“TRANSCRIPTOME SEQUENCING (RNA-SEQUENCING) HUMAN BRAIN, COMPARING FETAL AND ADULT USING NOVEL GENOMIC DATA SCIENCE APPROACH”

A thesis submitted in partial fulfillment of the requirements for the award of the degree of

Master of Science

Computer Science

by

Ajay Kumar

(18419CMP003)

Department of Computer Science
Institute of Science
Banaras Hindu University

CANDIDATE'S DECLARATION

I hereby certify that the work, which is being presented in the report/thesis, entitled **"Transcriptome sequencing (RNA-sequencing) human brain, comparing fetal and adult using novel genomic data science approach"**, in partial fulfillment of the requirement for the award of the Degree of **Master of Science** and submitted to the institution is an authentic record of my own work carried out during the period of January-2020 to May-2020 under the supervision of Dr. Manoj Kumar Singh. I also cited the reference about the text(s) /figure(s) /table(s) /equation(s) from where they have been taken.

The matter presented in this thesis as not been submitted elsewhere for the award of any other degree or diploma from any Institutions.

Date: Signature of the Candidate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: Signature of the Research Supervisor

The Viva-Voce examination of *Ajay Kumar*, M.Sc. Student has been held on _________________.

Signature of
External Examiner

Signature of
Head of the Department

DEDICATION

My thesis is dedicated to the most important persons in my life

“My Late Grandpa (Nana) and Grandma (Nani)”

Who was always there whenever I needed, without their constant money generation this THESIS would never have been completed and without constant compelling to make me study I would never have been admitted to this institution.

ABSTRACT

Human genome comprises of set of nucleic acid sequences, encoded as DNA (De-oxy ribose nucleic acid) and RNA (Ribose nucleic acid) within the 23 (22-autosomes and 1-allosome) chromosome pairs in cell nuclei, which is habitually recognized as a genetic material that passes from generation to generation and is, responsible for the variation in an organism.

The objective of this study is to understand the fundamental biological research question of how human being grows from fetus to an adult. As physical changes in human body is directly associated by the signals, sends by brain thus one way to do this is to study that how the human brain changes over time. To explore the answer for above question, we may have several ways, but here we are trying to practice novel genomic data science approaches and tools that became popular in recent researches some of them includes Galaxy Virtual Lab, Command line tools such as Bowtie2, tophat, HISAT, DESeq, BioString, STAR, Bioconductor, R-programming et cetra.

Last, we propose analytical and theoretical research findings along with their reproducible environment links to encourage the practical reproducibility of this thesis.

Keywords: Genomic Data Science, RNA-sequencing, GALAXY, brain-age group, Homo-sapiens (hg19).

ACKOWLEDGEMENTS

First, and most of all, I would like to thank **Dr. Manoj Kumar Singh**, for his expertise, assistance, guidance, and trust throughout the process of writing this thesis. Without his help, this THESIS would not have been possible.

Second, I will be thankful of **Dr. Andrew E Jeff**, **Dr. Kasper Daniel Hansen** and **Prof. Liliana Florea** from Liber Institute for Brain Development (LIBD) and Department of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health University.

Third, I would like to show my gratitude for open source online platforms together with National Center for Biotechnology Information (NCBI), Coursera, GitHub, Stack overflow, Biostars et-cetra for providing quality datasets, suggestions, guidelines that actually helped in implementation.

Last but not the least, I would like to thanks my friends who helped me to contribute in the THESIS; and to my best friend **Varad Srivastava**, thanks for keeping me company on long walks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	ADENINE
T	THYMINE
G	GUANINE
C	CYTOSINE
U	URACIL
AI	ARTIFICIAL INTELLIGENCE
DL	DEEP LEARNING
ML	MACHINE LEARNING
ANN	ARTIFICIAL NEURAL NETWORK
DNA	DEOXYRIBOSE NUCLEIC ACID
RNA	RIBONUCLEIC ACID
RNN	RECURRENT NEURAL NETWORK
SNP	SINGLE NUCLEOTIDE POLYMORPHISM
CNN	CONVOLUTIONAL NEURAL NETWORK
CAS9	CRISPR ASSOCIATED PROTEIN 9
CRISPR	CLUSTERED REGULARLY INTERSPACED SHORT PALINDROMIC REPEATS
ENCODE	ENCYCLOPEDIA OF DNA ELEMENT
HDSR	HARVARD DATA SIENCE REVIEW
GDS	GENOMIC DATA SCIENCE
SRA	SEQUENCE READ ARCHIVE
RIN	RNA INTEGRITY NUMBER (FACTOR)
LIBD	LIEBER INSTITUTE OF BRAIN DEVELOPMENT
NCBI	NATIONAL CENTER OF BIOTECHNOLOGY INSTITUTE
TSV	TABLE SEPARATED VALUE
GVL	GALAXY VIRTUAL ENVIRONMENT

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 COMPUTER SCIENCE

As Biology is not about building microscope, music is not about building musical instruments; astronomy is not about building telescopes. Similarly, Computer Science is not *just* about building computers or writing computer program[1].

Computer Science is the systematic study of the expression, feasibility, mechanization and structure of the methodological process (or algorithms) that underlie the acquisition, representation, processing, storage, communication of, and access of information, whether such information is encoded in bits or bytes in computer memory or transcribed in genes and protein structure in the human cell. In other terms, it could be understood as a development of the protocols required for automated processing and manipulation of the data[1].

The period from 1791-1871 was the prehistoric iconic phase for the *computing* when English mathematician Charles Babbage proposed the *Difference Engine No. 1* was the first successful *automatic calculator* or *calculating engine* and also, the finest example of precision engineering of that time. Since then, Charles Babbage (born on 26thDec, 1791), the son of Benjamin Babbage (a London Banker) became the "Father of Computing". Babbage became a popular sovereign mathematician because at his twenties he was elected as a Fellow of the Royal Society in 1816 and then grew the interest in machinery in 1820, which infers that the computer science originates from mathematics particularly calculus, algebra et cetra. Later, in 1832 due to suspension of funding for his Difference Engine and after an agonizing waiting period, he ended his project in 1842 with only fragments of Babbage's prototype Difference Engine. During his agonizing waiting period he met a woman named *Ada Lovelace* (1815-1852) in a party when Babbage was demonstrating the small working section of the Engine[2].

As *Ada* was often referred to as *"the first programmer",* she speculated that the Engine might act upon other things besides numbers…the Engine might compose elaborate and scientific pieces of music of any degree of complexity or extend. This idea of machine-manipulated symbols in accordance with rules and that number represented entities other than quantity mark from calculation to computation. Later, she has been referred to as "*Prophet of the computer age*". Concluded that, she was the first to express the potential for computers outside mathematics[2].

However, Babbage devoted most of his time and large fortune towards construction of his Analytical Engine, but he never succeeded in completing any of the several designs for it. Then George Scheutz (A Swedish Printer), successfully constructed a machine based on Babbage's Design in 1854. That machine printed mathematical, astronomical and actuarial tables with unprecedented accuracy and was used by the British and American governments. Later Babbage's work was continued by his son, Henry Prevost Babbage (1871)[2].

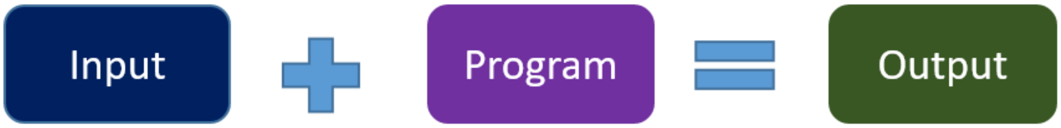
In 1985, the Science Museum in London began working on Difference Engine No. 2 using Babbage's original designs and ended the work with a calculating working device by 199. This was the initial starting of the computer science, later many scientists, professors, researchers contributed to extend it.

1.1.1. TRADITIONAL COMPUTER SCIENCE

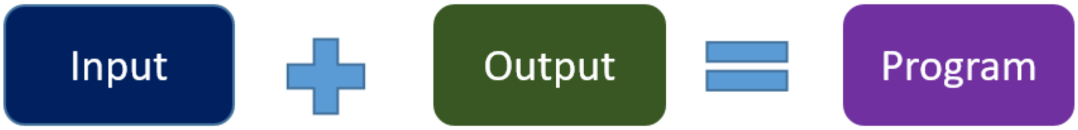
As far as, traditional computer and modern computer science is concern we cannot compare two entities based upon different parameters because parameters that we are assuming are changed after a certain time.

Although, we can compare traditional programming and modern programming. Traditional programming refers to manually creating a program that uses input data and runs on a computer to produce the output[3].

In this type of programming, we manually formulate the instructions/formula such as C, C++, FORTRAN, BASIC, COBOL, et cetra.



(Figure 1: Traditional programming)



(Figure 2: Modern programming)

1.1.2. MODERN COMPUTER SCIENCE

Computer Science is entering in a new generation from traditional programming to modern programming where we give the input and output to get the desired inferences or program. Here we automatically formulate the rules from the data that makes it more powerful.

It includes programming in Python, R-Programming, MATLAB et cetra.

The fields it includes are Data Science, Deep Learning, Machine Learning, Artificial Intelligence, et cetra.

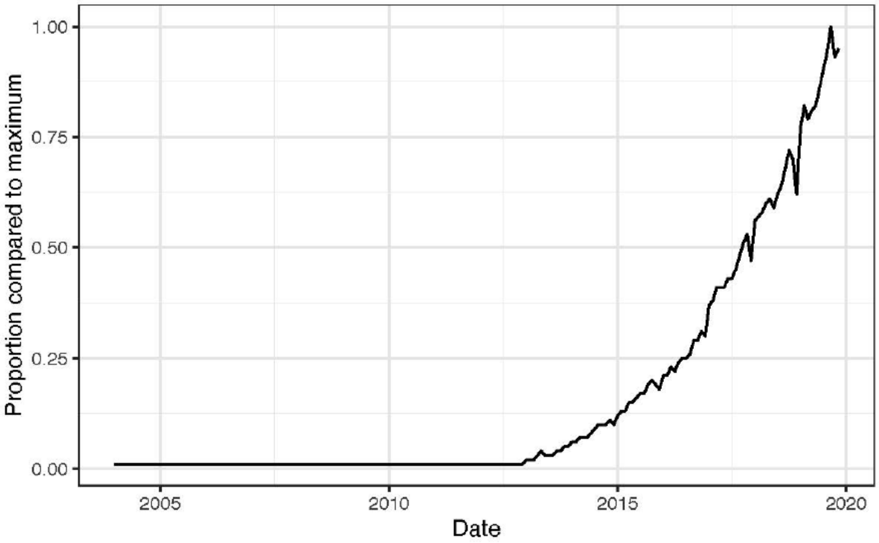
1.1.2.1 ARTIFICIAL INTELLIGENCE

As on date, there is no concrete definition for Artificial Intelligence (AI) (*Kirsh, 1991; Allen, 1998; Hearst and Hirsh, 2000; Brachman, 2006; Nilsson, 2009; Bhatnagar et al., 2018; Monett and Lewis, 2018*) because there could be many definitions of AI as per the practical or theorical application and people do not consider it a big problem[4]. AI term was first proposed by *John McCarthy* in 1956, for the basic understanding of AI, “the study and design of intelligent agents where an intelligent agent is a system that perceives its environment and takes actions which maximizes its chances of success”.

Artificial intelligence is also referred by other popular terms such as computational intelligence, machine intelligence or synthetic intelligence. AI is presently using for large industrial, research and academic applications such as, speech recognition, disease diagnosis, automate vehicle, image processing, AI games, personal finance, portfolio management, candidate’s selection for the job, heart sound analysis, et cetra.

1.1.2.2 DATA SCIENCE

Data Scientist: The Sexiest job of 21st century as per Harvard Business Review publication (*Davenport and Patil, 2012*). Data Science definition was first proposed in Harvard Data Science Review (HDSR), "the study of extracting value from data" (*Jeannette Wing, 2019*) and similar definition was proposed by American Statistical Association: "the science of learning from data and of measuring, controlling, and communicating uncertainty"[5].



(Figure 3 : Google Trends monthly term 'Data Science'[5]**)**

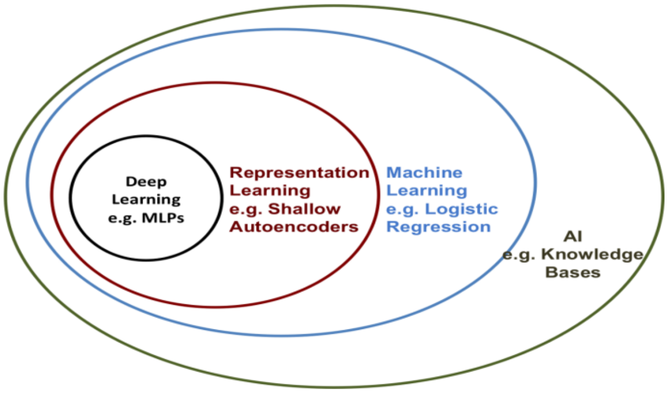
In the Review article, *"The Role of Academic in Data Science Education"* by *Rafael A. Irizarry* professor and chair department of Data Science *Dana-Farber* Cancer Institute and Department of Biostatistics Harvard, adapted the wing's definition and proposed that "*Data Science is an umbrella term to describe the entire complex and multistep processes used to extract value from data*", which is the best definition as on date.

1.1.2.3 MACHINE LEARNING

"*Machine Learning (ML) is most basic practice of using algorithms to parse data, learn from data and then make a prediction about the real world*" (*Definition by* NVIDIA).

"ML is an integral part of AI, and is the science of getting computers to act without being explicitly programmed" (Andrew Ng, Stanford).

ML is currently being use in vast fields like effective web search, understanding the human genome, audio, database mining, text understanding, building smart robots, anti-spam et cetra.



(Figure 4: Venn diagram representation)

1.1.2.4 DEEP LEARNING

Deep Learning (DL) is a class of machine learning methods capable of identifying highly complex patterns in large datasets[6]. DL is an integral part of artificial intelligence and used in the aggrandizement of ML.

DL is an umbrella that refers to the recent advances in neural networks and the corresponding training platforms (e.g. TensorFlow and PyTorch). The starting point of neural network is an artificial neuron, which makes as input a vector of real values and computes the weighted average of these values followed by a nonlinear transformation, which can be simple threshold. The weights are the parameter of the model that is learned during training. The power of neural networks stems from individual neurons being modular and composable, despite their simplicity. The output of one neuron can be directly fed as input into other neurons. By composing, neurons together, a neural network is created.

1.2 BIOINFORMATICS

Bioinformatics is an interdisciplinary field of biology and computer science concerned with the acquisition, storage, analysis, and dissemination of biological data, most often genetic material and amino acid sequences[7].

On the other way, it could be understood as a field of computational science that usually deals with the analysis of biological genes, Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA), or protein (Amino acids). It is particularly useful in comparing genes and other sequences like proteins between organisms or within an organism. It is also use to determine the evolutionary relationships between organisms, and many more[7].

1.2.1. GENOMIC SCINECE

Genomic Science is also referred as “Genomics”, which is an interdisciplinary study of whole genome of an organism, and incorporate elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes[8].

Genomics can be understood as the large-scale study of mass genes (mass genes can be all the genes of an organism or multiple organisms). It is different from traditional Genetics field, where we used to stick to the study of one gene or one gene product at a time. The field of genomics can be further divided into following categories[9]:

i. **Comparative genomics:** It can be used to define the important structural sequences that are identical in many genomes and to detect evolutionary changes across genomes.

ii. **Structural genomics:** It is a physical nature of the genome, includes sequencing and mapping of genomes.

iii. **Functional genomics:** It involves studying the expression and function of the genome.

(Table 1: **Difference between genetics and genomics)**

Biology and Genetics	***Genomics***
Target studies of one or few genes ↔	Studies considering all genes in the genome
Targeted, Low-throughput experiments ↔	Global, High-throughput experiments
Clever experimental design, Painstaking experimentation ↔	Tons of data, Uncertainly, Computation

As in above table, it is clearly explained that in traditional genetics field of study, we used to target a single gene but presently, we have developed a resource to work on entire genome of an organism.

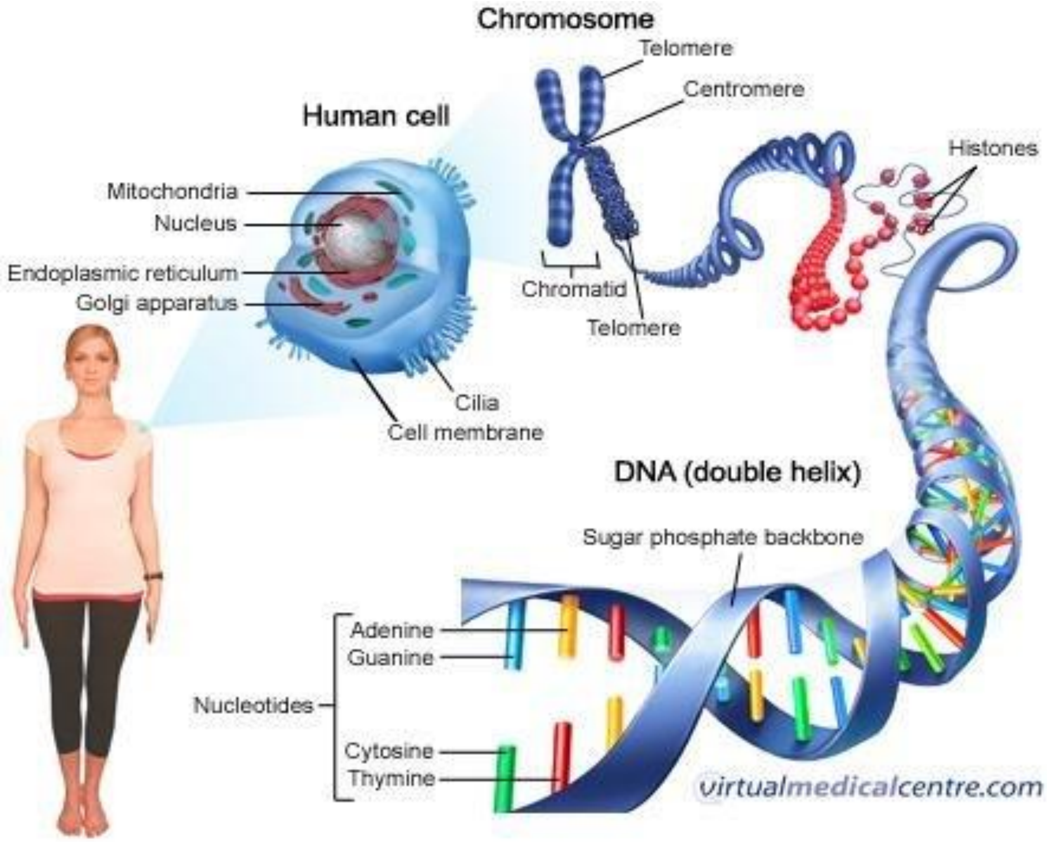
As far as we move forward towards better understanding of genomic sciences the challenges to deal with, is also increases like computing the high-throughput experiments is the biggest challenge genomic data scientists are facing on date.

1.2.2. HUMAN GENOME SCIENCE

The basic structural and functional unit of every living organism is a cell. [*Robert Hooke* 1665, *Anton van Leeuwenhoek* 1664, *Theodor Schwann and Matthias Schleiden*, 1838-39]. A single cell contains numerous elements including cell membrane, Cell wall, Golgi apparatus, Mitochondria, Nucleus, and Ribosome et-cetra. A solitary human body carries 3.72 x 10^{13} (37.2 Trillion) cells [10]. The information of characteristics passes from generation to generation, is stored inside the nucleus of the cell (known as nuclear DNA/genetic material). The nucleus of the cell contains 23 pairs of chromosomes, out of which 22 are acknowledged as an autosomal chromosome and the remaining one is termed as allosome chromosome or sex chromosome. Allosome is different from the other 22 pairs of chromosomes in the context of function and structural composition. Each chromosome is a long chain of nucleic acid, which approximates two-meter in length, recognized as a Deoxyribonucleic acid (DNA)/Ribonucleic acid (RNA) molecules. The major ration it contains is DNA, which is a bio-polymer made of two polynucleotide chains, often referred as a 'Double Helix' and, is a bridge connection between instructions contained in the chemical compound and physical activities of the organism. Each strand is made of four biological structures (Refer ***(Figure 5: Detailed diagrammatical representation of human cell to nitrogenous bases)*** known as nitrogenous bases, including Adenine (A), Guanine (G), Cytosine (C) and Thymine (T)/Uracil (U) [11].

In Ribonucleic acid (RNA), Uracil is present at the place of Thymine, where A and T/U links with double hydrogen bond while, C and G links with triple hydrogen bond inside the double-helical structure of the DNA.

In an analogy, as a set of alphabetical letters could combine and construct to form a meaningful word in the similar manner, a chain of predefined sequence of set of nitrogenous bases (A, T/U, G, and C) binds with their corresponding protein and performs a valid function.



(Figure 5: Detailed diagrammatical representation of human cell to nitrogenous bases)

Image credit: virtualmedicalcentre.com

The organism’s whole set of genetic material/DNA is called its Genome. The human’s nuclear genome comprises of approximately 3.2 Billion nucleotides of DNA [11]. All the information from replicating a single cell to formation of a complete new organism is stored in complex chemical molecule called a genetic material of an organism, which could be DNA/RNA.

In earlier era, DNA/RNA were both used for storing information but later it realized RNA is mutagenic and unstable thus it is not used to store information. Thus, DNA is use to information in an organism.

Transcriptome: The human genome is made up of DNA (deoxyribonucleic acid), a long, zigzagging molecule that contains the instructions needed to build and maintain cells. These instructions are spelled out in the form of "base pairs" of four different chemicals, organized into 20,000 to 25,000 genes. For the instructions to be carried out, DNA must be "read" and transcribed into RNA. These gene readouts are called transcripts, and its collection of all the gene readout in the cell is known as transcriptome [12].

Proteome: The term proteome was coined in 1994 by Marc Wilkins (then a postdoctoral fellow at Macquarie University, Sydney). In analogy to the term genome, the proteome represents the total protein repertoire able to be expressed from a given genome. The word has rapidly evolved to encompass diverse meanings; not just the proteome of an organism but also the proteome of a cell, tissue, or organ, referring to the set of proteins expressed in a particular cell, tissue, or organ at a particular time and under particular conditions. A proteome is the complete set of proteins expressed by an organism. The term can also be used to describe the assortment of proteins produced at a specific time in a particular cell or tissue type [13].

Out of the function and comparative genomics this thesis particularly focuses on functional genomics because, computational analysis of the sequences, using modern genomic techniques developed so far and we can use them to make a prediction of all the encoded proteins. We can also, use this strategy to observe difference between two haploid genomes of the same species and could develop a statistical picture to proof the analysis. Similarly, we can use also compare genomic expressions of different species to make infers that how two species are related to each other. This field of genomics is also referred as “comparative genomics”.

Population Genetics: Population genetics is the study of genetic variation within populations between individuals, and involves the examination and modelling of changes in the frequencies of genes and alleles in populations over space and time. Many of the genes found within a population will be **polymorphic** - that is, they will occur in a number of different forms (or **alleles**). Mathematical models are used to investigate and predict the occurrence of specific alleles or combinations of alleles in populations, based on developments in the molecular understanding of genetics, Mendel's laws of inheritance and modern evolutionary theory. The focus is the population or the species - not the individual.

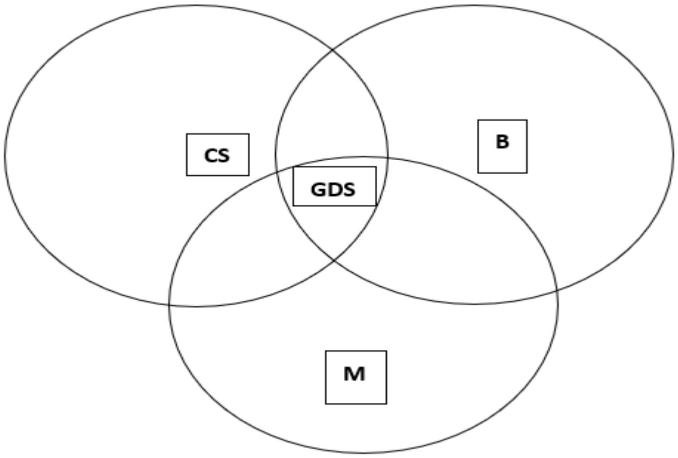
Example for use of population genetics is that it can be used to determine your ancestry from your DNA sequence, or for identifying, the mutations that help certain populations adapt to a new environment. It is an area of evolutionary science that has been around for more than century, and it has really benefitted in the last decade from an explosion of new DNA sequencing technologies.

NOTE: Genomics and population genetics are independent field of study as like astronomy and optics. Those who work on genetics are better known as **Geneticists**. They work on how different part of genomics work and they use genomics too. You can conclude with genetics and genomics are not separable

1.3 GENOMIC DATA SCIENCE

Genomic Data Science (GDS) is an interdisciplinary field of enormous number of fields of study that apply statistics and data science to the genome. Genome sparks a revolution in medical discoveries, it becomes imperative to be able to better understand the genome and be able to leverage the data and information from genomic datasets [14].

This interdisciplinary field is different from Bioinformatics (Bioinformatics is a large umbrella) and GDS focuses on the genome particularly.



(Figure 6: Venn-Diagram Representation of Genomic Data Science)

CS- Computer Science; B- Biology; M-Mathematics; GDS- Genomic Data Science

As mentioned in the Specialization of "*Genomic Data Science*" offered by *Johns Hopkins University* through *James Taylor, Jacob Pritt, Liliana Florea, Kasper Daniel Hansen, Mihaela Pertea, Steven Salzberg, Jeff Leek* and *Ben langmead* the genomic technologies include, Python, R-programming, Bioconductor, Galaxy, command line tools and more, for the genome analysis to answer fundamental biological research questions [14]. Genomic Data Science is sometimes also referred to as *Computational genomics, Computational Biology, Bioinformatics*, and *Statistical Genomics*. Those who work in genomic data science are known as *Genome Data Scientists.*

1.4 THESIS ORGANIZATION

This thesis/dissertation shows how a realistic real time project could help genomic data science learners to understand the performance characteristics of data science for genome analysis. The organization of this thesis is as follows:

Chapter 1 is about the quick overview of subject knowledge that you should have for further understanding of the thesis. The last section in this is an introduction to genomic data science.

Chapter 2 provides the highlights of the particular problem that is being focused as an instance along with the previous research work done.

Chapter 3 and 4, are for the core and in-depth analysis of the fundamental biological problem of identifying changes in brain signals in fetal and adult in RNA sequences. In this section, we have provided the link of the dataset to reproduce the results. As we strongly focus on reproducibility of the research insights by another person in another environment thus providing a link of all the software used, dataset used and algorithm for better analysis.

Chapter 5, we have provided the insights of the analysis along with the interpretation, which is again a challenging task to analyze biological data.

Chapter 6, In this section, we have summarized the results, inferenced the results and tried to support our answer for the mentioned fundamental biological research question.

References are mentioned in the end to support the theory building block of our thesis.

Plagiarism report is also attached for proving that work is genuine.

Chapter 2

LITERATURE REVIEW

2.1PROBLEM INTRODUCTION

The fundamental question in human biology is to understand how we develop from fetus to a grown human being. To search the answer for above question, we may have several ways, but here we see this problem as with a vision of genomic data scientist.

As physical changes in human being is directly associated by signals sends by brain thus one way to do this is to study how the human brain changes over time. In a recent study, Jaffe et al. measured gene expression* from different individuals across the human lifespan. They were looking for genes that showed patterns of expression that changed over time as people aged.

The primary data generated by this group is transcriptomic-sequencing data (RNA-Seq) from human post-mortem brains, sequenced on the Illumina 1.5 Sequencing platform.

Gene expression: It is the process by which the instructions in our DNA is converted into a functional product, such as protein.

RESEARCH QUESTION: WHAT IS THE DIFFERENCE BETWEEN GENE EXPRESSION OF FETAL AND ADULT BRAINS?

We try to answer this question in further sections of the thesis.

2.2 PROBLEM MOTIVATION

Most of the work described in this thesis was conducted at the Department of Computer Science, Banaras Hindu University and Center of Mathematics Sciences (CMS), Banaras Hindu University Varanasi, India.

The reason for conducting the research work at this university were: 1) Part of master’s thesis. 2) Involvement of eminent supervisor highly interested to explore the new emerging field of GDS. 3) I have been actively involved in the research associated with genomic data sciences inside or outside the university. 4) High-end computational resources at CMS. 5) Sponsorship of coursera courses for university students.

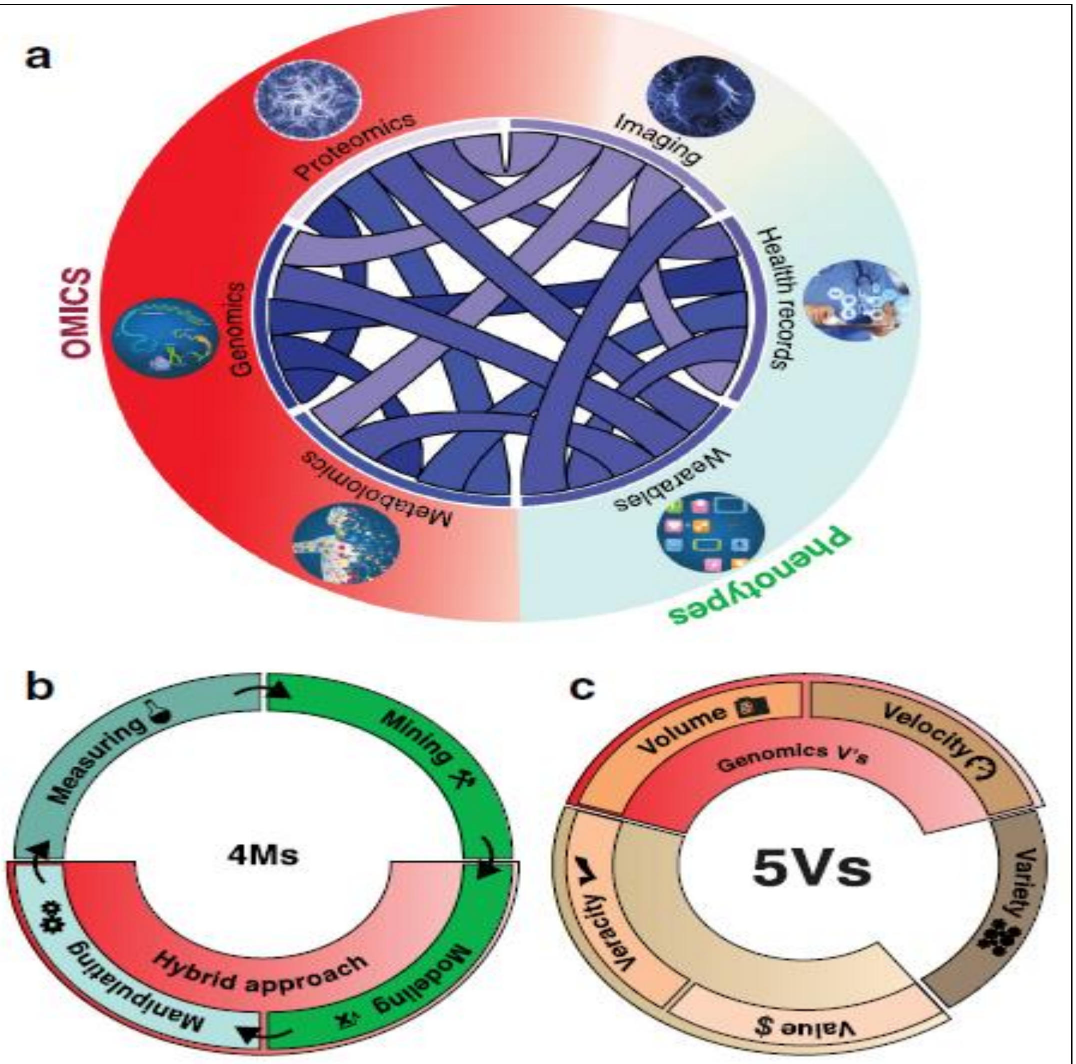
GDS always instigated me to explore it and contribute something great, even we had implemented a paper on “The discovery of transcription factors binding site in DNA” [6] : A primer on deep learning in genomics, during mini-project (August,2019 – December,2019).

2.3 RELATED WORKS

In a research paper, "*Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis*" Prof. Andre E. Jaffe et al explained that using developmental, genetic and illness-based RNA-seq expression analysis in human brain transcriptome around these loci and found enrichment for developmentally regulated genes with novel examples of shifting isoform usage across pre and post-natal life and also, provided resources to reproduce the results to encourage reproducibility [15].

In [16], Prof. Fábio C. P. Navarro at al, bring together the data science for genomics and contextualize data science as umbrella term, covering several subdomains. However, focused on how genomics fits as a specific application, in terms of renowned 3V and 4 M process frameworks (volume-variety-velocity and measurement-mining-manipulation-modeling respectively).

This project uses data generated during exploration of gene expression changes across human brain development, looking from fetal samples and adult age samples. This research is motivated by previous research that explored these changes and found widespread differences in gene expression comparing fetal to later in life samples. They used microarray technologies, which used pre-defined probe sequences to only query known gene sequence, and additionally existing RNA sequencing data sets like the BrainSpan project only has existing feature counts like genes and exons, which might also limit biological discovery for age related changes in expression. Thus, we opted to work on unbiased snapshot of the transcriptome RNA-sequencing data generated at the Lieber Institute for Brain Development (LIBD).



(Figure 7: Biological data science representation a. biological data science emerged at the confluence of connecting genomics, metabolomics and more to statistics and computer science. B. 4 M processes c. 5 V data framework)

Image credit: [8]

Chapter 3

DESIGN DETAILS

3.1 DESIGNED WORKFLOW

As implementation and strategy to analyze, the dataset is quite complex and also, it is irritating to deal with large high sequencing datasets. Thus, we have defined the workflow to execute it in an order.

Algorithmic steps for the analysis are as follows:

1. Download the raw sequence data and meta-data (phenotypic information and technological information) from the public database.
2. Preprocess the data.
3. Align the data with reference human genome (hg19).
4. Perform quality control on the aligned data.
5. Calculate express measurement at gene count level.
6. Perform exploratory analysis to identify major feature of the data and figure out which model is to build.
7. Fit this statistical model to identify genes that are different from fetal human brain to adult human brain.
8. Integrate results and answer the fundamental biological question.

3.2 CORRESPONDING ALGORITHMS

The designed workflow in previous section, involved all the necessary steps to be taken for the analysis. But going through the corresponding algorithms from the desired tool that we will be using is optional yet important step for deep understanding the behind descriptive statistical working of the analysis.

The following tools/algorithms/softwares are being use for the analysis:

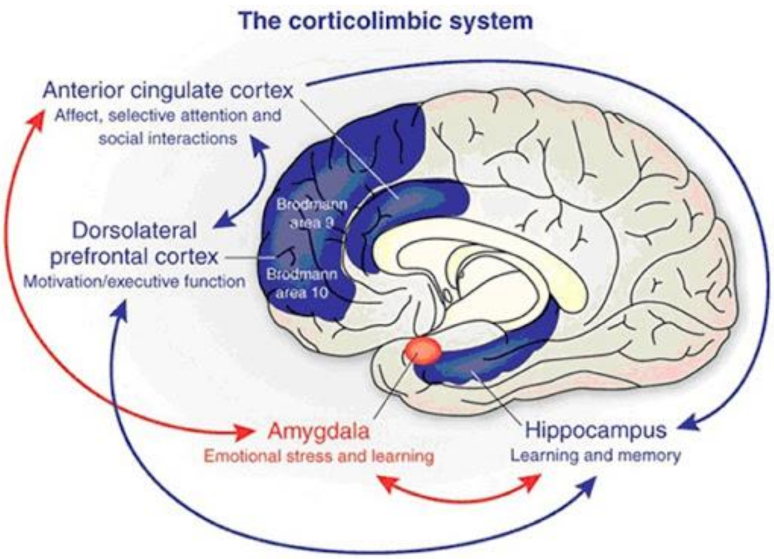
i. Galaxy Virtual Lab (GVL-GUI based workflow analysis environment)[17]
ii. Bowtie2 (For aligning the raw RNA-Seq data)[18]
iii. FastQC [19]
iv. FeatureCount [20]
v. R-programming and Bioconductors [21-22]
vi. DESeq [22]

3.3 DATASET

A data set is any permanently stored collection of information usually containing either case level data, aggregation of case level data, or statistical manipulations of either the case level or aggregated survey data, for multiple survey instances [23].

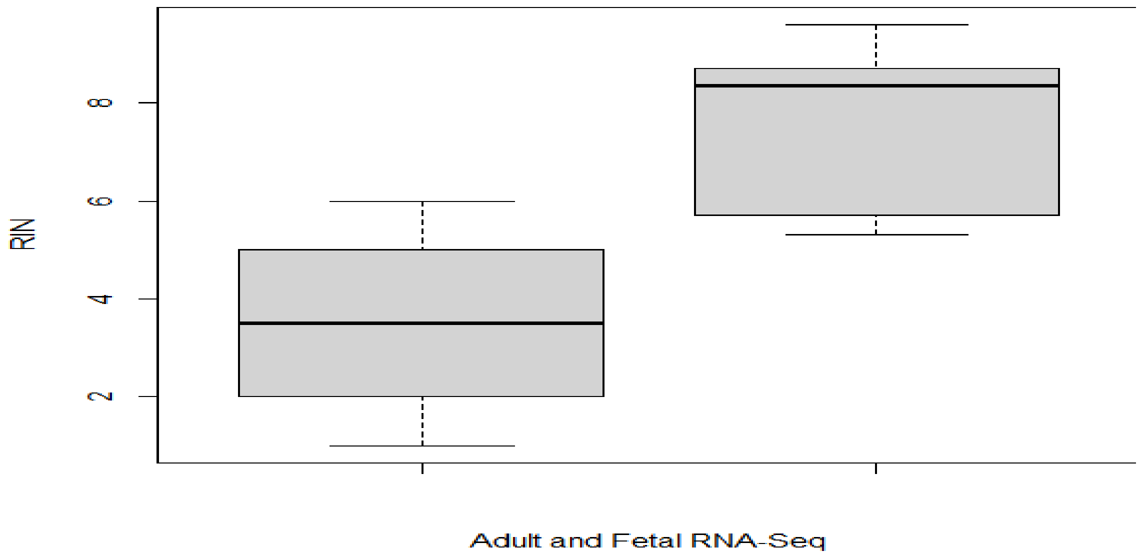
In our analysis, we have used Polyadenylated (PolyA+) RNA-transcripts human genome data of dorsolateral pre frontal cortex part of the brain and the data was sequenced by the next generation sequencing method at Illumnia platform.

We have taken six samples, three from the fetal and three from the adult brain. These datasets are very high quality dataset samples produced for the post-mortem human research because it is balanced and matched with potential confounding variables such as RNA integrity number (RIN), which measures the quality of the RNA and post mortem intervals, which is potentially also measure the quality of the sample.


The corticolimbic system
Anterior cingulate cortex
Affect, selective attention and social interactions
Dorsolateral prefrontal cortex
Motivation/executive function
Brodmann area 9
Brodmann area 10
Amygdala
Emotional stress and learning
Hippocampus
Learning and memory

(Figure 8: Dorsolateral prefrontal cortex view of the brain)

Image credit: Gerry Leisman,



(Figure 9: SRA and RIN Bar plot)

This project uses data generated during exploration of gene expression changes across the human brain development, looking from fetal samples to old age.

In the **(Figure 9: SRA and RIN Bar plot)**, the RIN of fetal is relatively larger than RIN of adult. The six samples taken for the analysis having corresponding short read archives (SRA):

1. SRR1554535 (Adult)
2. SRR1554536 (Adult)
3. SRR1554561 (Adult)
4. SRR1554541 (Fetal)
5. SRR1554537 (Fetal)
6. SRR1554567 (Fetal)

(As our end objective is to encourage the reproducibility of the dissertation. Thus, we have men-tioned the URL for downloading the data in the **reproduce results/implementation** subsections)

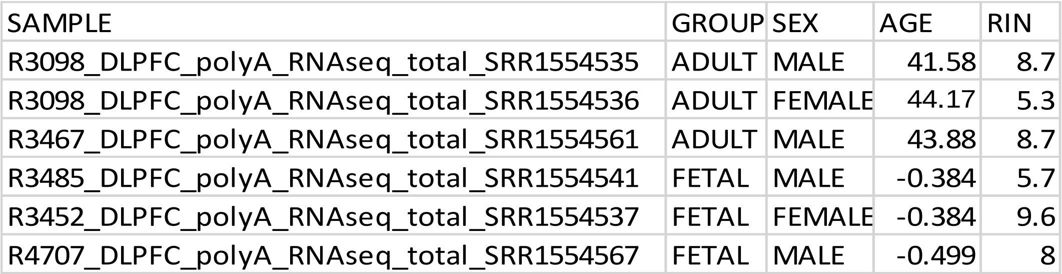
Chapter 4

IMPLEMENTATION

4.1 COLLECTING RAW DATA/METADATA/SAMPLES

The first step towards, implementation is “Download the raw sequence data and meta-data (phenotypic information and technological information) from the public database”. In genome data analysis, ‘collecting the data’ is one of the challenging task to obtain a consistent and high quality dataset.

As mentioned in Dataset subsection of Design details, we have collected a set of six data samples; Three from fetal and three from adult. For analysis, we are considering biological material dataset provided by Lieber Institute of Brain Development (LIBD) uploaded at National Center of Biotechnology Institute (NCBI) website. The dataset is uploaded after a certain period of post-mortem interval, which is another measure for the quality of the data.



SAMPLE	GROUP	SEX	AGE	RIN
R3098_DLPFC_polyA_RNAseq_total_SRR1554535	ADULT	MALE	41.58	8.7
R3098_DLPFC_polyA_RNAseq_total_SRR1554536	ADULT	FEMALE	44.17	5.3
R3467_DLPFC_polyA_RNAseq_total_SRR1554561	ADULT	MALE	43.88	8.7
R3485_DLPFC_polyA_RNAseq_total_SRR1554541	FETAL	MALE	-0.384	5.7
R3452_DLPFC_polyA_RNAseq_total_SRR1554537	FETAL	FEMALE	-0.384	9.6
R4707_DLPFC_polyA_RNAseq_total_SRR1554567	FETAL	MALE	-0.499	8

(Table 2: Phenotypic information of the sample datasets)

As per our objective is to, not only obtain the data and support the answer for the question but also to focus on reproduce the results. Thus, we have provided the links for the dataset below:

To download the dataset:

1. Link for first sample (SRR1554535-ADULT): https://www.ncbi.nlm.nih.gov/biosample?LinkName=sra_biosample&from_uid=956724

2. Link for second sample (SRR1554536-ADULT): https://www.ncbi.nlm.nih.gov/biosample?LinkName=sra_biosample&from_uid=956725

3. Link for third sample (SRR1554561-ADULT): https://www.ncbi.nlm.nih.gov/biosample?LinkName=sra_biosample&from_uid=956750

4. Link for fourth sample (SRR1554541-FETAL): https://www.ncbi.nlm.nih.gov/biosample?LinkName=sra_biosample&from_uid=956730

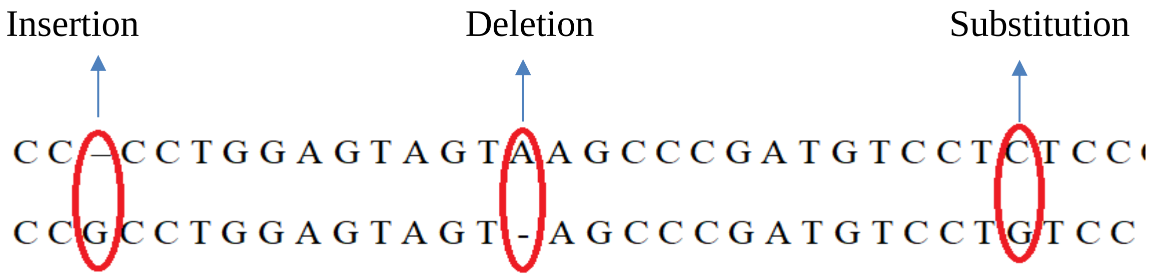
5. Link for fifth sample (SRR1554537-FETAL): https://www.ncbi.nlm.nih.gov/biosample?LinkName=sra_biosample&from_uid=956726

6. Link for sixth sample (SRR1554567-FETAL): https://www.ncbi.nlm.nih.gov/biosample?LinkName=sra_biosample&from_uid=956757

4.2 ALIGNING THE GENOME DATASET [16-17]

Once we have obtained the data, the next step is to perform alignment.

Alignment: It is a genomic feature that maps between the letters of the two sequences, with some spacers (indels)[7]. We perform alignment to inhibit polymorphism, or the induced error introduced while sequencing. The alignment will take into account differences such as polymorphism and sequencing errors, and introns (for genes).


Insertion
Deletion
Substitution
C C – C C T G G A G T A G T A A G C C C G A T G T C C T C T C C
C C G C C T G G A G T A G T - A G C C C G A T G T C C T G T C C

(Figure 10:Alignment in genomics (Indels))

In above sequences, sequence read is matching with its corresponding nitrogenous base but some places are empty known as insertions or deletions and some are wrongly filled known as substitution. These errors are corrected in the alignment phase.

For alignment, we can use spliced aligner including Tophat2, HISAT, STAR tools/algorithms But in this study, we have used **Bowtie2** alignment tool. Because Bowtie2 is, time consuming and memory-efficient tool for aligning reads to long sequences or high-sequencing throughput. It is predominantly better at aligning sequences of about 50s to 100s of characters to relatively long mammalian genomes and it uses 3.2 gigabytes of RAM that makes more feasible to work it on GVL.

Bowtie2 takes a bowtie2 index and a set of sequencing read files and outputs a set of alignments in SAM format.

For short, alignment is the process by which we determine how much genetic fragment is similar to reference genome.

4.2.1 END-TO-END ALIGNMENT VERSUS LOCAL ALIGNEMENT:

End-to-End	**Local-alignment**
It maps the alignments involving all of the read characters without trimming or clipping the alignment.	In this bowtie2 performs clipped or trimmed character read to enhance the alignment score.
Example:	Example:
Read: GACTGGGCGATCTCGACTTCG	Read: GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG	Reference: TAACTTGCGTTAAATCCGCCTGG

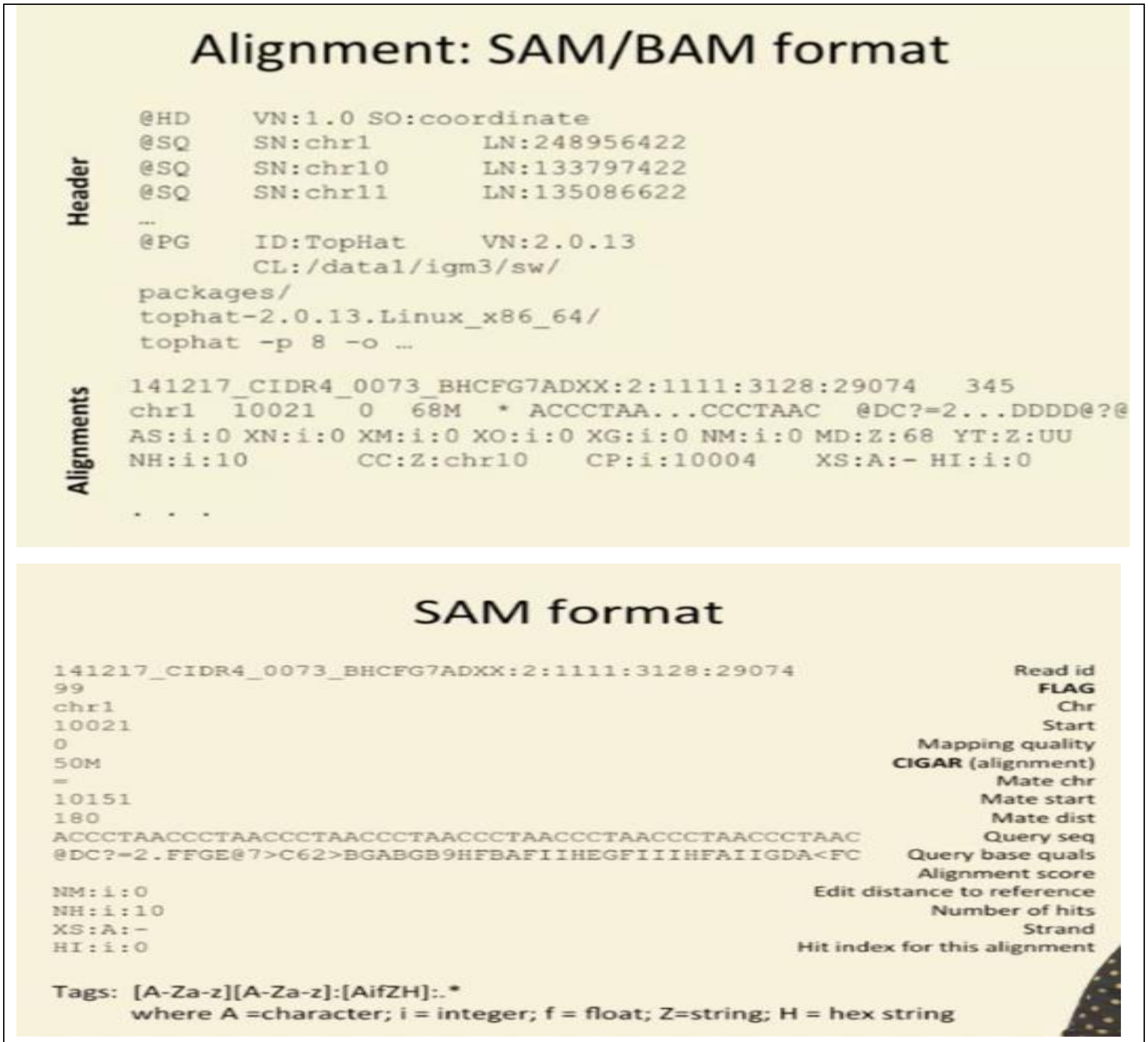
After Alignment:
`Read:          GACTGGGCGATCTCGACTTCG`
`               |||||  |||||||||| |||`
`Reference:     GACTG--CGATCTCGACATCG`
	After Alignment:
	`Read:      ACGGTTGCGTTAA-TCCGCCACG`
	`                ||||||||| ||||||`
	`Reference: TAACTTGCGTTAAATCCGCCTGG`

*No trimming in the reads.	*First 4 characters and last 3 characters are soft trimmed/soft clipped.

(Table 3: Difference between End-to-End and Local-Alignment)

Note: By default, Bowtie2 performs **End-to-End alignment**. Higher will be the alignment score, more similar reads will be.

After accomplishing the alignment procedure by Bowtie2, it will give the standard representation of Next Generation Sequencing (NGS) sequencing data format in SAM/BAM format.



Alignment: SAM/BAM format

Header

```
@HD     VN:1.0 SO:coordinate
@SQ     SN:chr1         LN:248956422
@SQ     SN:chr10        LN:133797422
@SQ     SN:chr11        LN:135086622
…
@PG     ID:TopHat       VN:2.0.13
        CL:/data1/igm3/sw/
packages/
tophat-2.0.13.Linux_x86_64/
tophat -p 8 -o …
```

Alignments

```
141217_CIDR4_0073_BHCFG7ADXX:2:1111:3128:29074   345
chr1  10021  0  68M  * ACCCTAA...CCCTAAC  @DC?=2...DDDD@?@
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:68 YT:Z:UU
NH:i:10       CC:Z:chr10    CP:i:10004    XS:A:- HI:i:0

. . .
```

SAM format

```
141217_CIDR4_0073_BHCFG7ADXX:2:1111:3128:29074                        Read id
99                                                                       FLAG
chr1                                                                      Chr
10021                                                                   Start
0                                                             Mapping quality
50M                                                         CIGAR (alignment)
=                                                                    Mate chr
10151                                                              Mate start
180                                                                 Mate dist
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC                Query seq
@DC?=2.FFGE@7>C62>BGABGB9HFBAFIIHEGFIIIHFAIIGDA<FC         Query base quals
                                                              Alignment score
NM:i:0                                                Edit distance to reference
NH:i:10                                                        Number of hits
XS:A:-                                                                 Strand
HI:i:0                                              Hit index for this alignment
```

Tags: [A-Za-z][A-Za-z]:[AifZH]:.*
where A =character; i = integer; f = float; Z=string; H = hex string

(Figure 11: BAM/SAM file format)

After performing alignment, the results are following:

1. For dataset sample one, SRR1554535
   91185969 reads; of these: 91185969 (100.00%) were unpaired; of these: 8945201 (9.81%) aligned 0 times 56816398 (62.31%) aligned exactly 1 time 25424370 (27.88%) aligned >1 times **90.19% overall alignment rate**

2. For dataset sample second, SRR1554536
   46971267 reads; of these: 46971267 (100.00%) were unpaired; of these: 2130239 (4.54%) aligned 0 times 17434738 (37.12%) aligned exactly 1 time 27406290 (58.35%) aligned >1 times **95.46% overall alignment rate**

3. For dataset sample third, SRR1554561
   89061863 reads; of these: 89061863 (100.00%) were unpaired; of these: 12590443 (14.14%) aligned 0 times 57693153 (64.78%) aligned exactly 1 time 18778267 (21.08%) aligned >1 times **85.86% overall alignment rate**

4. For dataset sample fourth, SRR1554541
   162403466 reads; of these: 162403466 (100.00%) were unpaired; of these: 17489969 (10.77%) aligned 0 times 112129877 (69.04%) aligned exactly 1 time 32783620 (20.19%) aligned >1 times **89.23% overall alignment rate**

5. For dataset sample fifth, SRR1554537
   121992326 reads; of these: 121992326 (100.00%) were unpaired; of these: 13500649 (11.07%) aligned 0 times 83156647 (68.17%) aligned exactly 1 time 25335030 (20.77%) aligned >1 times **88.93% overall alignment rate**

6. For dataset sample sixth, SRR1554567
   136312421 reads; of these: 136312421 (100.00%) were unpaired; of these: 14107633 (10.35%) aligned 0 times 94795954 (69.54%) aligned exactly 1 time 27408834 (20.11%) aligned >1 times **89.65% overall alignment rate**

NOTE: We have used Homo sapiens (hg19) genome for referencing the sample datasets.

The alignment is performed on GVL, to support the reusability or reproducibility we are providing the link of workflow for aligned sequence.

[Link: https://usegalaxy.org/u/ajay.ducs/h/genomic-data-science-capstone]

The output files are in BAM file format could be downloaded from above link to perform downstream analysis.

4.3 PERFORM QUALITY CONTROL ON ALIGNMENTS

After performing alignment, we get the BAM format, which further undergoes downstream analysis and we then, perform quality control using JAVA based tool, FastQC [19].

FastQC: It provides a simple way to do some quality control checks on a raw sequence data coming from high throughput sequencing pipelines. It provides set of analysis, which can use to give a quick impression of whether your data has any problems of which should be aware before doing any further analysis [19].

FastQC has following main functions:

- Import of data from BAM/SAM/FastQC file format.
- Provide a quick overview to tell you in which areas there may be problems.
- Summary graphs and tables to quickly access your data.
- Export of results to an HTML based permanent report.
- Offline operation to allow automated generation of reports without running the interactive e application.

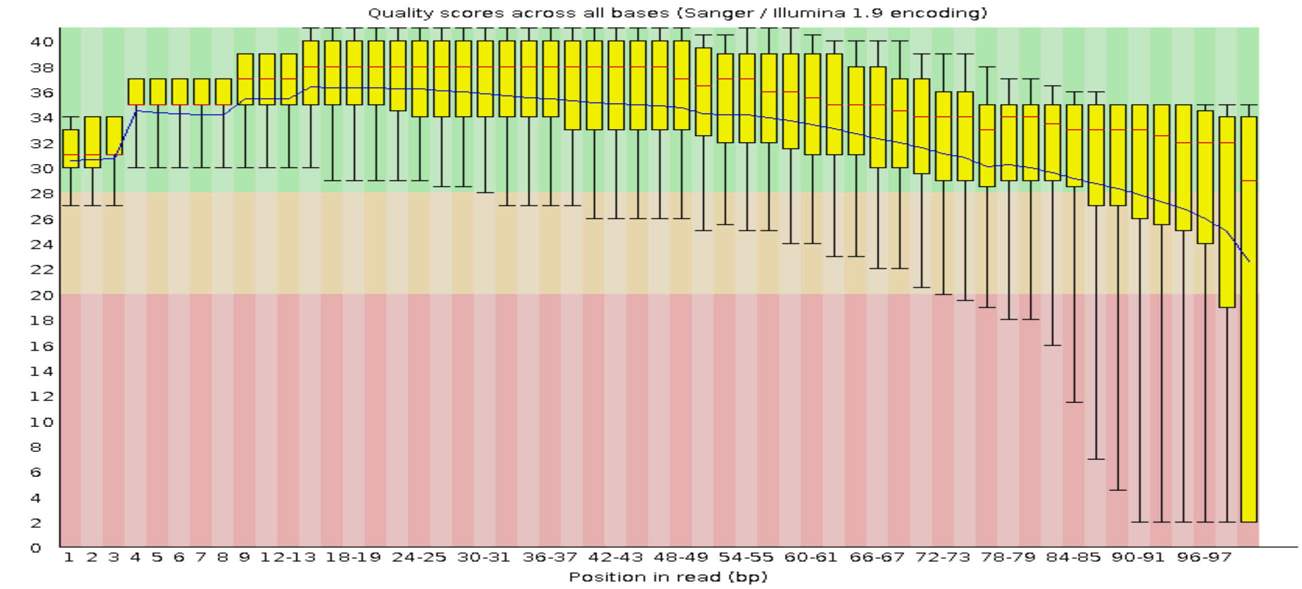
After performing FastQC operation, we have following reports for our sample datasets:

1. FastQC Report on First dataset sample: (SRR1554535)

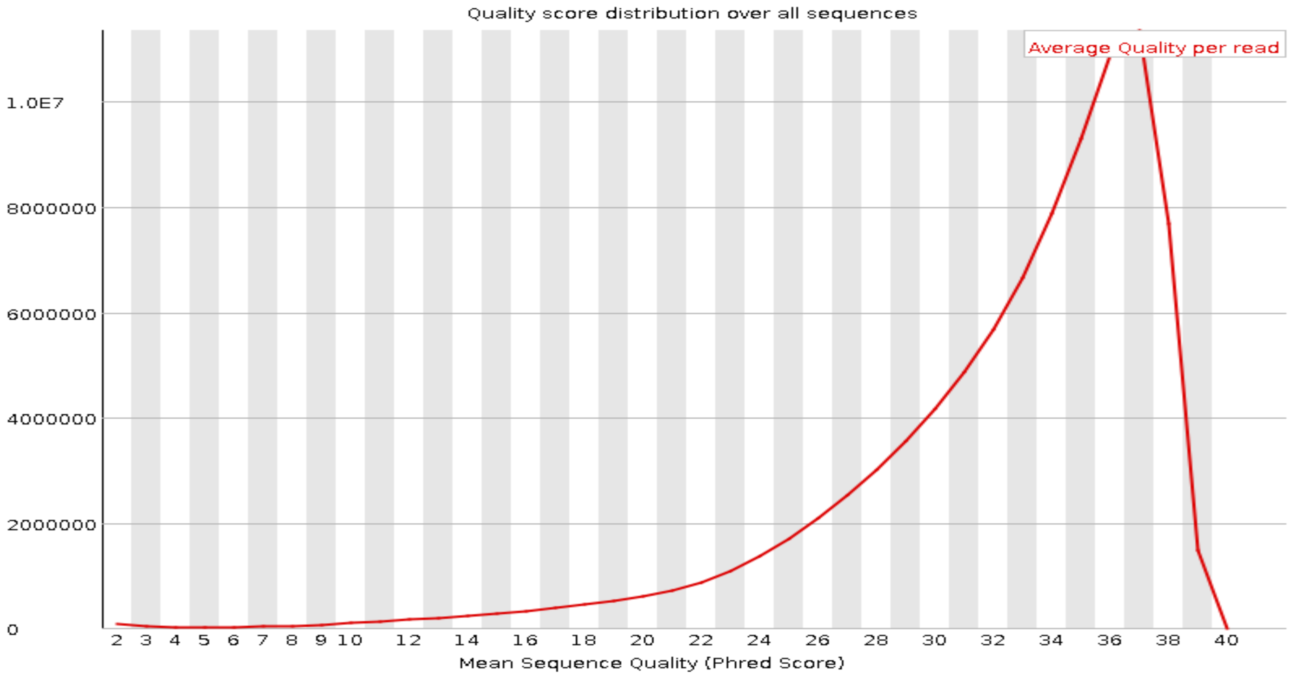
BASE STATISTICS

Measure	Value
Filename	Bowtie2 on data 5_ alignments
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	91185969
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

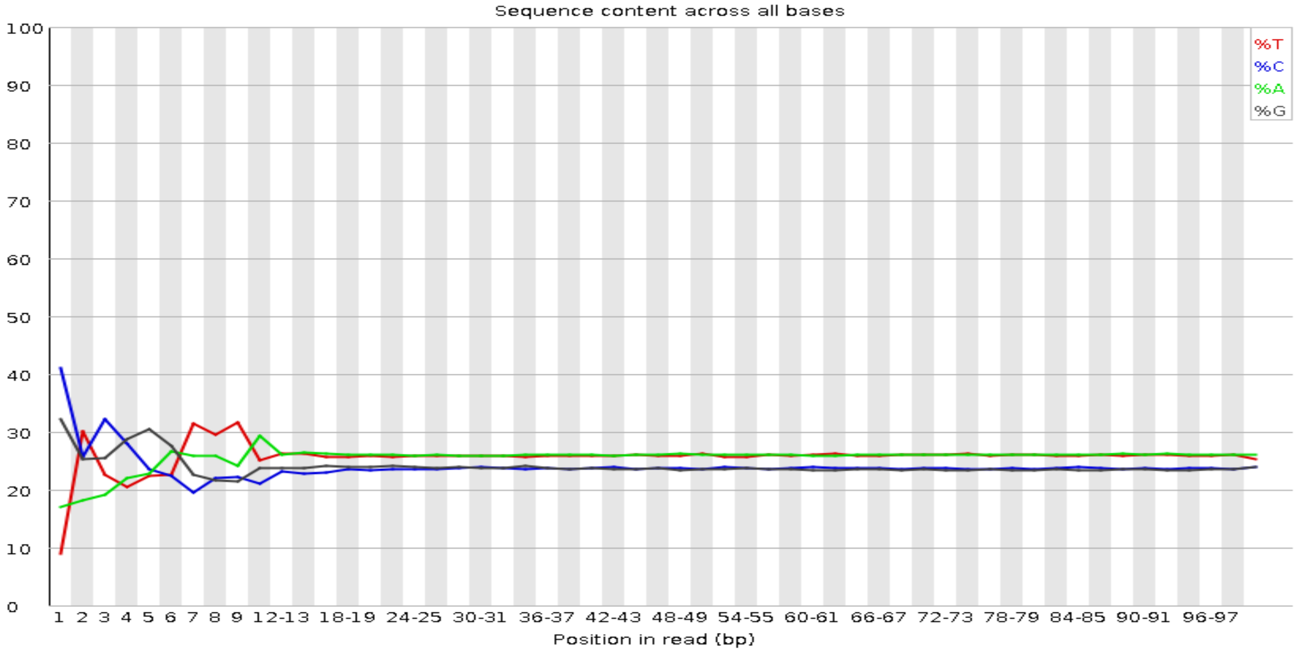
PER BASE SEQUENCE QUALITY



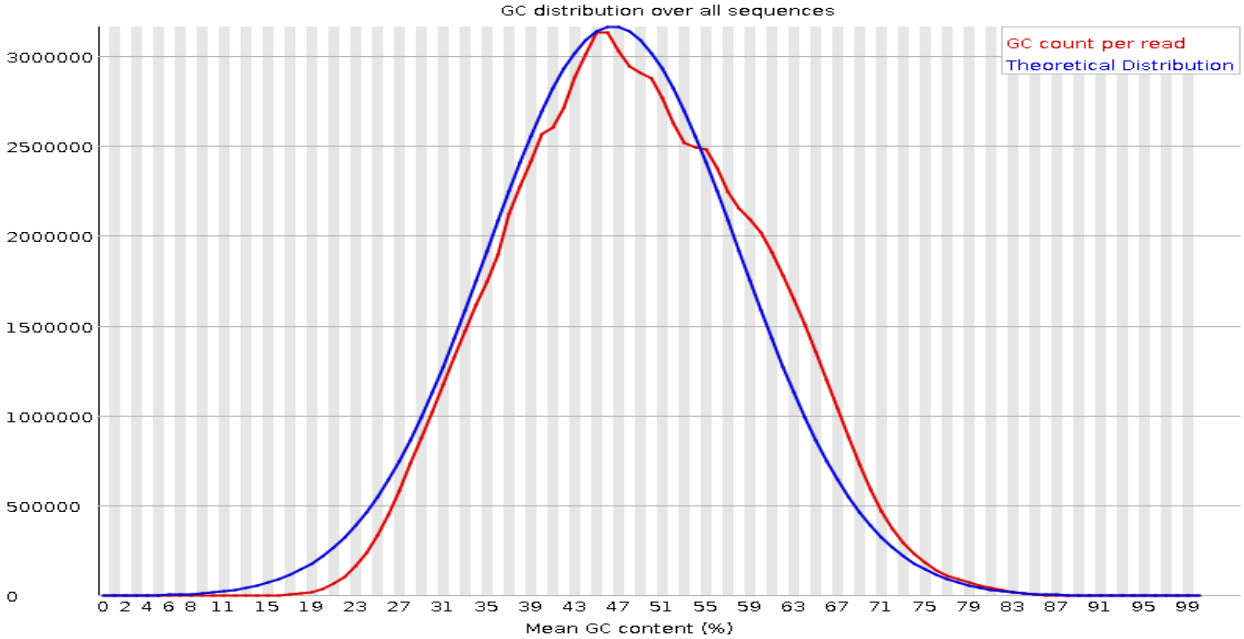
PER SEQUENCE QUALITY SCORES



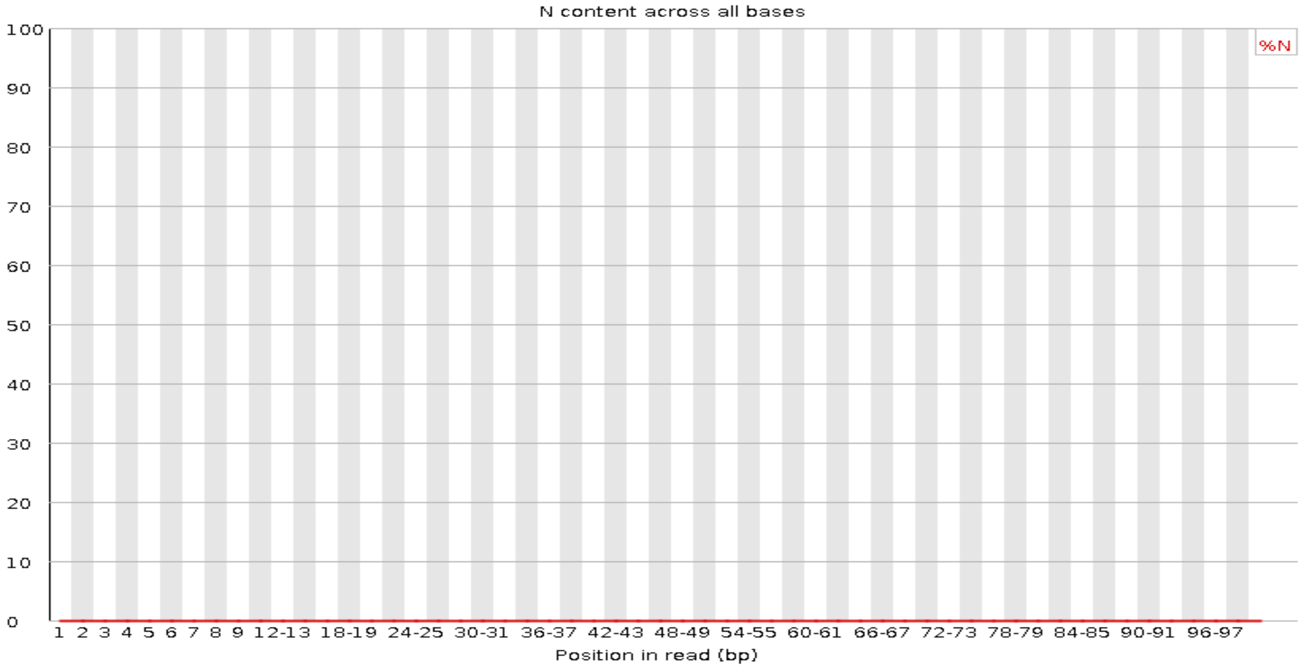
PER BASE SEQUENCE CONTENT



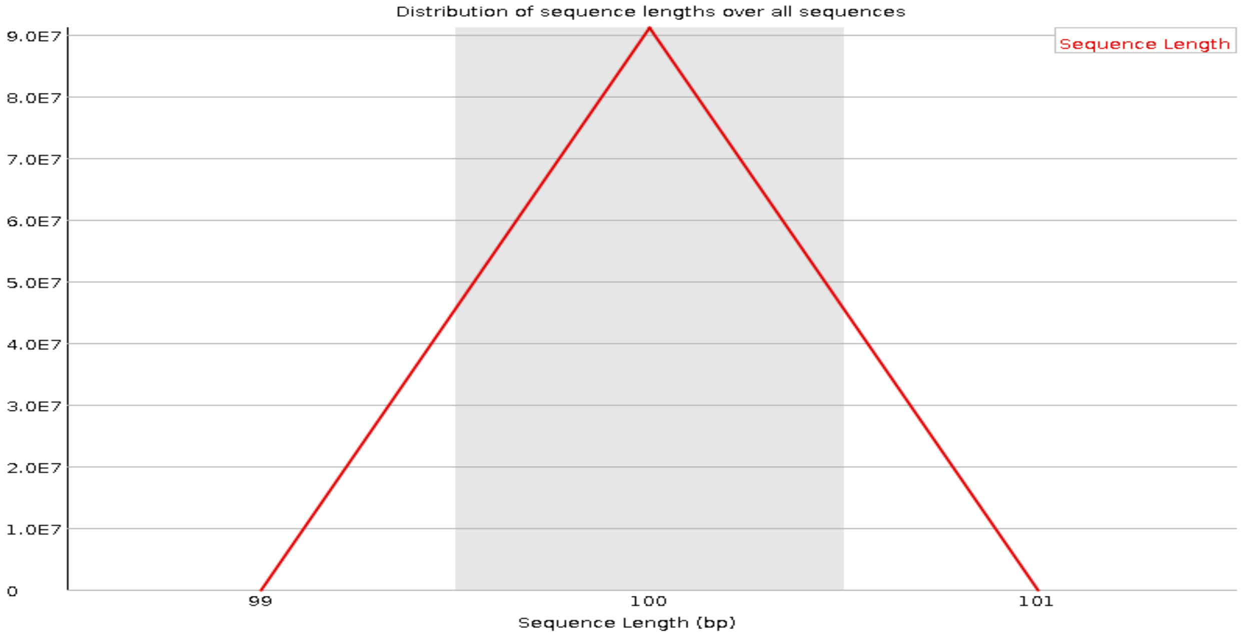
PER SEQUENCE GC CONTENT



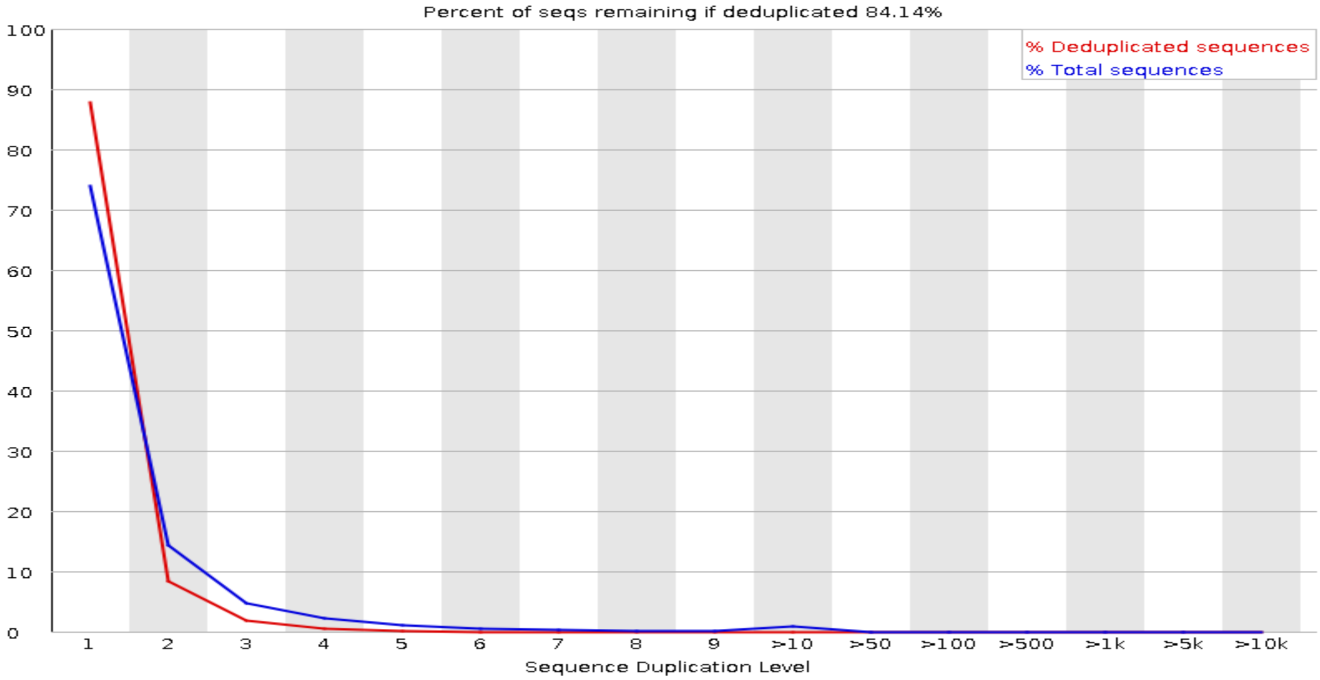
PER BASE N CONTENT



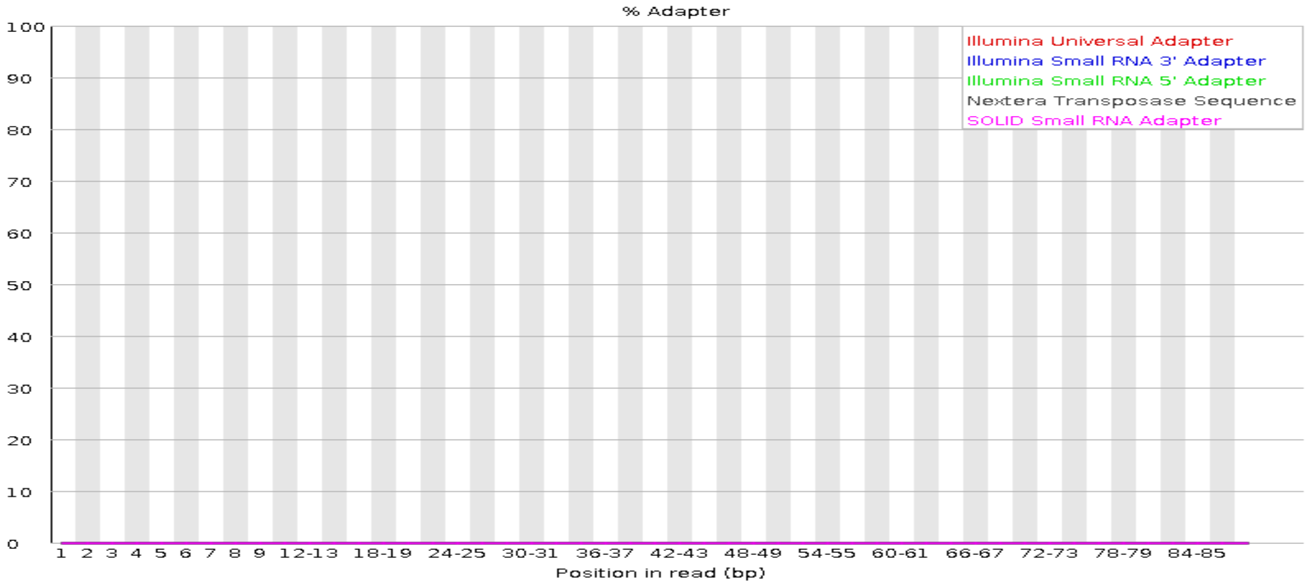
SEQUENCE LENGTH DISTRIBUTION



SEQUENCE DUPLICATION LEVELS



ADAPTER CONTENT



OVERREPRESENTED SEQUENCES

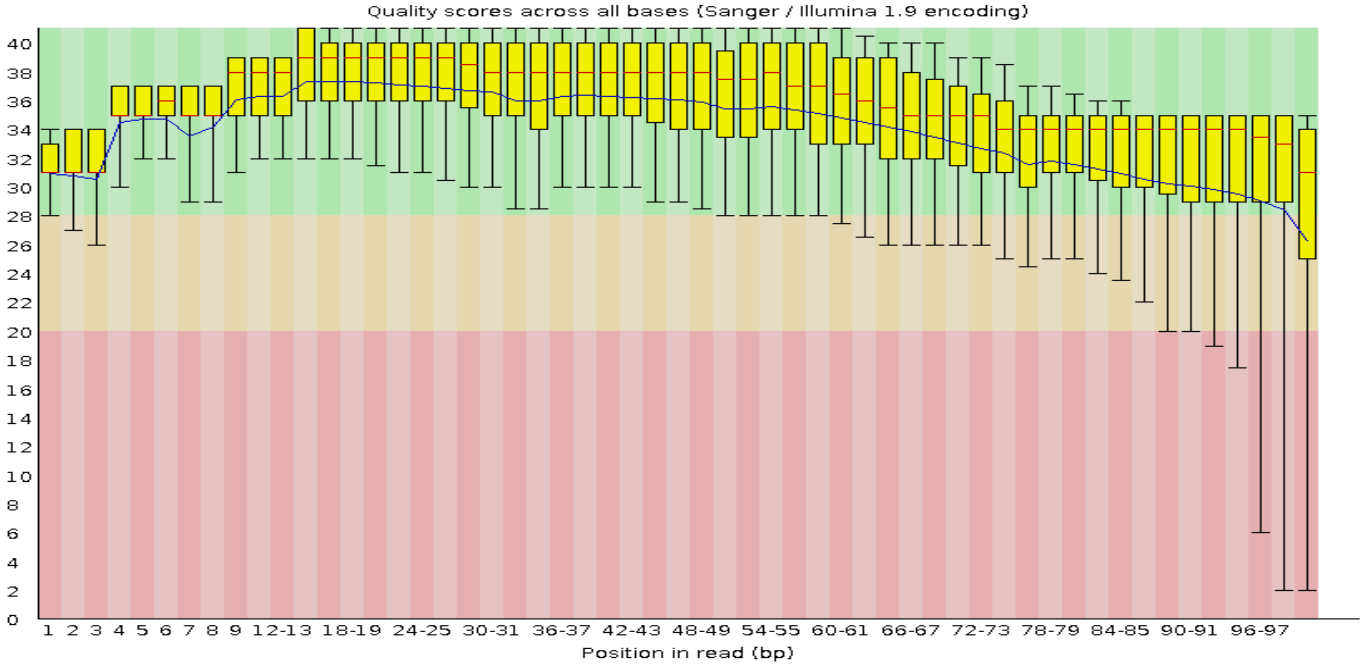
NO OVERREPRESENTED SEQUENCES FOUND.

2. FastQC Report on Second dataset sample: (SRA1554536)

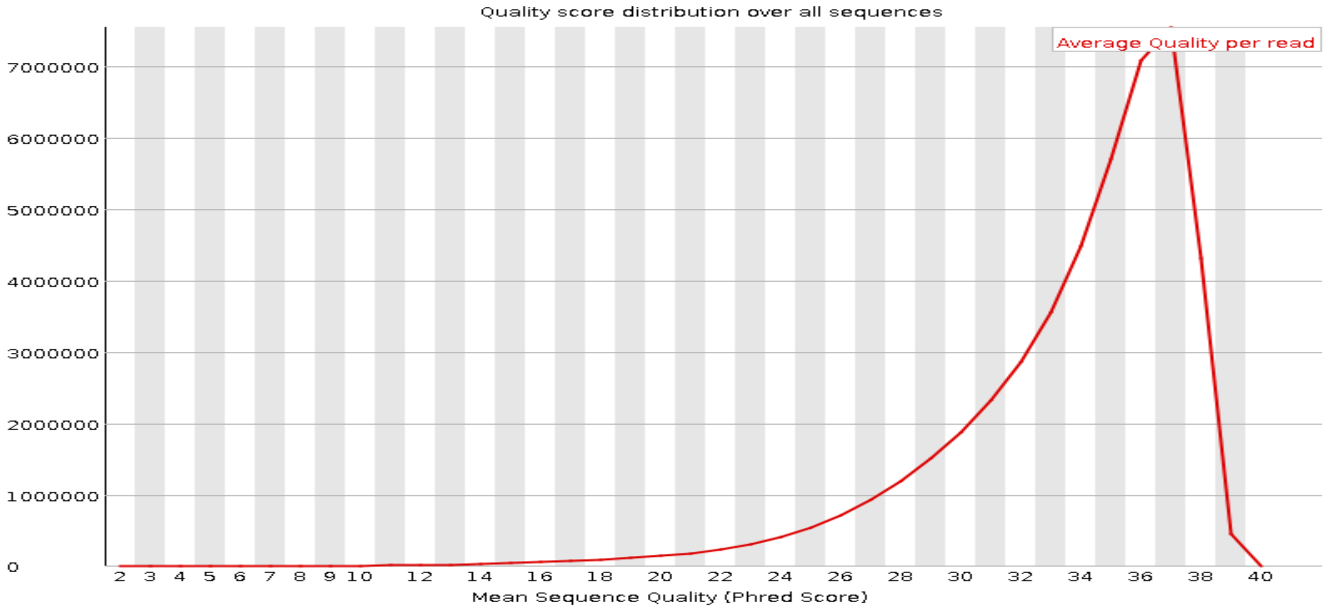
BASE STATISTICS

Measure	Value
Filename	Bowtie2 on data 6_ alignments
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	46971267
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

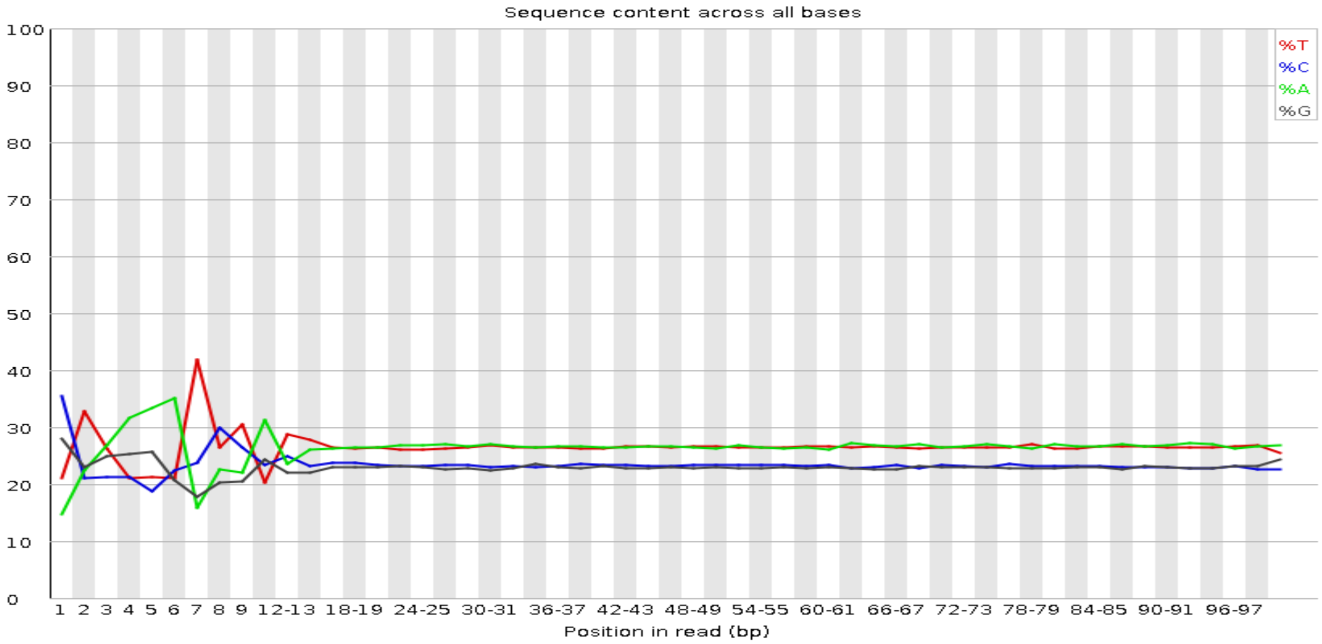
PER BASE SEQUENCE QUALITY



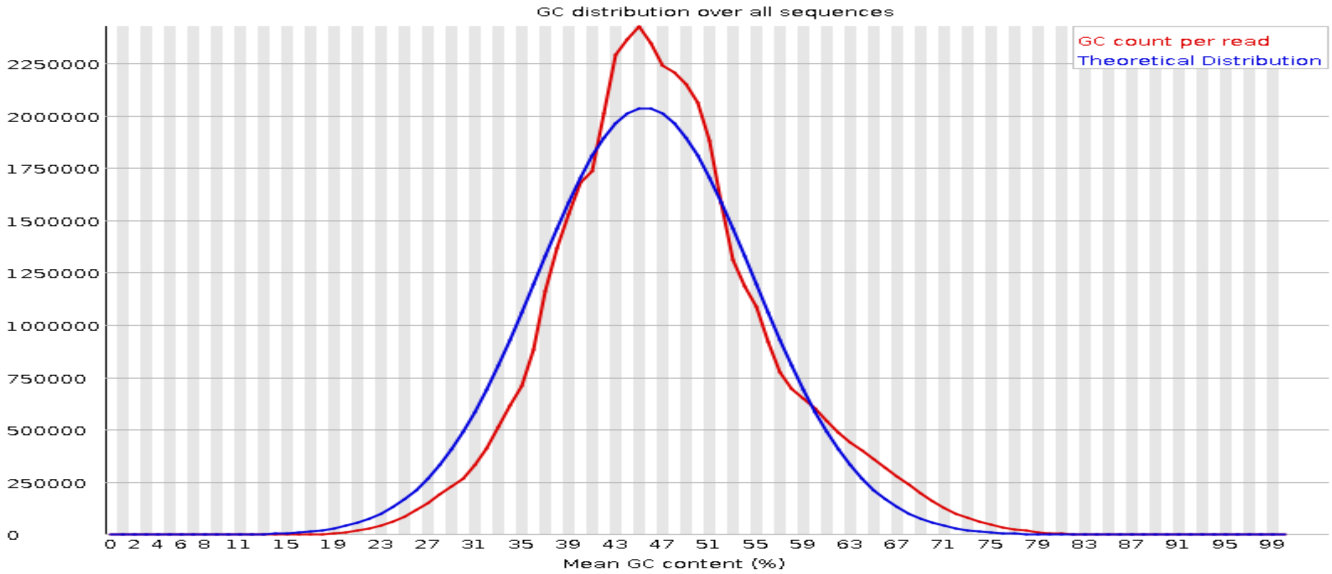
PER SEQUENCE QUALITY SCORES



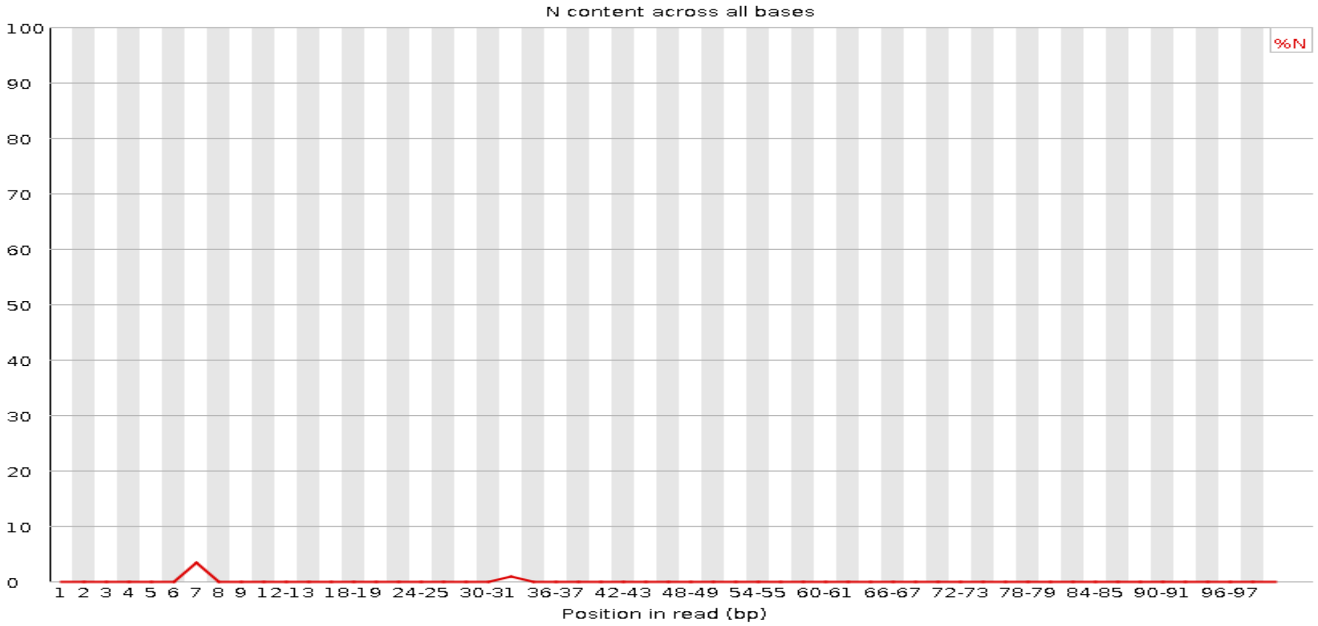
PER BASE SEQUENCE CONTENT



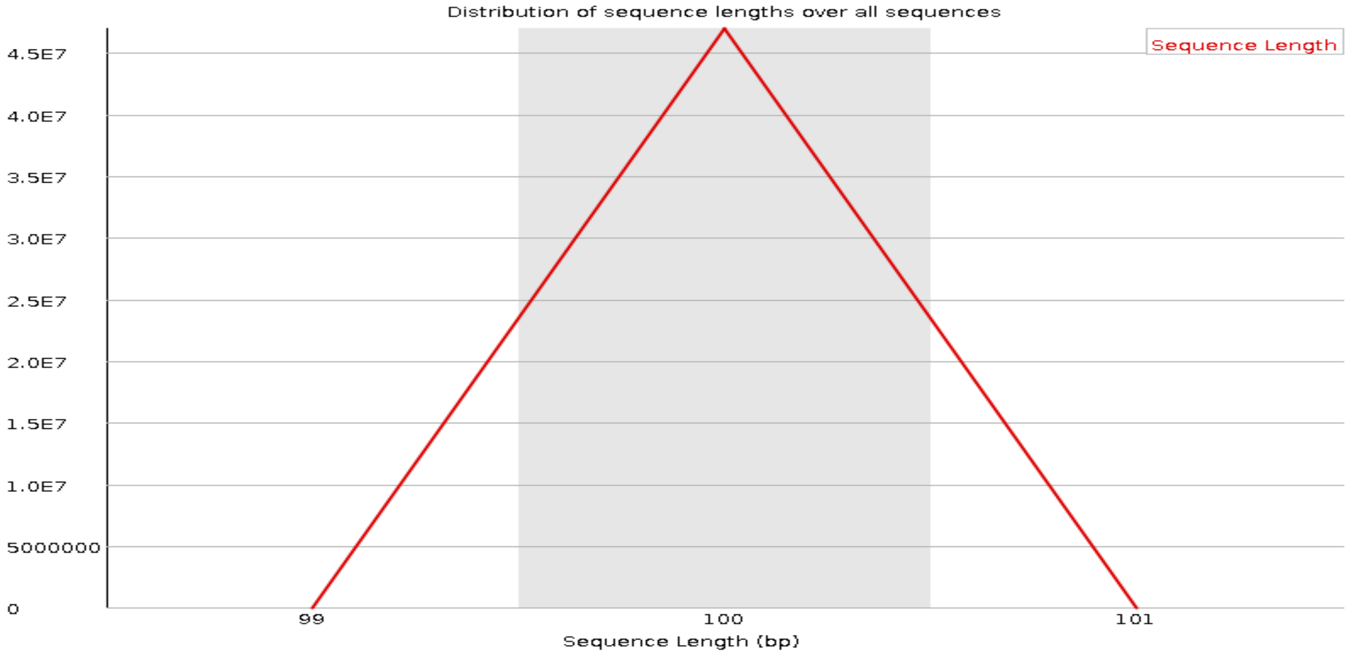
PER SEQUENCE GC CONTENT



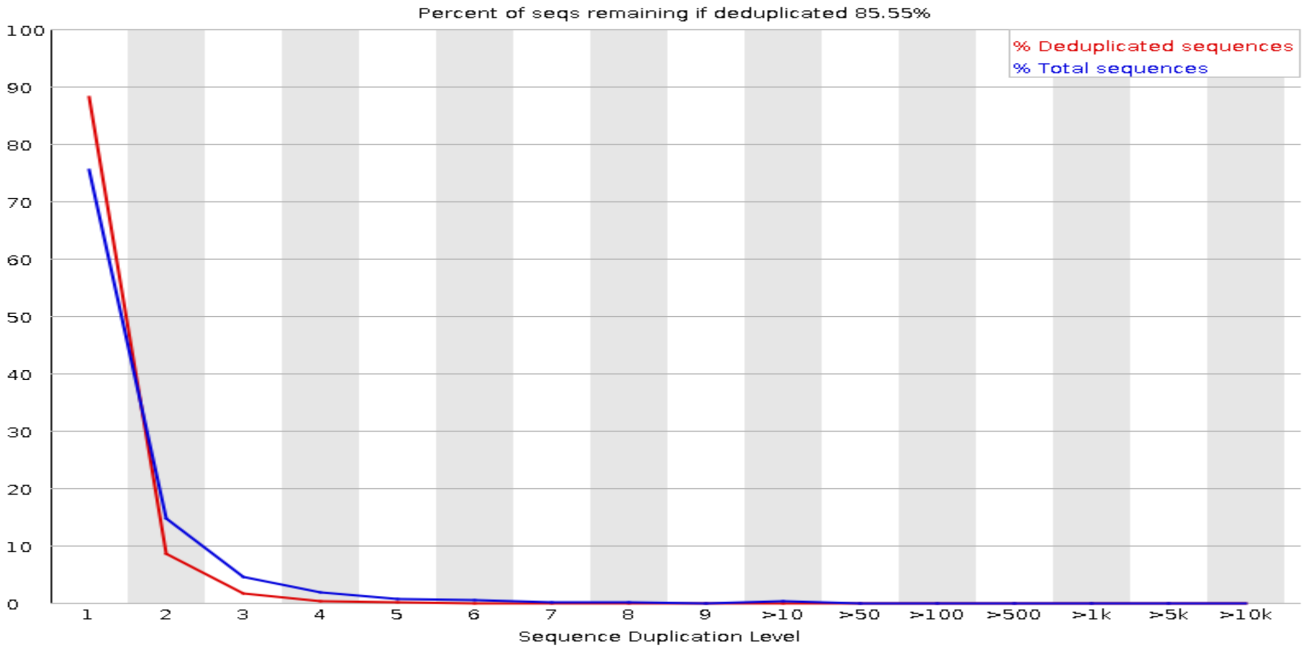
PER BASE N CONTENT



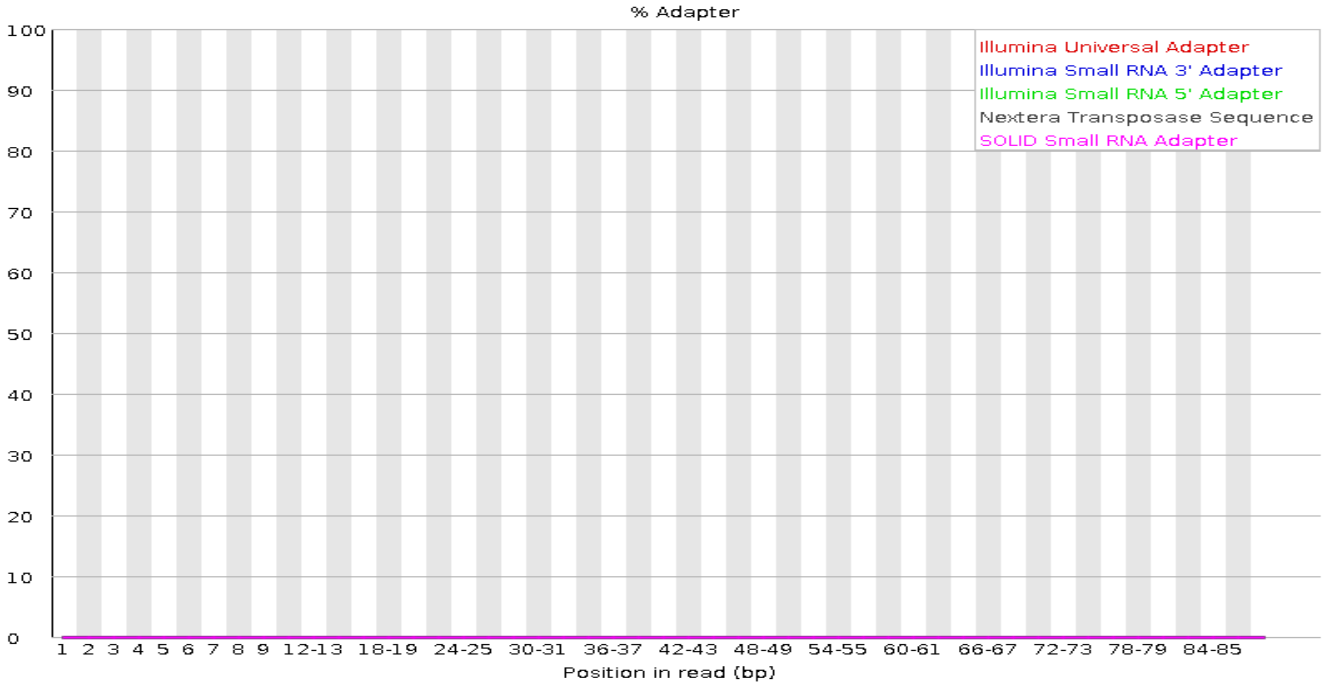
SEQUENCE LENGTH DISTRIBUTION



SEQUENCE DUPLICATION LEVELS



ADAPTER CONTENT



OVERREPRESENTED SEQUENCES

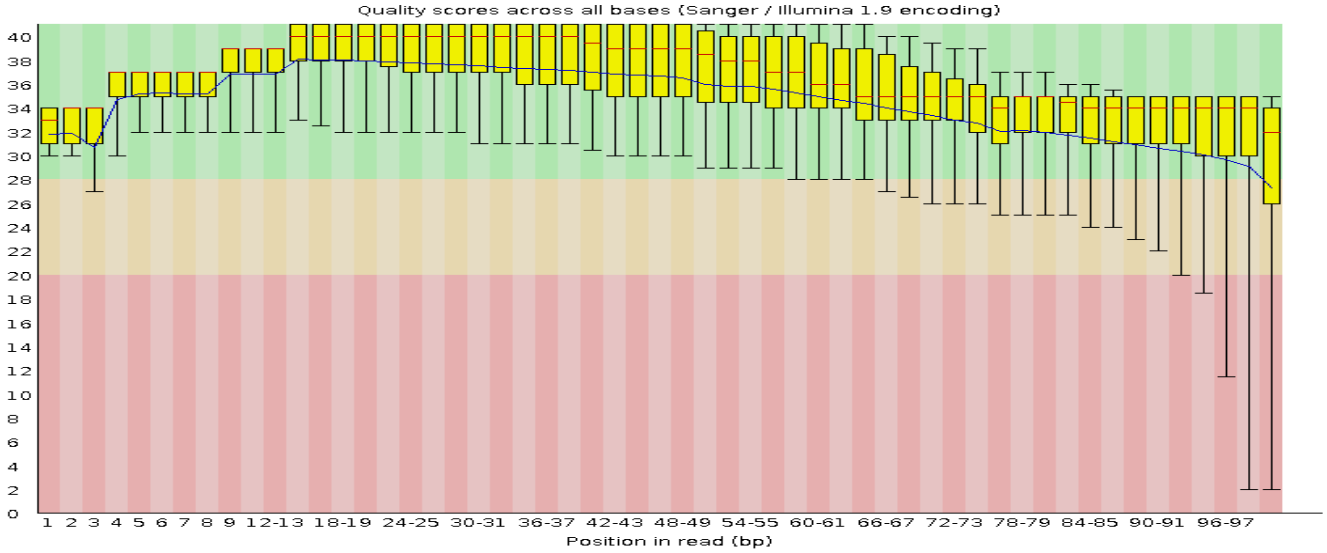
NO OVERREPRESENTED SEQUENCES FOUND.

3. FastQC Report on Third dataset sample: (SRR1554561)

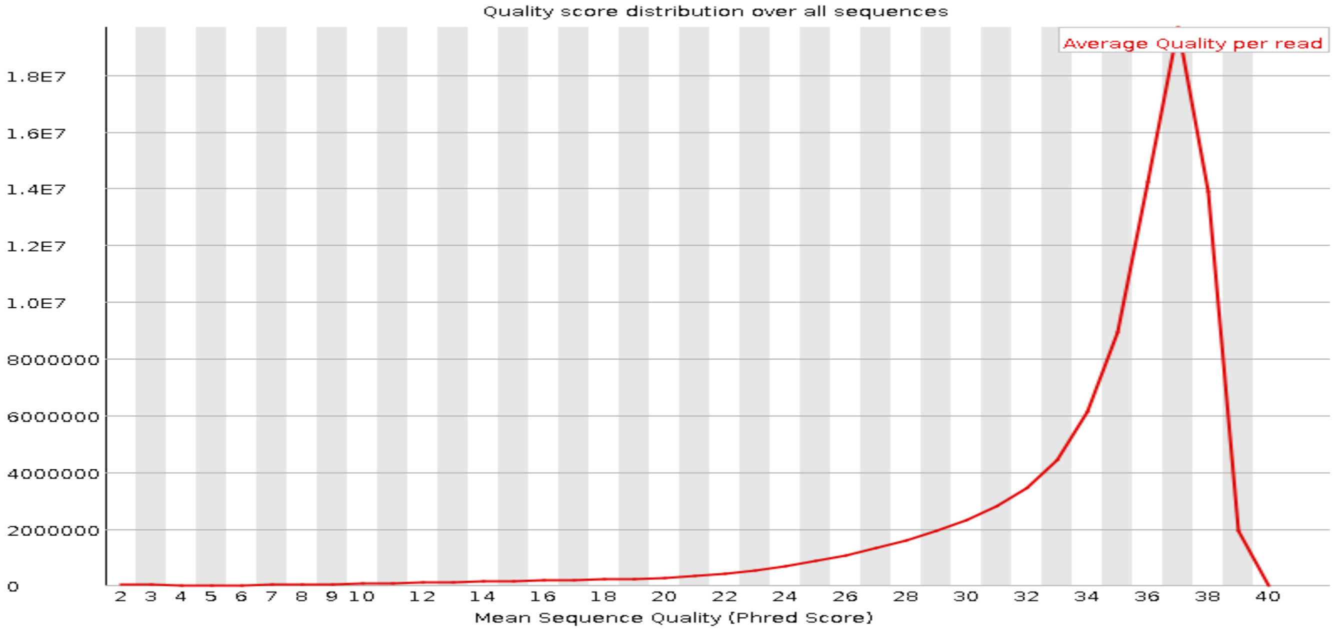
BASE STATISTICS

Measure	Value
Filename	Bowtie2 on data 7_ alignments
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	89061863
Sequences flagged as poor quality	0
Sequence length	100
%GC	52

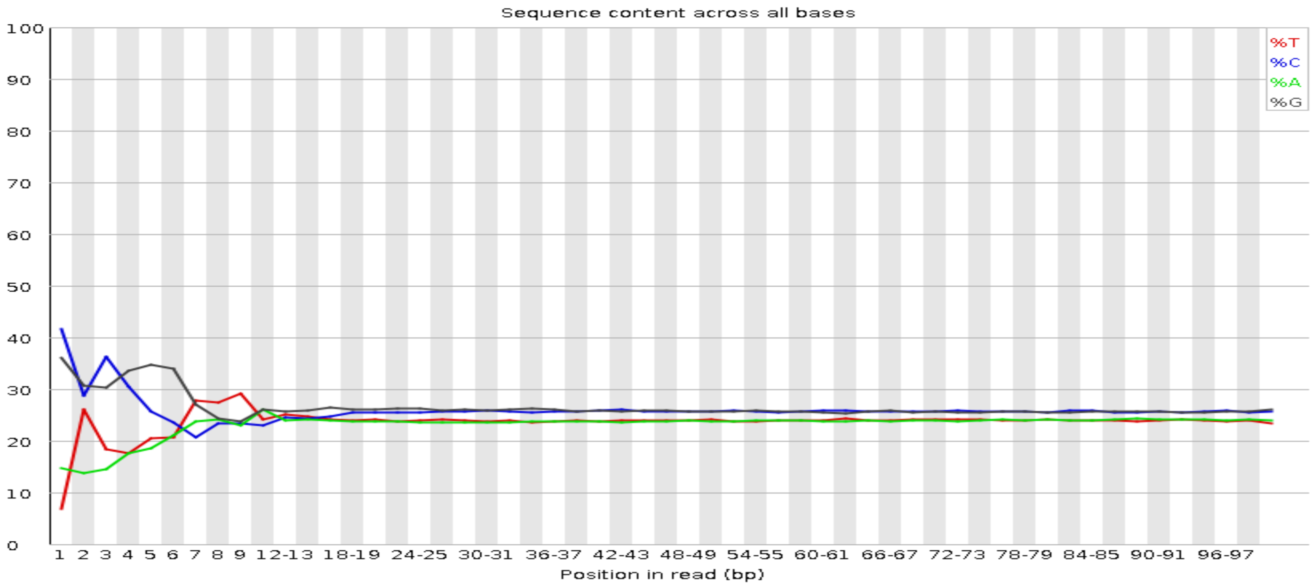
PER BASE SEQUENCE QUALITY



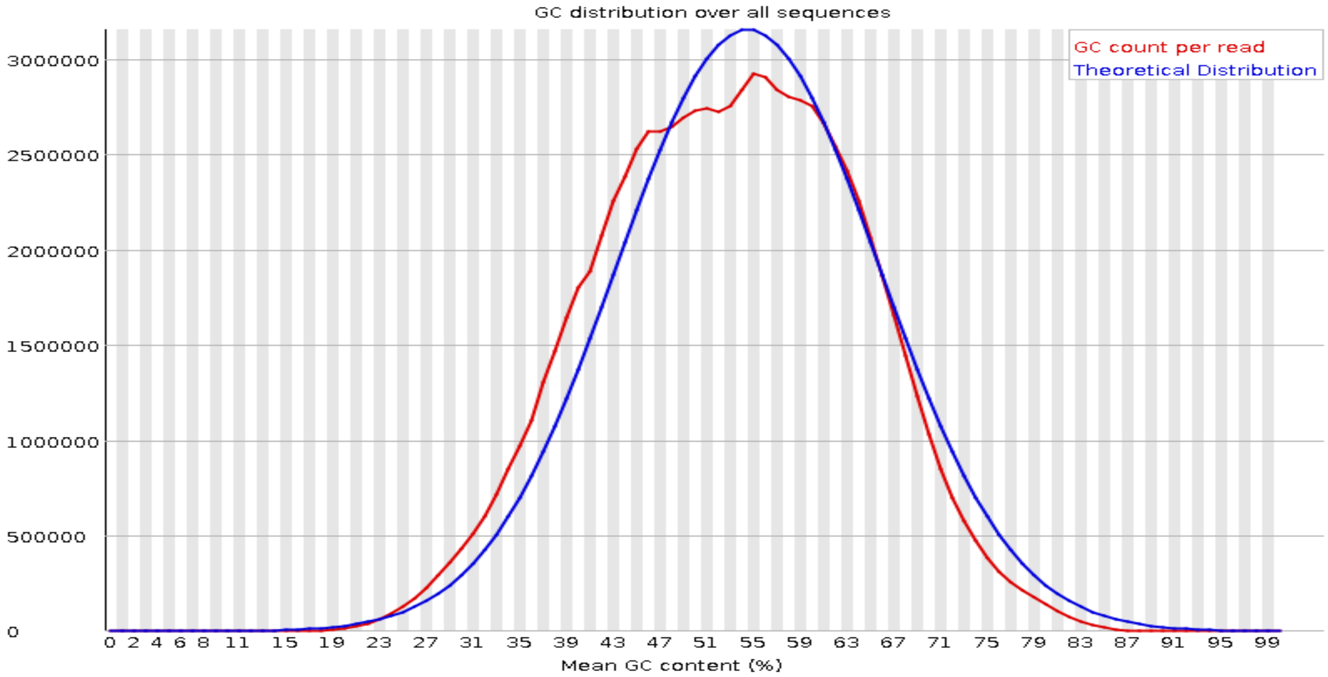
PER SEQUENCE QUALITY SCORES



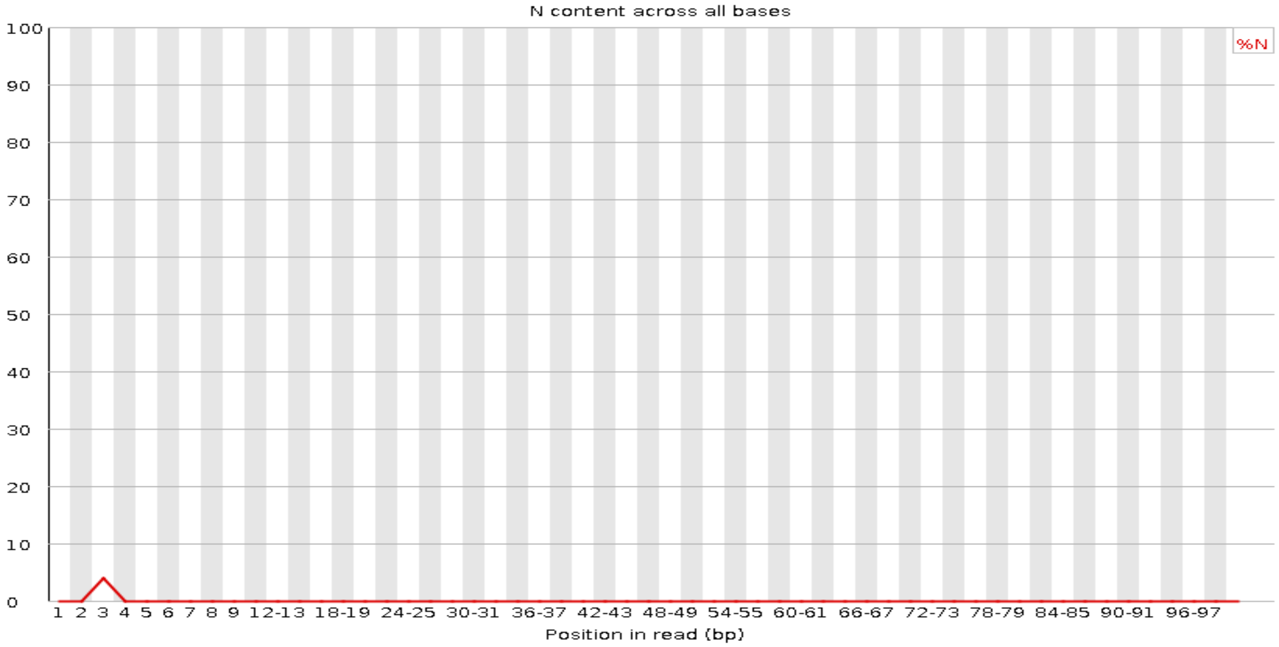
PER BASE SEQUENCE CONTENT



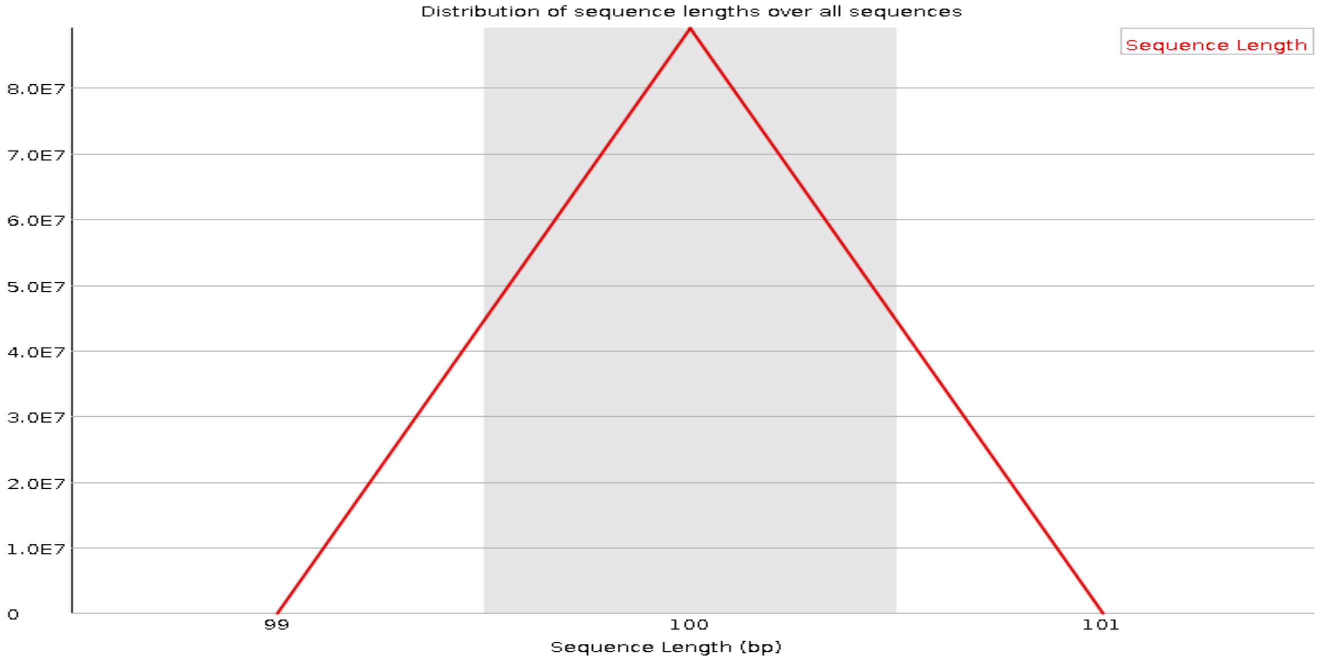
PER SEQUENCE GC CONTENT



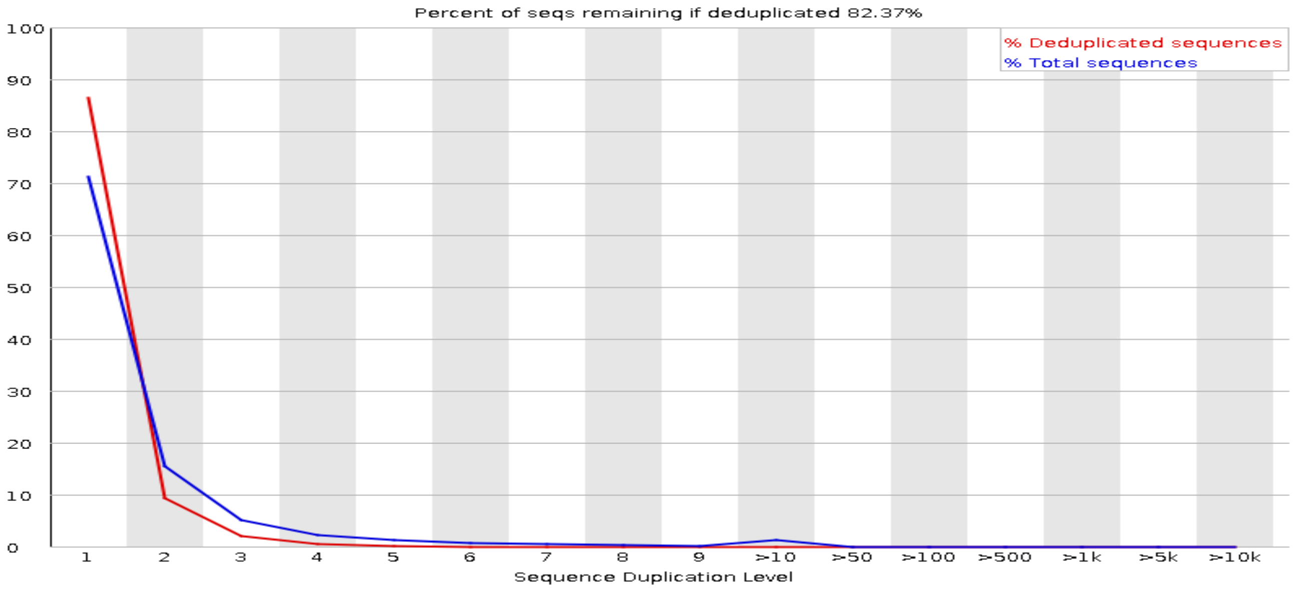
PER BASE N CONTENT



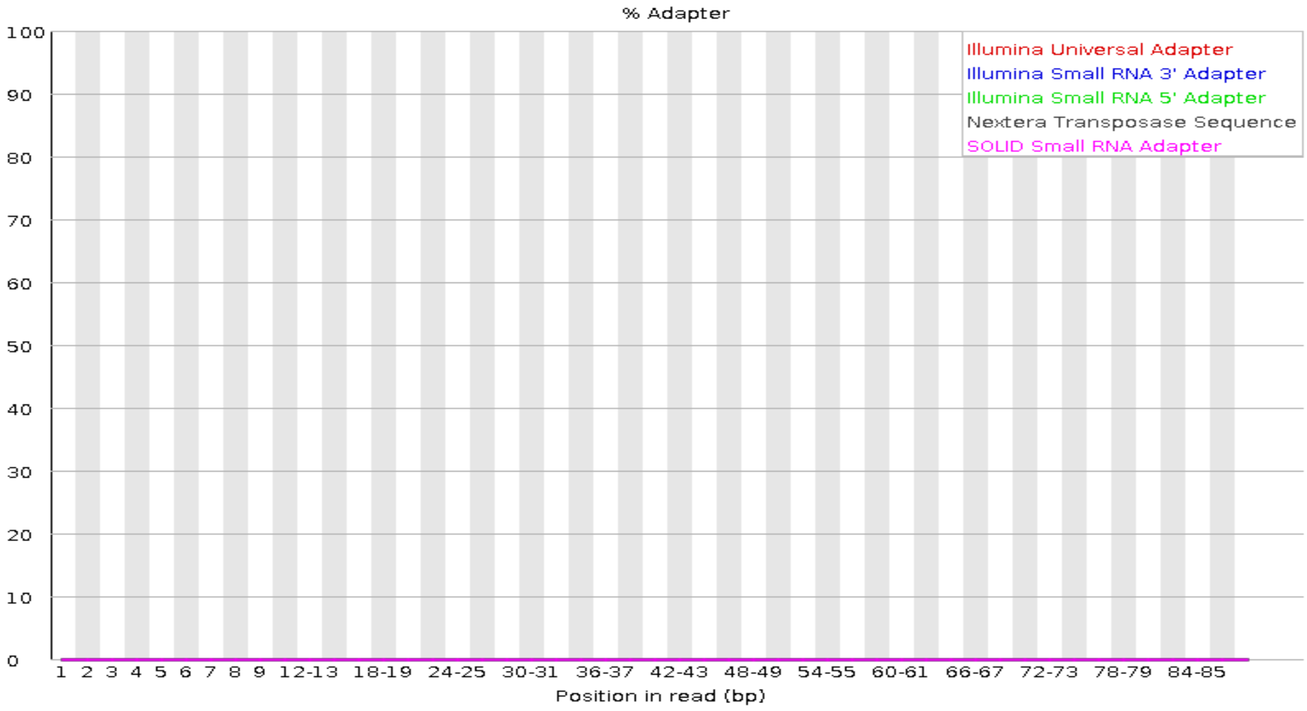
SEQUENCE LENGTH DISTRIBUTION



SEQUENCE DUPLICATION LEVELS



ADAPTER CONTENT



OVERREPRESENTED SEQUENCES

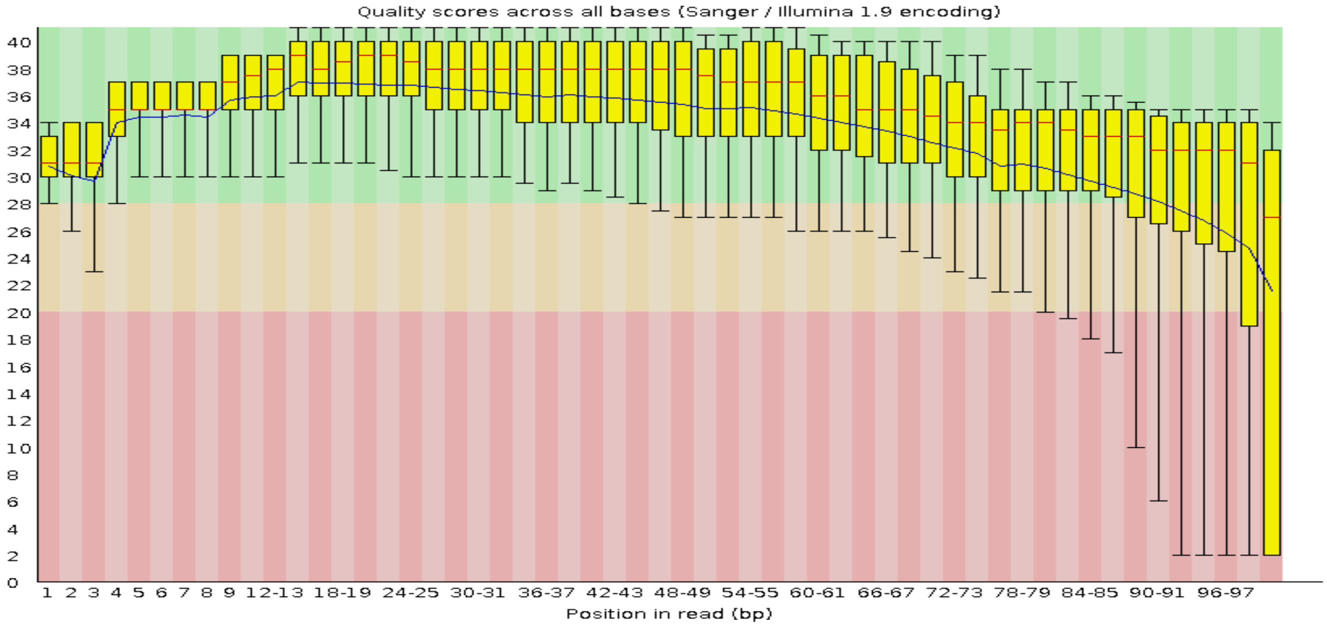
NO OVERREPRESENTED SEQUENCES FOUND.

4. FastQC Report on Fourth dataset sample: (SRR1554541)

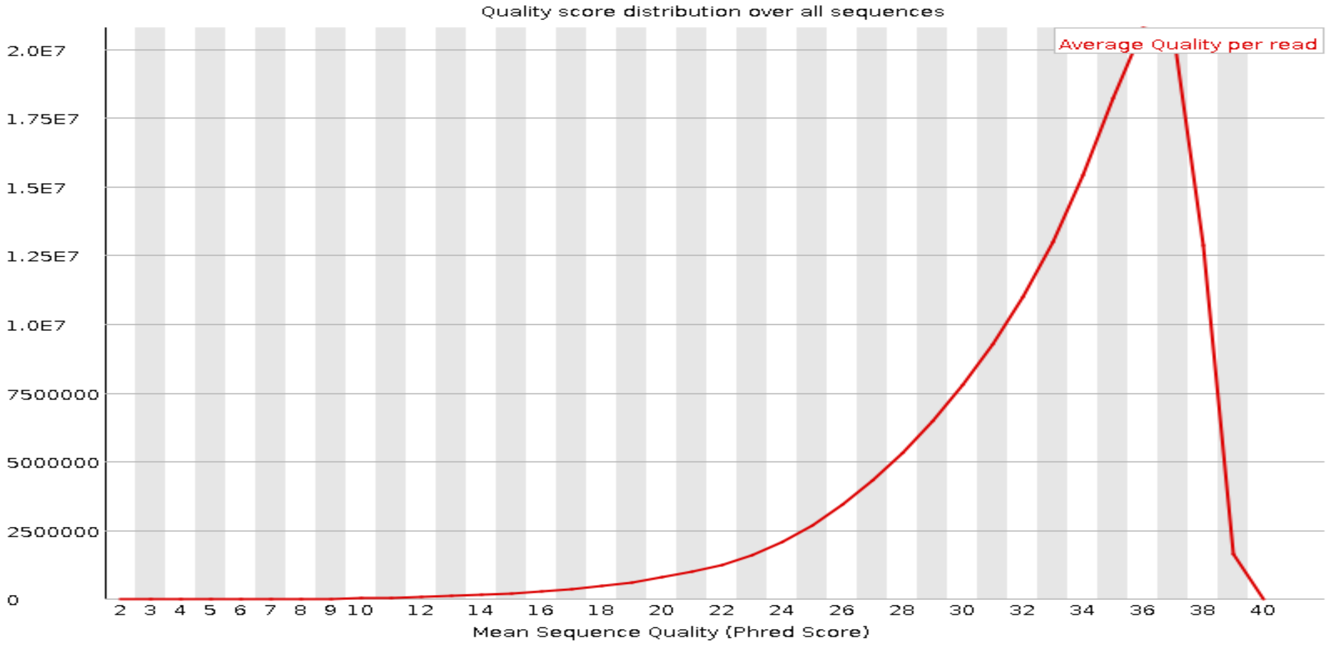
BASE STATISTICS

Measure	Value
Filename	Bowtie2 on data 8_ alignments
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	162403466
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

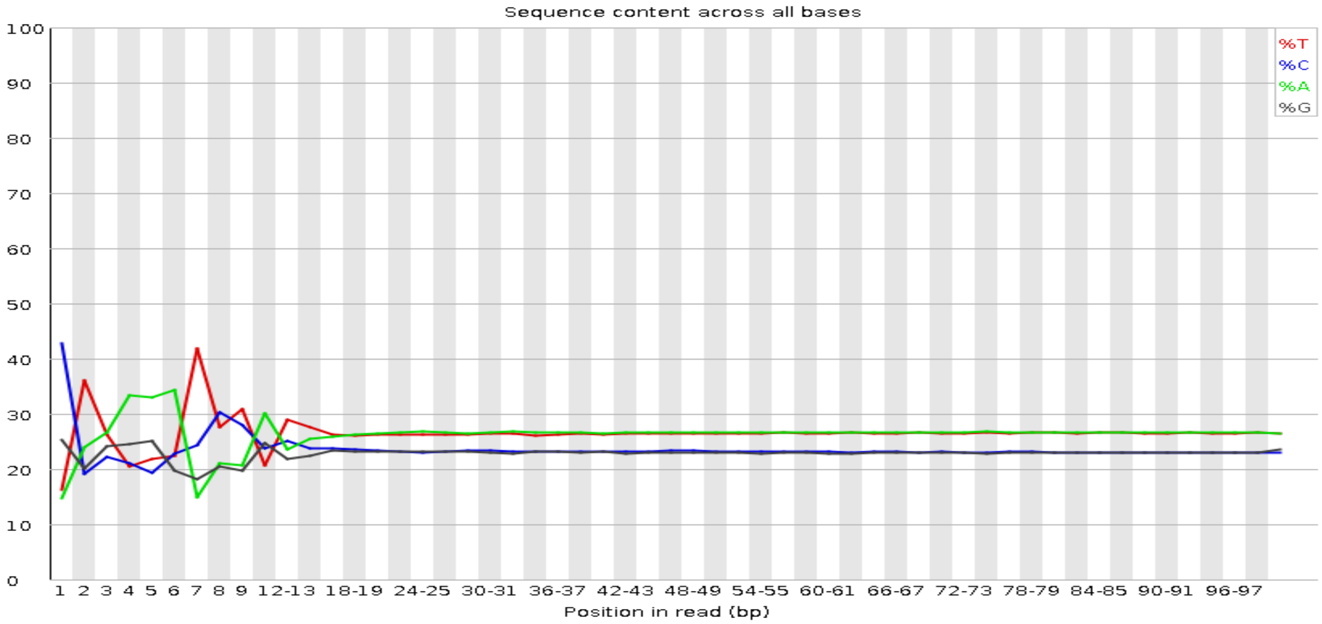
PER BASE SEQUENCE QUALITY



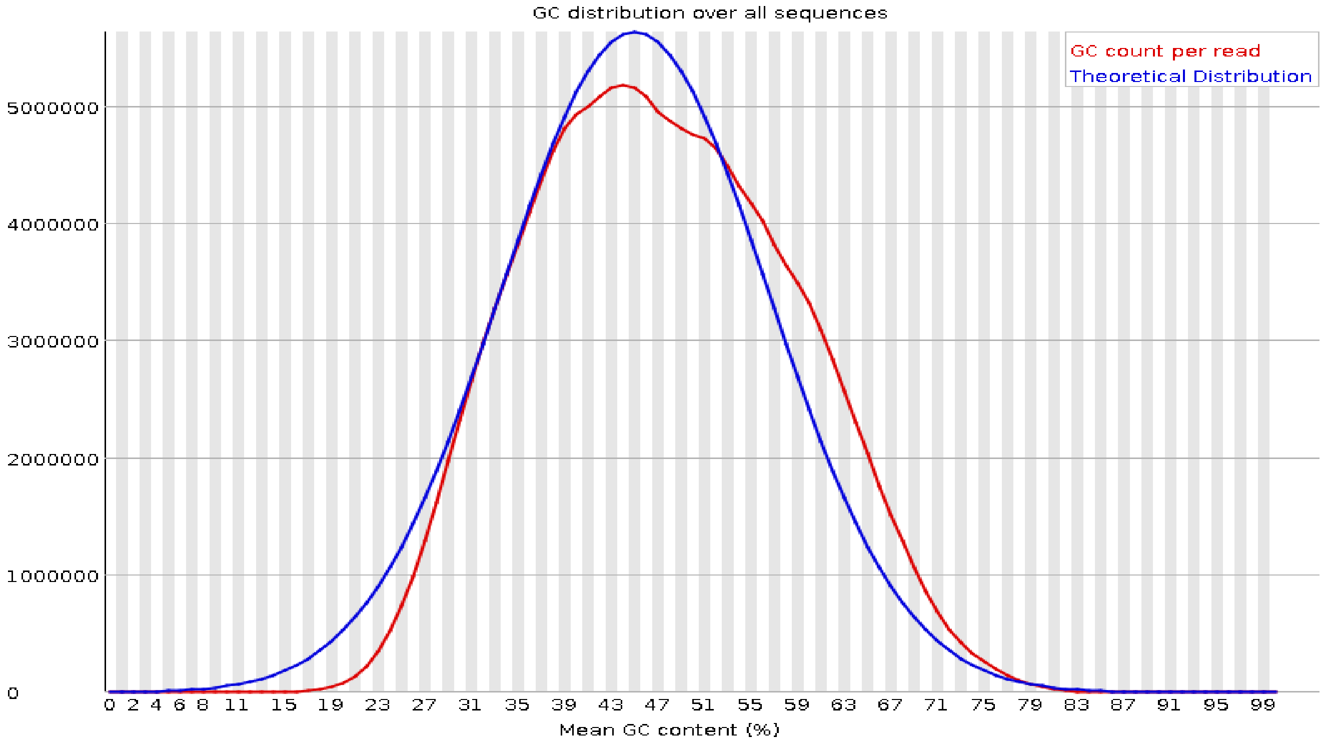
PER SEQUENCE QUALITY SCORES



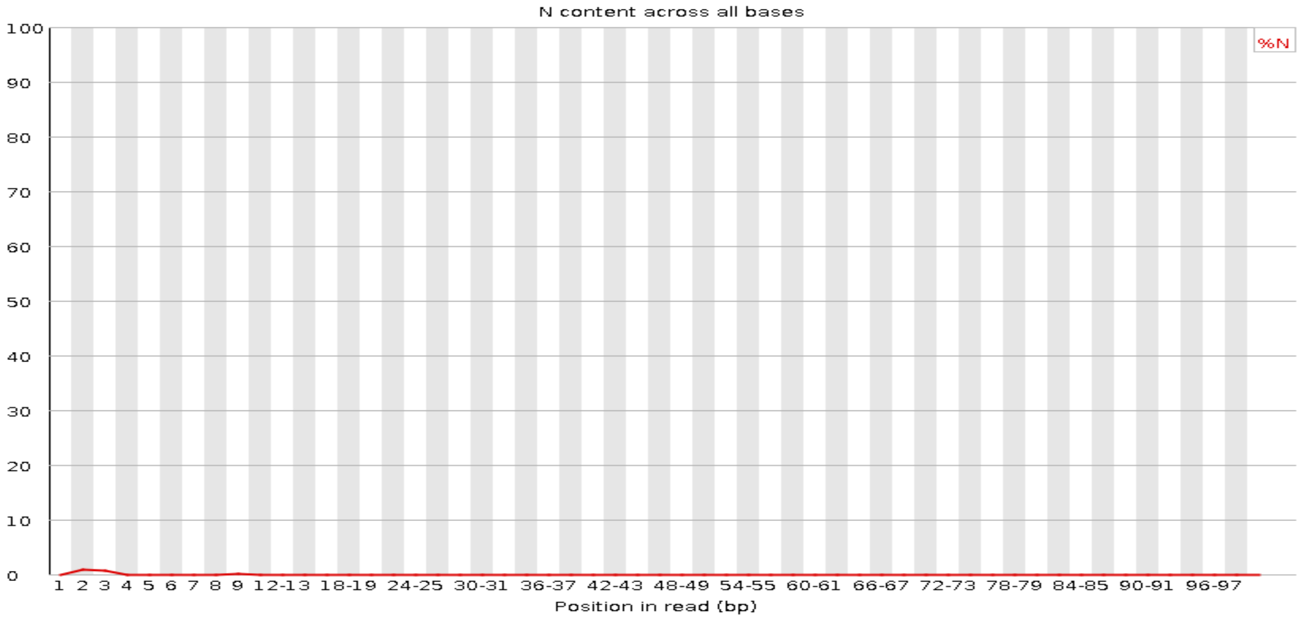
PER BASE SEQUENCE CONTENT



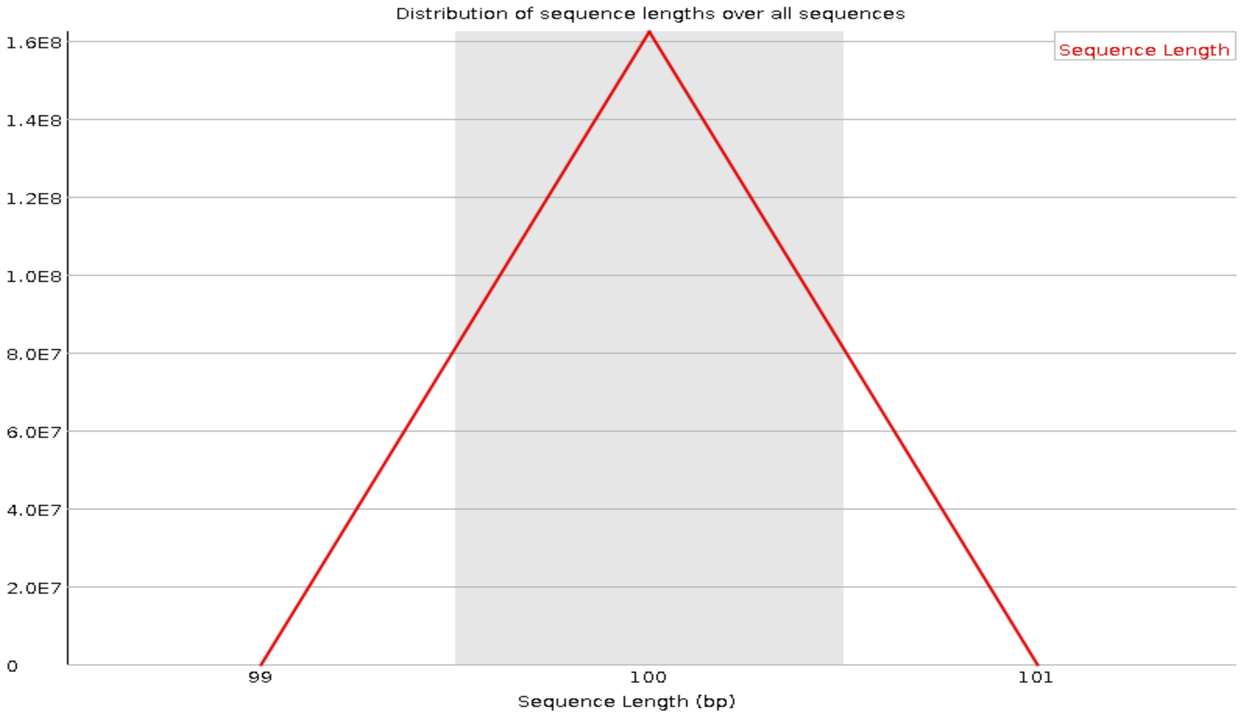
PER SEQUENCE GC CONTENT



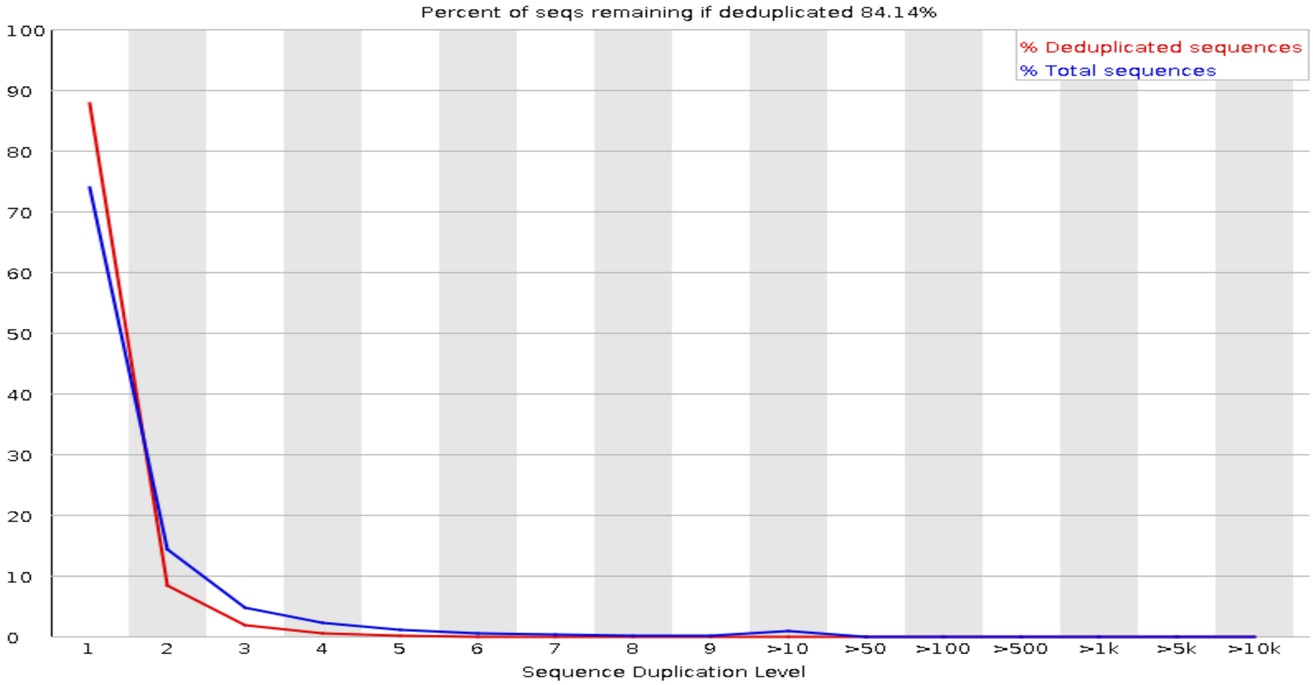
PER BASE N CONTENT



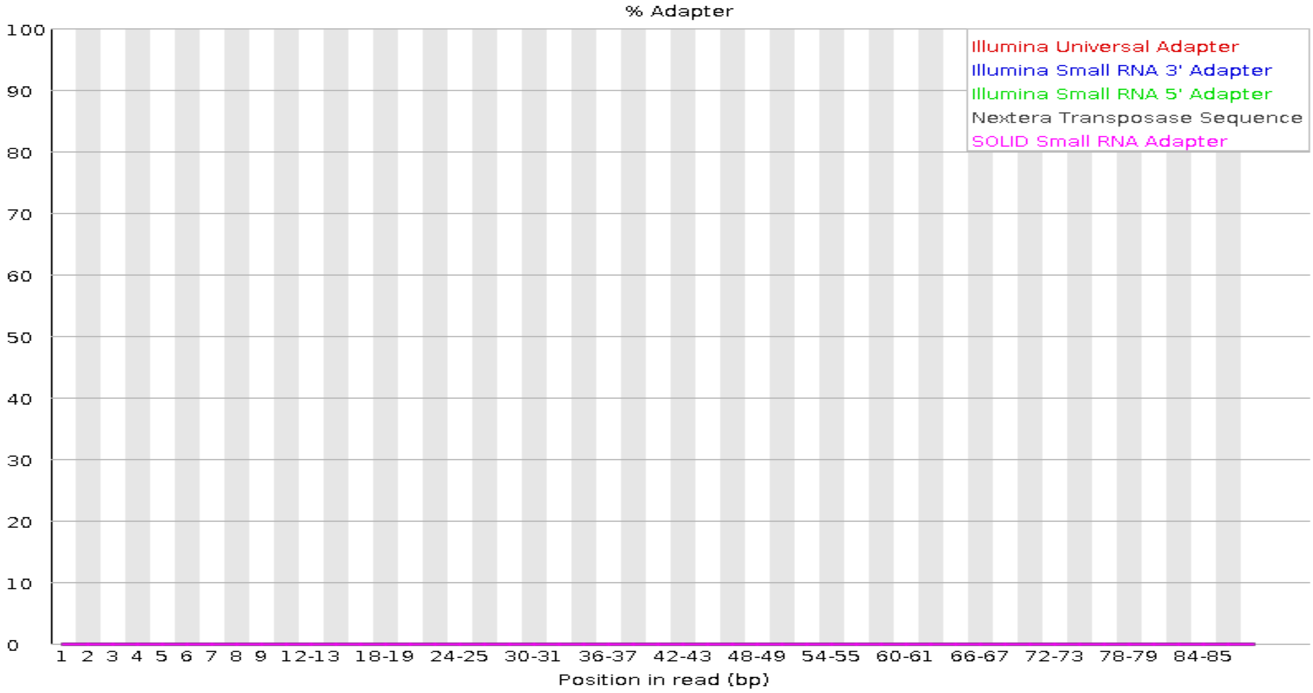
SEQUENCE LENGTH DISTRIBUTION



SEQUENCE DUPLICATION LEVELS



ADAPTER CONTENT



OVERREPRESENTED SEQUENCES

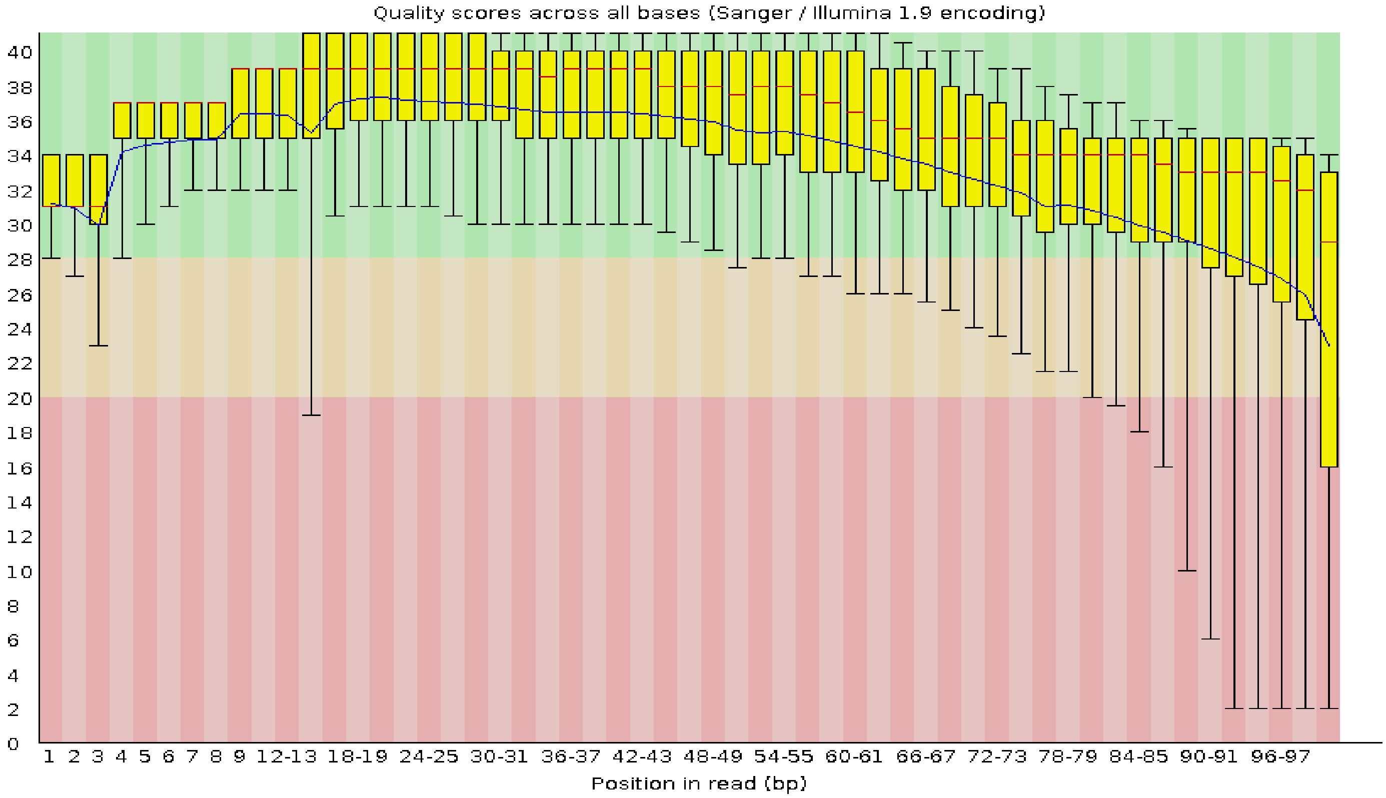
NO OVERREPRESENTED SEQUENCES FOUND.

5. FastQC Report on Fifth dataset sample: (SRR1554537)

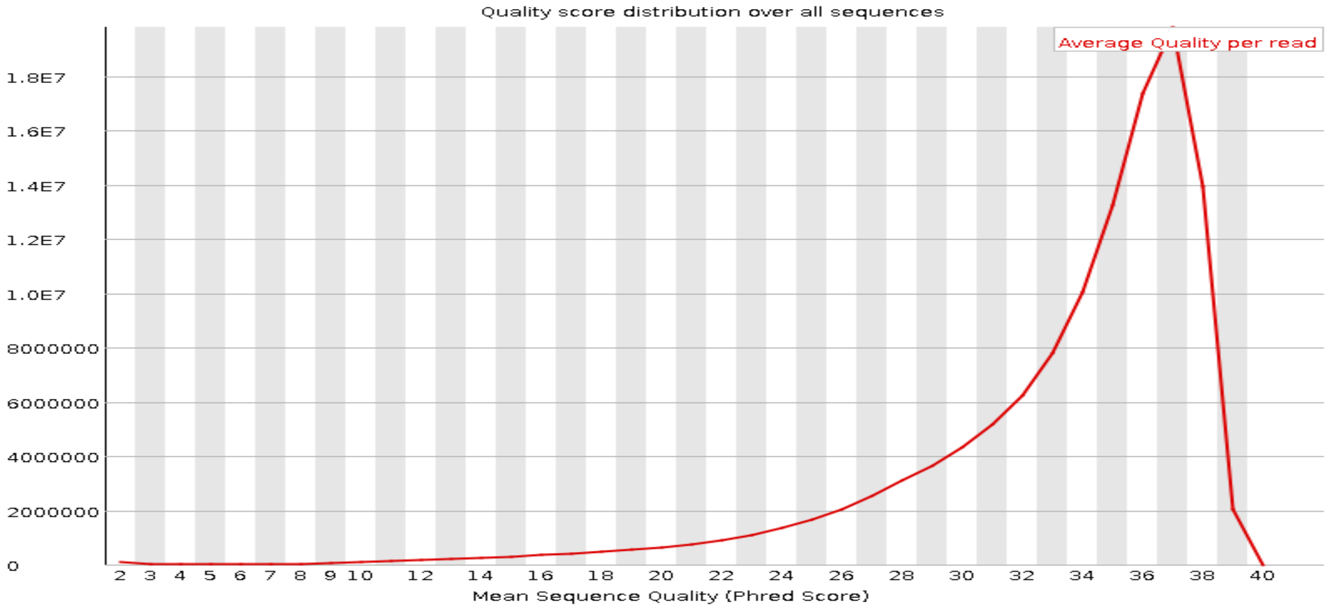
BASE STATISTICS

Measure	Value
Filename	Bowtie2 on data 9_ alignments
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	121992326
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

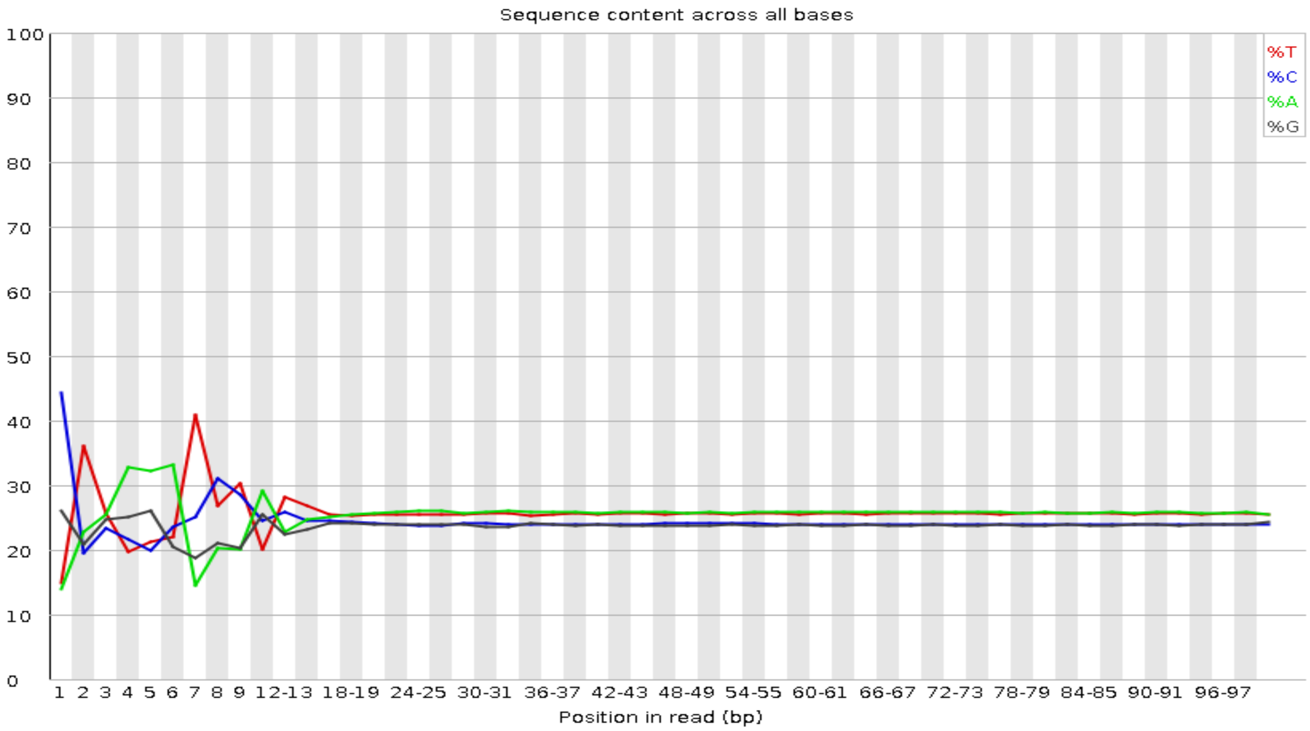
PER BASE SEQUENCE QUALITY



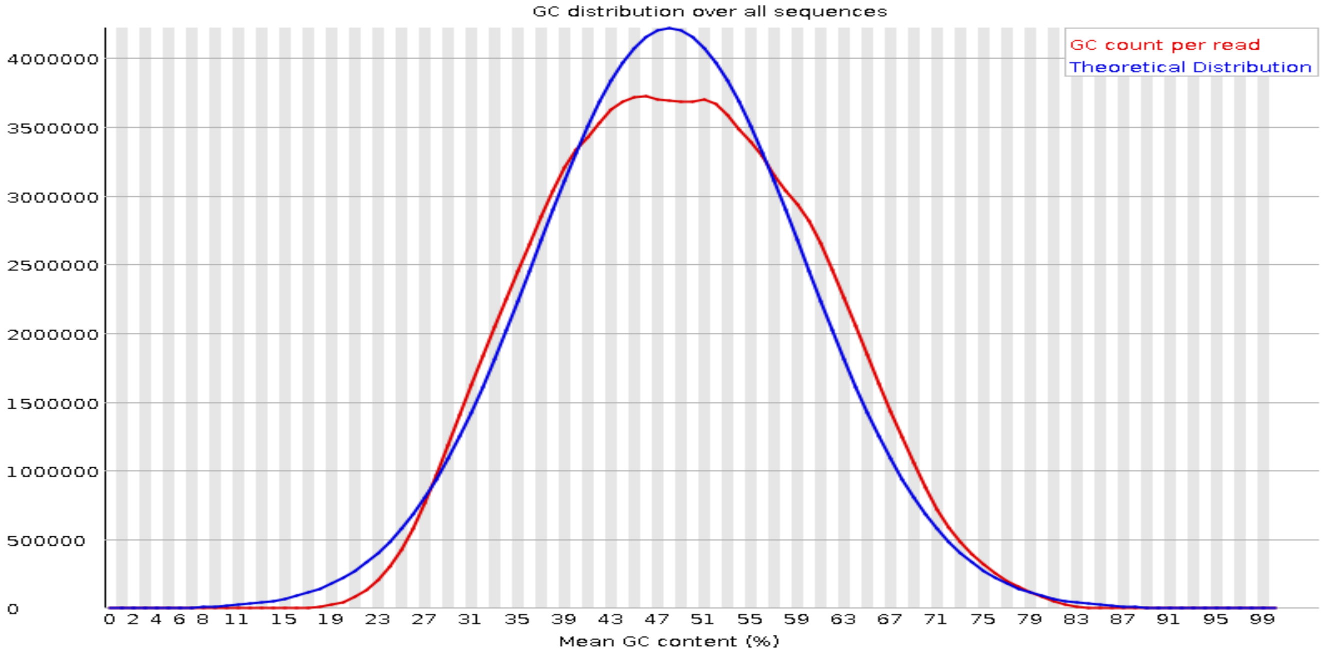
PER SEQUENCE QUALITY SCORES



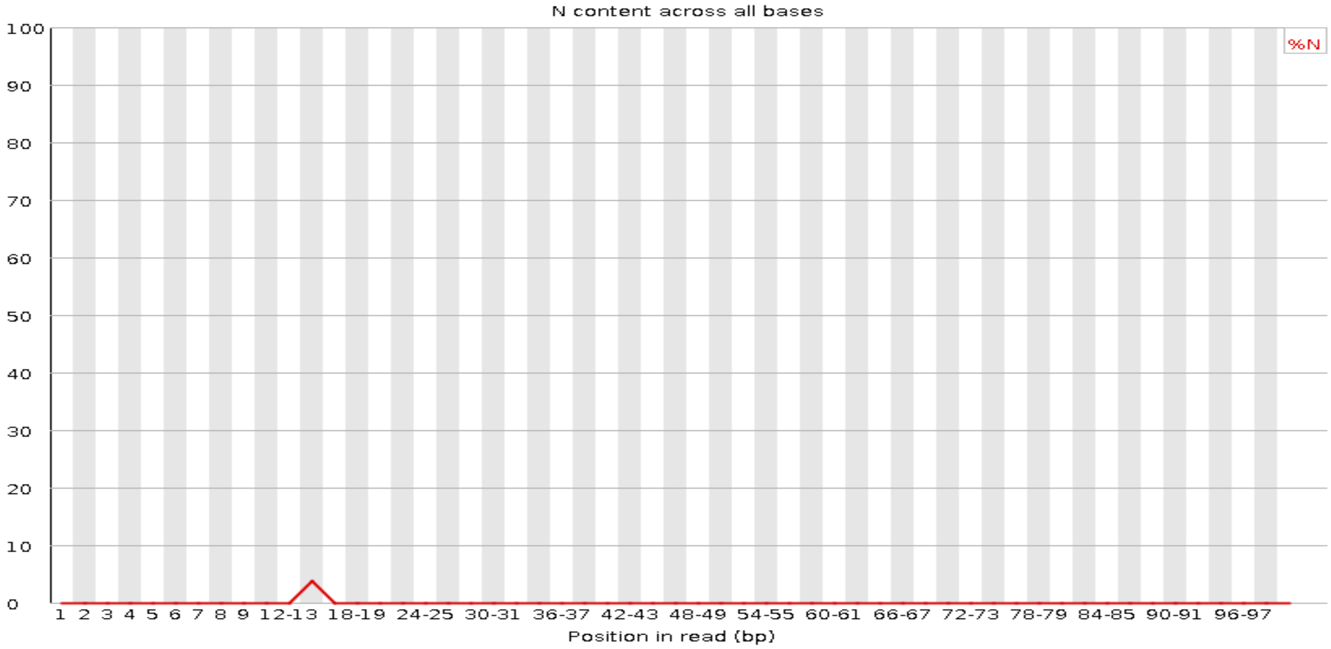
PER BASE SEQUENCE CONTENT



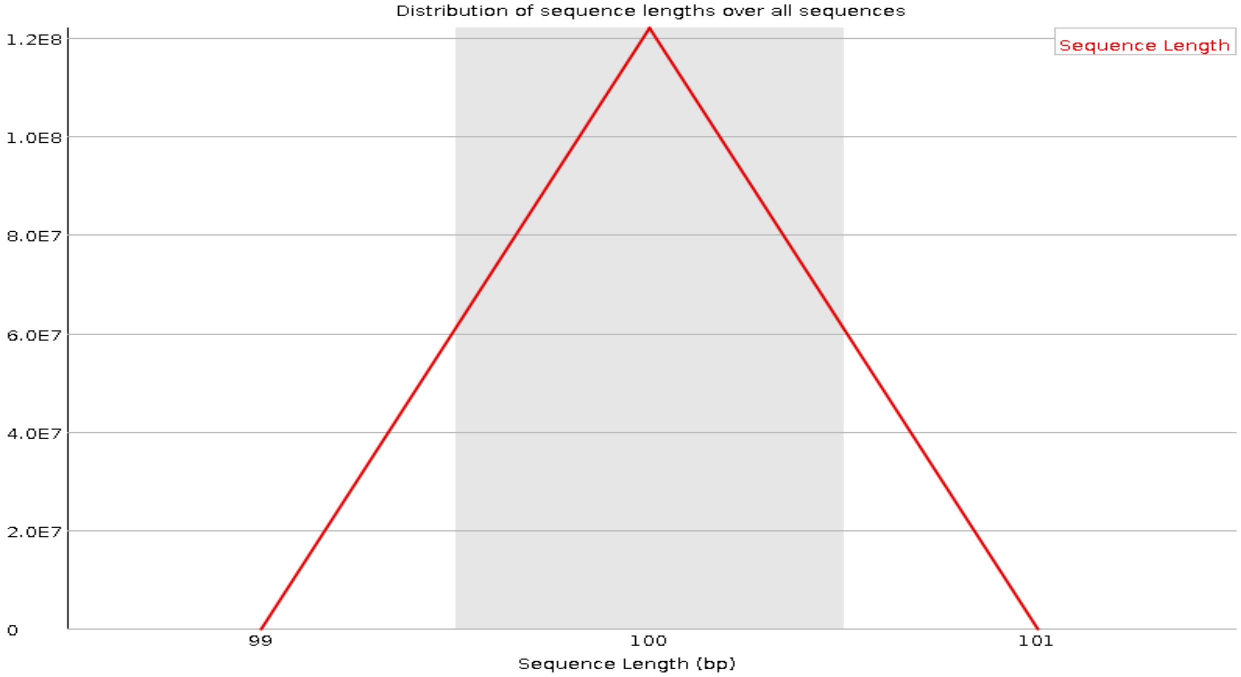
PER SEQUENCE GC CONTENT



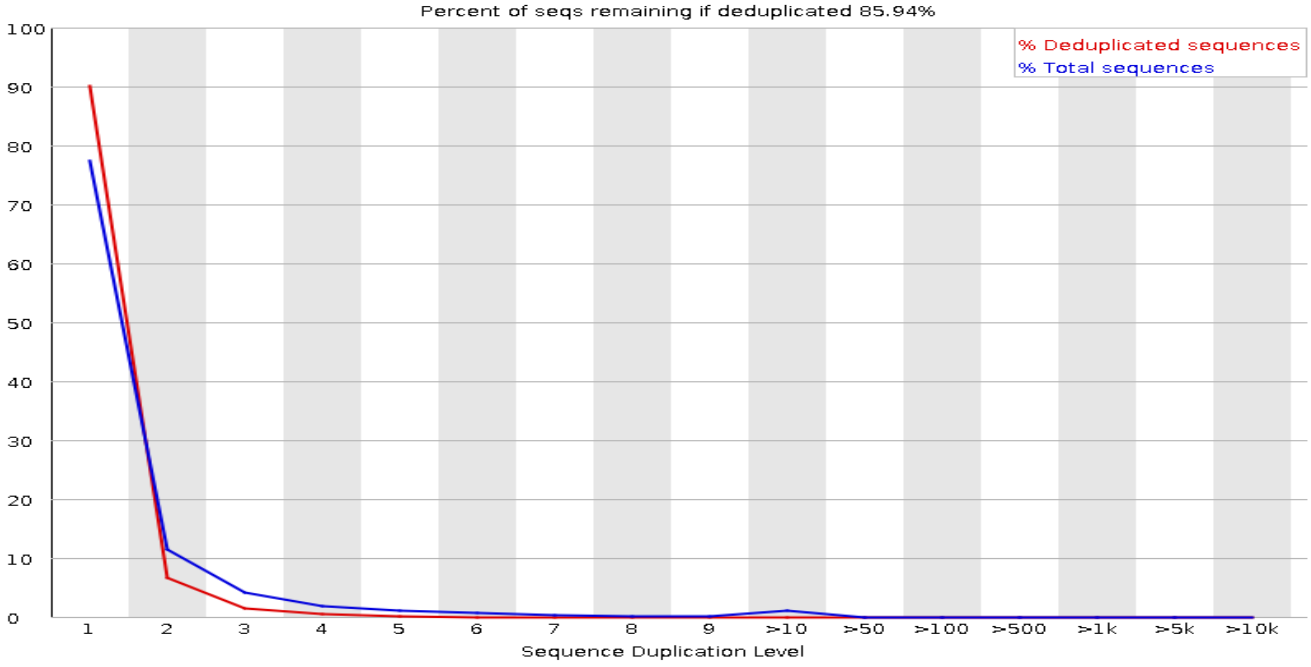
PER BASE N CONTENT



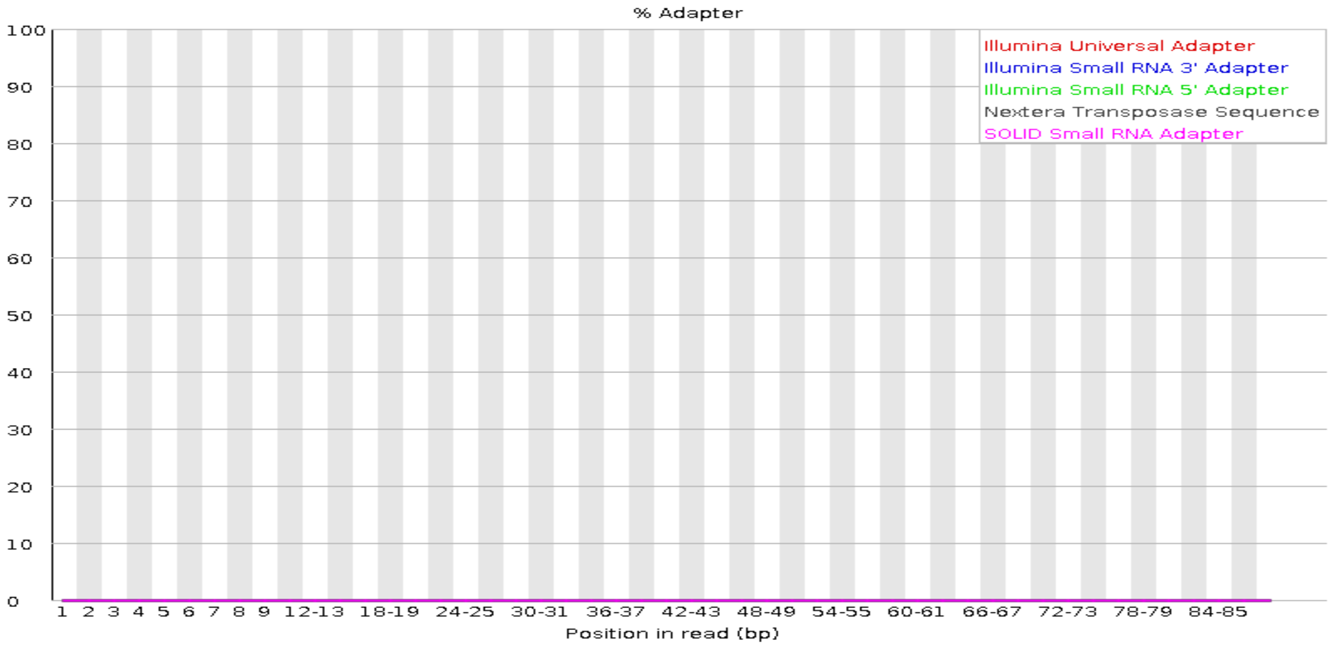
SEQUENCE LENGTH DISTRIBUTION



SEQUENCE DUPLICATION LEVELS



ADAPTER CONTENT



OVERREPRESENTED SEQUENCES

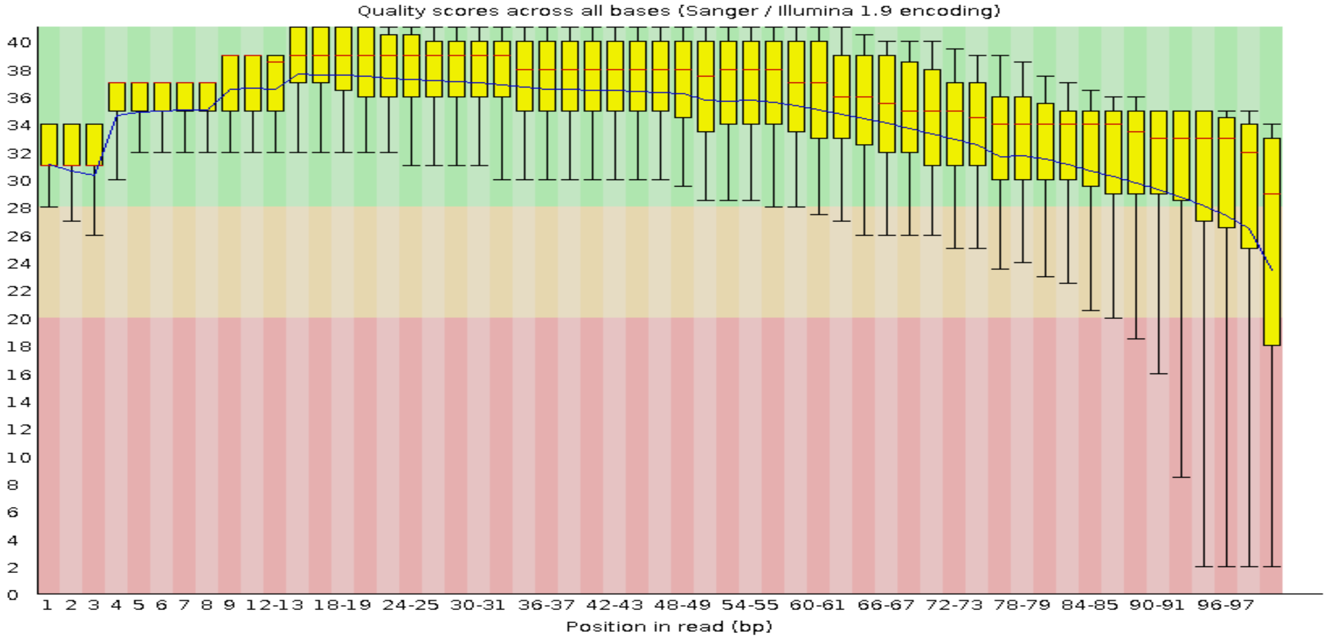
NO OVERREPRESENTED SEQUENCES FOUND.

6. FastQC Report on Sixth dataset sample: (SRR1554567)

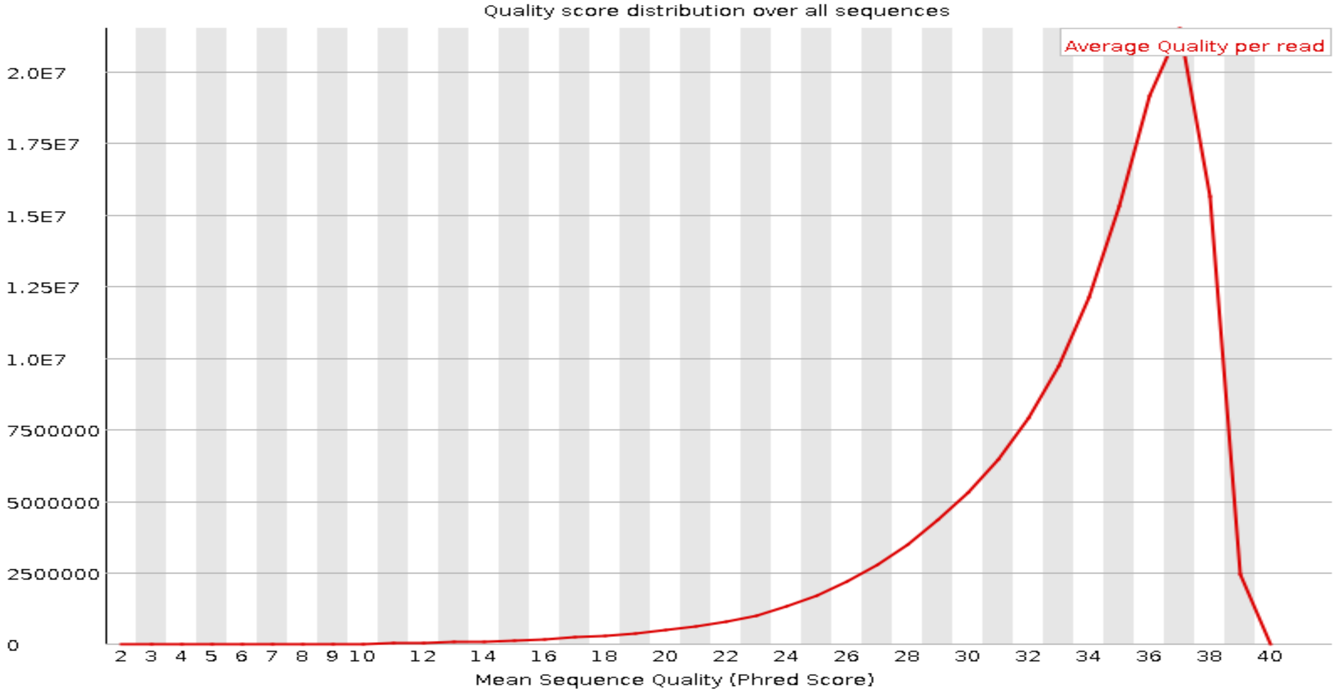
BASE STATISTICS

Measure	Value
Filename	Bowtie2 on data 10_ alignments
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	136312421
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

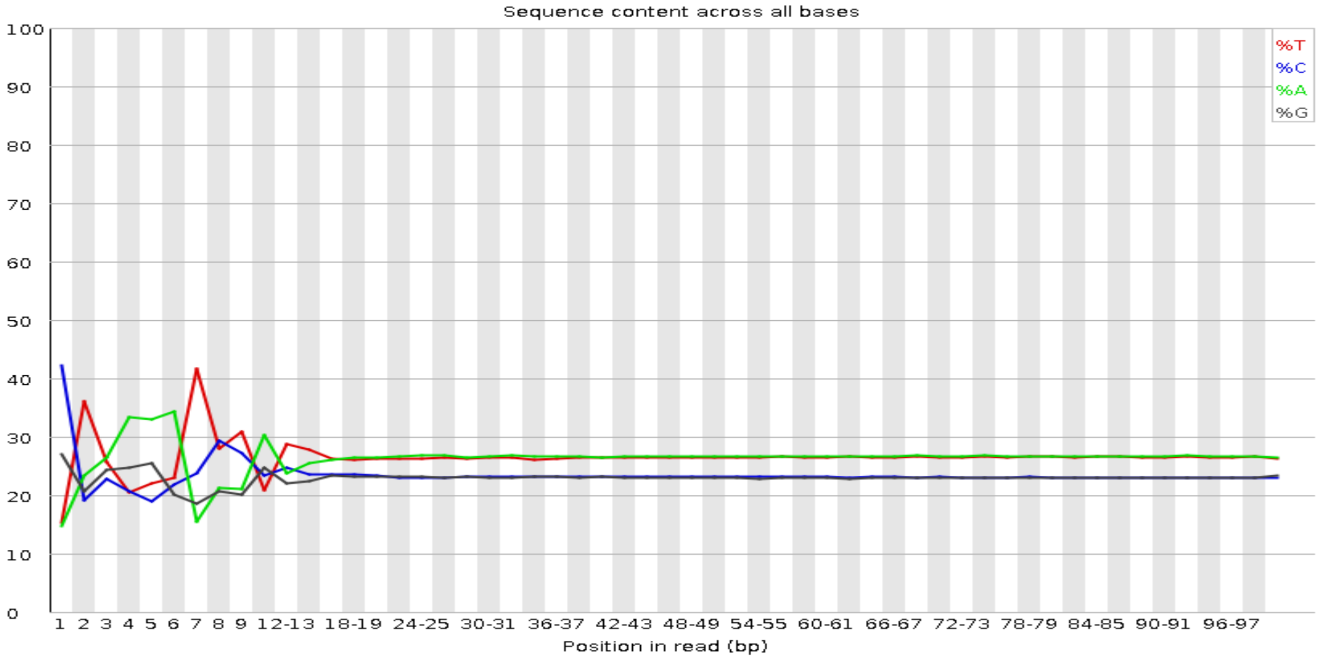
PER BASE SEQUENCE QUALITY



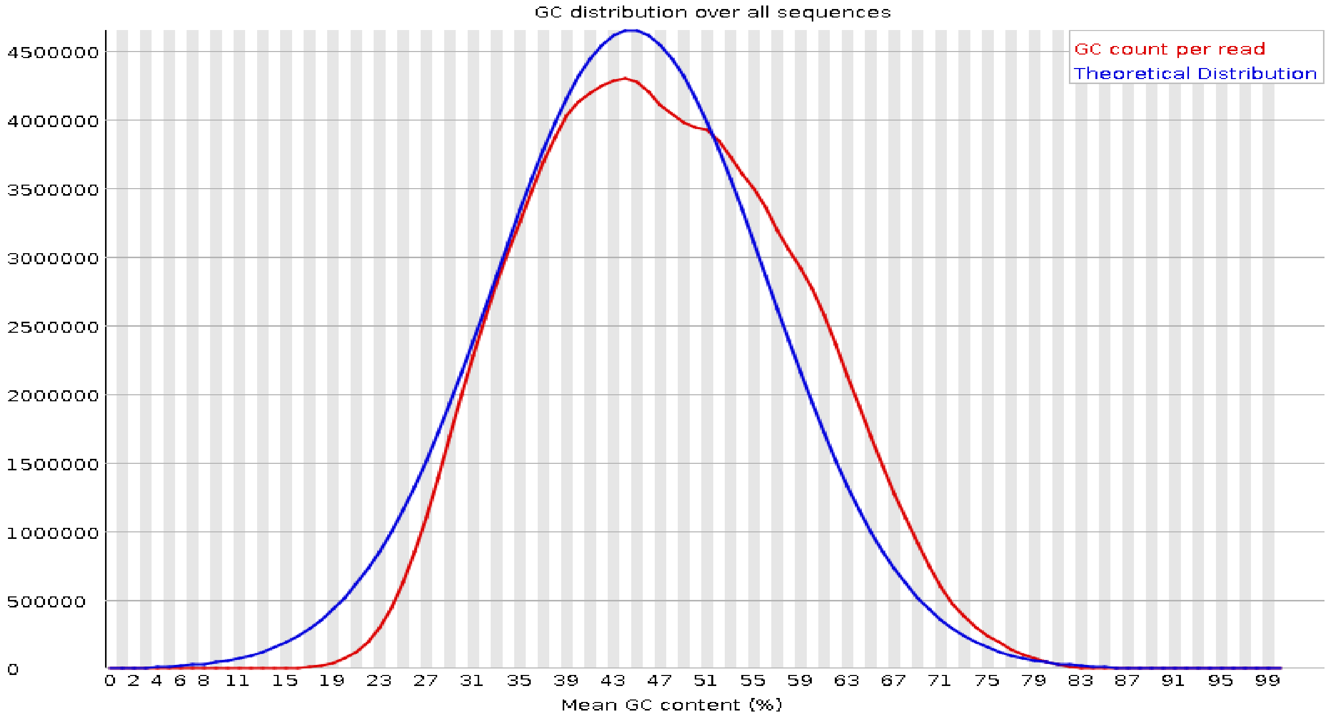
PER SEQUENCE QUALITY SCORES



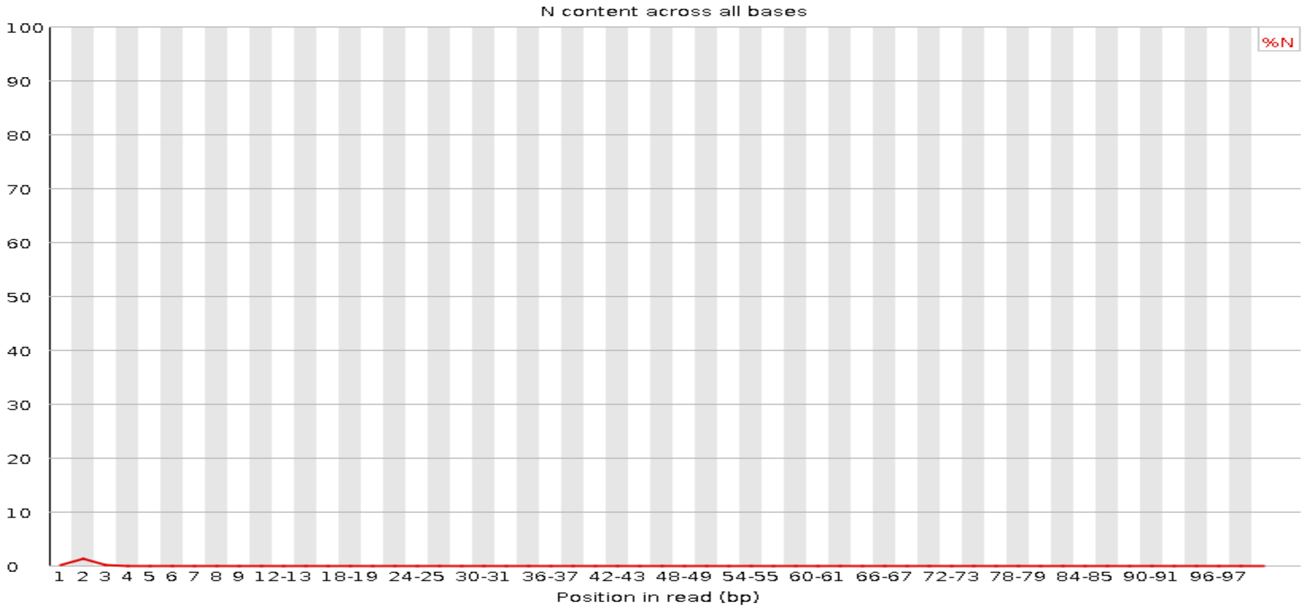
PER BASE SEQUENCE CONTENT



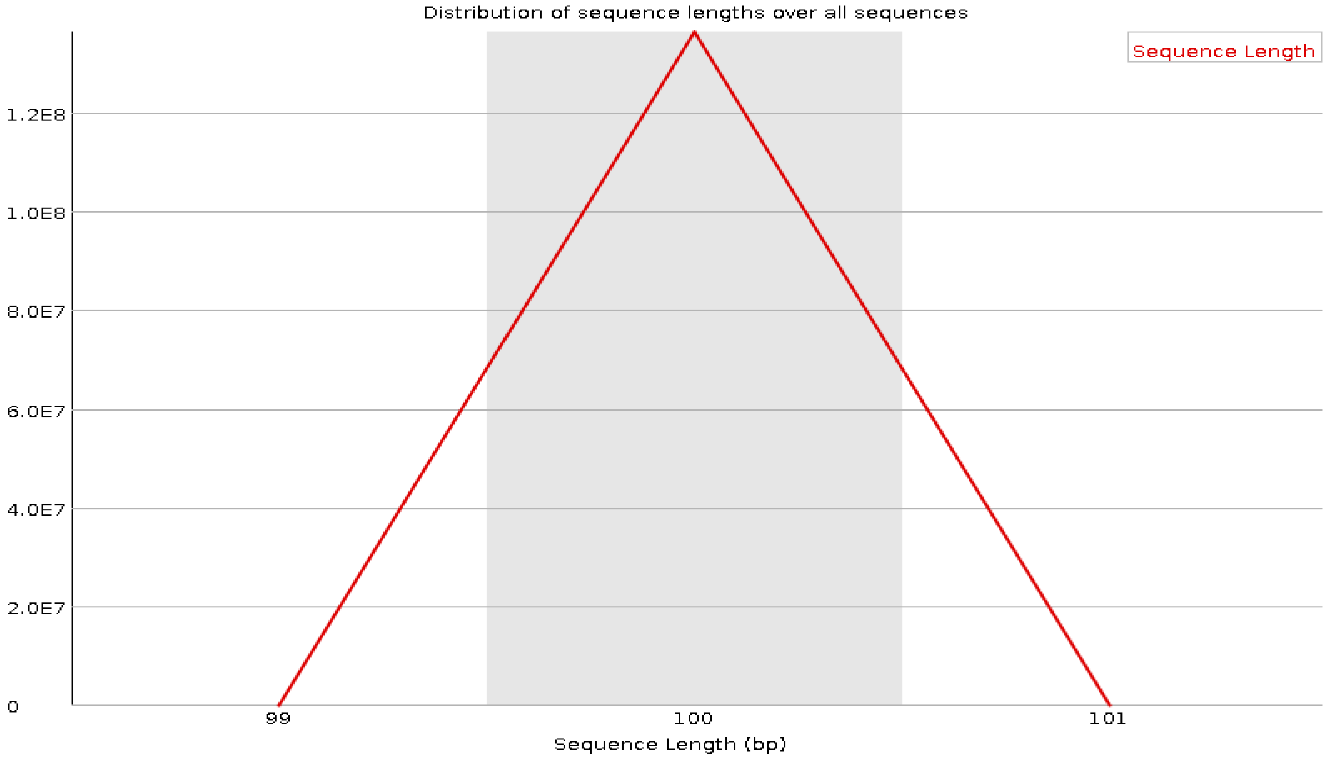
PER SEQUENCE GC CONTENT



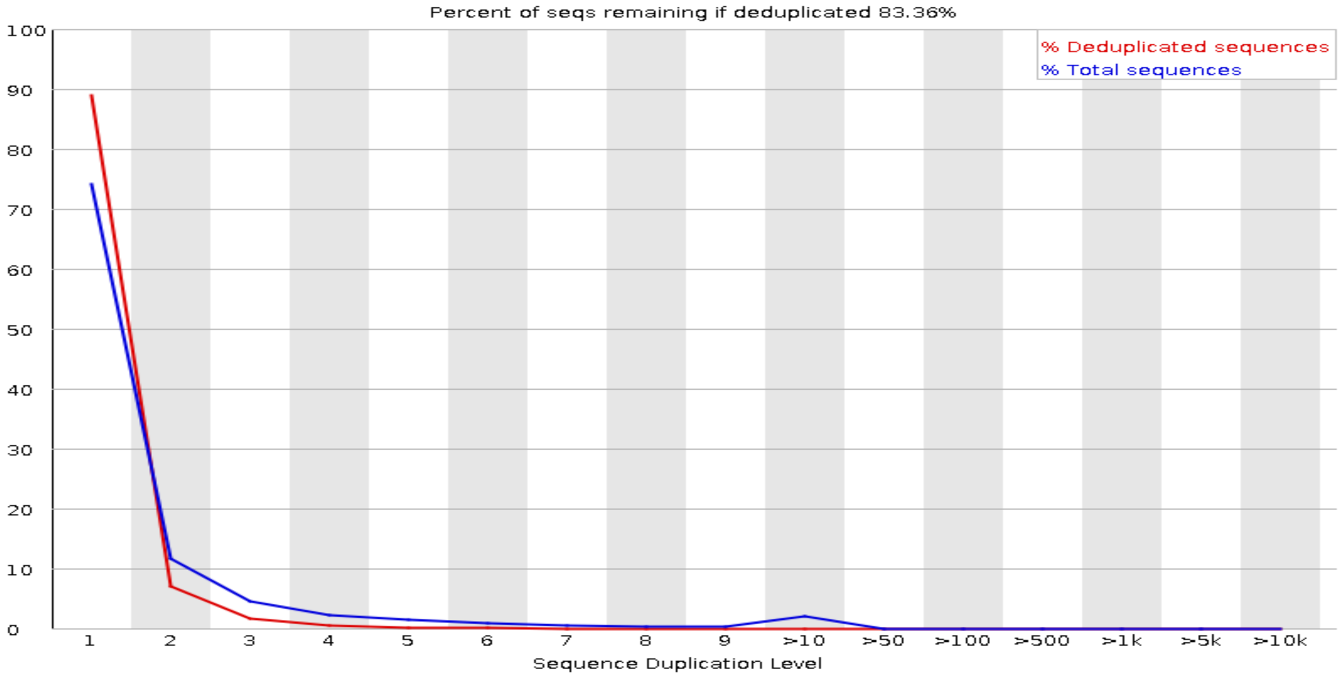
PER BASE N CONTENT



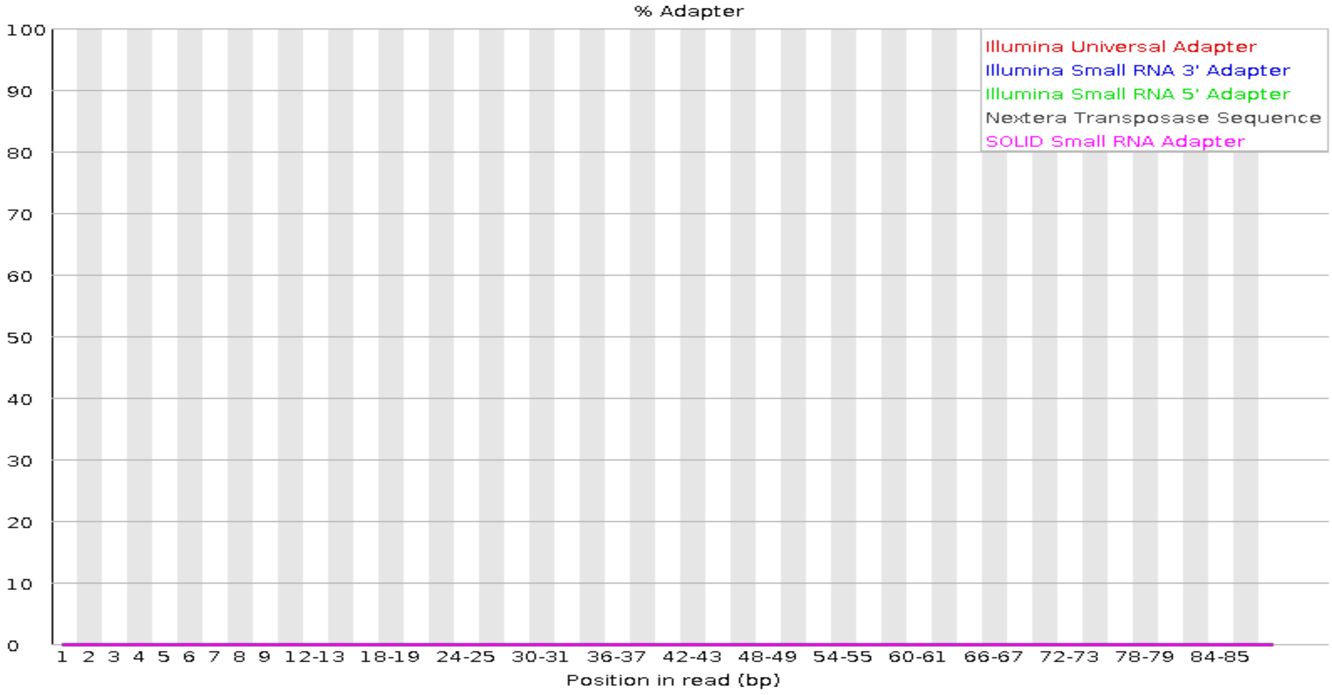
SEQUENCE LENGTH DISTRIBUTION



SEQUENCE DUPLICATION LEVELS



ADAPTER CONTENT

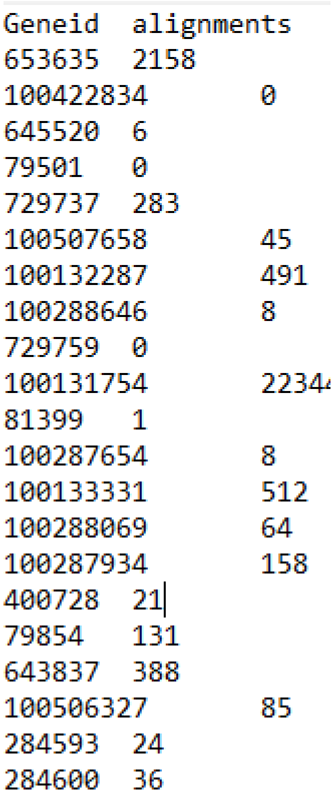


4.4 CALCULATE EXPRESSION MEASUREMENT AT GENE-COUNT LEVEL

After performing FastQC operation, we have to go obtaining gene expression at gene count level.

In order to obtain gene expression at gene count level, we have used featureCounts [20] package to find genes and their aligned reads. It gives a .tsv (tab-separated file) format file that would be further use for downstream analysis. We have performed this step via GVL and exported the data. Because for downstream analysis we will be using R-software [21].

The sample format of genes (.tsv-table separated file format) is shown below:

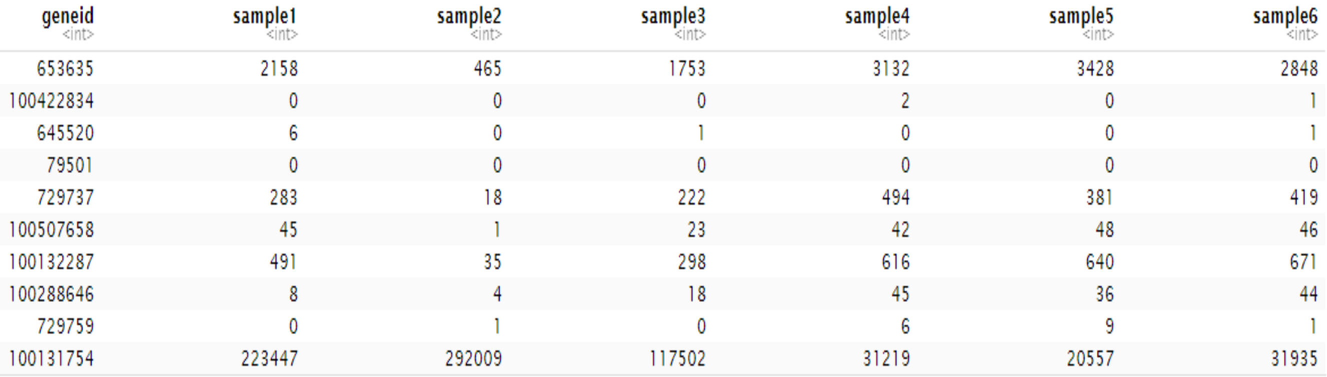


```
Geneid  alignments
653635  2158
100422834       0
645520  6
79501   0
729737  283
100507658       45
100132287       491
100288646       8
729759  0
100131754       2234
81399   1
100287654       8
100133331       512
100288069       64
100287934       158
400728  21
79854   131
643837  388
100506327       85
284593  24
284600  36
```

(Figure 12: .tsv file format (genes))

As shown in *(Figure 12: .tsv file format (genes)*, we have obtained six .tsv files for our six samples. For further downstream analysis, we have to combine these files into one single file and have to convert this single file into data frame using R programming.

Resultant data frame would look like this:



geneid <int>	sample1 <int>	sample2 <int>	sample3 <int>	sample4 <int>	sample5 <int>	sample6 <int>
653635	2158	465	1753	3132	3428	2848
100422834	0	0	0	2	0	1
645520	6	0	1	0	0	1
79501	0	0	0	0	0	0
729737	283	18	222	494	381	419
100507658	45	1	23	42	48	46
100132287	491	35	298	616	640	671
100288646	8	4	18	45	36	44
729759	0	1	0	6	9	1
100131754	223447	292009	117502	31219	20557	31935

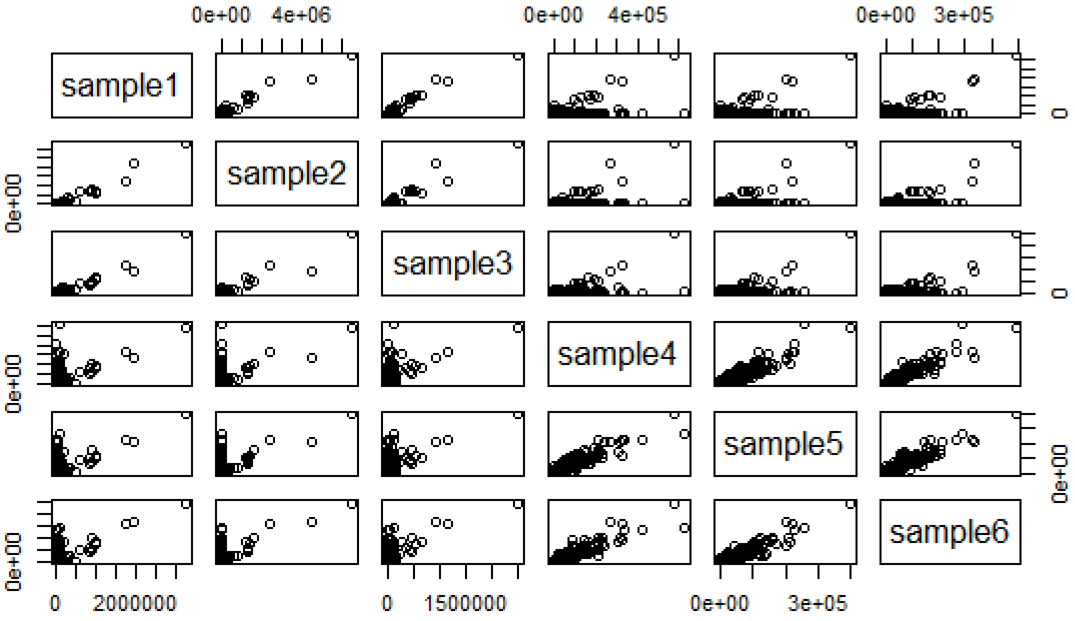
(Figure 13: Dataframe of six samples (.tsv files))

4.5 EXPLORATORY ANALYSIS

After obtaining single dataframe for our six samples from previous section, our data is ready for the actual graphical and statistical analysis.

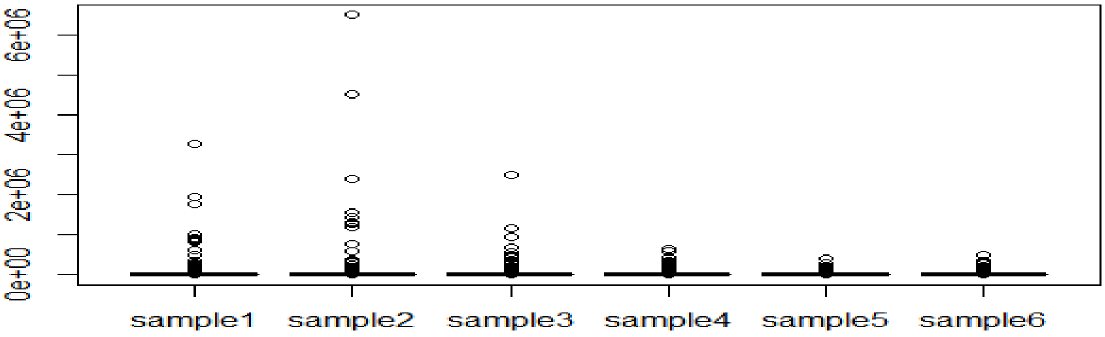
Herein, exploratory analysis we have several tools including Bar plot, plot, principle component analysis et cetra for analyzing graphically.

After performing simple **plot representation** on our data shows:



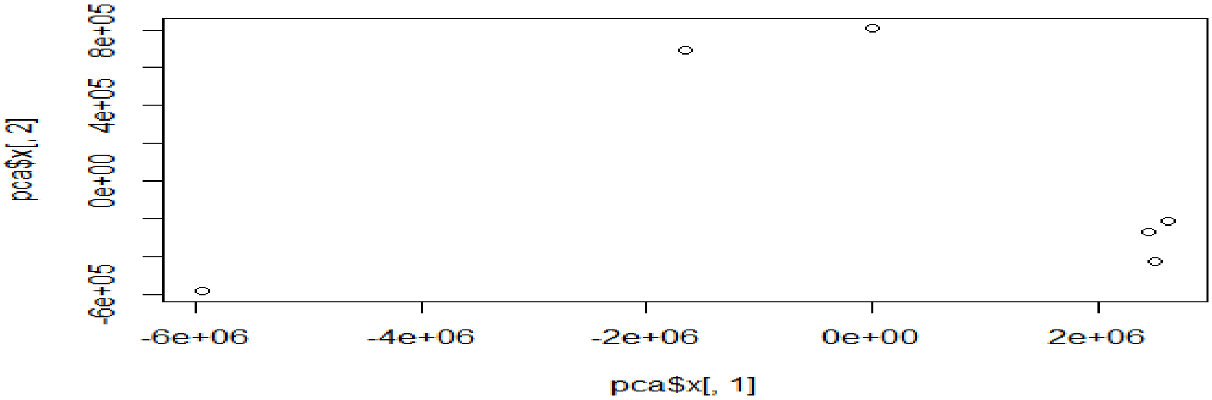
(Figure 14: Plot representation of six samples)

After performing simple **Bar plot representation** on our sample datasets:



(Figure 15: Barplot representation of our six samples)

After performing simple **PCA plot representation** on our data shows:



(Figure 16: PCA plot representation of our six samples)

Chapter 5

RESULTS

5.1 INTEGRATE FINDINGS

Transcriptome sequences (RNA-seq) of dorsolateral pre frontal cortex brain of fetal were **scattered** (*(Figure 14: Plot representation of six **samples*** ; *(Figure 15: Barplot representation of our six **samples*** ; *(Figure 16: PCA plot representation of our six **samples***) or could **not** make a compact cluster while on the other hand, adult sample's RNA-seq expression shows **compact or dense clustering.** There how, we can conclude that as human age changes from fetal to an adult the RNA-seq expression gets more compact or it converges towards a certain cluster.

In *(Figure 14: Plot representation of six **samples***, sample-1, sample-2 and sample-3 more lies on x-axis. While on the other hand sample-4, sample-5 and sample-6 (adult) having more data points on x=y line. That shows, sample-1,2 and 3 having something similar data and, sample-4,5 and 6 are also having similar data.

In *(Figure 15: Barplot representation of our six **samples***, herein it clearly shows that sample-1, sample-2 and sample-3 are scattered while on the other hand, sample-4, sample-5, and sample-6 are dense/compact.

In *(Figure 16: PCA plot representation of our six **samples***, we have further supported our finding with PCR (principal component analysis) sample 1, sample 2 and sample 3 can be seen as it scattered on the graph while in adult samples they are dense towards 2e+06.

5.2 SIMULATOR DETAILS

The software/tool that were used during the analysis:

i. Galaxy Virtual Lab (www.usegalaxy.org)
ii. Download and Extract Reads in FASTA/Q (extracted data from NCBI)
iii. Bowtie2 (map reads against reference genome – human genome (hg19))
iv. FastQC (Read Quality reports)
v. featureCounts (Measure gene expression in RNA-Seq experiments from SAM or BAM files)
vi. R and Bioconductor packages (plot, BioManager, boxplot, PCA plot, dataframe and more)

5.3 REPRODUCE RESEARCH

As our objective is to not only perform analysis but also to make it reproducible to encourage reproducibility of the thesis. We have published a galaxy workflow for reproduce the results.

[Galaxy workflow: https://usegalaxy.org/u/ajay.ducs/h/genomic-data-science-capstone]

Instructions:

1. Go to above link and run the results, it may take even days (up to 3 days) for executing of the commands.
2. After that, download featureCounts file (as per the experiment you will get 6 .tsv files).
3. Import the files in R-studio and then combine them using splicing to make a single file and perform exploratory analysis (code available at: https://github.com/Capriciousman/TRANSCRIPTOME-SEQUENCING-RNA-SEQUENCING-HUMAN-BRAIN-COMPARING-FETAL-AND-ADULT-USING-NOVEL-GENOMIC)

4. After performing exploratory analysis, get the insights from the data and support your answer.

The most common problem during reproducing the results, we face errors due to hardware/software incompatibility thus, we are enclosing hardware and software that were used during the analysis.

Hardware/Software details on which analysis performed:

MSI-GL63 9RCX + GALAXY VIRTUAL LAB
512GB SSD + 1TB HDD
RAM – 8GB DDR4
INTEL(R) Core i7-9750H CPU @2.60GHz
INTEL(R) UHD GRAPHICS 630 + NVIDIA GeForce GTX 1050 Ti
BOWTIE VERSION 2
R VERSION – 4.0.0

Chapter 6

SUMMARIZATION

This dissertation/thesis encourages the interdisciplinary study of genomic data science by using novel cutting edge data science tools/packages to analyze the RNA-seq and answer the biological research question support the answer by providing insights from the data. In this thesis, we have worked on RNA-seq data of dorsolateral pre-frontal cortex brain of fetal and adults to compare their age relation or how human being changes from fetal to an adult or how changes comes in human as age passes.

CONCLUSION

This dissertation demonstrated and supported the answer for the fundamental biological question that were asked in problem introduction section. As per the abstract, dissertation had focused on answering the question with proofs as insights from the data and also, encouraged the reproducibility of the thesis.

FUTURE WORK

We can use the similar experiment with more samples and also, with samples from different ages to see the pattern in RNA-seq data that how is it changing? If there is any pattern or particular gene responsible for growth of human being. If that happens, we would be able change the age of human being, using cutting edge CRISPR technology.

REFERENCES

[1] M. J. Streubel, “What is Computer Science? Computer,” 2003, [Online]. Available: http://www.cs.bu.edu/AboutCS/WhatIsCS.pdf.

[2] V. R. H. and H. D. Huskey, “‘Lady Lovelace and Charles Babbage,’ in Annals of the History of Computing,” *Ann. Hist. Comput.*, vol. 2, pp. 299–329, [Online]. Available: https://ieeexplore.ieee.org/document/4639398.

[3] O. Kharkovyna, “Machine Learning vs Traditional Programming.” https://towardsdatascience.com/machine-learning-vs-traditional-programming-c066e39b5b17.

[4] P. Wang, “On Defining Artificial Intelligence,” *J. Artif. Gen. Intell.*, vol. 10, no. 2, pp. 1–37, 2019, doi: 10.2478/jagi-2019-0002.

[5] R. A. Irizarry, “The Role of Academia in Data Science Education,” *Harvard Data Sci. Rev.*, pp. 1–8, 2020, doi: 10.1162/99608f92.dd363929.

[6] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, “A primer on deep learning in genomics,” *Nat. Genet.*, vol. 51, no. 1, pp. 12–18, 2019, doi: 10.1038/s41588-018-0295-5.

[7] M. D. Christopher P. Austin, “Bioinformatics.” https://www.genome.gov/genetics-glossary/Bioinformatics (accessed May 27, 2020).

[8] J. Leek, *Genomic Data Science*. 2015.

[9] Nature, “genomics,” 2014. https://www.nature.com/scitable/definition/genomics-126/.

[10] F. P. and others E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, “An estimation of the number of cells in the human body,” *Ann. Hum. Biol.*, vol. 40, pp. 463–471, 2013, [Online]. Available: https://www.tandfonline.com/doi/abs/10.3109/03014460.2013.807878.

[11] N. H. G. R. Institute, “A brief guide to genomics,” 2019. .

[12] N. H. G. R. Institute, “Transcriptome Fact Sheet,” 2015, [Online]. Available:

https://www.genome.gov/about-genomics/fact-sheets/Transcriptome-Fact-Sheet.

[13] S. B. and J. H. M. Garrels, J. I., “Proteome,” *Encyclopedia of Genetics*. pp. 1575–1578, 2001.

[14] kasper daniel hansen et al James Taylor, Jacob pritt, Liliana florea, “Genomic Data Science Specialization,” 2015. https://www.coursera.org/specializations/genomic-data-science.

[15] A. E. Jaffe *et al.*, “HHS Public Access transcriptome illuminate schizophrenia pathogenesis,” vol. 21, no. 8, pp. 1117–1125, 2019, doi: 10.7303/syn12299750.

[16] F. C. P. Navarro, H. Mohsen, C. Yan, S. Li, M. Gu, and W. Meyerson, “Genomics and data science : an application within an umbrella,” pp. 1–11, 2019.

[17] “GALAXY VIRTUAL LAB.” https://usegalaxy.org/.

[18] J. H. University, “BOWTIE2.” http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml (accessed Jul. 04, 2020).

[19] S. Andrews, “FastQC 1 . 1 What is FastQC 2 . Basic Operations 2 . 1 Opening a Sequence file,” 2010, [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[20] W. Shi and Y. Liao, “Rsubread / Subread Users Guide,” no. October, 2019.

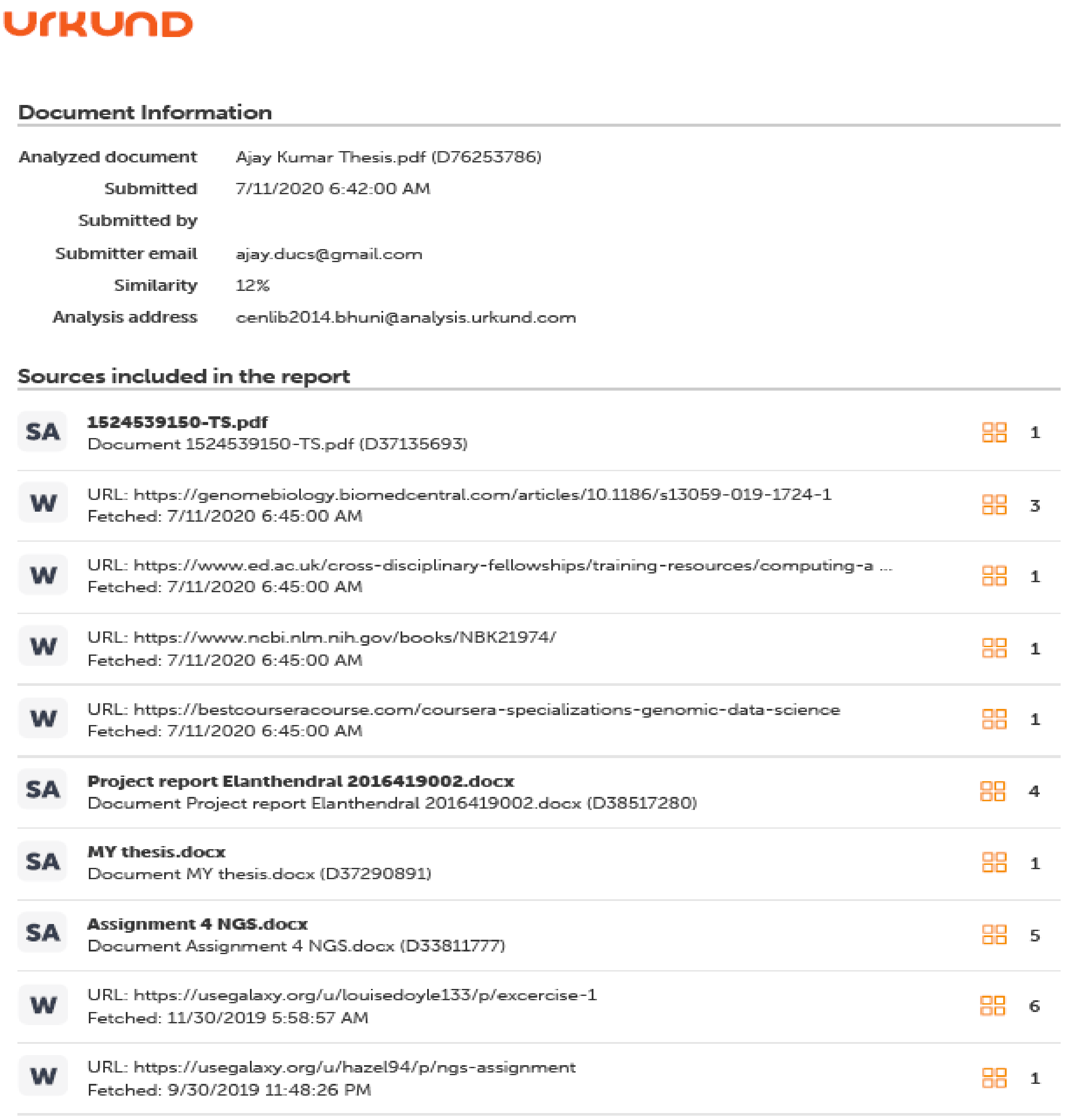
[21] T. Hothorn and B. S. Everitt, “- An Introduction to R,” *A Handb. Stat. Anal. using R*, vol. 2, pp. 32–55, 2020, doi: 10.1201/b17081-8.

[22] S. Anders and W. Huber, “Differential expression of RNA-Seq data at the gene level–the DESeq package,” *DESq Man.*, pp. 1–28, 2012, doi: 10.1016/S0926-3373(02)00165-0.

[23] S. S. D. United States Bureau of the Census, Software and Standards Management Branch, “Survey Design and Statistical Methodology Metadata,” no. section 3.3.7, p. 14, [Online]. Available: https://stats.oecd.org/glossary/detail.asp?ID=542.

[24] B. Langmead, C. Wilks, V. Antonescu, and R. Charles, “Scaling read aligners to hundreds of threads on general-purpose processors,” *Bioinformatics*, vol. 35, no. 3, pp. 421–432, 2019.

PLAGIARISM REPORT



URKUND

Document Information

Analyzed document	Ajay Kumar Thesis.pdf (D76253786)
Submitted	7/11/2020 6:42:00 AM
Submitted by	
Submitter email	ajay.ducs@gmail.com
Similarity	12%
Analysis address	cenlib2014.bhuni@analysis.urkund.com

Sources included in the report

SA	**1524539150-TS.pdf** Document 1524539150-TS.pdf (D37135693)	⊞	1
W	URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1724-1 Fetched: 7/11/2020 6:45:00 AM	⊞	3
W	URL: https://www.ed.ac.uk/cross-disciplinary-fellowships/training-resources/computing-a ... Fetched: 7/11/2020 6:45:00 AM	⊞	1
W	URL: https://www.ncbi.nlm.nih.gov/books/NBK21974/ Fetched: 7/11/2020 6:45:00 AM	⊞	1
W	URL: https://bestcourseracourse.com/coursera-specializations-genomic-data-science Fetched: 7/11/2020 6:45:00 AM	⊞	1
SA	**Project report Elanthendral 2016419002.docx** Document Project report Elanthendral 2016419002.docx (D38517280)	⊞	4
SA	**MY thesis.docx** Document MY thesis.docx (D37290891)	⊞	1
SA	**Assignment 4 NGS.docx** Document Assignment 4 NGS.docx (D33811777)	⊞	5
W	URL: https://usegalaxy.org/u/louisedoyle133/p/excercise-1 Fetched: 11/30/2019 5:58:57 AM	⊞	6
W	URL: https://usegalaxy.org/u/hazel94/p/ngs-assignment Fetched: 9/30/2019 11:48:26 PM	⊞	1