

*Genome Assembly and Annotation of Lactobacillus crispatus Lc116 isolated from
urine of Post Menopausal women*

Submitted in partial fulfillment of the requirements for the award of
Master of Science in Bioinformatics and Data Science

By

B.Sabarish (42739020)



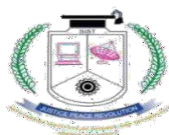
**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade “A++” by NAAC | 12B Status by UGC | Approved by
AICTE**

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119

APRIL – 2024



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A++" grade by NAAC I 12B Status by

UGC I Approved by AICTE Jeppiaar Nagar, Rajiv

Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of B.Sabarish (42739020), who carried out the final year project entitled "*Genome Assembly and Annotation of Lactobacillus crispatus Lc116 strain isolated from urine of postmenopausal women*" under our supervision from January 2024 to April 2024.

Dr. DANIEL ALEX ANAND

Internal Guide

Dr. H. JEMMY CHRISTY

Head of the Department

Submitted for Viva Voce Examination held on 20.04.2024

Internal Examiner

External Examiner

DECLARATION

I, B.Sabarish , hereby declare that the Project Report entitled *Genome Assembly and Annotation of Lactobacillus crispatus Lc116 isolated from the urine of postmenopausal women* done by me under the guidance of Dr. DANIEL ALEX ANAND at Sathyabama Institute of Science and Technology, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai-600119, Tamil Nadu, India, is submitted in partial fulfillment of the requirements for the award of Master of Science in Bioinformatics and Data Science.

DATE: 20.04.2024

PLACE:CHENNAI



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENTS

I am delighted to express my heartfelt acknowledgement to the Board of Management of SATHYABAMA for their kind encouragement and support throughout the duration of my project. Their guidance has been invaluable in ensuring the successful completion of this project, and I am sincerely grateful for their contribution.

I sincerely thank Dr. H. Jemmy Christy, Head of the Department, Dept. of Bioinformatics, for his unwavering support and timely provision of necessary details during the progressive reviews. His support has been instrumental in helping me to overcome obstacles and achieve success in this project.

I am deeply indebted to my Project Guide, Dr. Daniel Alex Anand, for his invaluable guidance, suggestions, and constant encouragement throughout the project. His expertise and insights have been instrumental in shaping my work and ensuring its successful completion.

I also wish to extend my gratitude to my family for their support and assistance in various ways, which have contributed to the successful completion of this project.

B.Sabarish



SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with “A++” grade by NAAC I 12B Status by

UGC I Approved by AICTE Jeppiaar Nagar, Rajiv

Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

CERTIFICATE

This is to certify that the dissertation entitled “*Genome Assembly and Annotation of Lactobacillus crispatus Lc116 isolated from the urine of post menopausal women*” was carried out by Mr. B.Sabarish (Reg. No. 42739020), Department of Bioinformatics and Data Science, Sathyabama Institute of Science and Technology, Chennai in partial fulfillment of the requirements of the award of the Degree of Master of Science in Bioinformatics and Data Science 2022-2024, is a bonafide record of the project work done under my guidance and supervision.

This thesis or part, therefore, has not been submitted elsewhere for any other degree or diploma.

Date:

Place:

Dr. DANIEL ALEX ANAND

ABSTRACT:

The human urinary microbiota performs a crucial role in maintaining urogenital health, compared to other human microbial niches our knowledge of urinary microbiota especially in females is in its infancy. It is a resident for microbial community of many adult women. *Lactobacillus crispatus* is a predominant species in the vaginal microbiota, but its presence and significance in urinary tract infections are not well understood. This project aimed to investigate the genomic characteristics of *Lactobacillus crispatus* strain Lc116 isolated from the urine of post-menopausal women through genome assembly and annotation. Genomic insights gained from this study serves as a foundational resource that offers opportunities for various research studies such as comparative genomics of different strains of *Lactobacillus crispatus*, host-microbe interactions, adaptations, therapeutic applications in the urinary tract. Furthermore, the integration of *Lactobacillus crispatus* genomic data with metagenomic studies of urinary microbiome holds promise for elucidating microbial community dynamics and their impact on urogenital health.

Keywords : Genome Assembly, Genome Annotation, Post-menopause, Urinary Microbiota

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	LIST OF FIGURES	iv
	LIST OF TABLES	v
1	INTRODUCTION	1
1.1	HUMAN MICROBIOME	1
1.2	URINARY TRACT INFECTIONS	1
1.3	LACTOBACILLUS CRISPATUS	2
1.4	GENOME ASSEMBLY	3
1.5	GENOME ANNOTATION	4
2	REVIEW OF LITERATURE	5
3	AIM AND SCOPE	8
3.1	AIM	8
3.2	SCOPE	8
4	METHODOLOGY	9
4.1	DATA ACQUISITION PRE-PROCESSING OF	9
4.2	RAW DATA	9

	4.3	DE-NOVO HYBRID ASSEMBLY	10
	4.4	QUALITATIVE ANALYSIS OF ASSEMBLY	10
	4.5	GENOME ANNOTATION	10
5		RESULT AND DISCUSSION	11
		QUALITY CONTROL	
	5.1	CHECK OF ILLUMINA READS	11
	5.1.1	BASIC STATISTICS	11
	5.1.2	PER BASE SEQUENCE QUALITY	12
	5.1.3	PER SEQUENCE QUALITY SCORES	13
	5.1.4	PER BASE SEQUENCE CONTENT	14
	5.1.5	PER SEQUENCE GC CONTENT	15
	5.1.6	PER BASE N CONTENT	16
	5.1.7	SEQUENCE LENGTH DISTRIBUTION	17
	5.1.8	OVERREPRESENTED SEQUENCES	18
	5.1.9	ADAPTER CONTENT	19
	5.1.10	BASIC STATISTICS	20
	5.1.11	PER BASE SEQUENCE QUALITY	21
		PER SEQUENCE	
	5.1.12	QUALITY SCORES	22

5.1.13	PER BASE SEQUENCE CONTENT	23
5.1.14	PER BASE GC CONTENT	24
5.1.15	PER BASE N CONTENT	25
5.1.16	SEQUENCE LENGTH DISTRIBUTION	26
5.1.17	OVERREPRESENTED SEQUENCES	27
5.1.18	ADAPTER CONTENT	28
	QUALITY CONTROL OF	29
5.2	ONT READS	
5.2.1	GENERAL SUMMARY	29
	Number, percentage and	
5.2.2	megabases of reads	29
	above quality cutoffs	
	Top 5 highest mean	
5.2.3	basecall quality scores	30
	and their read lengths	
	Top 5 longest reads and	
5.2.4	their mean basecall	30
	quality score	
5.2.5	GENERAL SUMMARY	31
	Number, percentage and	
5.2.6	megabases of reads	31
	above quality cutoffs	
	Top 5 highest mean	
5.2.7	basecall quality scores	32
	and their read lengths	
	Top 5 longest reads and	
5.5.8	their mean basecall	32
	quality score	

	5.3 ASSEMBLY	33
	5.3.1 ASSEMBLY STATISTICS	33
	5.3.2 GC CONTENT	35
	5.3.4 CUMULATIVE LENGTH	36
	5.3.5 ASSEMBLY GRAPH	37
	5.3.6 SCAFFOLDS	39
	5.3.7 LONG READ ASSEMBLY	40
	5.3.8 MERGED FINAL ASSEMBLY GRAPH	41
	5.4 GENOME ANNOTATION	43
	SUMMARY AND CONCLUSION	45
6	6.1 SUMMARY	45
	6.2 CONCLUSION	46
	REFERENCES	47

LIST OF FIGURES

FIGURE NO	DESCRIPTION	PAGE NO
5.1	PER BASE SEQUENCE QUALITY	12
5.2	PER SEQUENCE QUALITY SCORES	13
5.3	PER BASE SEQUENCE CONTENT	14
5.4	PER SEQUENCE GC CONTENT	15
5.5	PER BASE N CONTENT	16
5.6	SEQUENCE LENGTH DISTRIBUTION	17
5.7	OVERREPRESENTED SEQUENCES	18
5.8	ADAPTER CONTENT	19
5.9	BASIC STATISTICS	20
5.10	PER BASE SEQUENCE QUALITY	21
5.1.11	PER SEQUENCE QUALITY SCORES	22
5.1.12	PER BASE SEQUENCE CONTENT	23
5.1.13	PER SEQUENCE GC CONTENT	24
5.1.14	PER BASE N CONTENT	25
5.1.15	SEQUENCE LENGTH DISTRIBUTION	26
5.1.16	OVERREPRESENTED SEQUENCES	27
5.1.17	OVERREPRESENTED SEQUENCES	28
5.1.18	ADAPTER CONTENT	29
5.1.19	GC CONTENT	35
5.1.20	CUMULATIVE LENGTH	36
5.1.21	ASSEMBLY GRAPH	37
5.1.22	SCAFFOLDS	39
5.1.23	LONG READ ASSEMBLY GRAPH	40
5.1.24	MERGED FINAL ASSEMBLY GRAPH	41
5.1.25	FEATURE TYPE DISTRIBUTION	43

LIST OF TABLES

TABLE NO	DESCRIPTION	PAGE NO
5.2.1	GENERAL SUMMARY	29
5.2.2	NUMBER, PERCENTAGE AND MEGABASES OF READS ABOVE QUALITY CUTOFFS	29
5.2.3	TOP 5 HIGHEST MEAN BASECALL QUALITY SCORES AND THEIR READ LENGTHS	30
5.2.4	TOP 5 LONGEST READS AND THEIR MEAN BASECALL QUALITY SCORE	30
5.2.5	GENERAL SUMMARY	31
5.2.6	NUMBER, PERCENTAGE AND MEGABASES OF READS ABOVE QUALITY CUTOFFS	31
5.2.7	TOP 5 HIGHEST MEAN BASECALL QUALITY SCORES AND THEIR READ LENGTHS	32
5.2.8	TOP 5 LONGEST READS AND THEIR MEAN BASECALL QUALITY SCORE	32
5.3.1	ASSEMBLY STATISTICS	33

CHAPTER 1

INTRODUCTION:

1.1 HUMAN MICROBIOME

The human microbiome, is the aggregate of microbiota that reside on or within the human body, plays a pivotal role in maintaining human health and disease. Humans have co-evolved with the trillions of microbes that inhabit our bodies and that create complex, body-habit-specific, adaptive ecosystems that are finely aligned to relentlessly changing host physiology. The microbial cells that colonize the human body are as abundant as our somatic cells and certainly contain far more genes than our human genome. An estimated 500 – 1000 species of bacteria exist in the human body at any one time, although the number of subspecies could be orders of magnitude greater than this. Each bacterial strain has a genome containing thousands of genes, offering substantially more genetic diversity, and hence more flexibility than the human genome. Among the various microbial niches, the urinary tract harbors a diverse array of microorganisms collectively known as the urinary microbiota. While historically considered sterile, emerging evidence has revealed the presence of a complex microbial community in the urinary tract, particularly in females, where it is increasingly recognized as an integral component of urogenital health. These discoveries may provide new opportunities to address important women's health conditions, which include preterm labor and delivery, urinary tract infections and common forms of urinary incontinence.

1.2 URINARY TRACT INFECTIONS (UTI)

Urinary tract infections are the most common bacterial infections seen in the community, especially amongst females. Traditional diagnostic methods for urinary assessment rely on urine culture techniques, which often fail to detect the presence of slow-growing bacteria and the presence of fastidious microorganisms.

Consequently, a significant proportion of individuals with UTI symptoms may receive negative urine culture results, leading to underdiagnosis and suboptimal management of urinary tract-related conditions. In recent years, advances in high-throughput sequencing technologies have revolutionized our understanding of the urinary microbiota, enabling researchers to comprehensively profile microbial communities present in the urinary tract. Among the key microbial constituents of the urinary microbiota, *Lactobacillus* species, notably *Lactobacillus crispatus*, have garnered significant attention due to their predominant role in maintaining vaginal health but their role in preventing urogenital infections in women is not well understood.

1.3 LACTOBACILLUS CRISPATUS

Lactobacillus crispatus is a Gram-positive, rod-shaped bacterium belonging to the genus *Lactobacillus* and a beneficial microbiota located in both the vagina, through vaginal discharge and the urinary tract. The high prevalence of *Lactobacillus crispatus* is commonly associated with maintaining a healthy vaginal environment and in protecting this environment from several urogenital diseases such as bacterial vaginosis and viral infections. While *Lactobacillus crispatus* is extensively studied in the context of vaginal health, the presence and significance of *Lactobacillus crispatus* in maintaining the urinary tract health is not relatively well understood. Given its ability to inhibit growth of bacterial vaginosis associated bacteria and due to the production of various anti-microbial compounds, such as hydrogen peroxide, lactic acid and bacteriocin-like products, *Lactobacillus crispatus* holds potential implications for urinary tract health and the prevention of UTIs.

1.4 GENOME ASSEMBLY

Genome assembly is the computational process of reconstructing a genome from sequencing reads obtained through various high throughput sequencing technologies. This process is critical for elucidating the genetic blueprint of organisms, facilitating downstream analysis such as gene prediction, comparative genomics, functional annotation and many more. With advancements in sequencing technologies, including short read sequencing and long read sequencing, genome assembly has evolved to accommodate diverse data types, each with its unique strengths and challenges. Short read sequencing generates millions of short DNA fragments, enabling high throughput sequencing at a low cost but often resulting in fragmented assemblies, particularly in repetitive regions of the genome. Long read sequencing on the other hand, produces much longer reads, allowing for more contiguous assemblies that span complex genomic regions. Hybrid assembly approach which combines both short and long reads, leverage the complementary strength of each technology to produce high quality and perfect genome assembly. A perfect bacterial genome assembly is one where the assembled sequence is an exact match for the organisms genome-each replicon sequence is complete and contains no errors. There are two broad goals to consider with genome assembly: accuracy and completeness. Accuracy refers to the number of errors present in the assembled sequence. Such errors can be small in scale (e.g., an incorrect base) or larger in scale (e.g., the addition, removal, or inversion of hundreds of bases). Completeness refers to the length of the contigs relative to the corresponding genomics sequence, i.e., how fragmented the assembly is. Longer contigs are better, ideally each contig representing an entire replicon in the genome. Perfect assembly is defined as one with maximal completeness and with 100% accuracy. Genome assembly algorithms play a crucial role in stitching together sequencing reads into contiguous sequences such as contigs, scaffolds and ultimately chromosome-level assemblies. Assembling a genome involves several computational challenges, including run time and space occupied, necessitating the use of sophisticated bioinformatics tools and computational resources. Genome assembly serves as the foundation for downstream analysis, providing insights into genome structure, function and underpins various applications in fields such as genetics, molecular biology and medicine.

1.5 GENOME ANNOTATION

Genome annotation, a crucial step in deciphering the genetic content of an organism, involves identifying and characterizing the functional elements encoded within its DNA sequence. This process plays a pivotal role in elucidating gene structure, function and regulation, providing insights into the biological process underlying organismal phenotypes and traits. Genome annotation encompasses the prediction and annotation of diverse genomic features, including protein-coding genes, non-coding RNA genes (e.g., tRNA, rRNA ..), and other functional elements. Leveraging computational algorithms and bioinformatics tools, genome annotation integrates evidence from various sources, including sequence similarity, gene expression data and functional annotations from related organisms, to accurately predict gene models and annotate their biological roles. High throughput sequencing technologies have revolutionized genome annotation by providing vast amounts of sequencing data for comprehensive analysis. Genome annotation serves as the cornerstone for understanding the genetic basis of complex traits, evolutionary processes and disease mechanisms.

CHAPTER 2

REVIEW OF LITERATURE

Grin, Peter M., et al. investigates the efficacy of Lactobacillus probiotics in preventing recurrent Urinary Tract Infections(UTIs) in women. UTIs are common health issue often caused by uropathogenic bacteria such as Escherichia Coli. Lactobacillus are known for their potential to modulate the urogenital microbiota and maintain a healthy vaginal environment, which may confer protection against UTIs. This study synthesizes data from multiple clinical trials to assess the overall effect of Lactobacillus probiotic supplementation on UTI recurrence rates. The findings suggest that Lactobacillus probiotics have a beneficial effect in reducing the risk of recurrent UTIs in women.

Zhou, Xia, et al is studied to explore the vaginal microbiota of healthy adult women using cultivation independent methods. It aims to elucidate the composition and diversity of vaginal microbial communities beyond what traditional culture based techniques can provide. By employing 16sRNA gene sequencing and denaturing gradient gel electrophoresis(DGGE) approaches, researchers analyze the microbial profiles of vaginal samples collected from a cohort of healthy women. The findings reveal considerable interpersonal variation in the vaginal microbiota composition, with lactobacillus species predominating in many individuals but other bacterial taxa are also present. This resource highlights the importance of bacteria in shaping the vaginal microbiome and emphasizes the need for comprehensive, culture independent methods to fully understand microbial community dynamics in this niche environment.

Amabebe, Emmanuel, and Dilly OC Anumba provides an in-depth explanation of the physiological role of lactobacilli in the vaginal environment. Lactobacillus plays a crucial role in maintaining vaginal health by contributing to the acidic pH, providing antimicrobial compounds and modulating the immune response. This resource reviews the mechanisms by which lactobacilli promote a healthy vaginal environment, including the production of lactic acid, hydrogen peroxide and bacteriocins which inhibit the growth of pathogens. Additionally, it discusses the interactions between lactobacilli and other microorganisms in the vaginal ecosystem, highlighting the importance of microbial balance in preventing urogenital infections such as bacterial vaginosis and other urinary tract infections.

Eklblom, Robert, and Jochen BW Wolf. used to review the comprehensive resource for researchers navigating the complexities of whole genome sequencing projects. It provides a detailed overview of the various steps involved in WGS, including sample preparation, sequencing technologies, genome assembly and genome annotation. It discusses the advantages and limitations of different sequencing platforms and assembly algorithms offering practical recommendations for choosing the most suitable approach based on the research objectives and available resources.

Miller, Jason R., Sergey Koren, and Granger Sutton. presents an overview of the various algorithms employed in the assembly of next generation sequencing data. It discussed the fundamental challenges associated with the genome assembly, including the presence of repetitive sequences, sequencing errors and variations in coverage depth. The key algorithms on which de-novo genome assembly softwares work is categorized into two classes : Overlap Layout Consensus and De-Bruijn Graphs. Each class of algorithms is discussed in detail along with its principles, advantages and limitations. This is a valuable resource for researchers and bioinformaticians, offering comprehensive insights into the diverse array of algorithms employed in de-novo genome assembly

Wick, Ryan R., et al introduces Unicycler, a novel tool designed to resolve bacterial genome assemblies using both short and long sequencing reads. The need for such a tool arises from the limitations of traditional assembly methods, which struggle to handle the complexity of bacterial genomes, particularly those with repetitive elements or structural variations. Unicycler addresses these challenges by integrating multiple assembly strategies. By combining the short and long reads, unicycler achieves more accurate and contiguous genome assemblies, enabling us to better characterize bacterial genomes and study their genetic diversity and evolution. This resource demonstrates the effectiveness of unicycler through comprehensive evaluations on a diverse set of bacterial genomes, highlighting its ability to produce high-quality assemblies across a wide range of species and sequencing technologies.

CHAPTER 3

AIM AND SCOPE

3.1 AIM :

The present study aims to investigate the genomic characteristics of *Lactobacillus crispatus* strain Lc116, a strain isolated from the urine of post-menopausal women, through genome assembly and annotation. By elucidating the genomic repertoire of *Lactobacillus crispatus* strain Lc116, this research endeavors to provide insights into its genetic makeup and potential functional elements present in the genome. By utilizing advanced sequencing methods and bioinformatics software's, the goal is to produce a comprehensive annotation of the genome that can be utilized to serve as a foundational resource for future research endeavors, including comparative and functional genomics, systems biology investigations which may further elucidate adaptive strategies, metabolic capabilities, and potential interactions of *Lactobacillus crispatus* in maintaining the urinary tract environment.

3.2 SCOPE :

This study involves an extensive investigation into the genetic structure and functional capabilities of *Lactobacillus crispatus* Lc116, an organism of clinical significance in urogenital health. The scope of this study spans multiple facets, including data acquisition from cutting -edge technologies such as illumina paired-end sequencing and oxford nanopore long read sequencing. We employ a hybrid assembly approach to reconstruct the complete genome of *Lactobacillus crispatus* Lc116, integrating short read and long read sequencing data to achieve high contiguity and accuracy.

CHAPTER 4

METHODOLOGY

4.1 DATA ACQUISITION

Raw sequencing data for Lactobacillus strain Lc116 were retrieved from the Sequence Read Archive database using SRA tool kit and the fastq files were extracted from the dataset coming from distinct technologies: Illumina NextSeq 500 and Oxford Minlon platform respectively. The Illumina NextSeq 500 was used to generate 2 * 150 base pairs paired end reads. Long reads were obtained by sequencing on Minlon platform. A total of 1.08 Gb of raw data was generated on the Oxford Nanopore and Illumina platforms.

4.2 PRE-PROCESSING OF RAW DATA

The illumina paired end raw reads were checked for quality. FastQC v 0.12.1, a widely used quality assessment tool, was employed to analyze the raw illumina paired end reads. This comprehensive analysis evaluated various parameters such as basic statistics, Per base sequence quality, Per sequence quality scores, Per sequence GC content, Per base N content, Sequence length distribution, Overrepresented sequences and presence of Adapter content.

Subsequently, the fastp tool was utilized to preprocess the illumina reads. Fastp v 0.23.0 is an ultra-fast FASTQ preprocessor with useful quality and data filtering features including adapter trimming, quality filtering, removal of overrepresented sequences and many other operations with single scan of the FASTQ data.

The Nanopore raw reads were checked for quality and statistical properties of the raw Nanopore reads. NanoStats v 1.6.0 , a tool specifically designed for analyzing Oxford Nanopore sequencing data was employed. This analysis provided insights into the base Quality scores, Read length distribution and other relevant metrics.

NanoFilt v 2.8.0 , a versatile tools was employed to the Nanopore reads to remove low-quality reads and read length less than 200. This step aimed to enhance the overall quality of the Nanopore dataset by retaining reads with satisfactory quality score and read lengths.

4.3 DE-NOVO HYBRID ASSEMBLY

The *Lactobacillus crispatus* Lc116 genome was assembled using Unicycler v 0.5.0. Unicycler , a tool for assembling bacterial genomes from a combination of short and long reads, which produces assemblies that are accurate,complete and of high quality. Unicycler's bold mode was used to generate closed assembly.

4.4 QUALITATIVE ANALYSIS OF ASSEMBLY

Following genome assembly,the completeness and contiguity of the assembled genome were assessed using complementary tools. Bandage v0.8.1 , a graphical tool for visualizing genome assemblies from the graphs generated during the assembly by Unicycler, facilitated the exploration of assembly graphs to identify misassemblies and repetitive regions.

Qualitative Assessment tool was used for evaluating and comparing genome assemblies by computing various metrics.Assembly quality was evaluated without a reference genome.Metrics evaluated by QUAST are Contig sizes, Missassemblies and structural variations , Genome representation and its functional elements, Variations of N50 based on aligned blocks and Visualization.

4.5 GENOME ANNOTATION

The assembled genome was annotated using PROKKA, a rapid and automated annotation of prokaryotic genome,it integrates multiple bioinformatics tools and databases to determine the prediction of genes and its corresponding protein products, identification of functional elements such as mRNA,rRNA,tRNA and CDS based on homology searches and domain analysis.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 QUALITY CONTROL CHECK OF ILLUMINA READS

READ 1:



Basic Statistics

Measure	Value
Filename	SRR16002670_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2808394
Total Bases	417.4 Mbp
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	37

BASIC STATISTICS :

In the Fastqc analysis for read1(file name : SRR16002670_1.fastq) , a total of 28,08,394 sequences were processed, collectively constituting approximately 417.4 million base pairs. The sequence length of the reads within this file spans from 35-151 base pairs. The percentage of GC content in the sequences was determined to be 37%.Notably, there were no sequences flagged as poor quality.

PER BASE SEQUENCE QUALITY

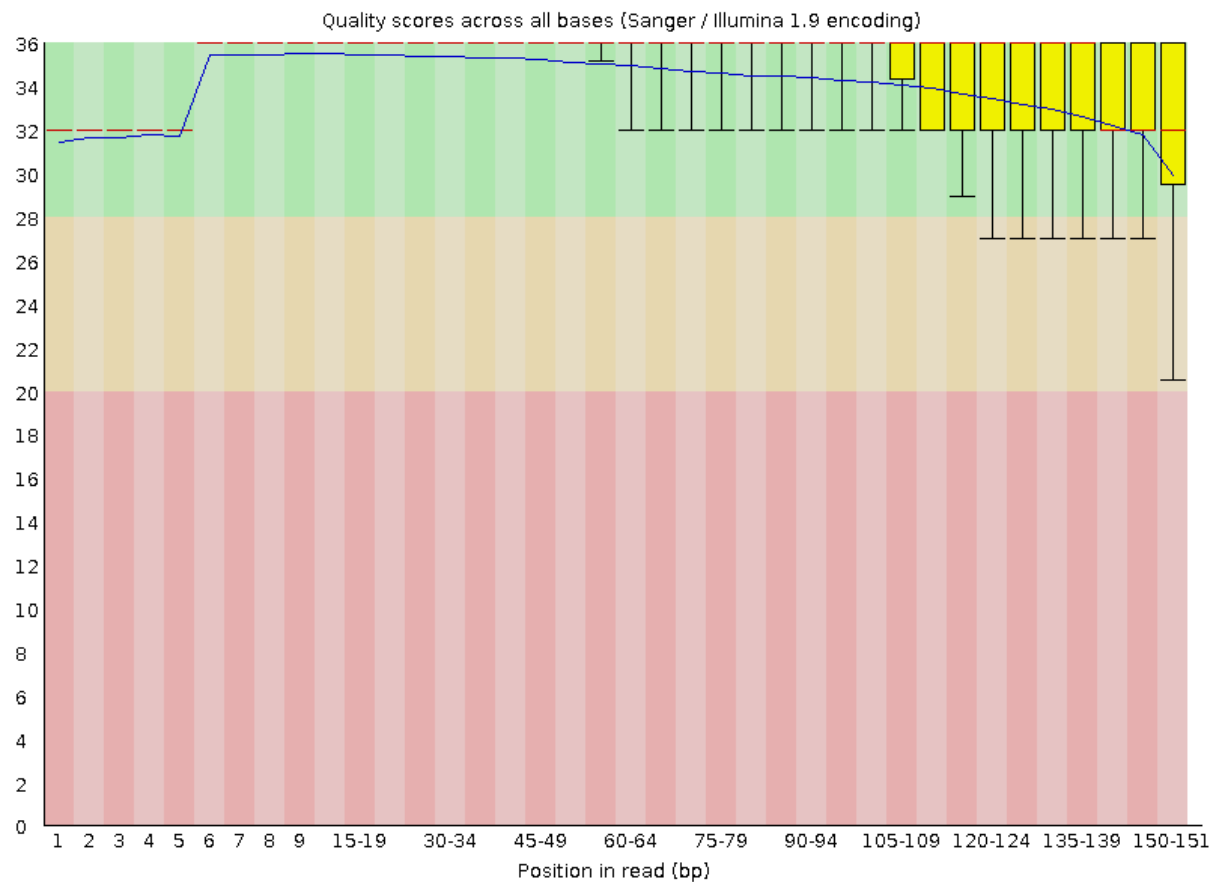


FIGURE 5.1 PER BASE SEQUENCE QUALITY SCORES

The per base sequence quality analysis have provided insights into the quality score distribution of bases in the reads across all the reads. In the case of read1 data the average quality score remains above 32 for the majority of the read, indicating high quality base calls. However, towards the end of the read, particularly beyond position 100, the presence of yellow boxes with whiskers indicated that there is a variability in the quality scores across different positions in the read.

PER SEQUENCE QUALITY SCORES

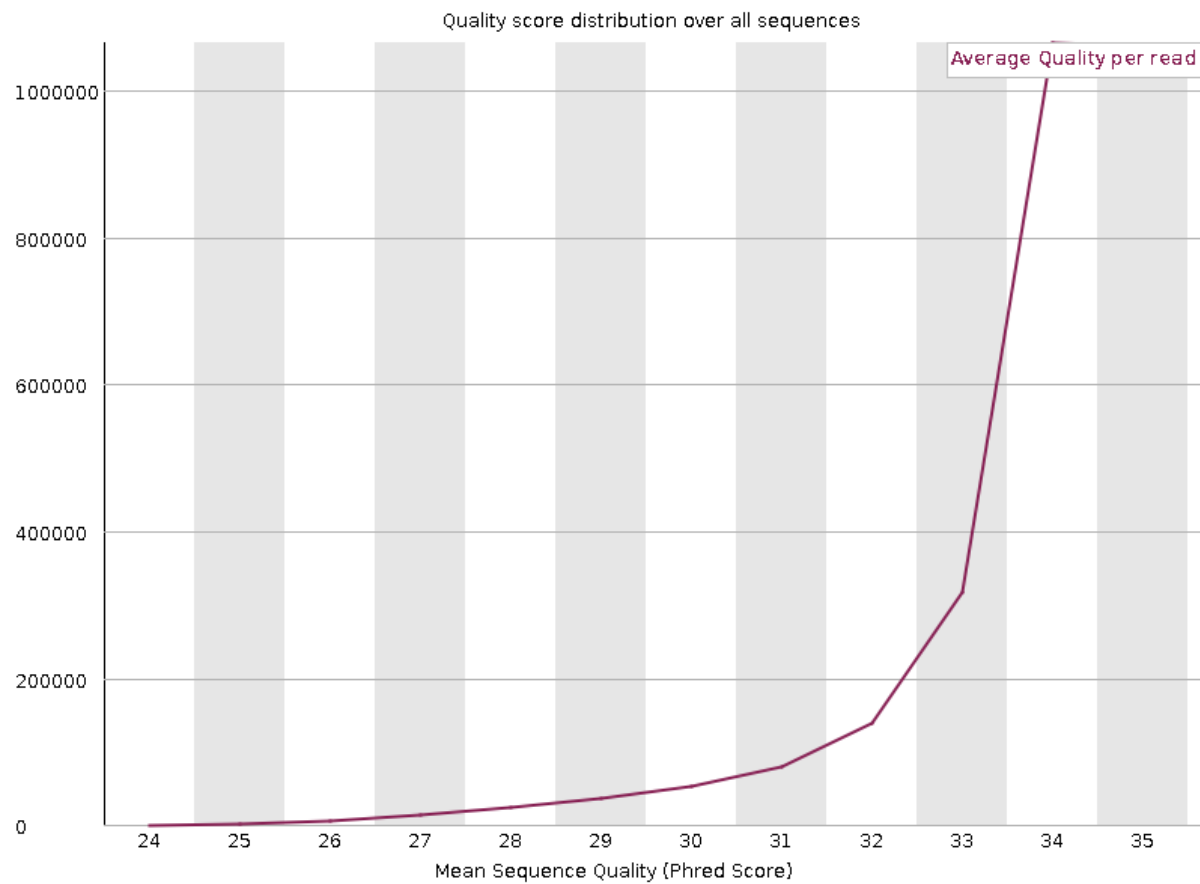


FIGURE 5.2 PER SEQUENCE QUALITY SCORES

The per sequence quality score analysis provides a comprehensive overview of the quality score distribution over all sequences in the dataset. In the case of read1, the mean sequence quality score exhibits a notable peak at 34, indicating a predominance of high-quality base calls across the majority of the reads. Notably, some reads exhibit quality scores 30, extending down to 26 particularly observed among reads with counts below 200000.

PER BASE SEQUENCE CONTENT

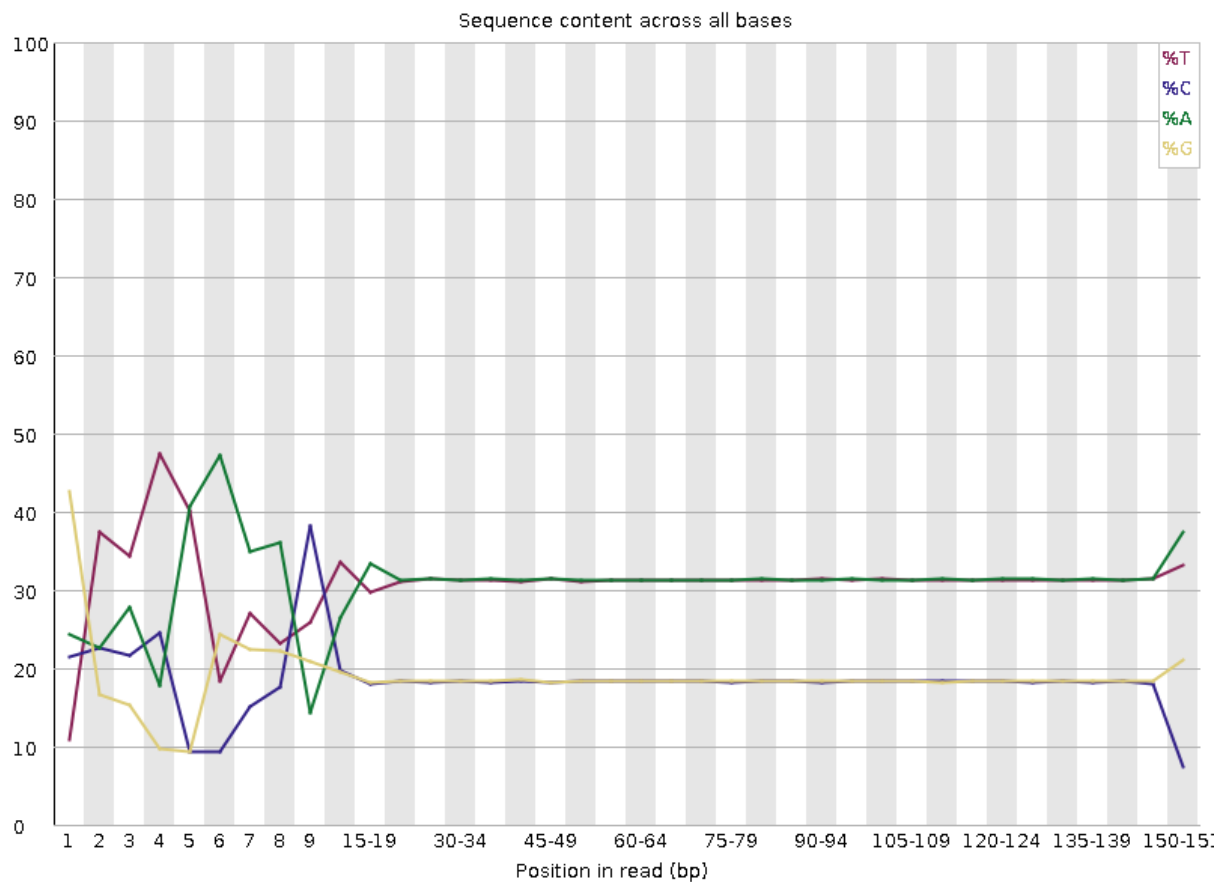


FIGURE 5.3 PER BASE SEQUENCE CONTENT

The per base sequence content analysis provides valuable insights into the distribution of nucleotides across each position within the sequencing reads. In this the x-axis represents the position within the reads, while the y-axis denotes the percentage of each nucleotide (A, T, G and C) at each position. Upon examination of the per base sequence content plot for read1, a notable deviation or periodic fluctuation in nucleotide distribution is observed particularly at the beginning of the graph spanning from 0 bp to 20 bp. Such deviations could arise due to various factors including PCR amplification biases or adapter contamination introduced during library preparation. Followed by this a uniform distribution of nucleotides is observed across most positions within the reads, indicating balanced nucleotide composition.

PER SEQUENCE GC CONTENT

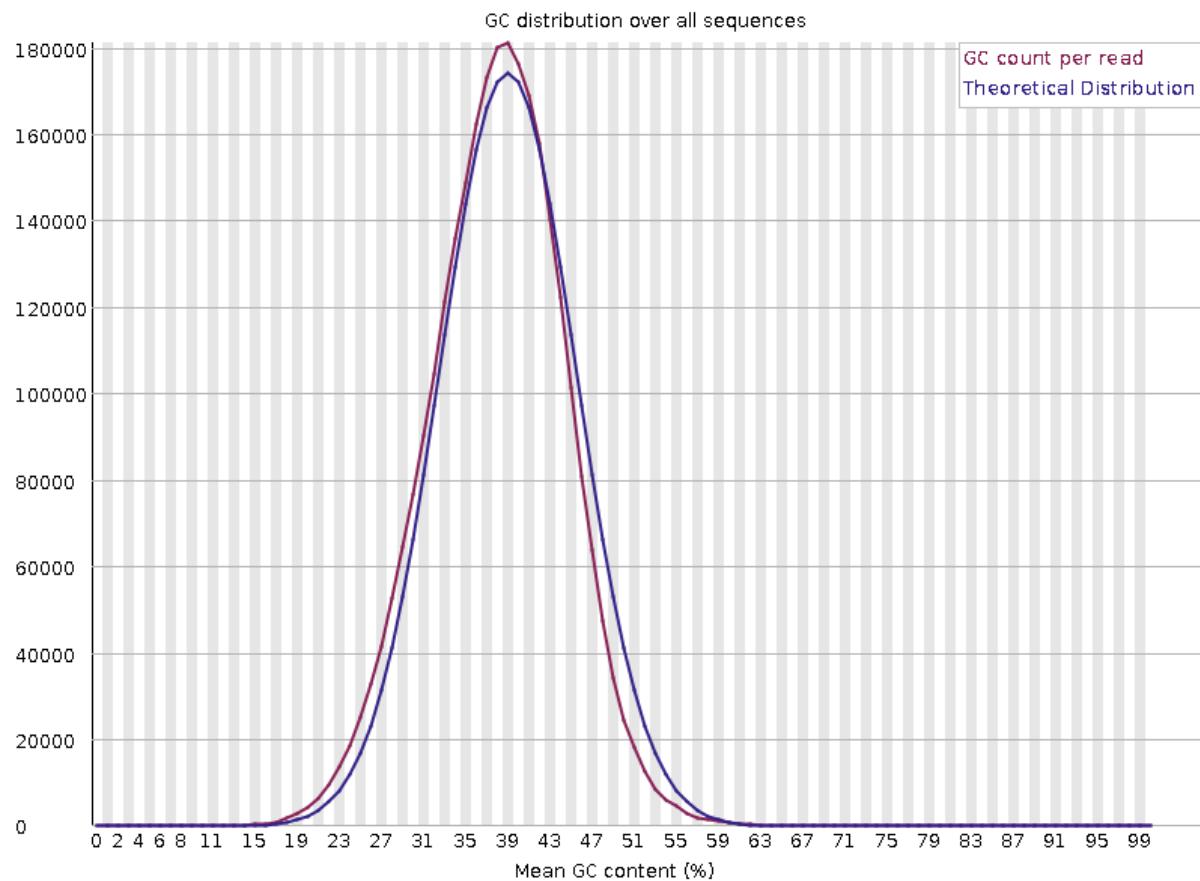


FIGURE 5.4 PER BASE GC CONTENT

In our analysis of per sequence GC content, we observed a distribution of sequences across a range of GC percentages. The distribution of GC content exhibited a peak around 37% indicating the presence of sequences with a predominant GC composition

PER BASE N CONTENT

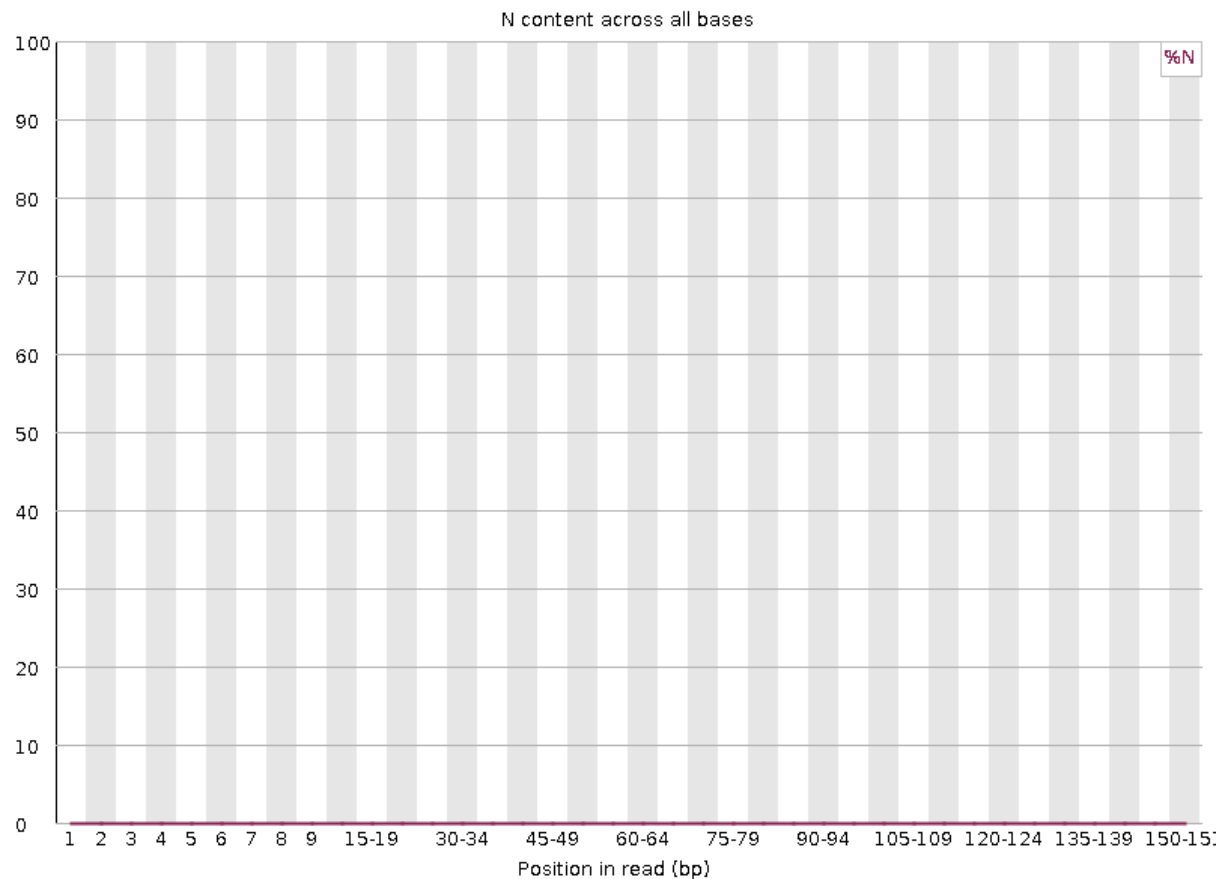


FIGURE 5.5 PER BASE N CONTENT

In our analysis of per base N content, we observed the absence of N bases across the sequencing reads. This indicates that the sequencing dataset is devoid of ambiguous bases, suggesting high quality sequencing and no sequence ambiguity. This suggests that the sequencing data is of high quality and that the nucleotide sequence can be confidently determined across all bases. This is indicative of reliable and accurate sequencing.

SEQUENCE LENGTH DISTRIBUTION

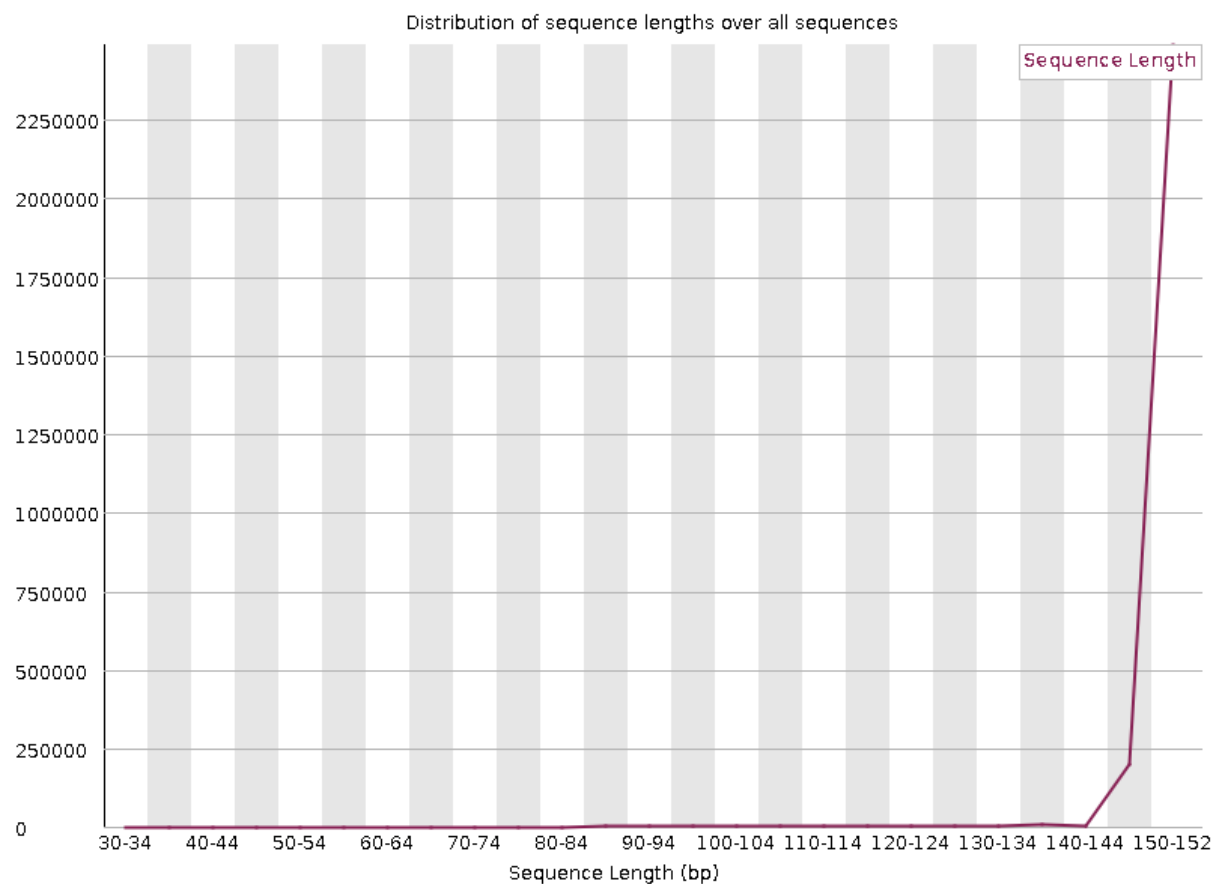


FIGURE 5.6 SEQUENCE LENGTH DISTRIBUTION

The distribution of sequence length provides valuable insights into the range of fragment sizes present in the sequencing dataset. In our analysis of sequence length distribution, the distribution of sequence length exhibited a predominant peak at 150-151 base pairs, indicating that a substantial proportion of sequencing reads in the dataset are of this length. Furthermore, a notable abundance of sequences with lengths ranging from 140 - 149 base pairs, particularly among sequences numbering less than 2,50,000.

OVERREPRESENTED SEQUENCES

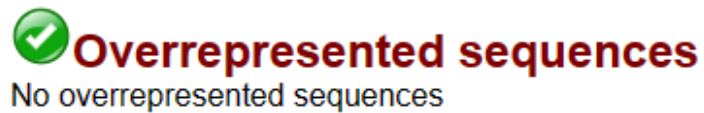


FIGURE 5.7 OVERREPRESENTED SEQUENCES

Overrepresented sequences, which are sequences that appear more frequently in the dataset than expected by chance, can provide insights into potential artifacts, contaminants, or highly abundant genomic regions present in the sequencing data. In our analysis we thoroughly examined the sequencing dataset for the presence of overrepresented sequences. Remarkably, no overrepresented sequences were detected in the dataset for read1.

ADAPTER CONTENT

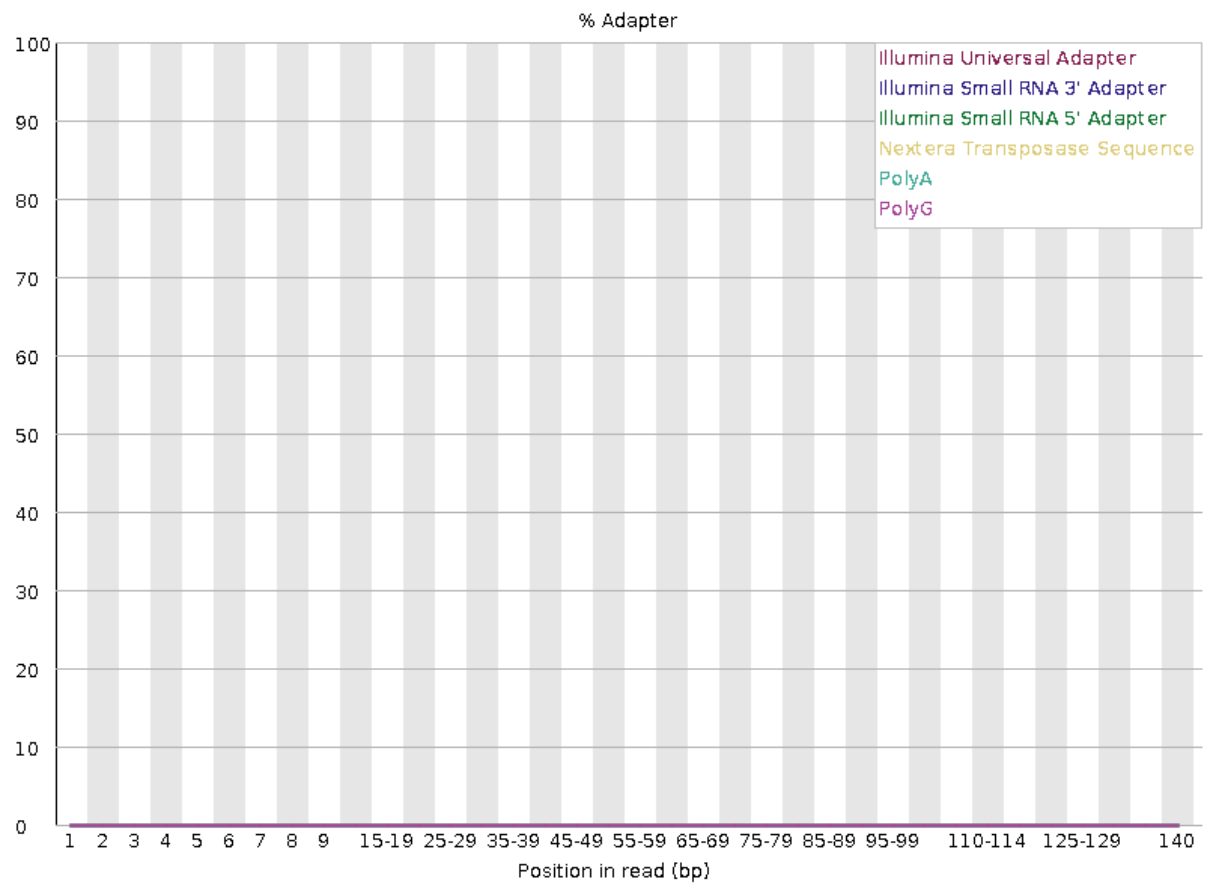


FIGURE 5.8 ADAPTER CONTENT

Adapter contamination, resulting from the presence of adapter sequences used during library preparation, can interfere with downstream analysis and affect the accuracy of sequencing data. In our analysis, we meticulously investigated the dataset for the presence of adapter contamination, notably no adapter sequences were identified

READ 2 :



Basic Statistics

Measure	Value
Filename	SRR16002670_2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2808394
Total Bases	417.4 Mbp
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	37

FIGURE 5.9 BASIC STATISTICS

BASIC STATISTICS

For read2, the filename associated is SRR16002670_2.fastq. This file contains a total of 28,08,394 sequences, comprising approximately 417.4 million base pairs. The sequence length ranges from 35-151 base pairs, with an average GC content of 37%. Notably for read2, no sequences were flagged as poor quality.

PER BASE SEQUENCE QUALITY

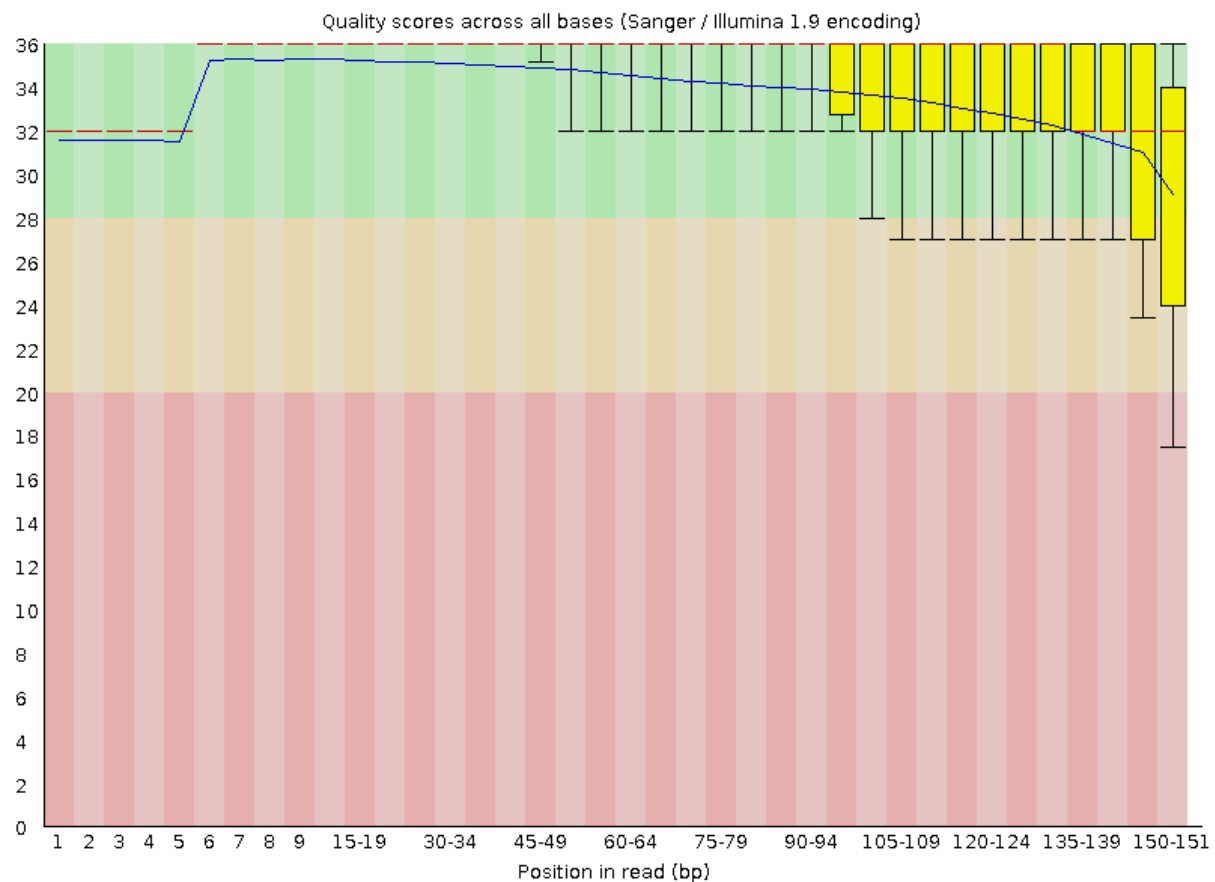


FIGURE 5.1.10 PER BASE SEQUENCE QUALITY

The per base sequence quality analysis have provided insights into the quality score distribution of bases in the reads across all the reads. In the case of read2 data the average quality score remains above 32 for the majority of the read, indicating high quality base calls. However, towards the end of the read, particularly from position 90, the presence of yellow boxes with whiskers indicated that there is a variability in the quality scores across different positions in the read, while most of the drastic variability coming from the reads starting from 140 base pair extending towards the end.

PER SEQUENCE QUALITY SCORES

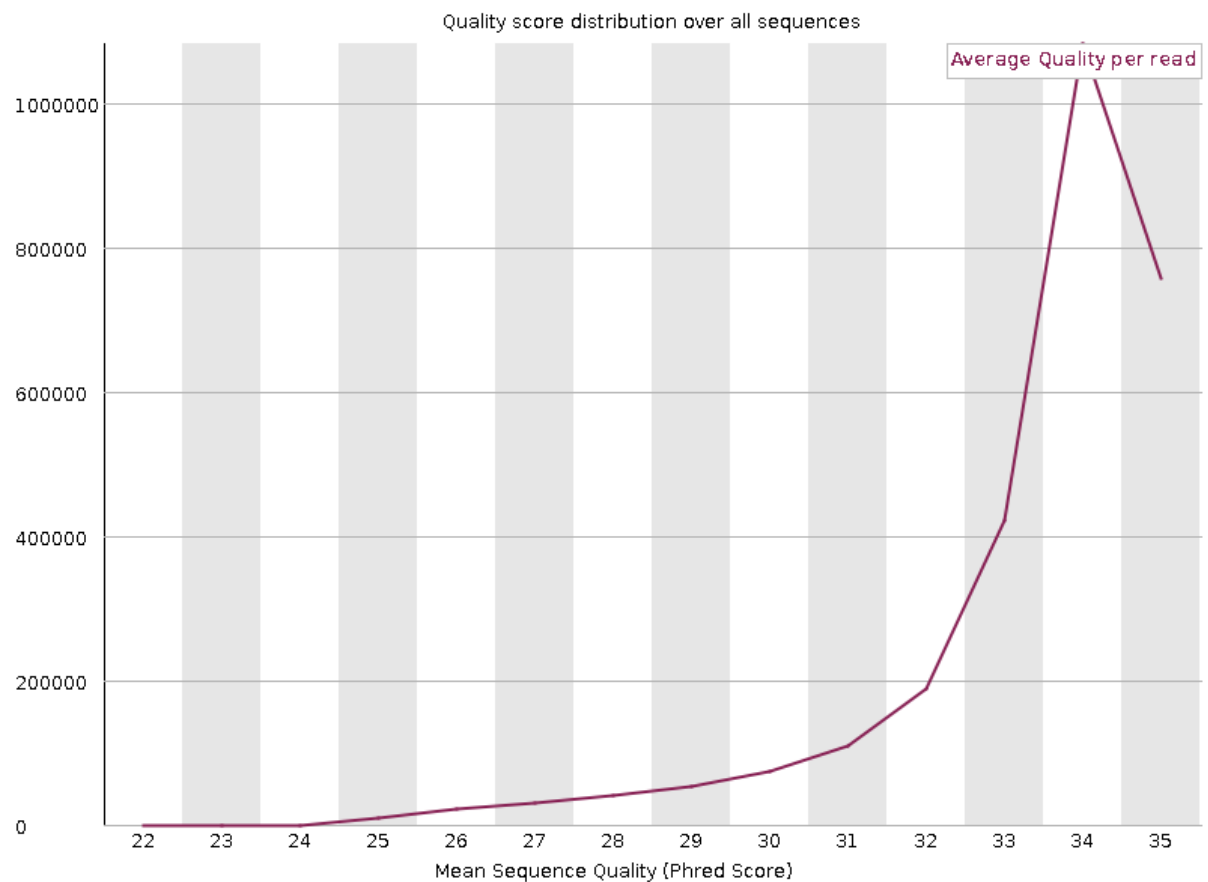
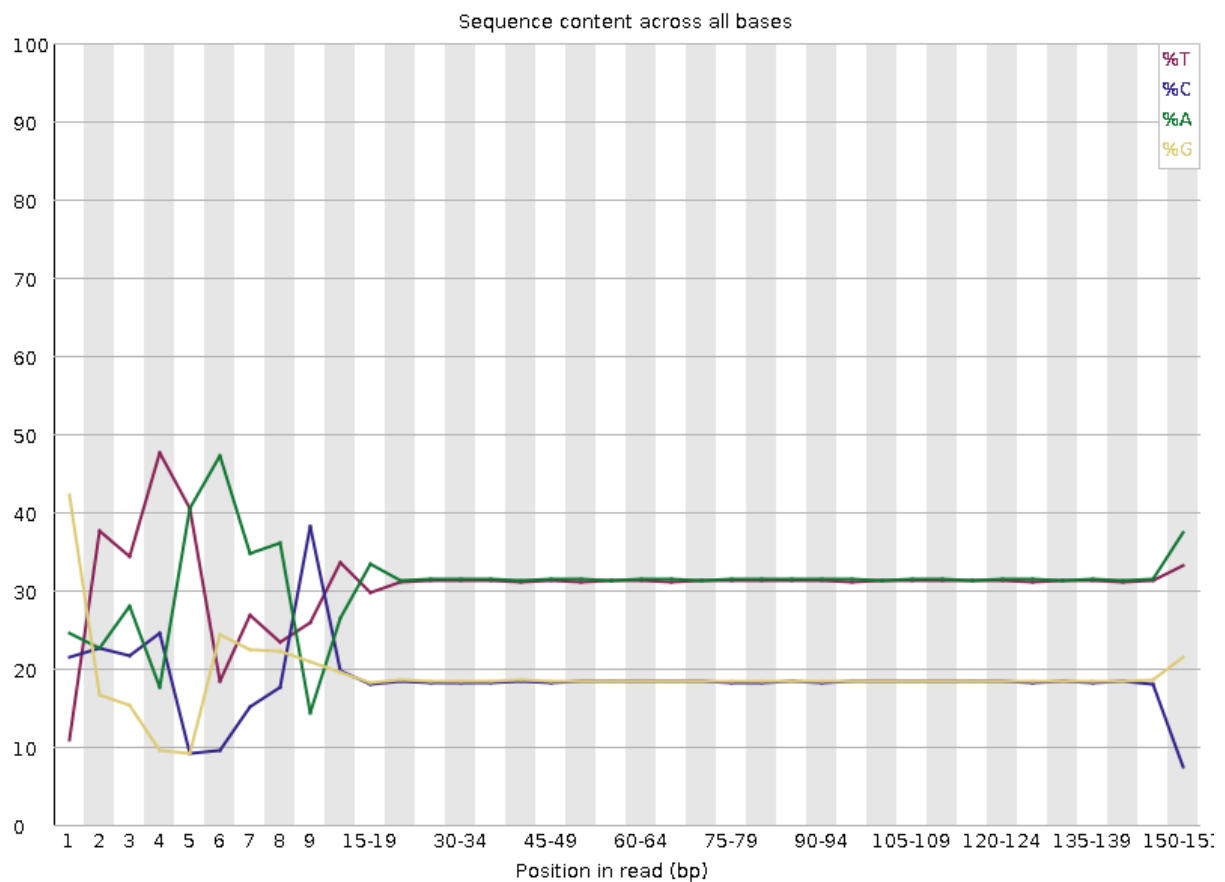


FIGURE 5.1.11 PER SEQUENCE QUALITY SCORES

The per sequence quality score analysis provides a comprehensive overview of the quality score distribution over all sequences in the dataset. In the case of read2, the mean sequence quality score exhibits a notable peak at 34, and also there is an increase in the quality score to 35 with reads counts below 10,00,000-8,00,000 indicating a predominance of high-quality base calls across the majority of the reads. Notably, some reads exhibit quality scores 30, extending down to 25 particularly observed among reads with counts below 200000.

PER BASE SEQUENCE CONTENT



The per base sequence content analysis provides valuable insights into the distribution of nucleotides across each position within the sequencing reads. In this the x-axis represents the position within the reads, while the y-axis denotes the percentage of each nucleotide (A, T, G and C) at each position. Upon examination of the per base sequence content plot for read1, a notable deviation or periodic fluctuation in nucleotide distribution is observed particularly at the beginning of the graph spanning from 0 bp to 20 bp. Such deviations could arise due to various factors including PCR amplification biases or adapter contamination introduced during library preparation. Followed by this a uniform distribution of nucleotides is observed across most positions within the reads, indicating balanced nucleotide composition.

PER BASE GC CONTENT

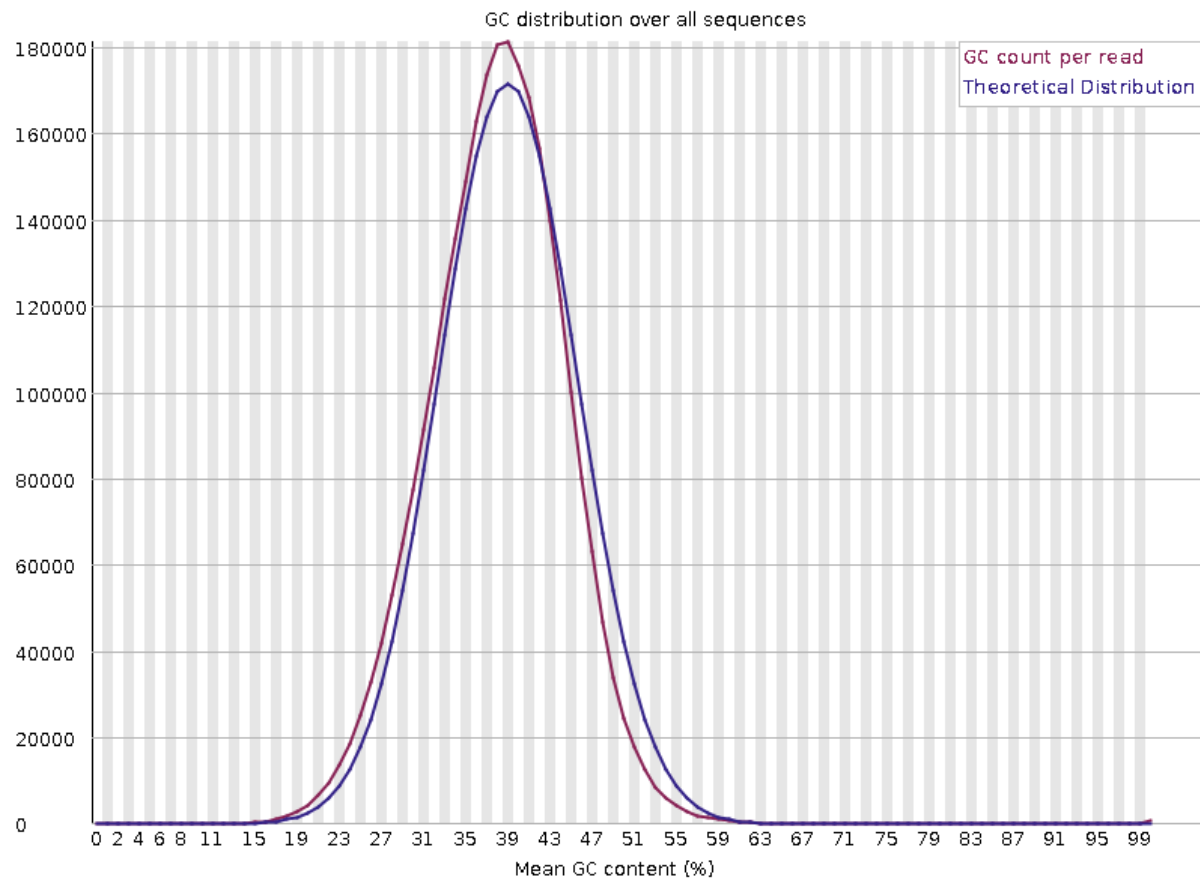


FIGURE 5.1.13 PER BASE GC CONTENT

In our analysis of per sequence GC content, we observed a distribution of sequences across a range of GC percentages. In our analysis, the distribution of GC content exhibited a peak around 37% indicating the presence of sequences with a predominant GC composition.

PER BASE N CONTENT

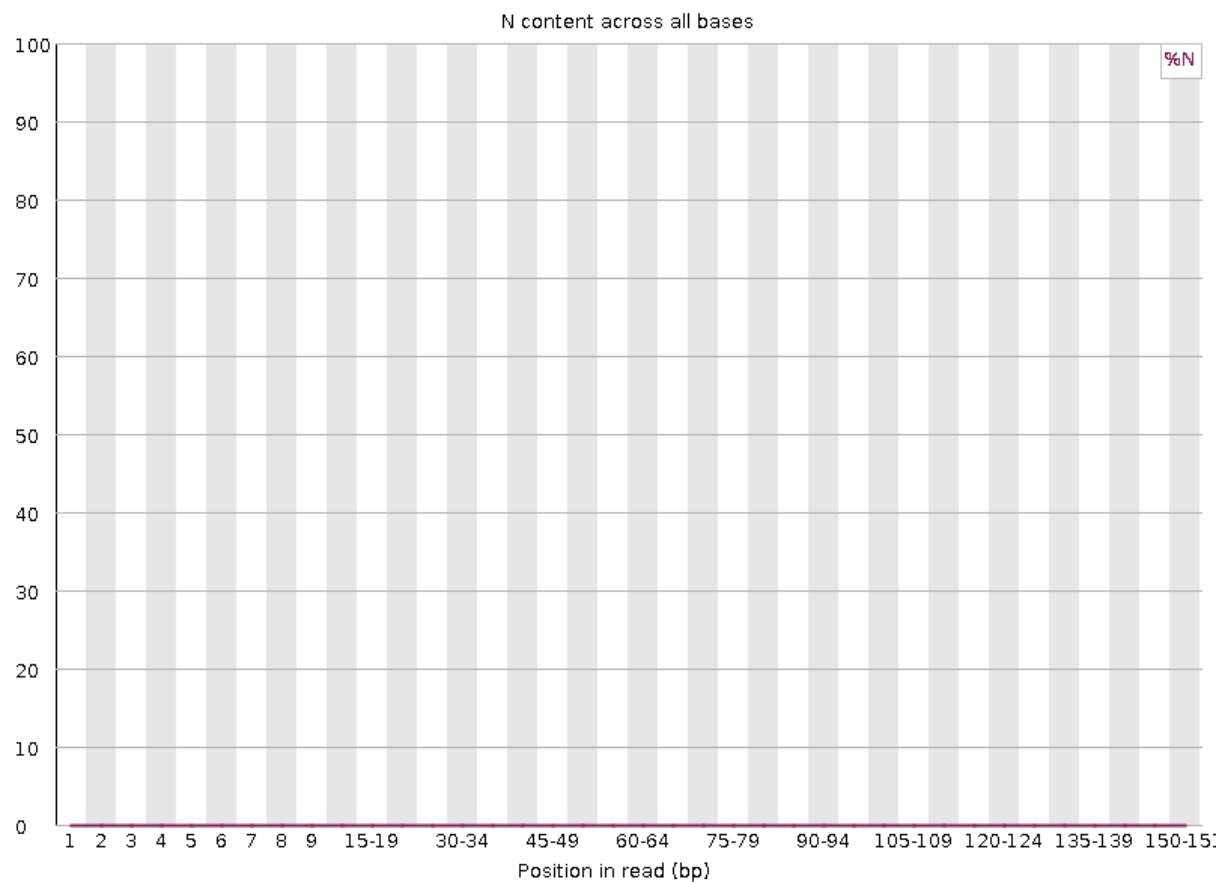


FIGURE 5.1.14 PER BASE N CONTENT

In our analysis of per base N content, we observed the absence of N bases across the sequencing reads. This indicates that the sequencing dataset is devoid of ambiguous bases, suggesting high quality sequencing and no sequence ambiguity. This suggests that the sequencing data is of high quality and that the nucleotide sequence can be confidently determined across all bases. This is indicative of reliable and accurate sequencing.

SEQUENCE LENGTH DISTRIBUTION

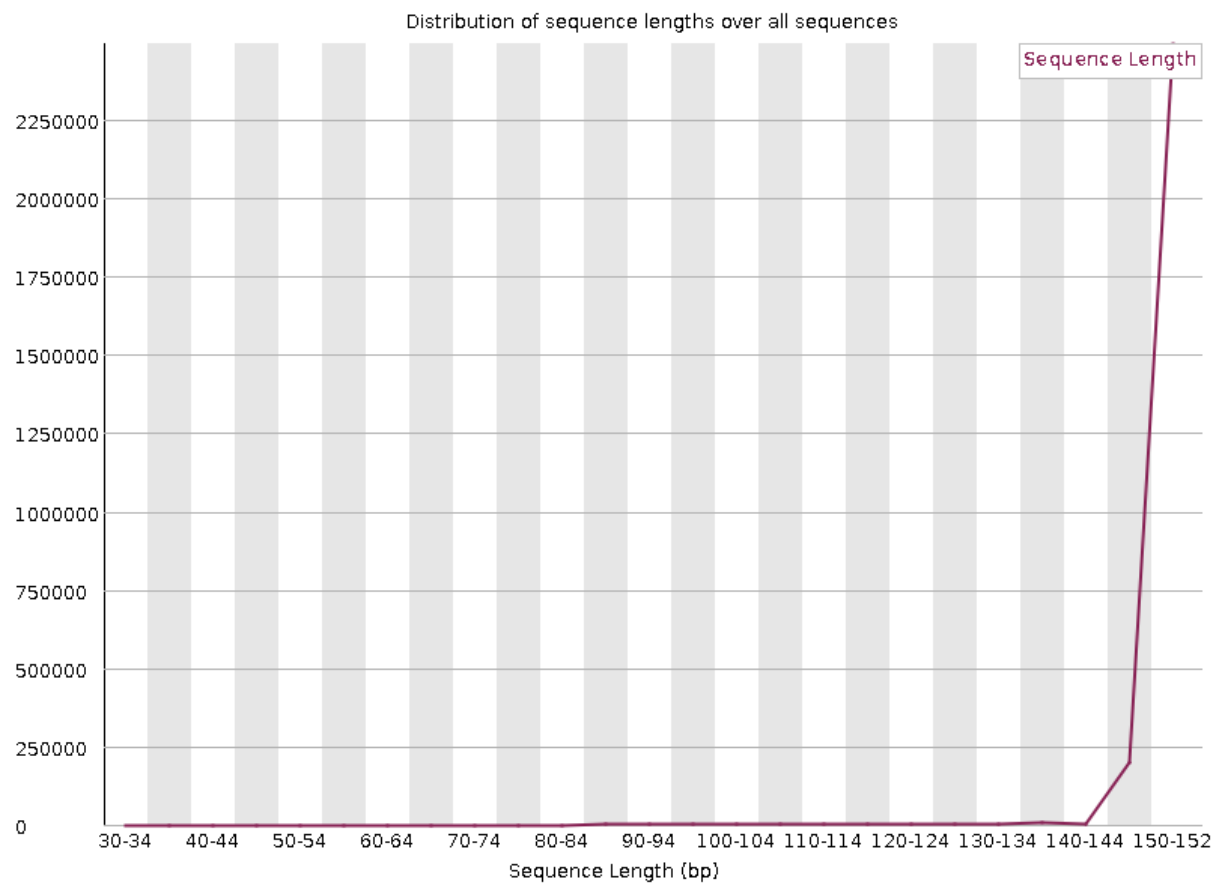


FIGURE 5.1.15 SEQUENCE LENGTH DISTRIBUTION

The distribution of sequence length provides valuable insights into the range of fragment sizes present in the sequencing dataset. In our analysis of sequence length distribution, the distribution of sequence length exhibited a predominant peak at 150-151 base pairs, indicating that a substantial proportion of sequencing reads in the dataset are of this length. Furthermore, a notable abundance of sequences with lengths ranging from 145 - 149 base pairs, particularly among sequences numbering less than 2,50,000.

BEFORE TRIM



GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG GGG
GG was observed with a count of 2954 reads, accounting for approximately 0.10% of the total reads. The source of this overrepresented sequence could not be identified.

AFTER TRIM



Following the trimming process the overrepresented sequence GG GG was successfully eliminated from the processed read data. As a result, no overrepresented sequences were detected in the trimmed dataset, indicating the effectiveness of the trimming procedure in removing the potential artifacts and improving the quality of the sequencing data

ADAPTER CONTENT

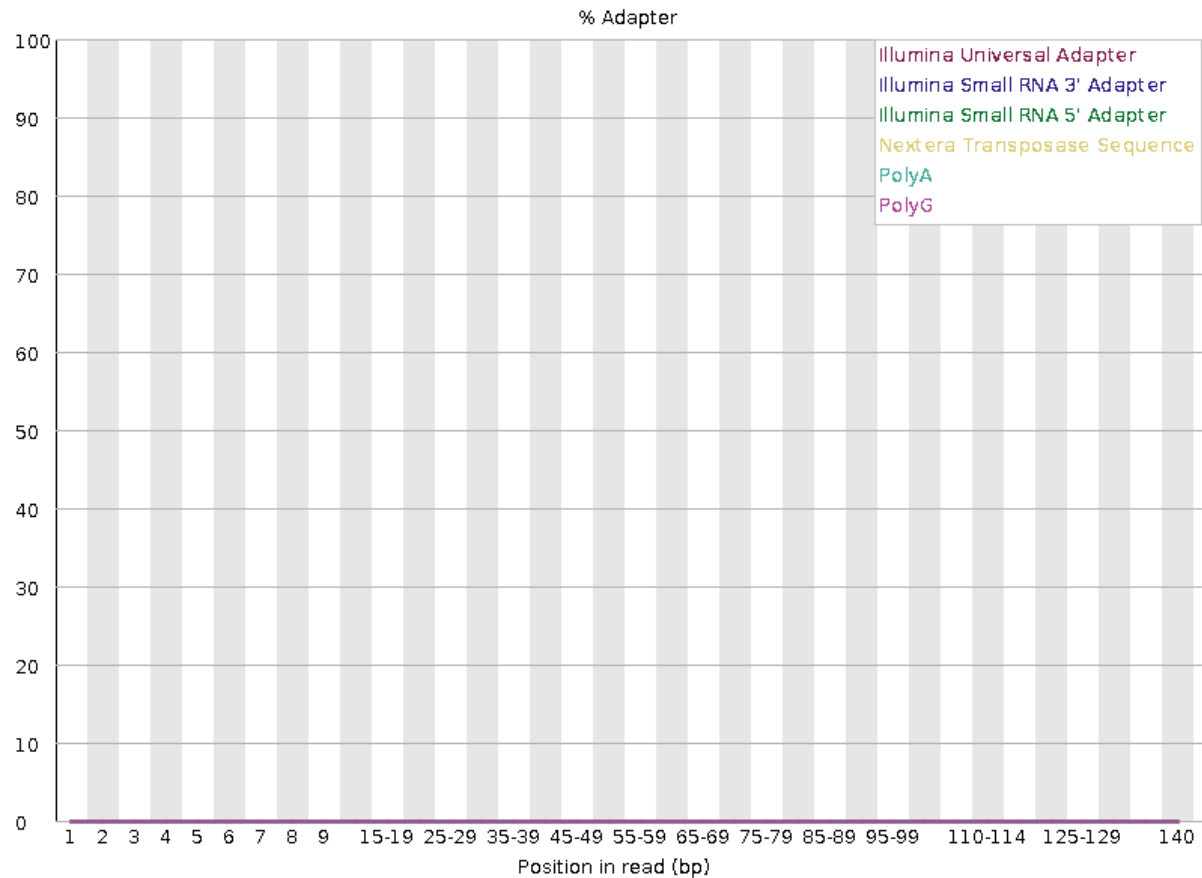


FIGURE 5.1.18 ADAPTER CONTENT

Adapter contamination, resulting from the presence of adapter sequences used during library preparation, can interfere with downstream analysis and affect the accuracy of sequencing data. In our analysis, we meticulously investigated the dataset for the presence of adapter contamination, notably no adapter sequences were identified

QUALITY CONTROL OF ONT READS

BEFORE TRIM

General Summary:

Mean read length	3,312.3
Mean read Quality	9.7
Median read length	2,275.0
Median read Quality	10.6
Number of reads	260,487
Read length N50	5427.0
STDEV read length	3296.8
Total bases	862,809,106.0

Number, percentage and megabases of reads above quality cutoffs:

>Q5	260487(100.0%)	862.8Mb
>Q7	260487(100.0%)	862.8Mb
>Q10	162742(62.5%)	548.6Mb
>Q12	42721(16.4%)	141.4Mb

Top 5 highest mean basecall quality scores and their read lengths

1	15.9
2	15.6
3	15.6
04	15.6
5	15.5

Top 5 longest reads and their mean basecall quality score

1	42510(10.1)
2	40335(7.3)
3	40277(7.8)
4	38697(11.1)
5	38036(8.1)

From the quality check analysis of Oxford Nanopore Sequencing data before pre processing, the statistics revealed several noteworthy metrics. The mean read length was found to be 3,312.3 base pairs, with a mean read quality of 9.7 followed by median read length and median read quality of 2,275 base pairs and 10.6 respectively. Furthermore, the dataset comprised a total of 2,60,487 reads collectively accounting for 862.8 megabases of sequencing data. The quality cutoff analysis revealed that all reads exceeded a quality score of Q5. Notably, over 62.5 % of the reads surpassed a quality threshold of Q10, indicating a substantial portion of high quality reads suitable for robust analysis.

AFTER TRIM

Mean read length	3,370.8
Mean read Quality	10.9
Median read length	2,377.5
Median read Quality	11.4
Number of reads	162,742
Read length N50	5,397.0
STDEV read length	3228.9
Total bases	548,571,541.0

Number, percentage and megabases of reads above quality cutoffs:

>Q5	162742(100.0%)	548.6Mb
>Q7	162742(100.0%)	548.6Mb
>Q10	162742(100.0%)	548.6Mb
>Q12	42721(26.3%)	141.4Mb

Top 5 highest mean basecall quality scores and their read lengths

1	15.9 (187)
2	15.6 (438)
3	15.6 (351)
4	15.6 (360)
5	15.5 (2410)

Top 5 longest reads and their mean basecall quality score

1	42510 (10.1)
2	38697 (11.1)
3	33721 (10.5)
4	33611 (10.1)
5	33594 (10.7)

From the quality check analysis of Oxford Nanopore Sequencing data after pre-processing, the statistics revealed several noteworthy metrics. The ONT reads exhibit a mean read length of 3,370.8 base pairs with a mean read quality of 10.9. The median read length is 2,377.5 with a median read quality of 11.4. A total of 1,62,742 reads were obtained , demonstrating a read length N50 of 5,397.0 base pairs. The standard deviation of the read length is 3,228.9 suggesting variability in read lengths. The total bases sequenced amount to 548,571,541.0 indicating substantial coverage of the genome. All reads pass quality cutoffs of Q5,Q7 and Q10, while 26.3 % of the reads meet the more stringent Q12 cutoff, implying high base quality. It's noteworthy that reads with quality score less than $< Q10$ were removed in the pre-processing step in order to retain high quality reads that are included in subsequent analysis. The top 5 highest mean base call quality scores range from 15.5 - 15.9, with corresponding reads lengths ranging from 187 base pairs and 2410 base pairs respectively. Similarly the top 5 longest reads range from 33,594 - 42,510 base pairs with mean baseball quality scores ranging from 10.7 - 10.1 .

ASSEMBLY

Assembly

# contig (>= 0bp)	2
#contig (>= 1000 bp)	2
#contig (>= 5000 bp)	2
#contig (>= 10000 bp)	2
#contig (>= 25000 bp)	2
#contig (>= 50000 bp)	2
Total length (>= 0 bp)	2447263
Total length (>= 1000 bp)	2447263
Total length (>= 5000 bp)	2447263
Total length (>= 10000 bp)	2447263
Total length (>= 25000 bp)	2447263
Total length (>= 50000 bp)	2447263
# contigs	2
Largest contig	2350946
Total length	2447263
GC (%)	37.7
N50	2350946
N90	2350946
auN	2262210.5
L50	1
L90	1

In our analysis of the genome assembly, the assembly consists of two contigs, with the largest one spanning 23,50,946 base pairs and the smaller contig spanning 96,317 base pairs was identified comprising a total length of 24,47,263 base pairs. The assembly metrics further validate the quality of the assembly. The N50 value, a measure of contiguity, indicates that 50% of the genome is contained within contigs of at least 23,50,946 base pairs, mirroring the size of the largest contig. Similarly, the N90 values align with the N50, confirming the assembly's coherence and minimal fragmentation. The absence of ambiguous bases, as indicated by the low N's per 100 kilo basepair values of 0.00, suggests that a high-quality assembly devoid of gaps or unresolved regions. Moreover, the calculated GC content of 37.17% provides insights into the nucleotide composition of the genome, contributing to our understanding of its genetic architecture.

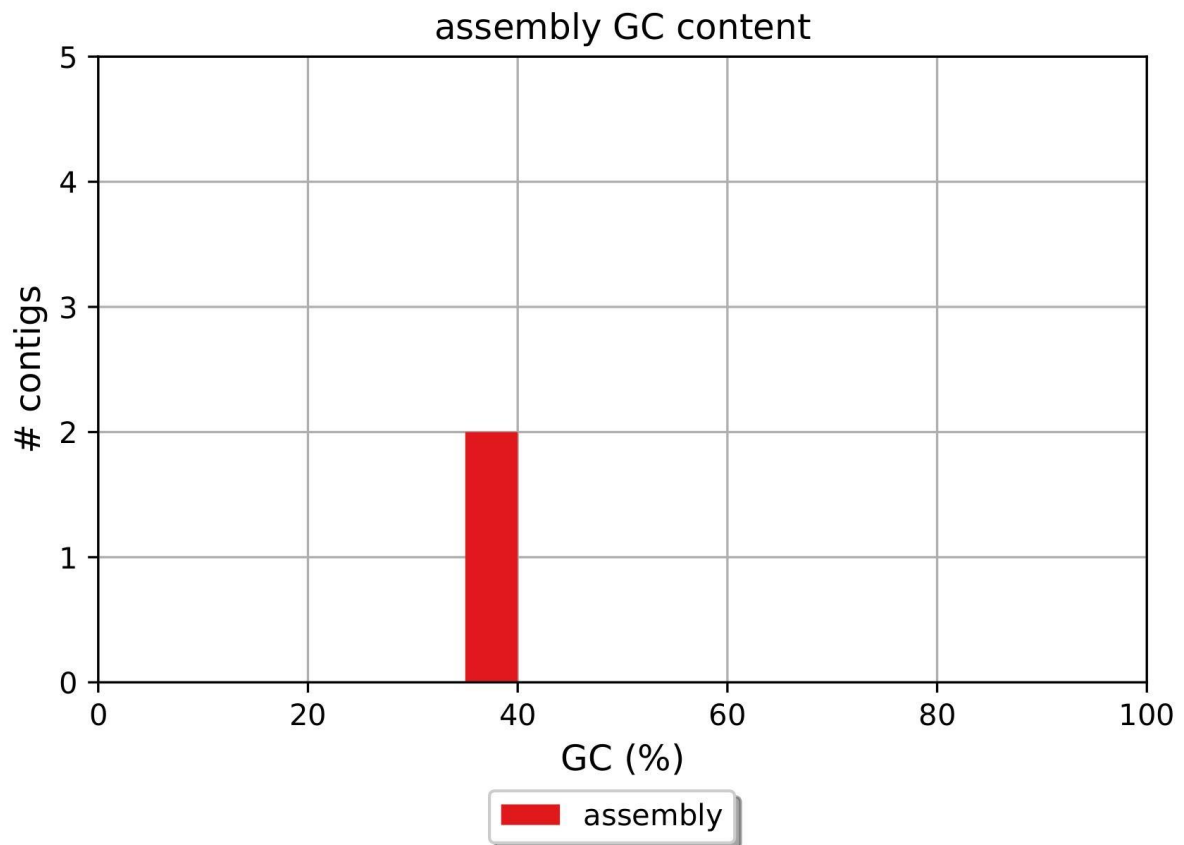


FIGURE 5.1.19 GC CONTENT

The assembly GC content refers to the percentage of Guanine (G) and Cytosine (C) bases in the assembled genome. In this case, the assembly consists of two contigs, and the GC percentage is approximately 37.7 %. A GC content of 37.7% suggests a moderate GC richness in the assembled genome.

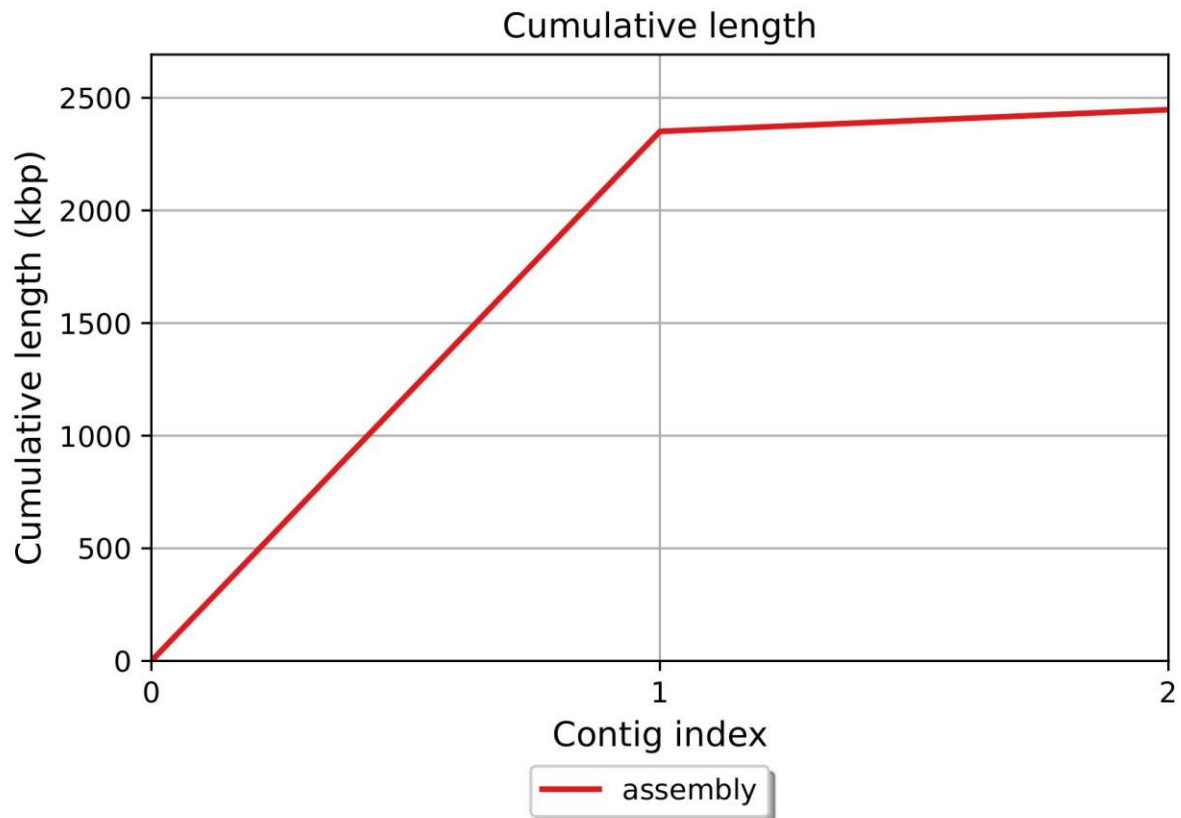


FIGURE 5.1.20 CUMULATIVE LENGTH

The cumulative length graph illustrates the total length of the assembled contigs in kilobase pairs(kbp) as we progress through the contig index.Each point on the graph represents the cumulative length of all contigs up to that index. In this case, the x-axis represents the contig index, with values ranging from 1 to 2 indicating the two contigs present in the assembly. The y-axis represents the cumulative length in kilobase pairs.

As we move along the x-axis, starting from contig index 1, the cumulative length increases gradually, reflecting the addition of the length of each contig to the total.At the first contig index, the cumulative length is approximately 2300 kbp, suggesting that the first contig alone contributes this length to the assembly. As we progress to the second contig index, the cumulative length continues to increase, reaching a value slightly below 2500 kbp, possibly around 2400 kbp. This indicates that the second contig adds additional length to the assembly, contributing to the cumulative total.

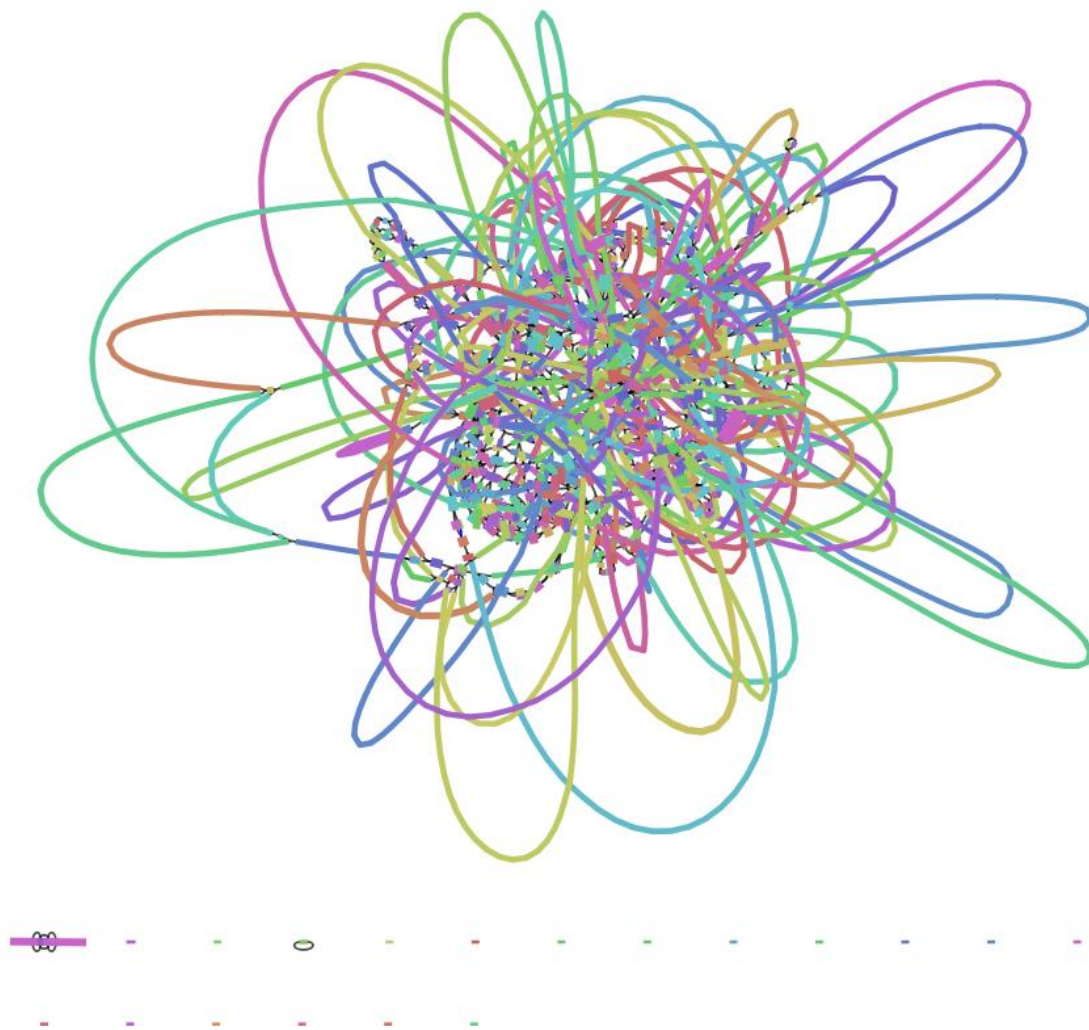


FIGURE 5.1.21 ASSEMBLY GRAPH

The figure represents the assembly graph with scaffolds. In an assembly graph, nodes represent sequences (such as contigs), and edges represent connections between these sequences. Scaffolds are constructed by linking contigs together based on paired-end or long-read sequencing data. This graph visualizes the structure of the assembly, showing how contigs are connected to form longer sequences. By examining this graph, researchers can gain insights into the overall structure and organization of the assembled genome, including potential mis assemblies, repetitive regions, and structural variations.

The graph consists of 779 nodes and 1037 edges, with edge overlaps totaling 127 base pairs. The total length of the graph, considering overlaps, is 2,339,252 base pairs, while the total length without overlaps is slightly less at 2,242,478 base pairs.

There are 34 dead ends in the graph, representing 2.18% of the total nodes. The graph contains 20 connected components, with the largest component spanning 2,333,488 base pairs, covering approximately 99.75% of the total graph length. Orphaned nodes, which are not connected to any other node, account for a total length of 5,253 base pairs.

The N50 value of the nodes is 19,393 base pairs, indicating that half of the total assembly length is contained in contigs or scaffolds equal to or longer than this size. The shortest node in the graph is 128 base pairs, while the longest node spans 55,447 base pairs. The lower quartile node size is 137 base pairs, the upper quartile node size is 793 base pairs, and the median node size is 234 base pairs. These metrics provide insights into the distribution of contig lengths within the assembly graph.

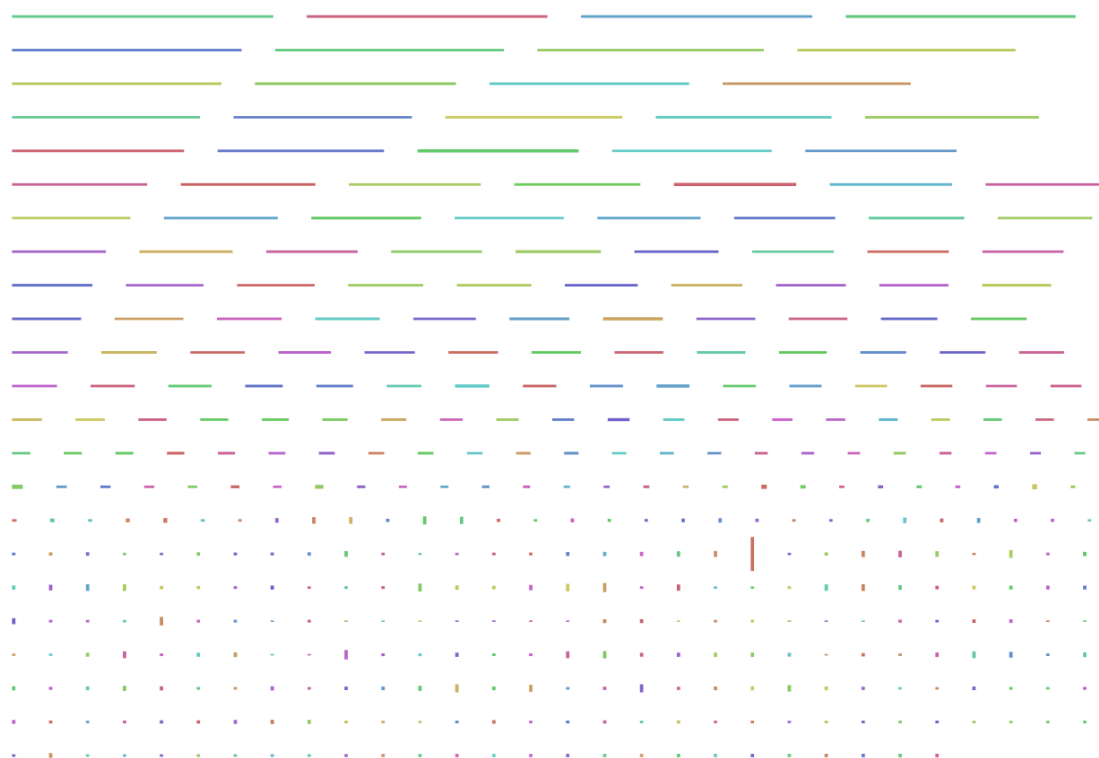


FIGURE 5.1.22 SCAFFOLDS

This figure represents the assembled scaffolds, which are larger contiguous sequences constructed from assembled contigs. Each scaffold sequence represents a connected sequence of contigs reflecting the continuity across long stretches of DNA. This figure contains 402 nodes, representing sequences or scaffolds, with a total length of 2,292,625 base pairs. Each node's length varies, with the N50 value indicating the length of the shortest sequence necessary to cover at least 50% of the assembly. In this case, the N50 value is 24,165 base pairs, suggesting that the assembly contains a mix of short and long scaffolds. The shortest scaffold in the assembly is 128 base pairs, while the longest scaffold spans 55,449 base pairs. The distribution of scaffold lengths across the assembly is further described by quartile values, including the lower quartile, upper quartile, and median node lengths.

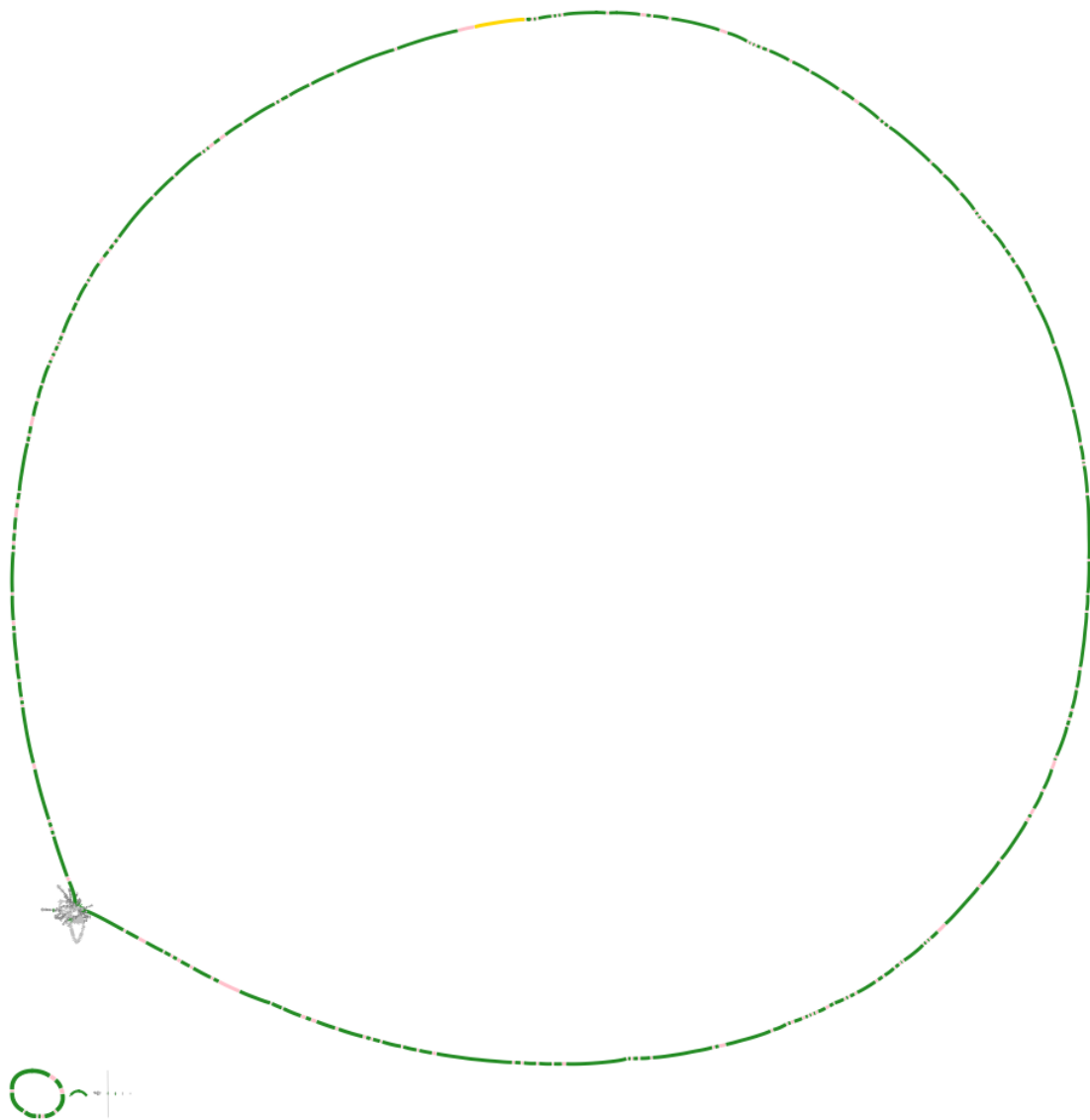


FIGURE 5.1.23 LONG READ ASSEMBLY GRAPH

This graph represents an intermediate stage in the assembly process where additional connections or bridges have been applied to improve the continuity and accuracy of the assembly. This graph may show connections between contigs that help bridge gaps or resolve ambiguities in the assembly, resulting in a more complete and contiguous representation of the genome. From our analysis, this graph indicates the presence of 690 nodes, each representing scaffold sequences and 790 edges denoting connections between these sequences. The total length of the graph spans 25,20,000 base pairs, without any overlapping edges observed, highlighting the integrity of the assembly.

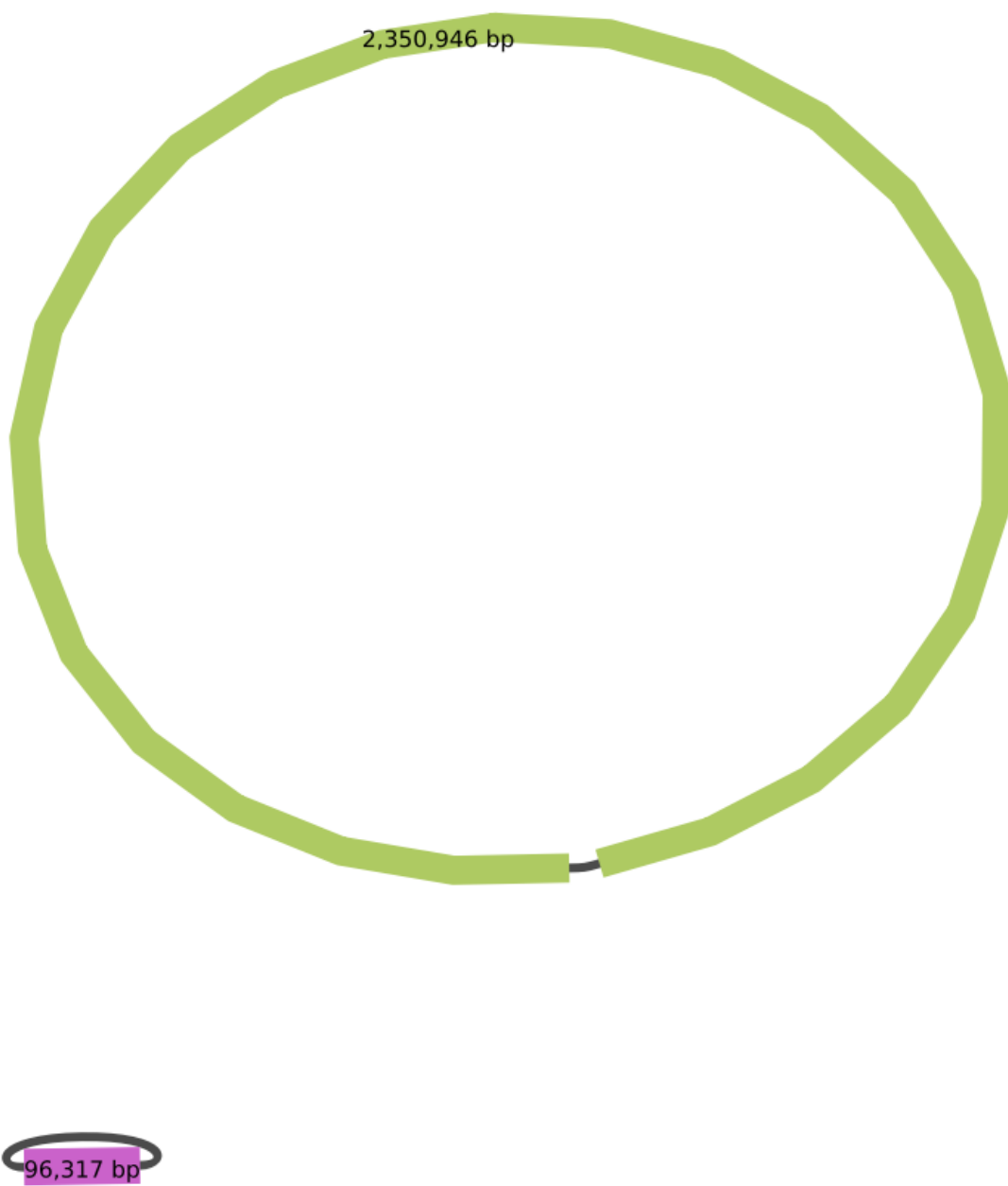


FIGURE 5.1.24 MERGED FINAL ASSEMBLY GRAPH

This graph represents the merged assembly graph, which combines the information from the previous steps of assembly refinement. This graph provides insights into the final structure of the assembled genome. The graph consists of two nodes and two edges in total, each representing a distinct genomic component, such as a chromosome and a plasmid. The edges between the nodes indicate connections or overlaps between the genomic components. From our analysis of the graph, there are two nodes and two edges in total. The total length of the graph is 24,47,263. The graph connectivity indicates that there are no dead ends in the graph, meaning that all nodes are connected in a continuous manner. The percentage of dead ends is 0.00%, indicating a highly connected graph structure. The graph consists of two connected components, with the largest component comprising 23,50,946 base pairs, representing the majority of the assembled genome. Node sizes provide additional information about the lengths of the genomic components represented by the nodes. The N50 values, which represents the length at which half of the total assembly length is contained in contigs or scaffolds of equal or greater length, is 23,50,946 base pairs. The shortest node in the graph is 96,317 base pairs long. The lower quartile node length is 6,59,974 base pairs, indicating the length below which 25% of the nodes fall. The upper quartile node length is 17,87,289 base pairs, indicating the length above which 25% of the nodes fall. The median node length is 12,23,632 base pairs, representing the midpoint of the node length distribution. Also, in this graph, the genomic components represented by nodes are rotated to improve the clarity and interpretability of the assembled graph.

Moreover, this graph offers a comprehensive view of the assembled genome, consolidating the information provided by the previous representations and is accompanied by a separate file containing the sequence of this assembly. This serves as a crucial resource for further genomic analysis, particularly for genome annotation.

GENOME ANNOTATION

The genome annotation process involved the comprehensive characterization of genetic elements within the assembled genome sequence. Utilizing the assembled sequence data obtained from the final assembly, genome annotation was conducted to identify and annotate various features. The annotation process began with the prediction and annotation of genomic features using PROKKA. PROKKA employs a combination of homology-based and de novo gene prediction algorithms to accurately identify putative protein coding genes within the genome. Each predicted gene was assigned functional annotations based on sequence similarity to known proteins available in public databases such as UniProt and NCBI's non-redundant protein databases.

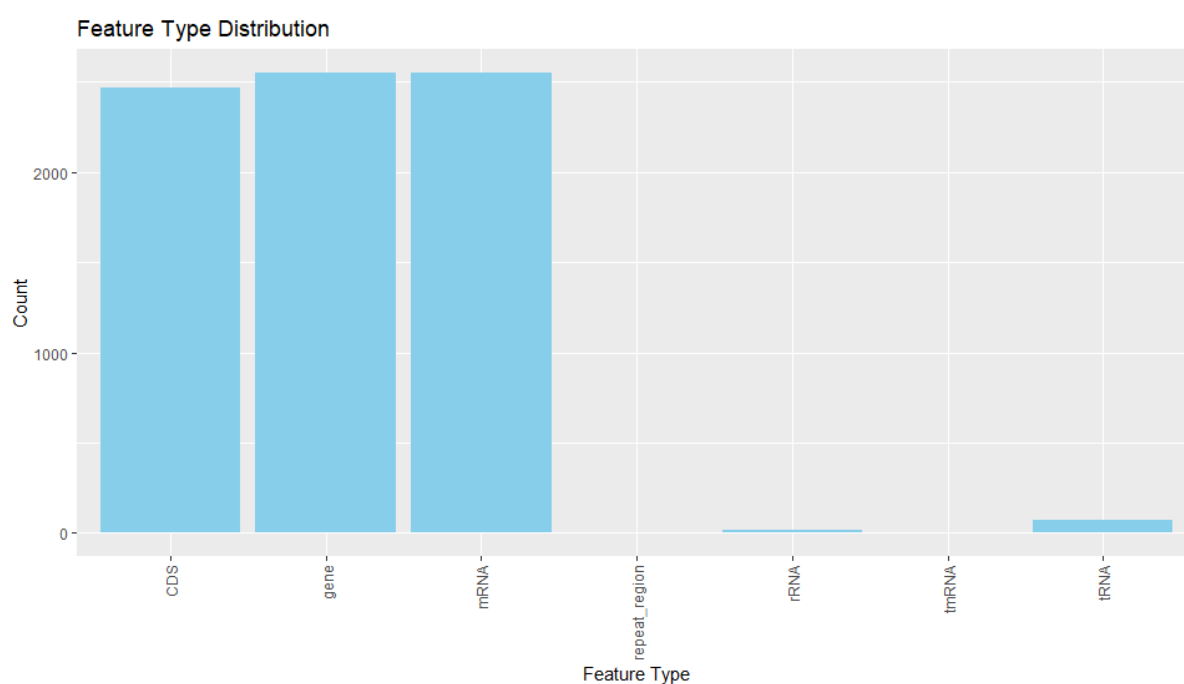


FIGURE 5.1.25 FEATURE TYPE DISTRIBUTION

From our analysis of this graph, which depicts the distribution of different feature types present in the genome annotation. The comprehensive annotation of *Lactobacillus crispatus* Lc116 genome reveals the presence of 15 rRNA coding genes which underscores the significance of ribosomal RNA in protein synthesis and cellular function, while the identification of 69 tRNA coding genes highlights the importance of transfer RNA molecules in mediating translation. Moreover, the detection of 2467 coding sequences suggests a vast protein coding genes within the genome, indicating a diverse number of proteins involved in essential cellular function. Overall, the annotated genome, with its delineated genetic elements and corresponding sequences, provide a valuable resource for a myriad of future studies and applications. The availability of gene sequences facilitates functional characterization, allowing for investigations into gene expression patterns, protein structure and function and regulatory mechanisms. Ultimately, the information encapsulated within the annotated genome opens avenues for diverse research endeavors, spanning fields such as microbiology, clinical medicine, biotechnology, thereby advancing our understanding of *Lactobacillus crispatus* Lc116 and its potential applications in health and disease.

CHAPTER 6

SUMMARY AND CONCLUSION

SUMMARY

This study offers a comprehensive exploration into the genomic landscape of *Lactobacillus crispatus* Lc116, a major urinary tract resident in adult women, but remains underexplored, particularly in terms of urogenital health. The project began with the retrieval of raw sequencing data from the NCBI SRA database, comprising both Illumina paired-end reads and Oxford Nanopore reads. Rigorous quality control procedures were meticulously applied to ensure the integrity and reliability of the data, including the identification and removal of low-quality reads. The Illumina reads were assessed for quality using FASTQC, followed by the removal of overrepresented sequences with FASTP, while Nanopore reads underwent quality evaluation with Nanostat and removal of low quality Nanopore reads with Nanofilt. Subsequently, a sophisticated hybrid assembly strategy, leveraging the strengths of both short-read and long-read sequencing technologies, was employed using the Unicycler assembler. The resulting assembly was subjected to comprehensive evaluation using the QUAST tool, revealing robust assembly metrics indicative of high-quality genomic reconstruction. This was further corroborated by visual inspection of assembly graphs generated using Bandage, which provided valuable insights into the assembly structure and integrity. Following assembly, the genome underwent thorough annotation using the PROKKA pipeline, facilitating the identification and annotation of various genomic features. The annotated genome serves as a rich resource for further investigations into the genetic basis of host-microbe interactions, adaptive strategies, and potential therapeutic applications in the context of urogenital health. This study not only contributes to advancing our understanding of the urinary microbiota but also lays the groundwork for future research endeavors aimed at unraveling the intricate dynamics of microbial communities and their impact on human health.

CONCLUSION

In conclusion, this study presents a thorough genomic analysis of *Lactobacillus crispatus* Lc116 strain, shedding light on its genetic makeup and potential implications for urogenital health. By employing meticulous sequencing and assembly techniques, a high quality genome assembly has been produced, laying the ground work for further exploration of the molecular mechanisms governing host-microbe relationships. The genome annotation has facilitated the recognition of crucial genomic characteristics, setting the stage for comprehending the functional roles of various genetic components within a bacterium. Additionally, this annotated genome acts as a valuable tool for upcoming investigations, serving as a point of departure for variety of genomic analysis including comparative genomics, evolutionary research and functional genomics. Furthermore, the availability of both chromosomal and plasmid sequences allows for targeted examination of specific genetic elements and their functions in regulating bacterial physiology and virulence. By studying these genes individually, scientists can acquire deeper insights into the biology of *Lactobacillus crispatus* Lc116 strain and its potential therapeutic applications in probiotics and interventions aimed at enhancing urogenital health. In essence, this study contributes to our comprehension of microbial diversity and adaptation in the urogenital region, opening avenues for future studies focused on enhancing human health.

REFERENCES

- 1) Grin, P. M., Kowalewska, P. M., Alhazzan, W., & Fox-Robichaud, A. E. (2013). Lactobacillus for preventing recurrent urinary tract infections in women: meta-analysis. *Can J Urol*, 20(1), 6607-6614.
- 2) Chee, W. J. Y., Chew, S. Y., & Than, L. T. L. (2020). Vaginal microbiota and the potential of Lactobacillus derivatives in maintaining vaginal health. *Microbial cell factories*, 19(1), 203.
- 3) Amabebe, E., & Anumba, D. O. (2018). The vaginal microenvironment: the physiologic role of lactobacilli. *Frontiers in medicine*, 5, 181.
- 4) Zhou, X., Bent, S. J., Schneider, M. G., Davis, C. C., Islam, M. R., & Forney, L. J. (2004). Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology*, 150(8), 2565-2573.
- 5) 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), 30.
- 6) Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics and Bioinformatics*, 14(5), 265-279.
- 7) Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884-i890.
- 8) De Coster, W., D'hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15), 2666-2669.
- 9) Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, 13(6), e1005595.

- 10) Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & 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.
- 11) Ekblom, R., & Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications*, 7(9), 1026-1042.
- 12) Del Angel, V. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., ... & Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, 7.
- 13) Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), 315-327.
- 14) Paszkiewicz, K., & Studholme, D. J. (2010). De novo assembly of short sequence reads. *Briefings in bioinformatics*, 11(5), 457-472.
- 15) Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature methods*, 9(4), 333-337.
- 16) Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072-1075.
- 17) Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20), 3350-3352.
- 18) Sharon, B. M., Hulyalkar, N. V., Nguyen, V. H., Zimmern, P. E., Palmer, K. L., & De Nisco, N. J. (2021). Hybrid de novo genome assembly for the generation of complete genomes of urinary bacteria using short-and long-read sequencing technologies. *JoVE (Journal of Visualized Experiments)*, (174), e62872.
- 19) Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068-2069.

