WHOLE GENOME SEQUENCING DATA ANALYSIS OF TWO UNKNOWN MULTI-DRUG RESISTANT BACTERIAL ISOLATES

DISSERTATION

SUBMITTED TO THE MAHATMA GANDHI UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

MASTER OF SCIENCE IN BIOINFORMATICS

By

EBIN SIBY

(200011019263)

SCHOOL OF BIOSCIENCES

Mar Athanasios College for Advanced Studies Tiruvalla (MACFAST)

Kerala - 689101

2022

DECLARATION

I, **EBIN SIBY**, do hereby declare that the dissertation titled '**Whole genome sequencing data analysis of two unknown multi-drug resistant bacterial isolates'**, submitted in partial fulfilment of the requirements for the award of the degree of **Master of Science in Bioinformatics** of the Mahatma Gandhi University, Kottayam, Kerala, is an authentic record of the studies and research work carried out by me under the supervision and guidance of **Dr Shijulal Nelson Sathi (RGCB Trivandrum)**, and that no part of this project work has been presented elsewhere for the award of any other degree and that there is no plagiarism in the written thesis.

Ebin Siby

RGCB
DISCOVERIES FOR A BETTER TOMORROW

Azadi Ka Amrit Mahotsav

Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram 695014, Kerala State, India.
An Autonomous National Institute for Discovery, Innovation & Translation in Biotechnology and Disease Biology,
Government of India, Ministry of Science & Technology, Department of Biotechnology.

राजीव गाँधी जैव प्रौद्योगिकी केन्द्र, तिरुवनन्तपुरम 695 014, केरल, भारत.
जैवप्रौद्योगिकी और रोग जीवविज्ञान में आविष्कार, नवीनता एवं अनुवाद की स्वायत्त राष्ट्रीय संस्थान,
भारत सरकार विज्ञान एवं प्रौद्योगिकी मंत्रालय, जैवप्रौद्योगिकी विभाग.

CERTIFICATE

This is to certify that the dissertation entitled, "**Whole genome sequencing data analysis of two unknown multi-drug resistant bacterial isolates**" submitted by **Mr. Ebin Siby,** Reg. No. 200011019263 to the Mahatma Gandhi University, in partial fulfillment of the requirements for the Degree of M.Sc. Bioinformatics is a record of bonafide work carried by him under my supervision (Dr. Shijulal Nelson Sathi, Scientist C, Transdisciplinary Biology, Rajiv Gandhi Centre for Biotechnology (RGCB), Trivandrum, Kerala). The contents of this dissertation, in full or in parts, have not been submitted to any other Institute or University for the award of any Degree or Diploma.

Date: 27.09.2022

Signature of the Guide
Dr. Shijulal Nelson Sathi
Scientist C, Transdisciplinary Biology

राजीव गांधी जैव प्रौद्योगिकी केन्द्र
जैव प्रौद्योगिकी विभाग
DEPT. OF BIOTECHNOLOGY
विज्ञान एवं प्रौद्योगिकी मंत्रालय
MINISTRY OF SCIENCE & TECHNOLOGY
भारत सरकार/GOVT. OF INDIA
तिरुवनन्तपुरम/THIRUVANANTHAPURAM
पिन/Pin-695 014
RAJIV GANDHI CENTRE FOR BIOTECHNOLOGY

तैक्काट पी.ओ., पूजप्पुरा, तिरुवनन्तपुरम - 695 014, केरल, भारत. फोण: 0471-2529400 (30 लाईन्स), 2347975, 2348104, 2348753, 2345899
फैक्स: 0471-2348096, ई-मेल: director@rgcb.res.in, info@rgcb.res.in, वेब: www.rgcb.res.in
Thycaud P.O., Poojappura, Thiruvananthapuram - 695 014, Kerala, India. Phone: 0471-2529400 (30 Lines), 2347975, 2348104, 2348753, 2345899
Fax: 0471-2348096, E-mail: director@rgcb.res.in, info@rgcb.res.in, Web: www.rgcb.res.in

CERTIFICATE

The thesis entitled '**Whole genome sequencing data analysis of two unknown multi-drug resistant bacterial isolates**' submitted by **Mr. Ebin Siby** (Reg. No. 200011019263) for the partial fulfilment of the requirements for the award of degree of **Master of Science in Bioinformatics** of Mahatma Gandhi University, Kottayam, Kerala is evaluated and approved.

Prof. Stephen James
Internal guide & Assistant Professor
School of Bioscience
MACFAST

Dr. Jenny Jacob
HOD
School of Bioscience
MACFAST

Signature, Name & Designation
of the Examiner 1

Signature, Name & Designation
of the Examiner 2

Date:

ACKNOWLEDGEMENTS

The satisfaction and euphoria that accompany the accomplishment of any work would be incomplete without mentioning the people who made it possible and whose consistent guidance and encouragement crown all the efforts. This project was not only a technical endeavour but also an interesting learning experience.

First of all, I would like to thank my mentor Dr Shijulal Nelson Sathi, RGCB Thiruvananthapuram for giving me the opportunity to do research and providing invaluable guidance throughout this research.

I am extremely greatful to the research scholar, Mrs Jiffy John for helping me learn all the techniques that were crucial for the project. I am also thankful to other scholars working in RGCB Trivandrum for their friendly and helpful behaviour.

I am obliged to Dr Jenny Jacob, HOD department of Bioscience, MACFAST and all other teachers including Mr Stephen James, Dr Blesson George for their invaluable advice and help during my studies and project work.

I would also like to express my thanks to my loving parents, seniors and friends for having made everything possible by giving me strength and confidence tom do this extraordinary work.

Over and above all, Thank you Jesus, for the blessings throughout my life

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Sl. No.	Abbreviation	Full Form
1	ARG	Antibiotic Resistance Gene
2	MIC	Minimum Inhibitory Concentration
3	BAM	Binary Alignment Map
4	SAM	Sequence Alignment Map
5	PCR	Polymerase Chain Reaction
6	EMBL	European Molecular Biology Laboratory
7	CARD	Comprehensive Antibiotic Resistance Database
8	tRNA	Transfer Ribonucleic Acid
9	rRNA	Ribosomal Ribonucleic Acid
10	tmRNA	Transfer Messenger Ribonucleic Acid
11	CDS	Coding Region
12	NCBI	National Center for Biotechnology Information

ABSTRACT

Antibiotic resistance is the ability of bacteria to resist the effects of an antibiotics. Bacteria become antibiotic resistant by either genetic mutations or by acquiring antibiotic resistance genes (ARGs). Multi-drug resistant bacteria are bacteria that are not destroyed by a number of different antibiotics. Two unknown bacterial isolates were obtained from the soil environment and named them as L1C1 and L2C10 respectively. Based on antibiotic resistance profiling, these microbes show high resistance to antibiotics. We performed whole genome sequencing data analysis of these two unknown multi-drug resistance bacterial isolates. First , the quality of the reads were checked using FASTQC followed by the whole genome assembly using spades assembler. Spades assembly gives 1,539 contigs and 2,126 contigs for L1C1 and L2C10 respectively. Using prokka annotation the genes present in the genomes were predicted and annotated. There are 1124 genes present in L1C1 and 1655 genes present in L2C10. All the antibiotic resistance genes present in the genomes of the organisms were predicted using card database. It was observed that L1C1 has 27 and L2C10 has 23 antibiotic resistant genes. The species were identified to be Staphylococcus gallinarum using 16S rRNA database for both the genomes.

Keyword: Antibiotic Resistance, Antibiotic resistance genes, Whole genome assembly, prokka annotation, card database, 16S rRNA database

1.INTRODUCTION

1.1 Antibiotic Resistance

Antibiotic resistance (Nji *et al*., 2021) is the ability of bacteria to resist the effects of an antibiotic which they were previously sensitive. Bacteria become antibiotic resistant by either genetic mutations or by acquiring antibiotic resistance genes (ARGs). Antibiotic resistance is a combination of germs exposed to antibiotics, and the spread of those germs and their resistance mechanisms. Antibiotic resistance does not mean our body is resistant to antibiotics. It means the bacteria or fungi causing the infection are resistant to the antibiotic or antifungal treatment.

Antibiotic resistant bacteria are bacteria that are not controlled or killed by antibiotics. They are able to survive and even multiply in the presence of an antibiotic. The presence of ARGs is the main factor for bacterial resistance. Pathogenic bacteria acquire ARGs (Preena *et al*., 2020) through plasmid exchange at the gene level and develop strong resistance to antibiotics. ARG-carrying plasmids, integrons (In), and transposons (Tn) in bacteria can undergo horizontal gene transfer (HGT) among strains of the same species and different species.

Some bacteria are multidrug resistant, meaning they are resistant to a variety of antibiotics. The spread of antibiotic resistance is facilitated by multidrug-resistant bacteria, which can be challenging to treat. Some multidrug resistant bacteria are Vancomycin resistant enterococcus (VRE), methicillin resistant staphylococcus aureus (MRSA), and Pseudomonas aeruginosa are all antibiotic-resistant bacteria. Next Generation Sequencing has increased our capacity to

track bacterial clones (such as the multidrug-resistant ones), to identify new antibiotic resistance genes (ARGs) and their genetic carriers, such as plasmids. Pyrosequencing, Ion torrent sequencing, illumina sequencing are some examples of NGS.

1.2 Antibiotic Resistance Profiling

Characterisation of antibiotic resistance properties of an organism through experimental or computational methods. Bacterial isolates can be tested for antibiotic resistance using several methods. Antimicrobial susceptibility (Juayang *et al*., 2014) is thereby expressed either as a Minimum Inhibitory Concentration (MIC), determined via serial dilution or E test, or as an inhibition zone diameter when determined via the disk diffusion.

1.2.1 Minimum Inhibitory Concentration

Minimum Inhibitory Concentration is the lowest drug concentration that prevents visible microorganism growth after overnight incubation.

1.2.2 Serial Dilution

Serial dilution is a laboratory technique, in which a stepwise dilution process is performed on a solution with an associated dilution factor.

1.2.3 Inhibition Zone Diameter

The Zone of inhibition is a circular area around the spot of the antibiotic in which the bacteria colonies do not grow. The zone of inhibition can be used to measure the susceptibility of the bacteria towards the antibiotic. The process of measuring the diameter of this Zone of Inhibition can be automated using Image processing is shown in figure 1.1.

Here, the two bacterial isolates L1C1 and L2C10 were chosen for the Antibiotic resistance profiling using disc diffusion method, it shows high resistance to various antibiotics. The antibiotics include Erythromycin, Oxacillin, Pencilin, Methicillin Azithromycin etc... Below figure shows the zone of inhibition of bacterial isolates shows that Antibiotics like Erythromycin, Oxacillin, Pencillin, Linezolid is highly resistant to this isolate.

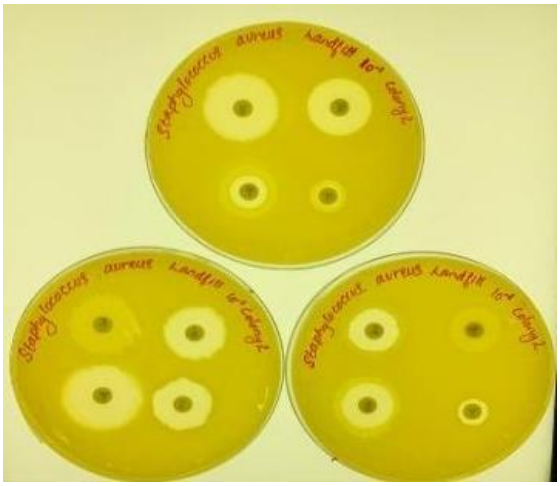


Fig 1.1: The zone of inhibition of bacterial isolates L1C1 and L2C10

Antibiotics.	Isolates 1	Isolates2
Erythromycin	R (00)	R (10)
Oxacillin	R (00)	R (00)
Pencilin	R (12)	R (22)
Linezolid	R (20)	R (20)
Methicilin	R (07)	R (00)
Kanamycin	R (13)	R (13)
Clindamycin	R (08)	I (18)
Chloramphenicol	I (17)	S (18)
Ciprofloxacin	S (21)	I (18)
Cefoxitin	S (20)	I (17)
Tetracyclin	I (16)	S (21)
Azithromycin	R (00)	R (00)
Vancomycin	-	S (15)

Table 1.2.1: Antibiotic Resistance Profiling

The above table 1.2.1 shows that unknown bacterial isolates are highly resistant to several antibiotics. Because of the highly resistance profiling my aim is to do "**Whole genome sequencing data analysis of two unknown multi-drug resistance bacterial isolates**".

2.REVIEW OF LITERATURE

2.1 Pre-Antibiotic Era

There are a number of studies that organized for the management of infections. Beginning with the discovery of antibiotics, they are apparently one of the most affluent forms of chemotherapy. Tracing of antibiotic exposure in the ancient population is more difficult to detect. Introduction to antibiotics in this era could be through the restorative use of traditional medicine (Aminov. 2010). One of the examples is the use of tetracycline has been found in human skeletal remains from ancient (Bassett *et al*., 1980). There were no any oral medicines available for the treatment of specific micro-organism infections during this period (SA shkenazi. 2013). The use of medicines comprises a striking effect in reducing the mortality rate of some life-threatening bacterial diseases when compared with the situation in the pre-antibiotic era. During the early 20^{th} century, treatment of microbe infections was based on ancient medicines. The different ancient culture was used mixtures congaing antimicrobial properties such as mold and plant extract to treat various infections.

2.2 Antibiotic Era

From the period of Second World War itself the development of sulfa drugs and penicillin was started. The introduction of these drugs improves the practice of medicine. The arrival of antibiotics follows a crucial reduces in bacterial infection-related to mortality and morbidity (Scott H. Podolsky. 2015).

2.3 Discovery of Antibiotics

Penicillin, the wonder drug was first naturally discovered antibiotic. It was discovered in 1928 by Alexander Fleming, a Scottish scientist working in St Mary's Hospital London. Penicillin was first released for extensive use in the early 1940's and it saved many lives during World War II. Antibiotics can also be derived from other bacteria. These include aminoglycosides and carbapenems. Because of the introduction of sulfonamides and penicillin, the infectious disease mortality got reduced. Late of 1960's certain new class of anti-bacterial drugs was discovered. During the 19th century, the use of medicinal plants also increases the value of medicine. Coal-tar, a rich by - product of the industrialization, contained many of the aromatic or aliphatic building blocks that became the toolkit of medicinal chemistry (E. Farber. 1952). By half of the 20th-century, drug research was enhanced by several areas. The successful antibiotic class is beta-lactam class of antibiotics. Most of the antibiotic class is discovered during the period of 1940's and 1950's. Successful class of antibiotics is bacteriocidal, including B-lactams and fluoroquinolones, while others are conditionally bacteriocidal such as rifampicins and aminoglycosides or bacteriostatic are tetracyclines and macrolides (D. Hughes and A. Karlén. 2014). The use of streptomycin for the treatment of tuberculosis started during 1947 resulted the mutant strains of Mycobacterium tuberculosis which shows resistance to the therapeutic attentiveness was originated during patient treatment. The availability of drugs has doubtless had a better effect on the entire medicinal field. Bacteria have invented an overabundance of mechanisms that cause resistance to antibiotics. Some of the main reasons include the direct modification of antibiotics; decrease the penetration of antibiotics through cell wall, efflux pump mechanisms, modification of active site and alteration of metabolic pathways.

2.4 Illumina Sequencing Technology

Illumina sequencing technology (Buermans & den Dunnen, 2014) leverages clonal array formation and proprietary reversible terminator technology for rapid and accurate large-scale sequencing. The innovative and flexible sequencing system enables a broad array of applications in genomics, transcriptomics, and epigenomics. The Illumina sequencing approach is built around a massive quantity of sequence reads in parallel. Deep sampling and uniform coverage is used to generate a consensus and ensure high confidence in determination of genetic differences. Deep sampling allows the use of weighted majority voting and statistical analysis, similar to conventional methods, to identify homozygotes and heterozygotes and to distinguish sequencing errors. Each raw read base has an assigned quality score so that the software can apply a weighting factor in calling differences and generating confidence scores.

2.5 Genome Assembly

Genome assembly is the computational process of deciphering the sequence composition of the genetic material (DNA) within the cell of an organism, using numerous short sequences called reads derived from different portions of the target DNA as input (Pop *et al*., 2004). The term genome is a collective reference to all the DNA molecules in the cell of an organism. Sequencing generally refers to the experimental process of determining the sequence composition of biomolecules such as DNA, RNA, and protein.

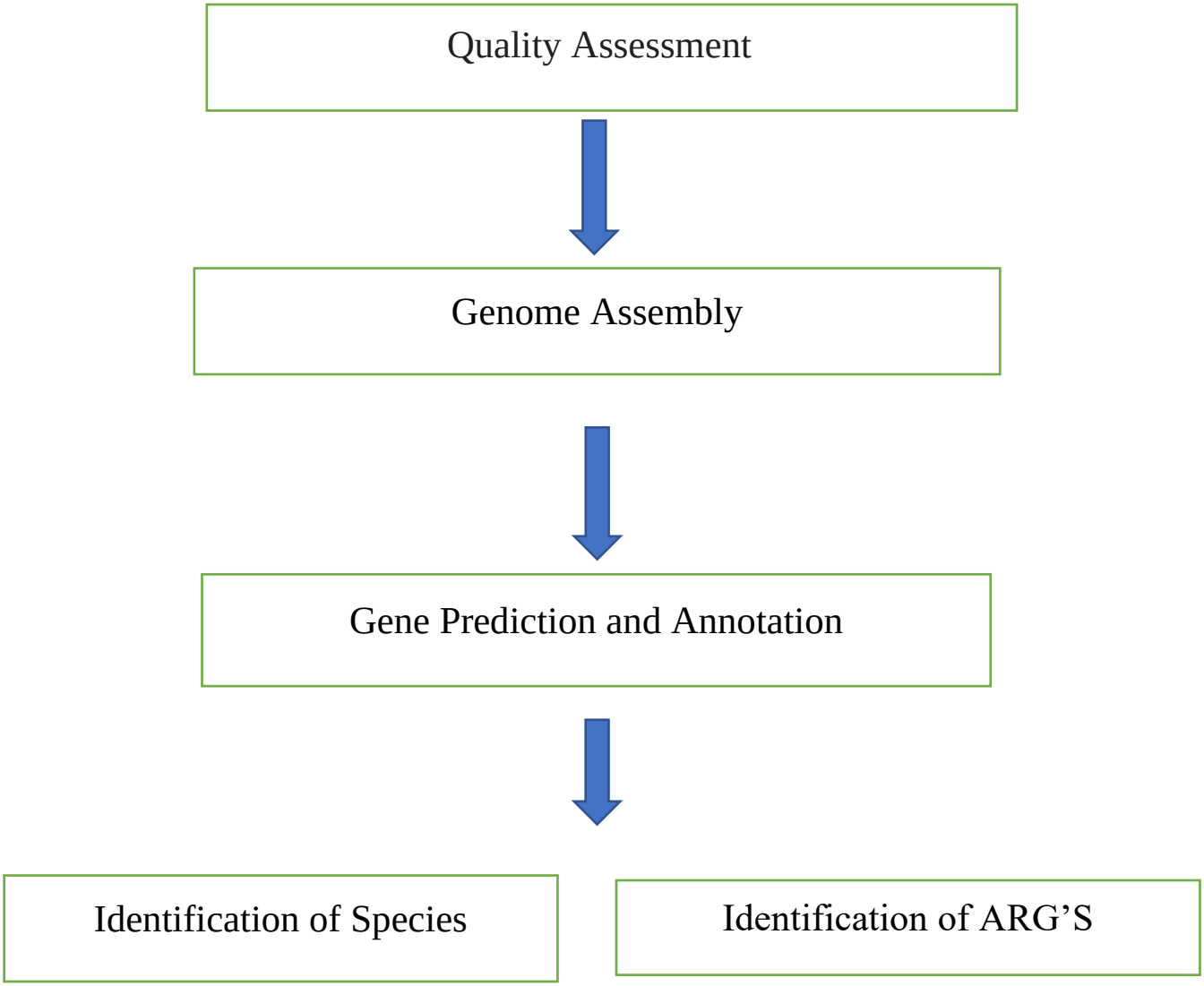
2.6 Gene Prediction and Annotation

The first version of the NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) combining HMM-based gene prediction algorithms with protein sequence similarity search methods was developed in 2001-2002. The initial pipeline used a combination of automatic

protein-coding gene model prediction via two prediction methods, GeneMarkS and Glimmer. These predictions were then augmented using information on evolutionarily conserved proteins from Clusters of Orthologous Groups or COGs and NCBI Prokaryotic Clusters. Proteins from these clusters were mapped to the genome in order to search for genes missed by the *ab initio* predictors. Genes of ribosomal RNAs were predicted either by the BLASTn sequence similarity search using the entries from the RNA sequence database as queries or by running specialized tools, such as Infernal and Rfam. Genes of transfer RNAs were predicted using tRNAscan-SE. The Standard Operating Procedure (SOP) for the first version of the NCBI genome annotation pipeline was published in 2008. (Tatusova *et al*., 2016)

3.MATERIALS AND METHODS

3.1 Work Flow of the Study



3.2 Quality Assessment

3.2.1 Fast QC

First, the quality of the reads were checked using FASTQC. FastQC (Brown *et al*., 2017) aims to provide a simple way to do some quality control checks on raw sequence data coming from

high throughput sequencing pipelines. It provides a modular set of analyses which we can use to give a quick impression of whether our data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are

- Import of data from BAM, SAM or FastQ files.
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application.

3.2.2 Data Pre-processing

Pre-processing of the data has done by Trimmomatic. Here the starting portion of the reads were trimmed using trimmomatic tool. Trimmomatic (Glazinska *et al*., 2017) is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters. These adapters can pose a real problem depending on the library preparation and downstream application. There are two major modes of the program: Paired end mode and Single end mode. The paired end mode will maintain correspondence of read pairs and also use the additional information contained in paired reads to better find adapter or PCR primer fragments introduced by the library preparation process. Trimmomatic works with FASTQ files.

3.2 Genome Assembly

3.2.1 Spades Assembler

Genome assembly is done by using spades assembler. Spades Genome Assembler (Bankevich *et al.*, 2012) is an open-source tool for de novo sequencing. This application is designed to assemble small genomes from MDA single-cell and standard bacterial data sets. Spades supports paired-end reads, mate-pairs and unpaired reads. Spades was initially designed for single-cell and standard bacterial data sets and is not intended for larger genomes. The four stages of Spades, which deal with issues that are particularly troublesome in SCS: sequencing errors; non-uniform coverage; insert size variation; and chimeric reads and bi reads:

Stage 1: assembly graph construction is addressed by every NGS assembler and is often referred to as de Bruijn graph simplification. We propose a new approach to assembly graph construction that uses the multisided de Bruijn graph, implements new bulge/tip removal algorithms, detects and removes chimeric reads, aggregates biread information into distance histograms, and allows one to backtrack the performed graph operations.

Stage 2: k-bimer adjustment derives accurate distance estimates between *k*-mers in the genome using joint analysis of distance histograms and paths in the assembly graph.

Stage 3: Constructs the paired assembly graph, inspired by the PDBG approach.

Stage 4: Contig construction was well studied in the context of Sanger sequencing. Since NGS projects typically feature high coverage, NGS assemblers generate rather accurate contigs.

Spades construct DNA sequences of contigs and the mapping of reads to contigs by backtracking graph simplifications

3.3 Gene Prediction and Annotation

Gene prediction and annotation is done by Prokka tool. Prokka (Seemann, 2014) finds and annotates features (both protein coding regions and RNA genes, i.e., tRNA, rRNA) present on a sequence. Prokka uses a two-step process for the annotation of protein coding regions: first, protein coding regions on the genome are identified using Prodigal; second, the function of the encoded protein is predicted by similarity to proteins in one of many protein or protein domain databases. Prokka is a software tool that can be used to annotate bacterial, archaeal and viral genomes quickly, generating standard output files in GenBank, EMBL and gff formats. Table 3.3.1 as follows

Suffix	Description of file contents
.fna	FASTA file of original input contigs (nucleotide)
.faa	FASTA file of translated coding genes (protein)
.ffn	FASTA file of all genomic features (nucleotide)
.fsa	Contig sequences for submission (nucleotide)
.tbl	Feature table for submission
.sqn	Sequin editable file for submission
.gbk	Genbank file containing sequences and annotations
.gff	GFF v3 file containing sequences and annotations
.log	Log file of Prokka processing output
.txt	Annotation summary statistics

Table 3.3.1: Description of file contents

3.4 Identification of Antibiotic Resistance Genes

Identification of antibiotic resistant genes is done by blast against CARD database.

The **Comprehensive Antibiotic Resistance Database (CARD)** (Alcock *et al*., 2020) is a biological database that collects and organizes reference information on antimicrobial resistance genes, proteins and phenotypes. CARD focuses on providing high quality reference data and molecular sequences within a controlled vocabulary, the Antibiotic Resistance Ontology (ARO), designed by the CARD biocuration team to integrate with software development efforts for resistome analysis and prediction.

3.5 Identification of the unknown Species

Identification of unknown species is done by blast against 16S rRNA database from the NCBI database.16S rRNA (Phumudzo *et al.*, 2013) is the standard for taxonomic identification. This is because current microbial taxonomies based on genomes have a number of limitations, including low phylogenetic resolution and a lack of absolute numbers. Compared with this usefulness, there are only a few public 16S databases available for microbial identification, because it requires a lot of effort to collect and maintain the most up-to-date taxonomic information. Here I use 16S rRNA database for the species identification.

4.RESULTS AND DISCUSSION

4.1 Quality Assessment of L1C1

4.1.1 Per Base Sequence Quality of L1C1

Per Base Sequence Quality shows an overview of the range of quality values across all bases at each position in the FastQ file. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25. A failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20. the per base sequence quality in which the quality score is above 20. It is a good quality score.

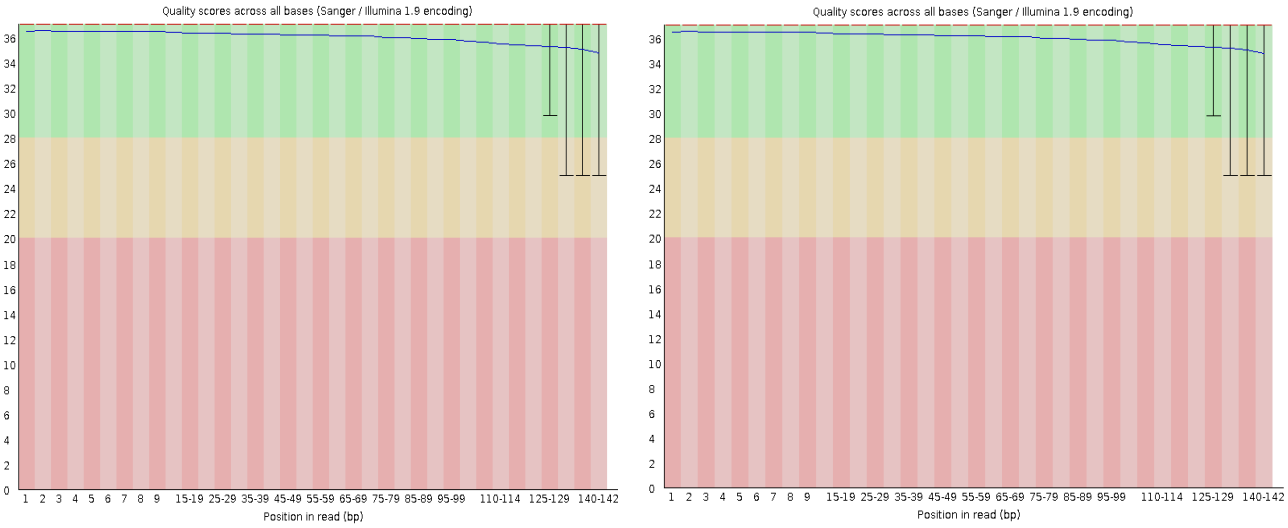


Figure 4.1.1: Per base sequence quality

4.1.2 Per Tile Sequence Quality of L1C1

Per Tile Sequence Quality graph allows to look at the quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell. This graph will issue a warning if any tile shows a mean Phred score more than 2 less than the mean for that base across all tiles and issue a failure if any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles. A good plot should be blue all over. This plot shows blue all over, so it is a good plot.

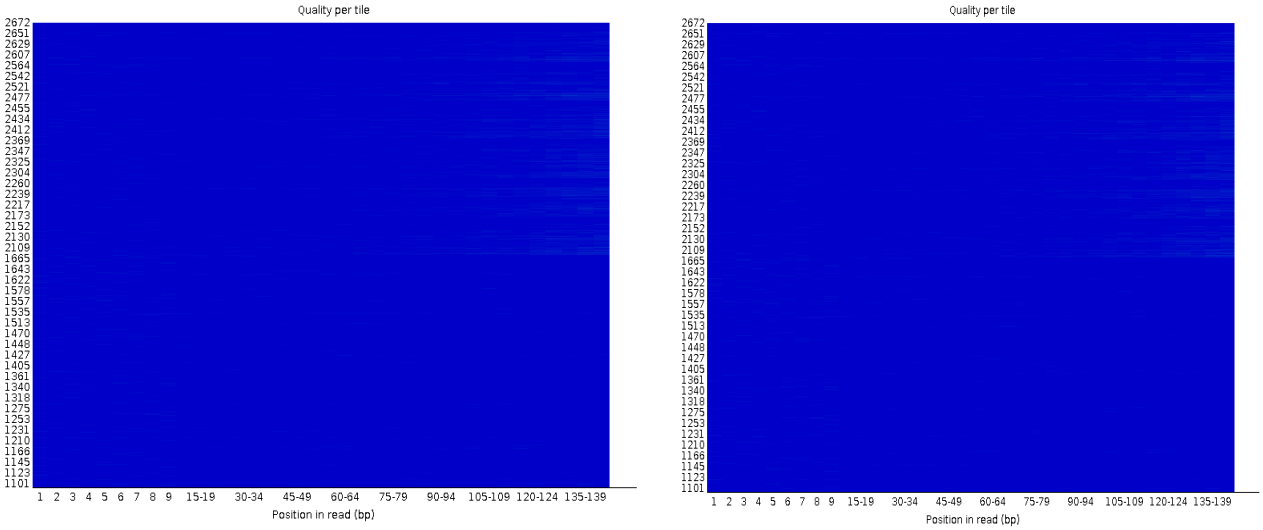


Figure 4.1.2: Per tile sequence quality

4.1.3 Per Sequence Quality Scores of L1C1

The per sequence quality score allows to see if a subset of your sequences have universally low-quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged. A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate and an error is raised if

the most frequently observed mean quality is below 20 - this equates to a 1% error rate. The averages quality score on the x-axis and the number of sequences with that average on the y-axis. Majority of our reads have a high average quality score.

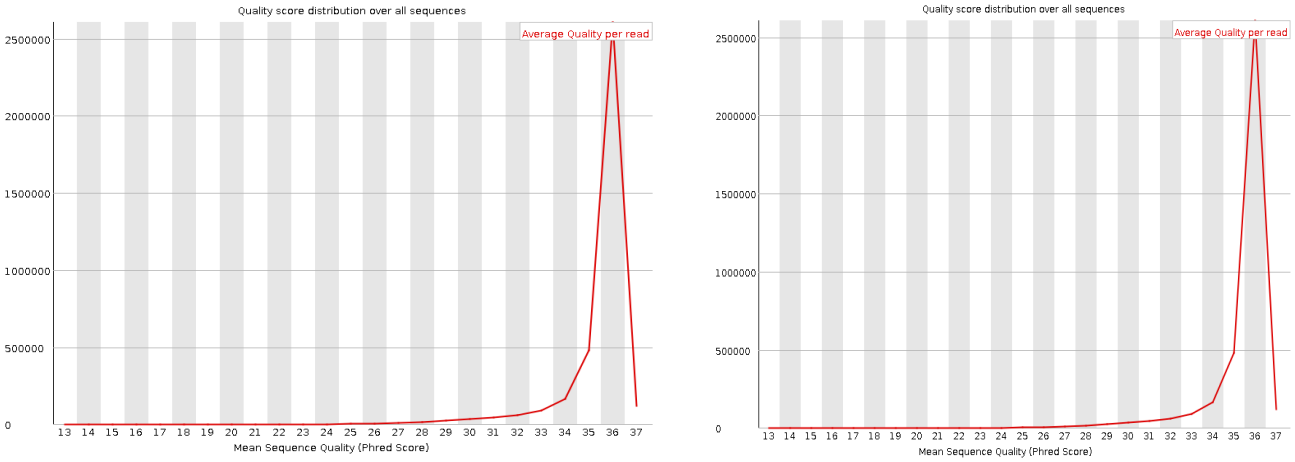


Figure 4.1.3: Per Sequence quality score

4.1.4 Per Base Sequence Content of L1C1

Per Base Sequence Content gives the proportion of each base position in a file for which each of the four normal DNA bases has been called. This graph shows a warning if the difference between A and T, or G and C is greater than 10% in any position and shows a fail if the difference between A and T, or G and C is greater than 20% in any position. The proportion of each of the four bases in our graph are equal which means %A=%T and %G=%C

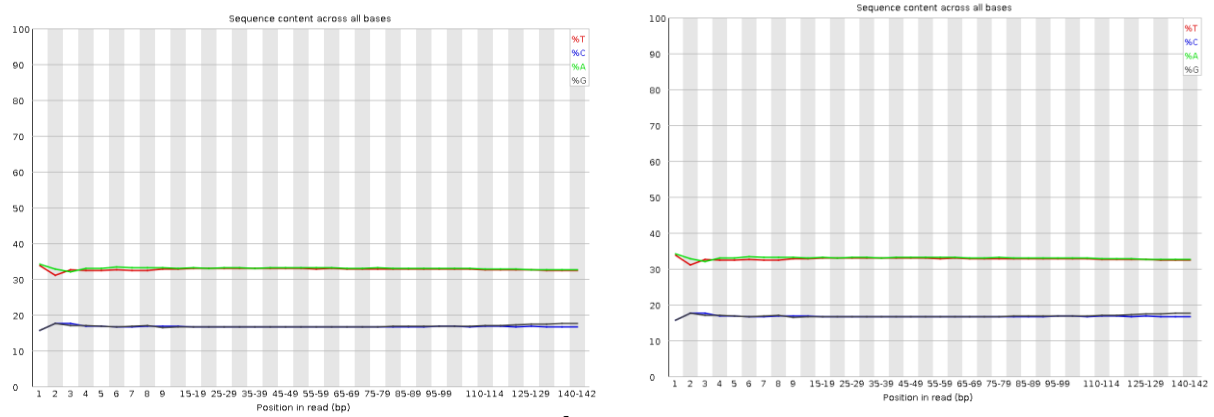


Figure 4.1.4: Per base sequence content

4.1.5 Per Sequence GC Content of L1C1

Per Sequence GC Content measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content. A warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads and this graph will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads**.** Here the mean GC content is 35

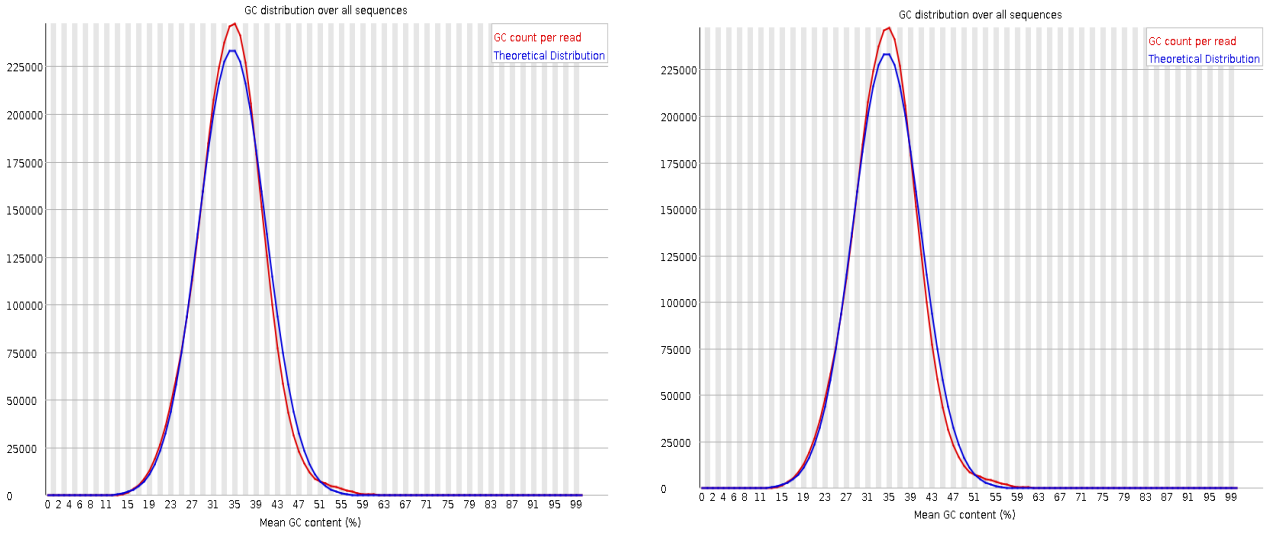


Figure 4.1.5: Per Sequence GC Content

4.1.6 Per Base N Content Of L1C1

Per Base N Content gives out the percentage of base calls at each position for which an N was called. This graph raises a warning if any position shows an N content of >5% and also the graph will raise an error if any position shows an N content of >20%. Here there is no base call i.e., N.

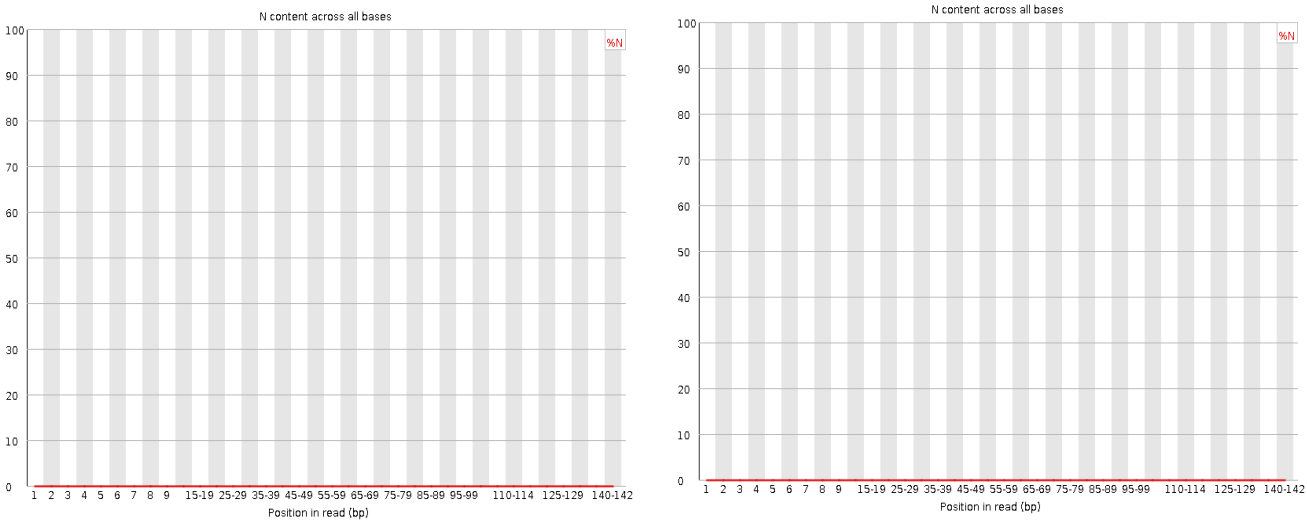


Figure 4.1.6: Per Base N Content

4.1.7 Sequence Length Distribution of L1C1

Sequence Length Distribution generates a graph showing the distribution of fragment sizes of the reads. This graph will raise a warning if all sequences are not the same length and the graph will raise an error if any of the sequences have zero length. Here the sequence length is 142 bp

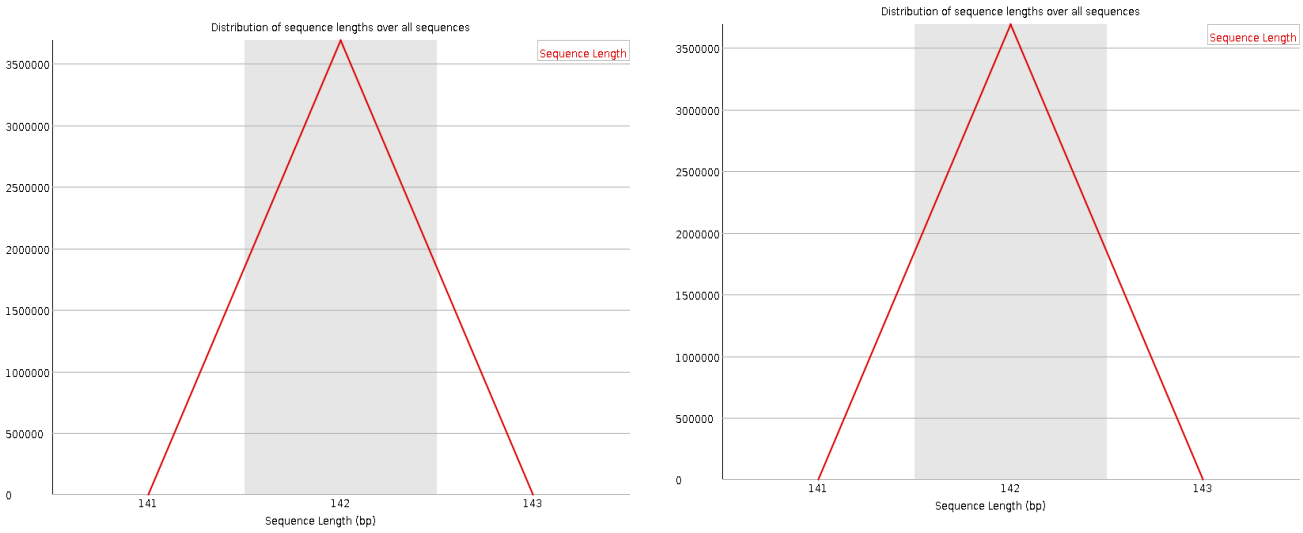


Figure 4.1.7: Sequence Length Distribution

4.1.8 Sequence Duplication Levels of L1C1

Sequence Duplication Levels counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication. This graph will issue a warning if non-unique sequences make up more than 20% of the total and will issue a error if non-unique sequences make up more than 50% of the total. There are two lines on the plot. Here the blue line takes the full sequence set and shows how its duplication levels are distributed. In the red plot the sequences are de-duplicated and the proportions shown are the proportions of the deduplicated set which come from different duplication levels in the original data

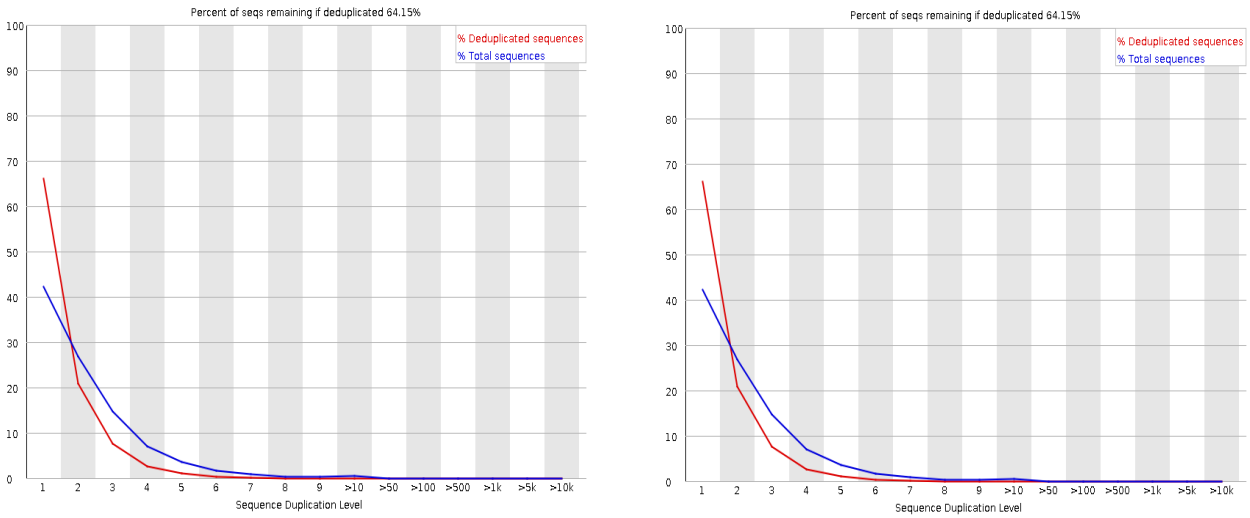


Figure 4.1.8: Sequence Duplication Levels

4.1.9 Adapter Content of L1C1

Adapter content shows a cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. This module will issue a warning if any sequence is present in more than 5% of all reads. This module will issue a warning if any sequence is present in more than 10% of all reads. Common reasons for warnings is that any library where a reasonable proportion of the insert sizes are shorter than the read length will trigger the reads. Here the graph shows a cumulative percentage count of the proportion

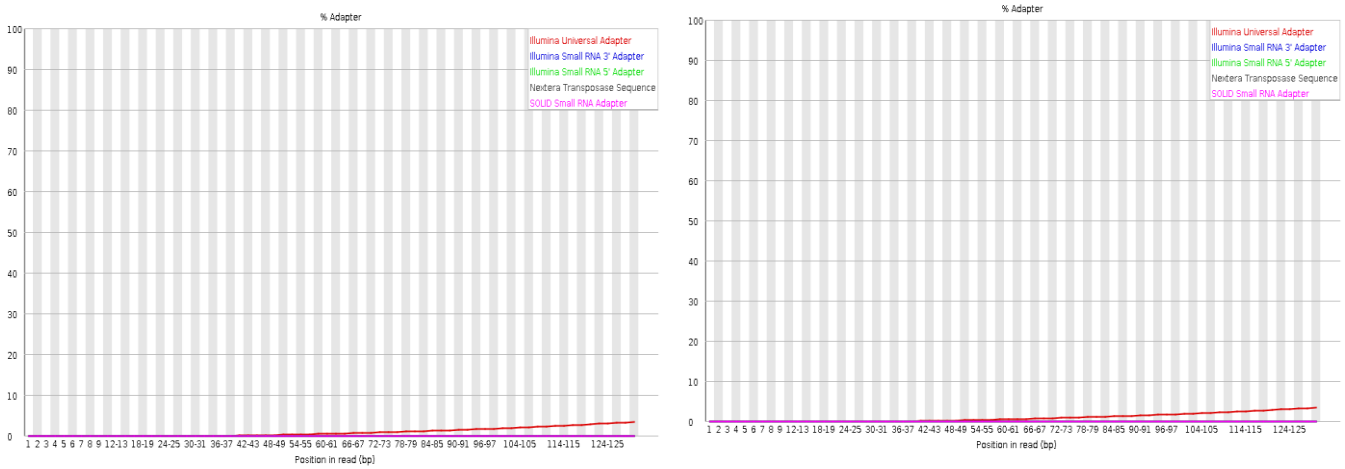


Figure 4.1.9: Adapter Content

These figures shows that L1C1 is a good data for further analysis

Quality assessment of L2C10

4.1.10 Per Base Sequence Quality of L2C10

Per Base Sequence Quality shows an overview of the range of quality values across all bases at each position in the FastQ file. The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25. A failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20. Here show the per base sequence quality in which the quality score is above 20. It is a good quality score.

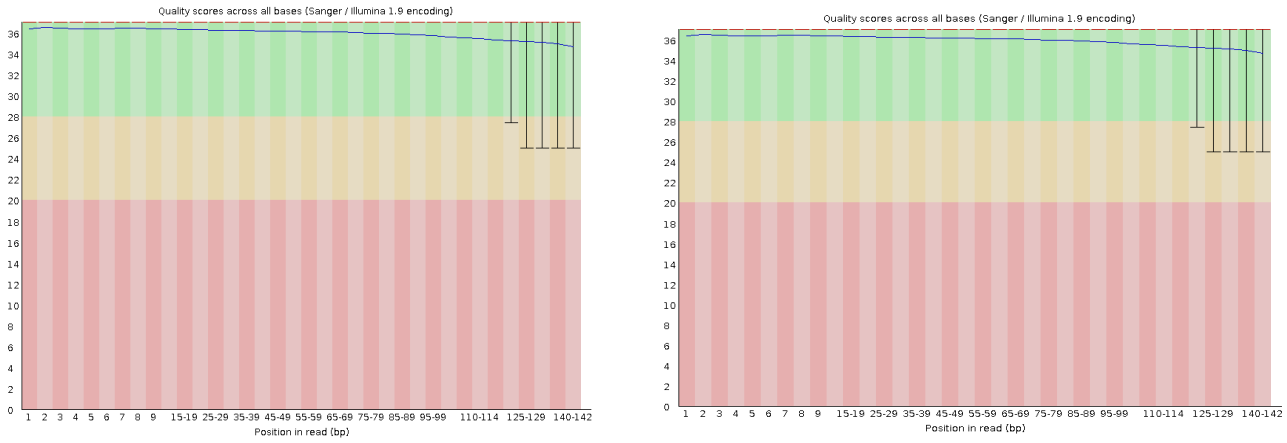


Figure 4.1.19: Per Base Sequence Quality

4.1.11 Per Tile Sequence Quality of L2C10

Per Tile Sequence Quality graph allows to look at the quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell. This graph will issue a warning if any tile shows a mean Phred score more than 2 less than the mean for that base across all tiles and issue a failure if any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles. . A good plot should be blue all over. This plot shows blue all over, so it is a good plot.

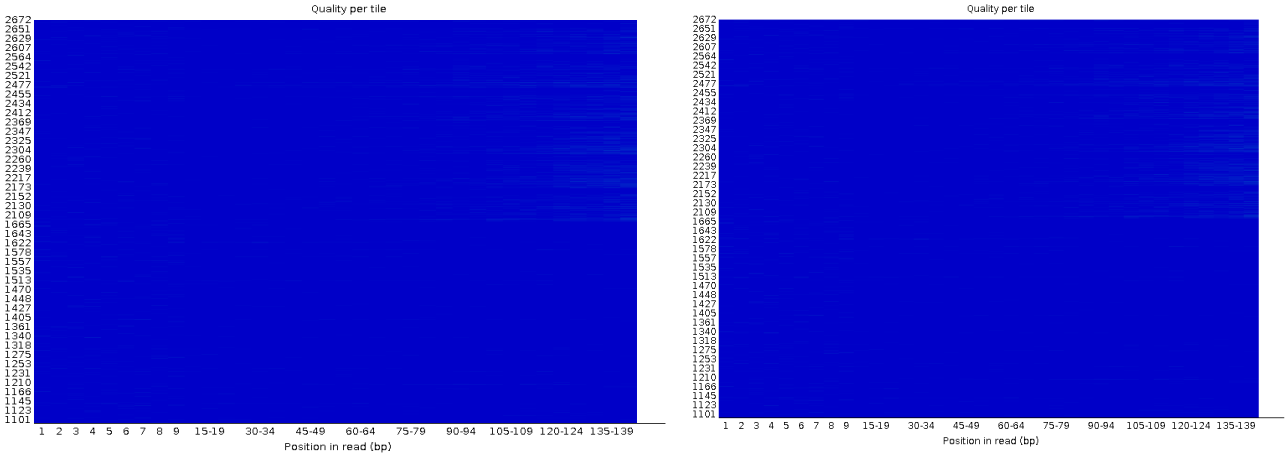


Figure 4.1.20: Per Tile Sequence Quality

4.1.12 Per Sequence Quality Score of L2C10

The per sequence quality score allows to see if a subset of your sequences have universally low-quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged. A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate and an error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate. Here the average quality score on the x-axis and the number of sequences with that average on the y-axis. Majority of our reads have a high average quality score.

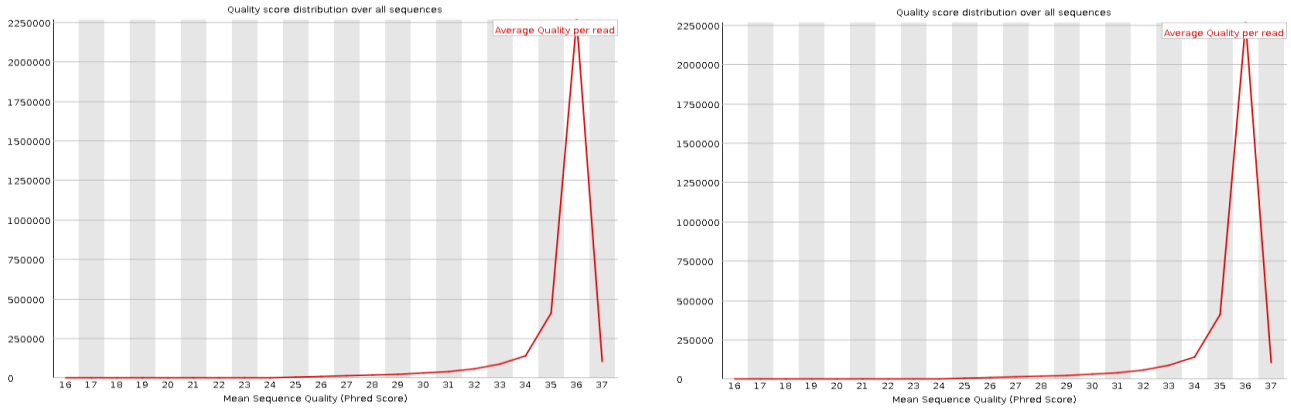
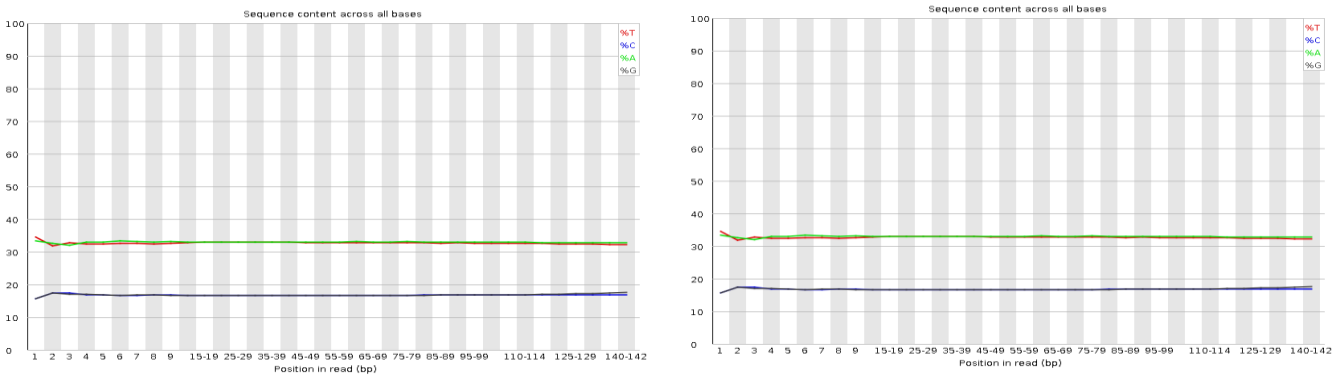


Figure 4.1.21: Per Sequence Quality Score

4.1.13 Per Base Sequence Content of L2C10

Per Base Sequence Content gives the proportion of each base position in a file for which each of the four normal DNA bases has been called. This graph shows a warning if the difference between A and T, or G and C is greater than 10% in any position and shows a fail if the difference between A and T, or G and C is greater than 20% in any position. The proportion of each of the four bases in our graph are equal which means %A=%T and %G=%C



4.1.13 Per Base Sequence Content

4.1.14 Per Sequence GC Content of L2C10

Per Sequence GC Content measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content. A warning is raised

if the sum of the deviations from the normal distribution represents more than 15% of the reads and this graph will indicate a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads. Here the GC content is 35

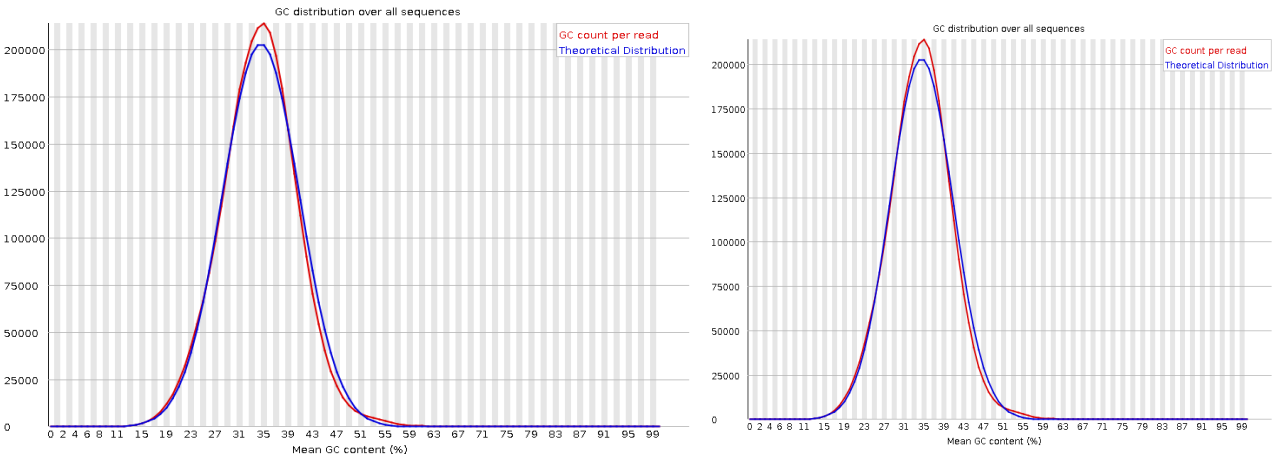


Figure 4.1.14: Per sequence GC content

4.1.15 Per base N content of L2C10

Per Base N Content gives out the percentage of base calls at each position for which an N was called. This graph raises a warning if any position shows an N content of >5% and also the graph will raise an error if any position shows an N content of >20%. Here there is no base call i.e., N

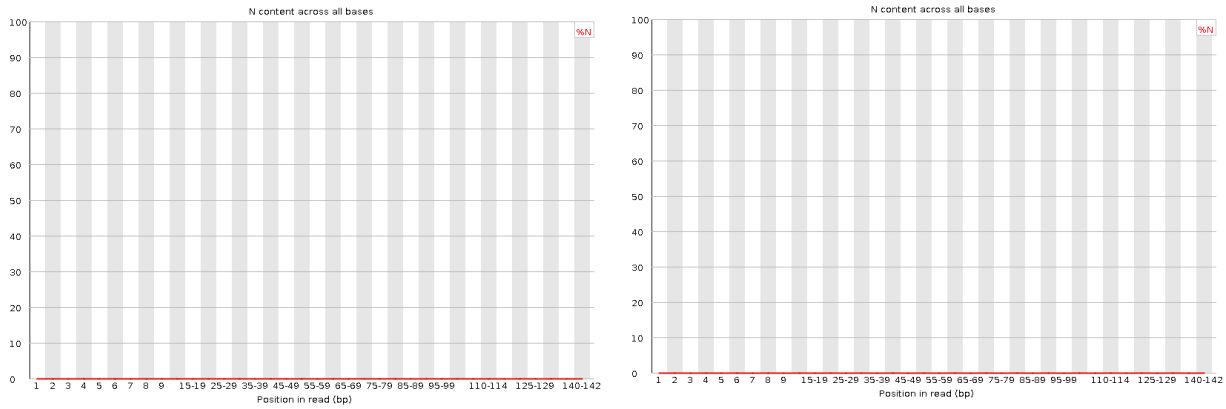


Figure 4.1.15: Per Base N Content

4.1.16 Sequence Length Distribution of L2C10

Sequence Length Distribution generates a graph showing the distribution of fragment sizes of the reads. This graph will raise a warning if all sequences are not the same length and the

graph will raise an error if any of the sequences have zero length. Here the sequence length is 142 bp

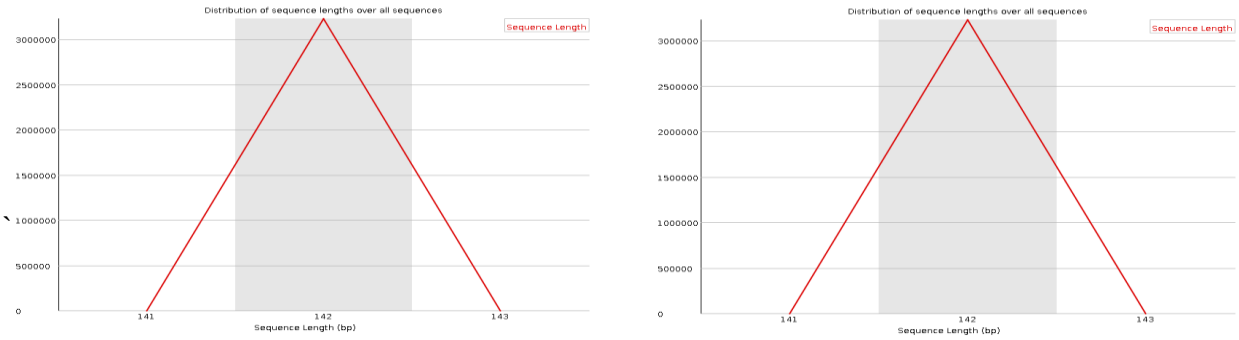


Figure 4.1.16: Sequence Length Distribution.

4.1.17 Sequence Duplication Levels of L2C10

Sequence Duplication Levels counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication. This graph will issue a warning if non-unique sequences make up more than 20% of the total and will issue an error if non-unique sequences make up more than 50% of the total. Here there are two lines on the plot. The blue line takes the full sequence set and shows how its duplication levels are distributed. In the red plot the sequences are de-duplicated and the proportions shown are the proportions of the deduplicated set which come from different duplication levels in the original data.

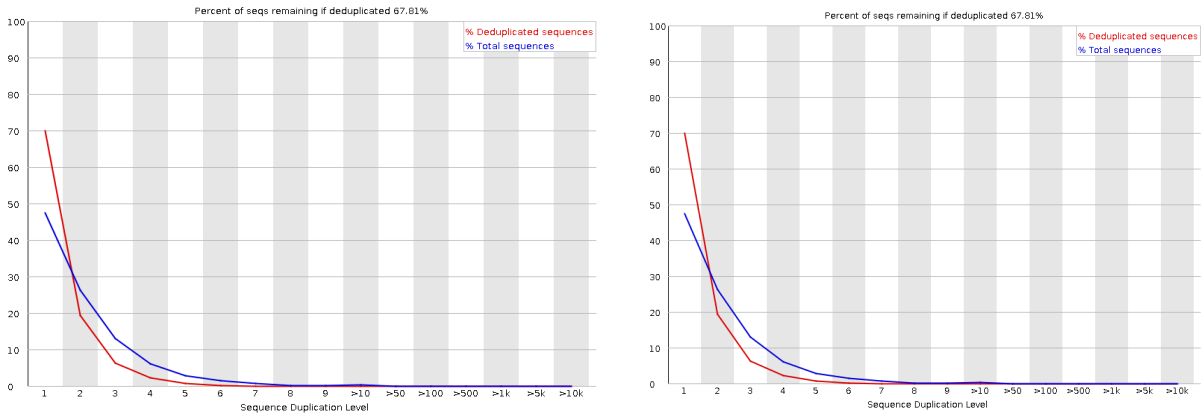


Figure 4.1.26 Sequence Duplication Levels

4.1.18 Adapter Content

Adapter content shows a cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. This module will issue a warning if any sequence is present in more than 5% of all reads. This module will issue a warning if any sequence is present in more than 10% of all reads. Common reasons for warnings is that any library where a reasonable proportion of the insert sizes are shorter than the read length will trigger the reads. The graph shows a cumulative percentage count of the proportion.

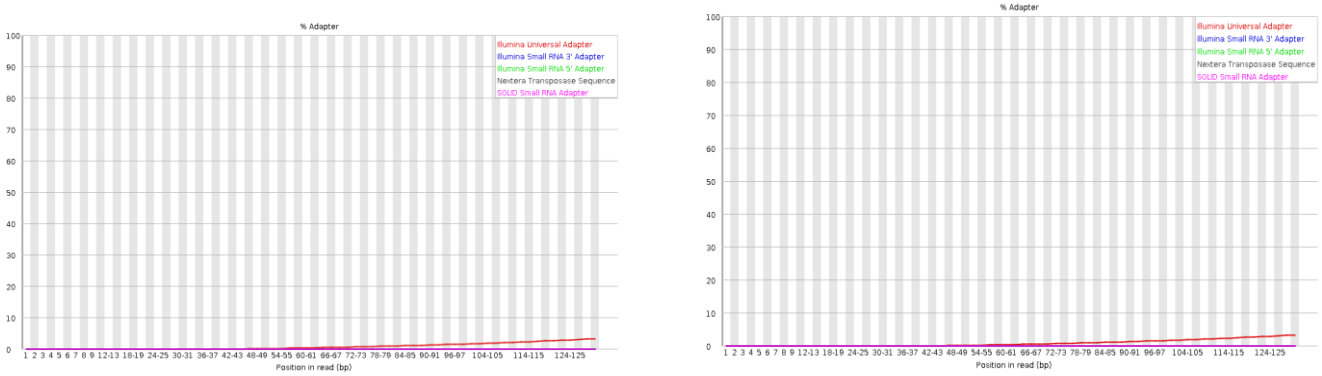


Figure 4.1.18: Adapter Content

These figures shows that L2C10 is a good data for further analysis

4.2 Genome Assembly

4.2.1 Spades Assembly of L1C1

Spades Assembly is an assembly for de novo sequencing. This application is designed to assemble small genomes from MDA single-cell and standard bacterial data sets. Spades was initially designed for single-cell and standard bacterial data sets and is not intended for larger genomes. Spades assembly of L1C1 gives 5722408 Reads ,17824 N50s and 1539 contigs. The largest contig is 195085. This table 4.2.1 shows that Assembly of L1C1.

Assembly L1C1	
Reads	5722408
N50	17824
Largest Contig	195085
Contigs	1539

Table 4.2.1 Spades Assembly of L1C1

4.2.2 Spades Assembly of L2C10

Spades Assembly is an assembly for de novo sequencing. This application is designed to assemble small genomes from MDA single-cell and standard bacterial data sets. Spades was initially designed for single-cell and standard bacterial data sets and is not intended for larger genomes. Spades assembly of L2C10 gives 7842828 Reads, 9871 N50s and 2126 contigs. The largest contig is 70317.

Assembly L2C10	
Reads	7842828
N50	9871
Largest Contig	70317
Contigs	2126

Table 4.2.2 Spades Assembly of L2C10

4.3 Gene Prediction and Annotation

4.3.1 L1C1

Prokka finds and annotates features (both protein coding regions and RNA genes, i.e., tRNA, rRNA) present on a sequence. Prokka uses a two-step process for the annotation of protein coding regions: first, protein coding regions on the genome are identified using Prodigal; second, the function of the encoded protein is predicted. The Sample L1C1 has 1539 contigs,5722408 bases and 1124 genes. Among them 1086 are coding regions,28 are tRNAs,9 are rRNAs and 1 is tmRNA. This result in table 4.3.1 show the

Contigs	1539
Bases	5722408
tRNA	28
CDS	1086
rRNA	9
tmRNA	1
Gene	1124

4.3.1 Prokka Annotation of L1C1

4.3.2 L2C10

Prokka finds and annotates features (both protein coding regions and RNA genes, i.e., tRNA, rRNA) present on a sequence. Prokka uses a two-step process for the annotation of protein coding regions: first, protein coding regions on the genome are identified using Prodigal; second, the function of the encoded protein is predicted. The Sample L2C10 has 2126 contigs,7842828 bases and 1655 genes. Among them 1609 are coding regions,36 are tRNAs,7 are rRNAs and 3 are tmRNAs. This result in table 4.3.2 show the prokka annotation of L2C10.

Contigs	2126
Bases	7842828
tRNA	36
CDS	1609
rRNA	7
tmRNA	3
Gene	1655

4.3.2 Prokka Annotation of L1C1

4.4 Identification of Antibiotic Resistance Genes

L1C1

L1C1 has 27 antibiotic resistance genes present in it. Antibiotic genes have several functions include antibiotic efflux, antibiotic target alteration, antibiotic target protection for the sample L1C1. Genes patA, cdeA, norA, norB, drfE, Saur Lmrs, Saur norA has antibiotics efflux. Genes rpoB2, ImrD, vanR in vanG cl, vanR in vanD cl, vanR in vanG cl has antibiotic target alteration and norC, sepA, sdrM, salD has antibiotic target proteins functions.

Gene	Function
patA	antibiotic efflux
cdeA	antibiotic efflux
norA	antibiotic efflux
norB	antibiotic efflux
rpoB2	antibiotic target alteration & replacement
rpoB2	antibiotic efflux
arlR	antibiotic efflux

arlS	antibiotic target protection
vgaALC	antibiotic target protection
vgaD	antibiotic target replacement
dfrC	antibiotic target replacement
dfrE	antibiotic efflux
lmrD	antibiotic target alteration
vanR_in_vanC_cl	antibiotic target alteration
vanR_in_vanD_cl	antibiotic target alteration
vanR_in_vanG_cl	antibiotic efflux
efrA	antibiotic efflux
Saur_LmrS	antibiotic efflux
Saur_norA	antibiotic efflux
norC	antibiotic efflux
norC	antibiotic target protection
sepA	antibiotic target protection
sdrM	antibiotic target protection
salD	antibiotic target protection

Table 4.4.1: Antibiotic Resistant Genes of L1C1 and its function

L2C10

L1C1 has 23 antibiotic resistance genes present in it. Antibiotic genes have several functions include antibiotic efflux, antibiotic target alteration, antibiotic target protection for the sample L2C10. Genes IsaA, rpoB2, vgaALC, vgaD, salD, salE, has antibiotic target proteins. norA, norB, arlS, arlR, ImrD, emeA, efrA, norC, sepA genes have antibiotic efflux and salD, salE, msr(G), vgaD has antibiotic target protection function. Table 4.4.2 is on the following gives the Antibiotic gene and its function

GENE	FUNCTION
lsaA	antibiotic target protection
norA	antibiotic efflux
norB	antibiotic efflux
rpoB2	antibiotic target alteration &replacement
arlR	antibiotic efflux
arlS	antibiotic efflux
vgaALC	antibiotic target protection
vgaD	antibiotic target protection
dfrC	antibiotic target replacement
dfrE	antibiotic target replacement
lmrD	antibiotic efflux
vanR gene in vanC cluster	antibiotic target alteration
vanR gene in vanD cluster	antibiotic target alteration
vanR gene in vanG cluster	antibiotic target alteration
emeA	antibiotic efflux
efrA	antibiotic efflux
Staphylococcus aureus LmrS	antibiotic efflux
Staphylococcus aureus norA	antibiotic efflux
norC	antibiotic efflux
sepA	antibiotic efflux
salD	antibiotic target protection

salE	antibiotic target protection
msr(G)	antibiotic target protection

Table 4.4.2: Antibiotic Resistant Genes of L1C1 and its function

4.5 Identification of Species

4.5.1 L1C1

As the blast against 16S rRNA database showed 99.367 percentage identity with query coverage of 100 and e value 0.0. From this result, it is confirmed that the organism is Staphylococcus *gallinarum*. It is a gram-positive bacteria.

Name of Species	Staphylococcus *gallinarum*
Query Coverage	100
E Value	0
Bit score	2289
Percentage Identity	99.367

Table 4.5.1 Properties of L1C1

4.5.2 L2C10

As the blast against 16S rRNA database showed 98.589 percentage identity with query coverage of 100 and e value 0.0. From this result, it is confirmed that the organism is Staphylococcus *gallinarum*. It is a gram-positive bacteria.

Name of Species	Staphylococcus *gallinarum*
Query Coverage	100
E Value	0
Bit score	1003
Percentage identity	98.589

Table 4.5.2 Properties of L2C10

5.SUMMARY AND CONCLUSION

The ability of bacteria to fight off antibiotic effects is known as antibiotic resistance. Bacteria can develop antibiotic resistance through genetic mutations or by acquiring antibiotic resistance genes (ARGs). From the study of "Whole genome sequencing data analysis of two unknown multi-drug resistant bacterial isolates" we came to a conclusion that the quality assessment of reads done using FASTQC were acceptable hence the analysis was followed by genome assembly using spade assembler. Spades assembly gives 1539 and 2126 contigs for L1C1 and L2C10 respectively. Using prokka annotation we also predicted 1124 and 1655 genes for L1C1 and L2C10 respectively. L1C1 and L2C10, blasted against the Card database reveals 23 and 27 antibiotic-resistant genes, respectively. According to the 16S rRNA database, Staphylococcus *gallinarum* is the predicted species for both the genomes.

6.REFERENCE

1.Ashkenazi, S. (2013). Beginning and possibly the end of the antibiotic era. *Journal of Paediatrics and Child Health*, *49*(3), 179–182. https://doi.org/10.1111/jpc.12032

2.Baker, M. (2012). De novo genome assembly: What every biologist should know. *Nature Methods*, *9*(4), 333–337. https://doi.org/10.1038/nmeth.1935

3.Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

4.Hughes, D., & Karlén, A. (2014). Discovery and preclinical development of new antibiotics. *Upsala Journal of Medical Sciences*, *119*(2), 162–169. https://doi.org/10.3109/03009734.2014.896437

5. Kitts, P. (2002). Genome assembly and annotation process. McEntyre J, Ostell Jeditors. *The NCBI Handbook.* Bethesda: National Center for Biotechnology Information.

6.Miller, D. R. (2006). A tribute to Sidney Farber - The father of modern chemotherapy. *British Journal of Haematology*, *134*(1), 20–26. https://doi.org/10.1111/j.1365-2141.2006.06119.x

7.Peterson, E., & Kaur, P. (2018). Antibiotic resistance mechanisms in bacteria: Relationships between resistance determinants of antibiotic producers, environmental bacteria, and clinical pathogens. *Frontiers in Microbiology*, *9*(NOV), 1–21. https://doi.org/10.3389/fmicb.2018.02928

8.Podolsky, S. H., Bud, R., Gradmann, C., Hobaek, B., Kirchhelle, C., Mitvedt, T., Santesmases, M. J., Thoms, U., Berild, D., & Kveim Lie, A. (2015). History Teaches Us That Confronting Antibiotic Resistance Requires Stronger Global Collective Action.

Journal of Law, Medicine and Ethics, *43*(S3), 27–32. https://doi.org/10.1111/jlme.12271

9. Roe, M. T., & Pillai, S. D. (2003). Monitoring and identifying antibiotic resistance mechanisms in bacteria. *Poultry Science*, *82*(4), 622–626. https://doi.org/10.1093/ps/82.4.622

10. Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569

11. Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., … McArthur, A. G. (2020). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *48*(D1), D517–D525. https://doi.org/10.1093/nar/gkz935

12. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

13. Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

14. Glazinska, P., Wojciechowski, W., Kulasek, M., Glinkowski, W., Marciniak, K., Klajn, N., Kesy, J., & Kopcewicz, J. (2017). De novo transcriptome profiling of flowers, flower pedicels and pods of lupinus luteus (Yellow lupine) reveals complex expression changes during organ abscission. *Frontiers in Plant Science*, *8*(May). https://doi.org/10.3389/fpls.2017.00641

15. Juayang, A. C., De Los Reyes, G. B., De La Rama, A. J. G., & Gallega, C. T. (2014). Antibiotic resistance profiling of Staphylococcus aureus isolated from clinical specimens in a tertiary hospital from 2010 to 2012. *Interdisciplinary Perspectives on Infectious Diseases*, *2014*. https://doi.org/10.1155/2014/898457

16. Nji, E., Kazibwe, J., Hambridge, T., Joko, C. A., Larbi, A. A., Damptey, L. A. O., Nkansa-Gyamfi, N. A., Stålsby Lundborg, C., & Lien, L. T. Q. (2021). High prevalence of antibiotic resistance in commensal Escherichia coli from healthy human sources in community settings. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-82693-4

17. Phumudzo, T., Ronald, N., Khayalethu, N., & Fhatuwani, M. (2013). Bacterial species identification getting easier. *African Journal of Biotechnology*,

12(41), 5975–5982. https://doi.org/10.5897/ajb2013.12057

18. Pop, M., Phillippy, A., Delcher, A. L., & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in Bioinformatics*, *5*(3), 237–248. https://doi.org/10.1093/bib/5.3.237

19. Preena, P. G., Swaminathan, T. R., Kumar, V. J. R., & Singh, I. S. B. (2020). Antimicrobial resistance in aquaculture: a crisis for concern. *Biologia*, *75*(9), 1497–1517. https://doi.org/10.2478/s11756-020-00456-4

20. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

21. Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107