined and less error-prone contigs to be used in ABRicate later.

2.4.2 Antimicrobial Resistance Genes & Virulence Factor Detection

ABRicate detected both antimicrobial resistance (AMR) genes and virulence factors (VF) in the polished assembled contigs using the NCBI Bacterial Antimicrobial Resistance Reference Gene and VFDB databases respectively. Antimicrobial resistance gene (AMR) indicates that the resistance of antimicrobial action of the species in metagenome whereas virulence factor allows the microorganism to cause disease, hence it is the main contributing factor for pathogenicity. ABRicate was used to detect the antimicrobial resistance gene by screening the assembled datasets from Flye by pairing it up with NCBI Bacterial Antimicrobial Resistance Reference Gene Database and input of collection output from medaka consensus pipeline. Similarly, ABRicate is paired up with VFDB database with the same collection output to detect the presence of virulence factors. The data were very insightful in detecting specific pathogenic traits without relying on the taxonomy profiling of the microbiome, thus can be used as a secondary measure to detect the presence of pathogens in samples. Figure below showcased the interface of ABRicate on the public server of Galaxy platform on *usegalaxy.eu.*

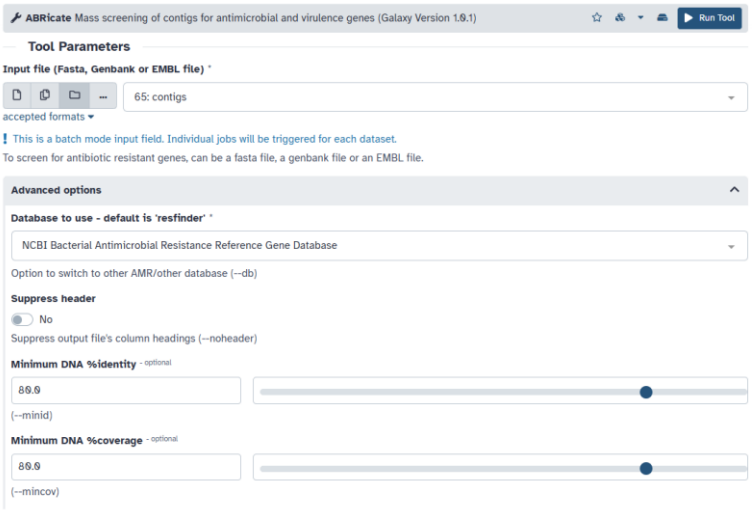


***Fig. 4*:** Interface of ABRicate

2.5 Allele-Based Identification

Allele or single-nucleotide polymorphism-based pathogen identification compares the reads with a specific reference genome, typically one that is commonly seen contaminating the sample and identifying the differences between sample reads and reference genome [12]. The differences or variants in specific positions on the genome would be used to determine whether the variations would indicate pathogenicity or not, thus identifying foodborne pathogens in the sample. The coverage and depth of the reference genome to the sample reads can also be found out by how distinct it is.

2.5.1 Variant Calling and SNP Analysis

Minimap2 mapped reads to reference genomes, including *Salmonella* [7], *Paenibacillus sp* [13]., and *Acinetobacter sp* [14], for tutorial datasets, potato samples spiked with *S.aureus* and whole milk samples respectively. Clair3 identified genetic variants and indels, which were normalized and processed with BCFtools. SnpSift filtered and tabulated the results for clarity. The results showed the number of allele variants found when comparing both the sample reads and reference genome data, indicating the potential case of mutation if there happened to be a mutation. This step was to ensure the pathogenicity of the commonly mutating strains could be detected and identified.

2.6 Pathogen Tracking and Visualisation

Pathogen tracking and visualisation refers to the aggregation of the results from ABRicate to further aid in the tracking of potential pathogens in the sample. The workflow continues from gene-based pathogen detection which creates heatmaps and phylogenetic trees based on the tabular results of VF genes detected by ABRicate. Due to server errors of the Galaxy platform, some of the tools were not able to be properly used hence it was necessary to follow the predetermined workflow of the tutorial for Pathogen Tracking and Visualisation by inputting with final output of the previous steps.

2.7 Validation of Pipelines

The pipelines' effectiveness was verified through using different but similarly sourced samples. For example, the spiked pathogens in potato samples, *Staphylococcus aureus* should be found and verified in both samples, thus proving that the identification of said strain using the pipeline was consistent and replicable across similar samples. To validate the effectiveness in detection of different pathogens, different samples from different sources were used, including the simulated datasets spiked with *Salmonella sp* and whole milk samples without any spiked microbes.

Progress in Engineering Application and Technology Vol. 6 No. 1 (2025) p. 574-588

3. Results and Discussion

This topic covers the analysis and results obtained from the previous section and only the potato samples data would be shown out of the three sample collections due to the restriction of space. The analysis emphasized the principles behind the results and what the findings could bring.

3.1 Preprocessing Results

3.1.1 Before Preprocessing

The weighted histograms of read lengths after log transformations for both potato samples, SRR26410753 and SRR26410754 were obtained from Nanoplot. The distributions of the reads for SRR26410753 and SRR26410754 were both right-skewed, reflecting a similar trend observed in other studies [15]. For SRR26410753, the mode approached 210 kilobases at approximately 400 base pairs read length, while in SRR26410754, the mode was slightly lower, at 193 kilobases, occurring around 390 base pairs read length. In both cases, the number of bases decreased as the read length increased beyond these points. This phenomenon was attributed to the lower signal-to-noise ratio introduced during long-read sequencing, which affects the ability to read longer sequences accurately [16].

(a) **(b)**

***Fig. 5*:** Weighted Histogram of Read Lengths after Log Transformation Before Preprocessing

(a) SRR26410753; (b) SRR26410754

The mean quality scores of both potato samples were obtained from the MultiQC tool. The graph showed that the average Phred quality score was around 22 with the lowest being 4.66 and the highest being 40.90. Quality of the data started extremely low then stabilised at a Phred score above 22, which was in the acceptable range for reliable base calling at the middle region of 500 to 45k bp. After that point, the quality score of SRR26410753 sharply plummeted into the red zone (Phred score <20) then rocketed back up into green zone (Phred score >30) whereas the quality score of SRR26410754 declined slowly into red zone before increasing back to the yellow zone. This degradation was typical of long-read sequencing, where error rates increased as reads extended further [17]. Both of the reads experienced sharp decline before recovery towards the end of the sequencing which signified poor signal accuracy as the sequencing chemistry struggled with long-read fidelity.

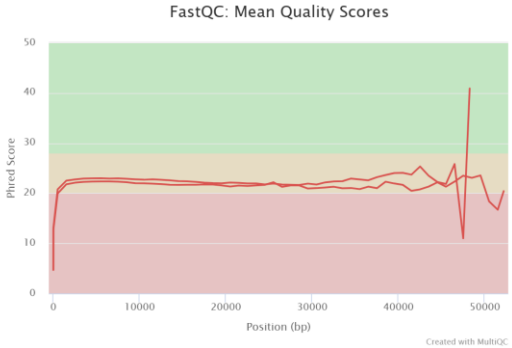


***Fig. 6*:** Mean Quality Scores of the Potato Samples spiked with S. aureus Before Preprocessing

3.1.2 After Preprocessing

When comparing the weighted histograms of read lengths after log transformation before and after preprocessing, the curve distributions for both figures remained relatively unchanged, although the number of bases decreased in both cases. In SRR26410753 sample, the mode shifted from above 200k bases to around 150k bases, while in SRR26410754 sample, the mode shifted from approaching 200k bases to around 130k bases. These shifts indicated that a substantial number of low-quality bases had been trimmed and filtered during preprocessing. The overall distribution shapes remained consistent because preprocessing primarily filtered out lower-quality data rather than altering the inherent read-length trends [18]. As evidence of the effectiveness of preprocessing, the overall Phred scores improved. In SRR26410753, at least 54% of the reads had scores higher than Q15, while in SRR26410754, at least 44% of the reads exceeded Q15. This demonstrated the enhancement in data quality following preprocessing while preserving the original distribution characteristics.

(a) **(b)**

***Fig.7*:** Weighted Histogram of Read Lengths after Log Transformation After Preprocessing

(a) SRR26410753; (b) SRR26410754

The graph showed that the average Phred quality score was around 22 with the lowest being 11.57 and the highest being 40.92. Quality of the data started high, stabilizing at a Phred score above 22, which was in the acceptable range for reliable base calling at the middle region of 500 to 45k bp. Similar trend that appeared before preprocessing could be observed again but with the average Phred score being higher than before, signifying an improvement of read quality. The results of the preprocessing were deemed sufficient to proceed towards the next step of the processing.

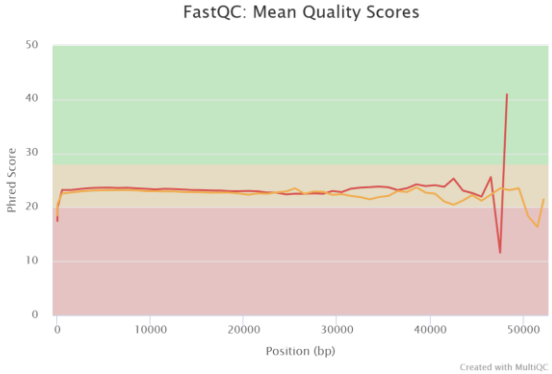


***Fig. 8*:** Mean Quality Scores of the Potato Samples spiked with S. aureus After Preprocessing

Progress in Engineering Application and Technology Vol. 6 No. 1 (2025) p. 574-588

3.2 Taxonomy Profiling Results

3.2.1 Krona Pie Charts

The Krona Pie Charts illustrated the taxonomic compositions of the microbial communities in SRR26410753 and SRR26410754. Both datasets contained a significant proportion of unclassified reads, with 38% and 39% unclassified, respectively, indicating a notable portion of the taxonomy remained unlabeled and potentially leading to inconsistencies.

The taxonomy for both datasets was predominantly from the Bacteria kingdom. In SRR26410753, the largest segment (60%) belonged to the Bacilli class (59%), primarily the Bacillales order (59%). Similarly, SRR26410754 showed 59% Bacilli class (58%) and Bacillales order (58%). At the genus level, *Paenibacillus* was the most abundant genus in both datasets, comprising 51% in SRR26410753 and 50% in SRR26410754, followed by *Viridibacillus* (6% in both) and *Staphylococcus* (0.7% in both).

The spiked species, *Staphylococcus aureus*, was identified within the *Staphylococcus* genus, accounting for 30% of the genus in SRR26410753 (0.2% of the total root) and 26% in SRR26410754 (also 0.2% of the total root). Other notable species included *Paenibacillus sp*. G2S3 (18%), *Viridibacillus sp*. (6%), and *Paenibacillus sp*. FSL H7-0737 (5%) across both datasets. The greater abundance of *Paenibacillus sp*. compared to the spiked *Staphylococcus aureus* was likely due to the natural presence of *Paenibacillus sp*. in potatoes, which skewed the taxonomic distribution significantly [19].

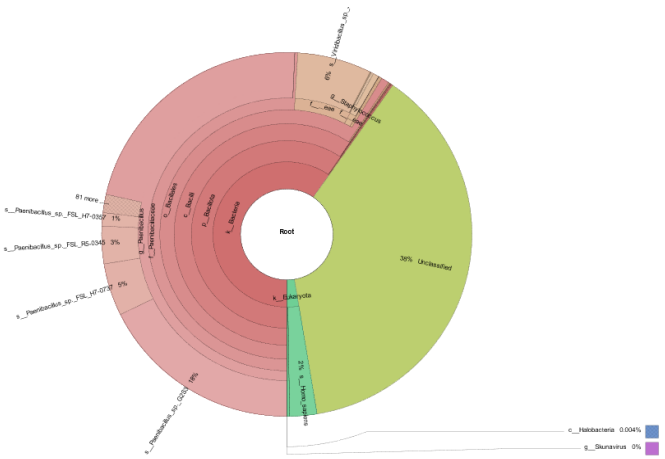


Fig. 9: Krona Pie Chart for SRR26410753

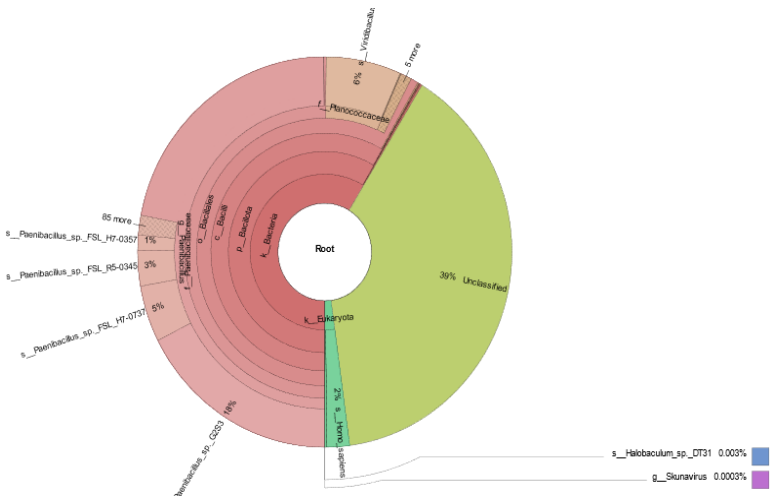
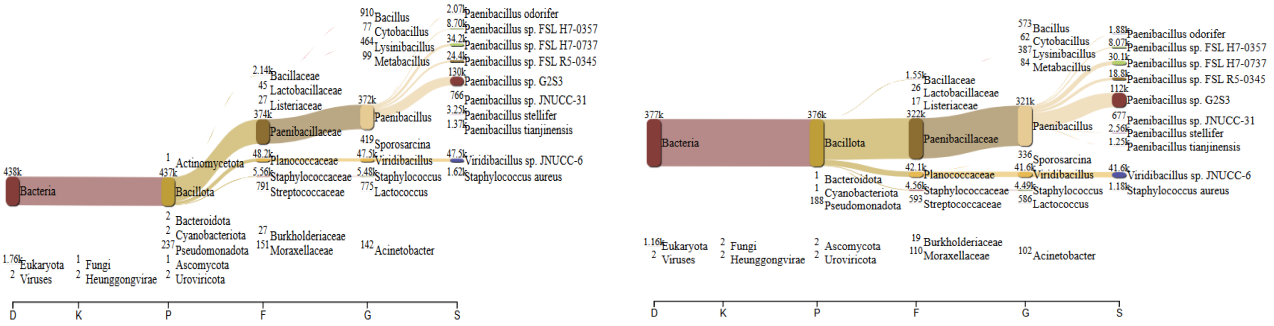


Fig. 10: Krona Pie Chart for SRR26410754

3.2.2 Sankey Visualisations

Sankey diagrams were generated by Pavian Visualisation based on the report output of the Kraken2. In both SRR26410753 and SRR26410754, the Bacteria domain dominated, comprising 438k and 377k reads, respectively, with the majority classified under the Bacillota phylum, accounting for 437k and 376k reads. Within Bacillota, the family Paenibacillaceae was the most abundant, with 374k reads in SRR26410753 and 322k in SRR26410754. The other families within Bacillota included Planococcaceae (48.2k reads in SRR26410753 and 42.1k in SRR26410754), Staphylococcaceae (5.56k in SRR26410753 and 4.56k in SRR26410754), Bacillaceae (2.14k in SRR26410753 and 1.55k in SRR26410754), Streptococcaceae (791 in SRR26410753 and 593 in SRR26410754), Lactobacillaceae (45 in SRR26410753 and 26 in SRR26410754), and Listeriaceae (27 in SRR26410753 and 17 in SRR26410754). Paenibacillaceae primarily consisted of several *Paenibacillus* species, with *Paenibacillus sp. G2S3* being the most abundant, showing 130k reads in SRR26410753 and 112k in SRR26410754. Other notable species included *Paenibacillus sp. H7-0737* (34.2k in SRR26410753 and 30.1k in SRR26410754) and *Paenibacillus sp. FSL R5-0345* (24.4k in SRR26410753 and 18.8k in SRR26410754). The second most abundant family, Planococcaceae, was largely represented by *Viridibacillus sp. JNUCC-6*, which accounted for 47.5k reads in SRR26410753 and 41.6k in SRR26410754. Additionally, the spiked species *Staphylococcus aureus* was observed in the Staphylococcaceae family, comprising 1.62k reads in SRR26410753 and 1.18k in SRR26410754. These statistics highlight the prominent presence of Paenibacillaceae and Planococcaceae in both datasets, as well as the diverse microbial species within the Bacillota phylum.



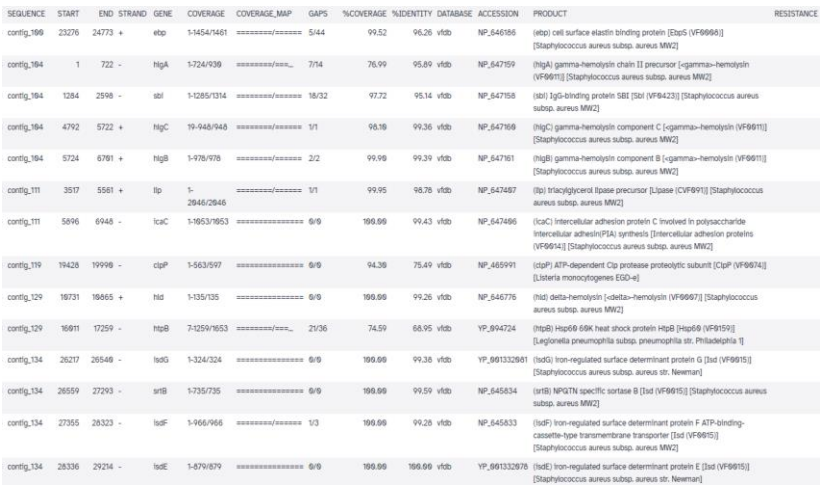
(a) **(b)**

***Fig. 11*:** Sankey Visualisations (a) SRR26410753; (b) SRR26410754

3.3 Gene-based Identification Results

3.3.1 Virulence Factors

In the SRR26410753 sample, 68 different virulence factors (VFs) were identified, several of which had 100% coverage, including genes such as icaC, hld, isdG, srtB, isdF, and isdE. For instance, the icaC gene encodes intercellular adhesion protein C, which plays a key role in various processes such as growth phase variation, migration, transposon insertion, PNAG modification, and biofilm formation, all of which contribute to the bacterium's survivability [20]. In the SRR26410754 sample, 49 VFs were detected, with several genes, such as sspA, sspB, sspC, geh, cap8C, cap8D, and cap8E, also showing 100% coverage. For example, the geh gene encodes glycerol ester hydrolase, which is involved in hydrolyzing host lipids at the infection site, providing fatty acids for membrane biogenesis, substrates for oleate hydratase, and inhibiting immune cell activation, all contributing to the pathogen's virulence [21]. These findings highlight the extensive coverage and functional diversity of virulence factors in both samples, reflecting their potential roles in bacterial pathogenicity.



SEQUENCE	START	END	STRAND	GENE	COVERAGE	COVERAGE_MAP	GAPS	%COVERAGE	%IDENTITY	DATABASE	ACCESSION	PRODUCT	RESISTANCE
contig_169	23276	24773	+	ebp	1-1454/1461	========/=======	5/44	99.52	96.26	vfdb	NP_646186	(ebp) cell surface elastin binding protein [EbpS (VF9696)] [Staphylococcus aureus subsp. aureus MW2]	
contig_164	1	722	-	higA	1-724/939	========/===_	7/14	76.99	95.89	vfdb	NP_647159	(higA) gamma-hemolysin chain II precursor [<gamma>-hemolysin (VF9611)] [Staphylococcus aureus subsp. aureus MW2]	
contig_164	1284	2598	-	sbl	1-1285/1314	========/=======	18/32	97.72	95.14	vfdb	NP_647158	(sbl) IgG-binding protein SBI [Sbl (VF8423)] [Staphylococcus aureus subsp. aureus MW2]	
contig_164	4792	5722	+	higC	19-948/948	========/=======	1/1	98.18	99.36	vfdb	NP_647160	(higC) gamma-hemolysin component C [<gamma>-hemolysin (VF9611)] [Staphylococcus aureus subsp. aureus MW2]	
contig_164	5724	6781	+	higB	1-978/978	========/=======	2/2	99.98	99.39	vfdb	NP_647161	(higB) gamma-hemolysin component B [<gamma>-hemolysin (VF9611)] [Staphylococcus aureus subsp. aureus MW2]	
contig_111	3517	5561	+	lip	1-2646/2646	========/=======	1/1	99.95	98.78	vfdb	NP_647487	(lip) triacylglycerol lipase precursor [Lipase (CVF891)] [Staphylococcus aureus subsp. aureus MW2]	
contig_111	5896	6948	-	icaC	1-1953/1953	==========	9/9	100.00	99.43	vfdb	NP_647496	(icaC) intercellular adhesion protein C involved in polysaccharide intercellular adhesin(PIA) synthesis [Intercellular adhesion proteins (VF9614)] [Staphylococcus aureus subsp. aureus MW2]	
contig_119	19428	19999	-	clpP	1-563/597	==========	9/9	94.39	75.49	vfdb	NP_465991	(clpP) ATP-dependent Clp protease proteolytic subunit [ClpP (VF9674)] [Listeria monocytogenes EGD-e]	
contig_129	16731	16665	+	hld	1-135/135	==========	9/9	100.00	99.26	vfdb	NP_646776	(hld) delta-hemolysin [<delta>-hemolysin (VF9667)] [Staphylococcus aureus subsp. aureus MW2]	
contig_129	16911	17259	-	htpB	7-1259/1653	========/===_	21/36	74.59	68.95	vfdb	YP_994724	(htpB) Hsp60 60K heat shock protein HtpB [Hsp60 (VF9159)] [Legionella pneumophila subsp. pneumophila str. Philadelphia 1]	
contig_134	26217	26549	-	lsdG	1-324/324	==========	9/9	100.00	99.38	vfdb	YP_991332981	(lsdG) Iron-regulated surface determinant protein G [Isd (VF9615)] [Staphylococcus aureus subsp. aureus str. Newman]	
contig_134	26559	27293	-	srtB	1-735/735	==========	9/9	100.00	99.59	vfdb	NP_645834	(srtB) NPQTN specific sortase B [Isd (VF9615)] [Staphylococcus aureus subsp. aureus MW2]	
contig_134	27355	28323	-	lsdF	1-966/966	========/=======	1/3	100.00	99.28	vfdb	NP_645833	(lsdF) Iron-regulated surface determinant protein F ATP-binding-cassette-type transmembrane transporter [Isd (VF9615)] [Staphylococcus aureus subsp. aureus MW2]	
contig_134	28336	29214	-	lsdE	1-879/879	==========	9/9	100.00	100.00	vfdb	YP_991332978	(lsdE) Iron-regulated surface determinant protein E [Isd (VF9615)] [Staphylococcus aureus subsp. aureus str. Newman]	

Fig. 12: Subset of the Tabular Dataset of Virulence Factors for SRR26410753

SEQUENCE	START	END	STRAND	GENE	COVERAGE	COVERAGE_MAP	GAPS	%COVERAGE	%IDENTITY	DATABASE	ACCESSION	PRODUCT	RESISTANCE
contig_182	19639	20497	-	see	1-769/774	========/=======	5/14	98.45	71.76	vfdb	AAA26617	(see) staphylococcal enterotoxin E precursor [SE (VF9629)] [Staphylococcus aureus]	
contig_182	21382	22677	+	sed	1-777/777	========/=======	1/1	99.87	99.61	vfdb	AAB96195	(sed) staphylococcal enterotoxin D precursor [SE (VF9629)] [Staphylococcus aureus RN4220]	
contig_187	26782	21816	+	sspA	1-964/984	========/=======	1/45	100.00	94.85	vfdb	NP_645749	(sspA) serine protease; V8 protease; glutamyl endopeptidase [V8 protease (VF9623)] [Staphylococcus aureus subsp. aureus MW2]	
contig_187	21892	23673	+	sspB	1-1182/1182	==========	9/9	100.00	99.75	vfdb	NP_645748	(sspB) staphopain cysteine proteinase SspB [Staphopain (VF9996)] [Staphylococcus aureus subsp. aureus MW2]	
contig_187	23111	23449	+	sspC	1-339/339	==========	9/9	100.00	99.76	vfdb	NP_645747	(sspC) Staphostatin B [Staphopain (VF9996)] [Staphylococcus aureus subsp. aureus MW2]	
contig_11	26565	27637	-	cap8P	1-1053/1146	========/====_	8/14	91.27	65.94	vfdb	NP_644954	(cap8P) capsular polysaccharide synthesis enzyme Cap8P [Capsule (VF9663)] [Staphylococcus aureus subsp. aureus MW2]	
contig_19	23462	25537	+	geh	1-2073/2073	========/=======	1/3	100.00	99.13	vfdb	NP_645114	(geh) glycerol ester hydrolase [Lipase (VF9612)] [Staphylococcus aureus subsp. aureus MW2]	
contig_27	48699	50078	-	spa	1-1479/1479	========/=======	5/6	99.80	97.17	vfdb	NP_644899	(spa) Immunoglobulin G binding protein A precursor [SpA (VF9617)] [Staphylococcus aureus subsp. aureus MW2]	
contig_28	9782	10265	+	cap8A	1-512/516	========/=======	13/18	96.71	94.69	vfdb	NP_644939	(cap8A) capsular polysaccharide synthesis enzyme [Capsule (VF9663)] [Staphylococcus aureus subsp. aureus MW2]	
contig_28	10458	11142	+	cap8B	1-687/687	========/=======	5/6	99.42	97.97	vfdb	NP_644940	(cap8B) capsular polysaccharide synthesis enzyme Cap8B [Capsule (VF9663)] [Staphylococcus aureus subsp. aureus MW2]	
contig_28	11145	11919	+	cap8C	1-765/765	========/=======	1/1	100.00	99.61	vfdb	NP_644941	(cap8C) capsular polysaccharide synthesis enzyme Cap8C [Capsule (VF9663)] [Staphylococcus aureus subsp. aureus MW2]	
contig_28	11939	13753	+	cap8D	1-1824/1824	==========	9/9	100.00	99.73	vfdb	NP_644942	(cap8D) capsular polysaccharide synthesis enzyme Cap8D [Capsule (VF9663)] [Staphylococcus aureus subsp. aureus MW2]	
contig_28	13743	14771	+	cap8E	1-1629/1629	==========	9/9	100.00	99.22	vfdb	NP_644943	(cap8E) capsular polysaccharide synthesis enzyme Cap8E [Capsule (VF9663)] [Staphylococcus aureus subsp. aureus MW2]	
contig_28	16635	17749	-	lap	62-1766/2661	========/==_	29/59	64.44	68.99	vfdb	NP_465159	(lap) Listeria adhesion protein Lap [Lap (VF8444)] [Listeria monocytogenes EGD-e]	

Fig. 13: Subset of the Tabular Dataset of Virulence Factors for SRR26410754

3.3.2 Antimicrobial Resistance Gene

In the SRR26410753 sample, 7 different antibiotic resistance genes (AMRs) were identified, each contributing to the pathogen's resistance to specific antibiotics. For example, the mecA gene encoded PBP2a, a family beta-lactam-resistant peptidoglycan transpeptidase, which confers resistance to methicillin. Similarly, the blaZ gene encoded penicillin-hydrolyzing class A beta-lactamase (BlaZ), which provides resistance to beta-lactam antibiotics. In the SRR26410754 sample, 7 different AMRs were also found, with each gene linked to resistance against specific antibiotics. The fosB-Saur gene encoded FosB1/FosB3 family fosfomycin resistance bacillithiol transferase, which confers resistance to fosfomycin, while the blaZ gene similarly provided resistance to beta-lactam antibiotics through its beta-lactamase activity. These findings highlight the presence of several antibiotic resistance mechanisms in both samples, emphasizing their potential role in the pathogens' survival against antibiotic treatments.

SEQUENCE	START	END	STRAND	GENE	COVERAGE	COVERAGE_MAP	GAPS	%COVERAGE	%IDENTITY	DATABASE	ACCESSION	PRODUCT	RESISTANCE
contig_113	1566	2154	-	cat-TC	1-652/717	========/======_	2/5	99.38	99.81	ncbi	NG_647562.1	type A-9 chloramphenicol O-acetyltransferase Cat-TC	CHLORAMPHENICOL
contig_147	19639	12645	-	mecA	1-2997/2997	==========	0/0	100.00	99.95	ncbi	NG_647949.1	PBP2a family beta-lactam-resistant peptidoglycan transpeptidase MecA	METHICILLIN
contig_147	12745	13719	+	mecR1	1-975/1758	========___	0/0	55.46	100.00	ncbi	NG_651163.1	beta-lactam sensor/signal transducer MecR1	METHICILLIN
contig_162	14779	15624	-	blaZ	1-846/846	==========	0/0	100.00	99.65	ncbi	NG_647532.1	penicillin-hydrolyzing class A beta-lactamase BlaZ	BETA-LACTAM
contig_162	15731	17487	+	blaR1	1-1758/1758	========/=======	1/1	99.94	99.29	ncbi	NG_651774.1	beta-lactam sensor/signal transducer BlaR1	BETA-LACTAM
contig_162	17477	17857	+	blaI_of_Z	1-381/381	==========	0/0	100.00	99.74	ncbi	NG_647499.1	penicillinase repressor BlaI	BETA-LACTAM
contig_34	42259	43611	+	tet(38)	1-1353/1353	==========	0/0	100.00	99.93	ncbi	NG_648133.1	tetracycline efflux MFS transporter Tet(38)	TETRACYCLINE

(a)

SEQUENCE	START	END	STRAND	GENE	COVERAGE	COVERAGE_MAP	GAPS	%COVERAGE	%IDENTITY	DATABASE	ACCESSION	PRODUCT	RESISTANCE
contig_162	11669	12585	-	blaZ	1-846/846	==========	0/0	100.00	99.65	ncbi	NG_647532.1	penicillin-hydrolyzing class A beta-lactamase BlaZ	BETA-LACTAM
contig_162	12612	14368	+	blaR1	1-1758/1758	========/=======	1/1	99.94	99.29	ncbi	NG_651774.1	beta-lactam sensor/signal transducer BlaR1	BETA-LACTAM
contig_162	14358	14738	+	blaI_of_Z	1-381/381	==========	0/0	100.00	99.74	ncbi	NG_647499.1	penicillinase repressor BlaI	BETA-LACTAM
contig_13	9784	10289	-	fosB-Saur	1-426/429	========/======	5/6	100.00	98.12	ncbi	NG_065844.1	FosB1/FosB3 family fosfomycin resistance bacillithiol transferase	FOSFOMYCIN
contig_28	28449	29796	-	tet(38)	1-1353/1353	========/=======	2/4	100.00	99.63	ncbi	NG_648133.1	tetracycline efflux MFS transporter Tet(38)	TETRACYCLINE
contig_59	19254	21269	-	mecA	1-2997/2997	==========	0/0	100.00	99.95	ncbi	NG_647948.1	PBP2a family beta-lactam-resistant peptidoglycan transpeptidase MecA	METHICILLIN
contig_59	21369	22334	+	mecR1	1-975/1758	========___	0/0	55.46	100.00	ncbi	NG_651163.1	beta-lactam sensor/signal transducer MecR1	METHICILLIN

(b)

***Fig. 14*:** Subset of the Tabular Dataset of Antimicrobial Resistance Gene (a) SRR26410753; (b) SRR26410754

3.4 Allele-based Identification Results

3.4.1 Variant Identification

The subset showcased a small portion of the allele variation in specific chromosomes of the sample reads when mapped against the reference genome. The POS ID indicated the position where the allele had changed, while ALT represented the change in allele from its original allele, REF. A total of 117,677 variants were found when mapped to the reference genome in SRR26410753, while 115,336 variants were identified in SRR26410754.

CHROM	POS ID	REF	ALT	FILTER
NZ_CP126995.1	414	A	G	PASS
NZ_CP126995.1	453	A	T	PASS
NZ_CP126995.1	1062	C	T	PASS
NZ_CP126995.1	1104	A	G	PASS
NZ_CP126995.1	1125	G	A	PASS
NZ_CP126995.1	3745	A	G	PASS
NZ_CP126995.1	4786	T	C	PASS
NZ_CP126995.1	4969	C	T	PASS
NZ_CP126995.1	5119	G	A	PASS
NZ_CP126995.1	5359	T	C	PASS
NZ_CP126995.1	5437	C	T	PASS
NZ_CP126995.1	5482	T	C	PASS
NZ_CP126995.1	5533	A	T	PASS
NZ_CP126995.1	5602	T	C	PASS
NZ_CP126995.1	5761	A	G	PASS
NZ_CP126995.1	5791	A	T	PASS
NZ_CP126995.1	5794	A	C	PASS
NZ_CP126995.1	5806	T	C	PASS

***Fig. 15*:** Subset of the Tabular Dataset of Variants for SRR26410753

Progress in Engineering Application and Technology Vol. 6 No. 1 (2025) p. 574-588

CHROM	POS	ID	REF	ALT	FILTER
NZ_CP126995.1	564		T	C	PASS
NZ_CP126995.1	657		T	C	PASS
NZ_CP126995.1	1125		G	A	PASS
NZ_CP126995.1	1901		T	C	PASS
NZ_CP126995.1	1949		T	C	PASS
NZ_CP126995.1	1953		A	C	PASS
NZ_CP126995.1	1991		T	C	PASS
NZ_CP126995.1	2561		C	T	PASS
NZ_CP126995.1	2591		C	T	PASS
NZ_CP126995.1	4786		T	C	PASS
NZ_CP126995.1	4969		C	T	PASS
NZ_CP126995.1	5119		G	A	PASS
NZ_CP126995.1	5359		T	C	PASS
NZ_CP126995.1	5374		A	G	PASS
NZ_CP126995.1	5437		C	T	PASS
NZ_CP126995.1	5821		A	G	PASS
NZ_CP126995.1	5833		C	T	PASS
NZ_CP126995.1	5872		C	T	PASS

***Fig. 16*:** Subset of the Tabular Dataset of Variants for SRR26410754

3.5 Pathogen Tracking and Visualisation

Pathogen tracking and visualisation refers to the aggregation of the results from ABRicate to further aid in the tracking of potential pathogens in the sample. The workflow continues from gene-based pathogen detection which creates heatmaps and phylogenetic trees based on the results by ABRicate. Thus, any error in the previous process might ultimately affect the results of this part.

Due to heatmaps being incredibly large image, it is unavoidable for the heatmap to be shown clearly. Additionally, due to limitations of the Galaxy platform, the phylogenetic trees could not be properly constructed in one image but were instead split into several tens to hundreds of branches, thus cannot be shown entirely and properly in this project too. As such, only two out of the hundreds branches of the tree will be shown and analysed.

3.5.1 Heatmap Visualisation

The upper section of the heatmap indicates the virulence factor of the SRR26410754 and the lower section indicates those of SRR26410753. Heatmap below showed the abundance or presence of specific genes attributing to the virulence factor across both SRR26410753 and SRR26410754. Darker colour (bright red) indicates great abundance of the gene, with examples of them being hlgA & ukF−PVin SRR26410754. On the other hand, lighter colour of red indicates lower presence, with descending order according to the gradient of colour, in this case, darker shade of red-orange have cap8D & cap8E in SRR26410753.

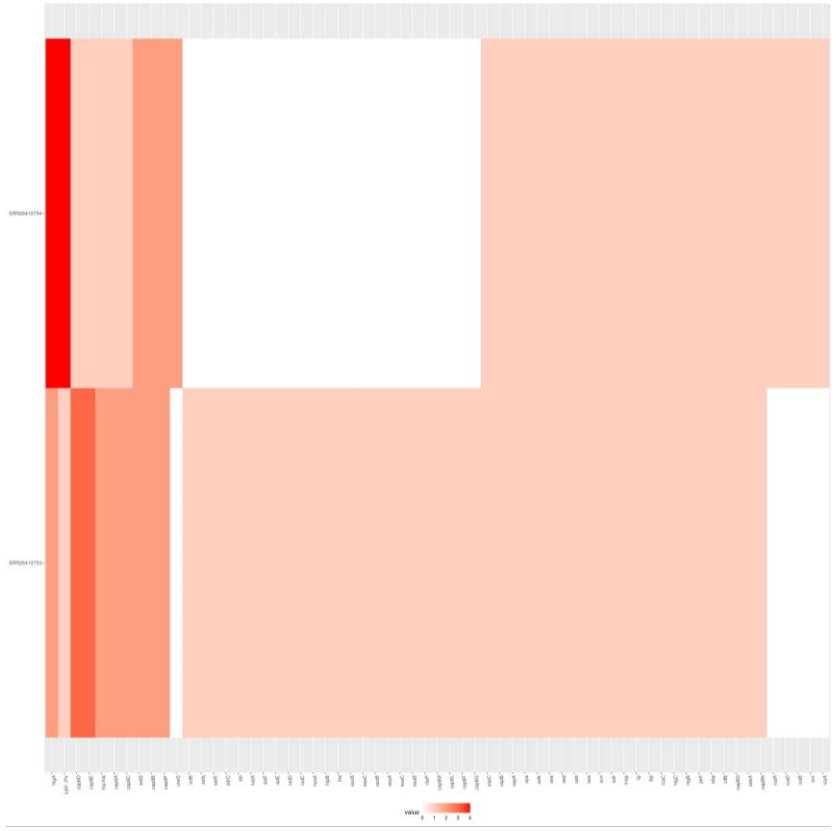
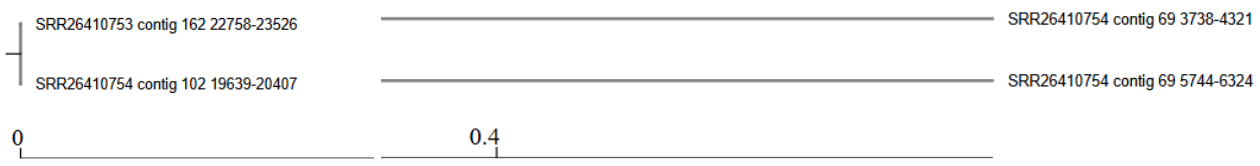


Fig. 17: Heatmap of the Virulence Factors for the Potato Samples spiked with S. aureus

3.5.2 Phylogenetic Tree

The phylogenetic tree could not be constructed properly within the Galaxy platform hence it was split into numerous branches. For the potato samples, the tree was split into 34 branches. For example, AAA26617 is the ascension id of the enterotoxin type E precursor [22] and it was found in contig 162 of SRR26410753 and contig 102 of SRR26410754. Another example of the branches of the tree would be NP_206868, the ascension id of the Urease accessory protein UreG [23] and it was found in contig 69 of SRR26410754 but absent in the other sample.



(a) **(b)**

Fig. 18: Ascension IDs of the Phylogenetic Tree of the Potato Samples (a) AAA26617; (b) NP_206868

4. Conclusion

In this article, it is shown that a comprehensive bioinformatics pipeline for the analysis of direct Nanopore sequencing data within the Galaxy platform was developed, addressing our primary objective. The pipeline's development included meticulous configuration and integration of various bioinformatics tools to ensure robust detection and characterization of foodborne pathogens.

Utilizing both simulated and real-world datasets containing known foodborne pathogens, the pipeline's accuracy and efficiency were confirmed. The pipeline demonstrated high sensitivity and specificity, accurately identifying and characterizing pathogen sequences amidst complex food sample matrices. This validation underscores the reliability and practical applicability of our bioinformatics solution. It was also able to accurately tell between contaminated samples and uncontaminated samples as showcased for Whole Milk samples' (unpublished data) lack of virulence factor in the heatmap.

The research project is not without its limitations as the entire process was reliant on an external server for data processing which might be interfered if server overload happens. It is more recommended to use a localised version but the technical expertise it requires might prove to be a huge obstacle for those inexperienced. Other than that, the phylogenetic tree was not able to be built properly as the tree's branches were separated instead of aligned. Tools such as Interactive Tree of Life (ITOL) are not available in the web browser version of Galaxy thus were unable to be used.

In conclusion, the bioinformatics pipeline developed in this study provides a powerful, flexible, and user-friendly tool for foodborne pathogen detection using direct Nanopore sequencing data within the Galaxy platform. This report lays the foundation for future improvements and adaptations, potentially expanding its use to broader pathogen surveillance and diagnostics contexts.

Acknowledgement

The authors would like to express the sincerest gratitude and appreciation towards all the individuals involved and Universiti Tun Hussein Onn for contributing and supporting the completion of this research paper.

References

[1] World Health Organisation. (2019). Estimating the burden of foodborne diseases. https://www.who.int/activities/estimating-the-burden-of-foodborne-diseases

[2] Alahi, M., & Mukhopadhyay, S. (2017). Detection methodologies for pathogen and toxins: A review. Sensors, 17(8), 1885. https://doi.org/10.3390/s17081885

[3] Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. Nature, 265(5596), 687-695. https://doi.org/10.1038/265687a0

[4] Lopes, M., Louzada, S., Gama-Carvalho, M., & Chaves, R. (2021). Genomic tackling of human satellite DNA: Breaking barriers through time. International Journal of Molecular Sciences, 22(9), 4707. https://doi.org/10.3390/ijms22094707

[5] Yakun, X., Yue, M., Xiaoxi, H., & Jun, W. (2019). Analysis of prospective microbiology research using third-generation sequencing technology. Biodiversity Science, 27(5), 534-542. https://doi.org/10.17520/biods.2018201

[6] Silverman, J. D., Bloom, R. J., Jiang, S., Durand, H. K., Mukherjee, S., & David, L. A. (2019). Measuring and mitigating PCR bias in microbiome data. https://doi.org/10.1101/604025

[7] Batut, B., Nasr, E., & Zierep, P. (2024, September 27). Pathogen detection from (direct Nanopore) sequencing data using galaxy - Foodborne edition. Galaxy Training Network. https://training.galaxyproject.org/topics/microbiome/tutorials/pathogen-detection-from-nanopore-foodborne-data/tutorial.html

[8] Sciensano. (2023, October 17). ID 1029022 - BioProject - NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1029022

[9] Fera Science Limited. (2021, September 27). PromethION sequencing; Ready-to-eat food, metagenome long-read sequencing: whole milk. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/sra/ERX6221067

[10] University of Florida. (2024, January 25). Guides @ UF: Genomic data processing: Long read processing: Intermediate. Guides @ UF at University of Florida. https://guides.uflib.ufl.edu/c.php?g=1325624&p=9810272

[11] Seeman, T., & Grüning, B. (2019, May 12). Tseemann/abricate: :mag_right: Mass screening of contigs for antimicrobial and virulence genes. GitHub. https://github.com/tseemann/abricate

[12] Bogaerts, B., Van den Bossche, A., Verhaegen, B., Delbrassinne, L., Mattheus, W., Nouws, S., Godfroid, M., Hoffman, S., Roosens, N. H., De Keersmaecker, S. C., & Vanneste, K. (2024). Closing the gap: Oxford Nanopore technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. Journal of Clinical Microbiology, 62(5). https://doi.org/10.1128/jcm.01576-23
[13] Technical University of Denmark. (2023, May 27). Paenibacillus Sp. G2S3 genome assembly ASM3012310v1. NCBI. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_030123105.1/
[14] Universidad Tecnologica Metropolitana (UTEM). (2019, October 25). Acinetobacter Sp. genome assembly ASM936032v1. NCBI. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009360325.1/
[15] De La Cerda, G. Y., Landis, J. B., Eifler, E., Hernandez, A. I., Li, F., Zhang, J., Tribble, C. M., Karimi, N., Chan, P., Givnish, T., Strickler, S. R., & Specht, C. D. (2023). Balancing read length and sequencing depth: Optimizing Nanopore long‐read sequencing for monocots with an emphasis on the Liliales. Applications in Plant Sciences, 11(3). https://doi.org/10.1002/aps3.11524
[16] Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A. O., Liang, F., Li, J., Sandoval, J. R., Cooper, D. N., Ye, K., Ruan, J., Xiao, C., Wang, D., Wu, D., & Wang, S. (2024). NextDenovo: An efficient error correction and accurate assembly tool for noisy long reads. Genome Biology, 25(1). https://doi.org/10.1186/s13059-024-03252-4
[17] Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biology, 21(1). https://doi.org/10.1186/s13059-020-1935-5
[18] Senol Cali, D., Kim, J. S., Ghose, S., Alkan, C., & Mutlu, O. (2018). Nanopore sequencing technology and tools for genome assembly: Computational analysis of the current state, bottlenecks and future directions. Briefings in Bioinformatics, 20(4), 1542-1559. https://doi.org/10.1093/bib/bby017
[19] Buytaers, F. E., Verhaegen, B., Van Nieuwenhuysen, T., Roosens, N. H., Vanneste, K., Marchal, K., & De Keersmaecker, S. C. (2024). Strain-level characterization of foodborne pathogens without culture enrichment for outbreak investigation using shotgun metagenomics facilitated with nanopore adaptive sampling. Frontiers in Microbiology, 15. https://doi.org/10.3389/fmicb.2024.1330814
[20] Prabu, R., Mohanty, A., Balakrishnan, S. S., & Sundar, K. (2021). Molecular docking and simulation of ICAC protein as O-succinyltransferase function in Staphylococcus Epidermidis Biofilm formation. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3995107
[21] Subramanian, C., Frank, M. W., Yun, M., & Rock, C. O. (2023). The Phospholipase A1 activity of glycerol ester Hydrolase (Geh) is responsible for Extracellular 2-12(S)-methyltetradecanoyl-Lysophosphatidylglycerol production in staphylococcus aureus. mSphere, 8(2). https://doi.org/10.1128/msphere.00031-23
[22] Subramanian, C., Frank, M. W., Yun, M., & Rock, C. O. (2023). The Phospholipase A1 activity of glycerol ester Hydrolase (Geh) is responsible for Extracellular 2-12(S)-methyltetradecanoyl-Lysophosphatidylglycerol production in staphylococcus aureus. mSphere, 8(2). https://doi.org/10.1128/msphere.00031-23
[23] Damas, M. S., Ferreira, R. L., Campanini, E. B., Soares, G. G., Campos, L. C., Laprega, P. M., Soares da Costa, A., Freire, C. C., Pitondo-Silva, A., Cerdeira, L. T., Cunha, A. F., & Pranchevicius, M. D. (2022). Whole genome sequencing of the multidrug-resistant Chryseobacterium indologenes isolated from a patient in Brazil. Frontiers in Medicine, 9. https://doi.org/10.3389/fmed.2022.931379